MedSyn: Text-Guided Anatomy-Aware Synthesis of High-Fidelity 3D CT Images

Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich

***Abstract*— This paper introduces an innovative methodology for producing high-quality 3D lung CT images guided by textual information. While diffusion-based generative models are increasingly used in medical imaging, current state-of-the-art approaches are limited to low-resolution outputs and underutilize radiology reports' abundant information. The radiology reports can enhance the generation process by providing additional guidance and offering fine-grained control over the synthesis of images. Nevertheless, expanding text-guided generation to high-resolution 3D images poses significant memory and anatomical detail-preserving challenges. Addressing the memory issue, we introduce a hierarchical scheme that uses a modified UNet architecture. We start by synthesizing low-resolution images conditioned on the text, serving as a foundation for subsequent generators for complete volumetric data. To ensure the anatomical plausibility of the generated samples, we provide further guidance by generating vascular, airway, and lobular segmentation masks in conjunction with the CT images. The model demonstrates the capability to use textual input and segmentation tasks to generate synthesized images. Algorithmic comparative assessments and blind evaluations conducted by 10 board-certified radiologists indicate that our approach exhibits superior performance compared to the most advanced models based on GAN and diffusion techniques, especially in accurately retaining crucial anatomical features such as fissure lines and airways. This innovation introduces novel possibilities. This study focuses on two main objectives: (1) the development of a method for creating images based on textual prompts and anatomical components, and (2) the capability to generate new images conditioning on anatomical elements. The advancements in image generation can be applied to enhance numerous downstream tasks.**

***Index Terms*— Diffusion Model, Text-guided image generation, 3D image generation, Lung CT, Volume Synthesis with Radiology Report, Controllable Synthesis.**

This work is equally contributed by Y. Xu, L. Sun and W. Peng and was partially supported by NIH Award Number 1R01HL141813-01, NSF 1839332 Tripod+X, and SAP SE. We were also grateful for the computational resources provided by Pittsburgh SuperComputing grant number TG-ASC170024.

Y. Xu, L. Sun, S. Jia and K. Batmanghelich are with Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 (email: {yanwuxu,lisun,brucejia,batman}@bu.edu)

W. Peng are with Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305 (email: wepeng@stanford.edu)

K. Morrison, A. Perer and M. Eslami are with Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213 (email: kcmorris@cs.cmu.edu, adamperer@cmu.edu, meslami@andrew.cmu.edu)

A. Zandifar is with the University of Pittsburgh Medical Center, Pittsburgh, PA 15213 (email: zandifara@upmc.edu)

S. Visweswaran is with the Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206 (email: shv3@pitt.edu)

I. Introduction

Denoising Diffusion Probabilistic Models (DDPM) [1], also known as score-based generative models [2], have emerged as powerful tools in both computer vision and medical imaging due to their stability during training and exceptional generation quality. State-of-the-art image generation tools, such as Imagen [3] and latent Diffusion Models (LDMs) [4], can employ text prompts to provide fine-grained guidance during image generation. This capability is promising for medical imaging, encompassing applications like privacy-preserving data generation, image augmentation, black box uncertainty quantification, and explanations. Although methods such as RoentGen [5] have illustrated the potential of 2D cross-modality generative models conditioned on text prompts, there are no known text-guided volumetric image generation techniques for medical imaging. Extending such an approach to 3D presents challenges, including high memory demands and preserving crucial anatomical details. This paper aims to address these challenges.

To enhance the image resolution for generative models, increasing the voxel count within a fixed field of view is memory-intensive [6]. Synthesizing 3D volumetric images at high resolutions (e.g., $256 \times 256 \times 256$) demands substantial memory since the neural network must store intermediate activations for backpropagation. To preserve essential anatomical details and condition on input text prompts, a high-capacity network is essential, further compounding the issue. Although GANs [6]–[9] and conditional Denoising Diffusion Probabilistic Models (cDPMs) [10]–[12] have set benchmarks in volumetric medical imaging synthesis, each has its limitations. GANs, despite their inference efficiency, can compromise sample diversity and sometimes produce anatomically implausible artifacts. Conversely, diffusion models, built to boost sample diversity through iterative denoising, grapple with memory constraints, primarily due to their resource-intensive 3D attention UNet. The iterative denoising coupled with the sequential sub-volume generation in cDPMs makes them time-intensive at the inference stage. As a result, many diffusion models are relegated to 2D or low-resolution volumetric applications. Incorporating text prompts, such as radiology reports, into image generation presents another challenge. Generative models utilizing text prompts as conditions demand

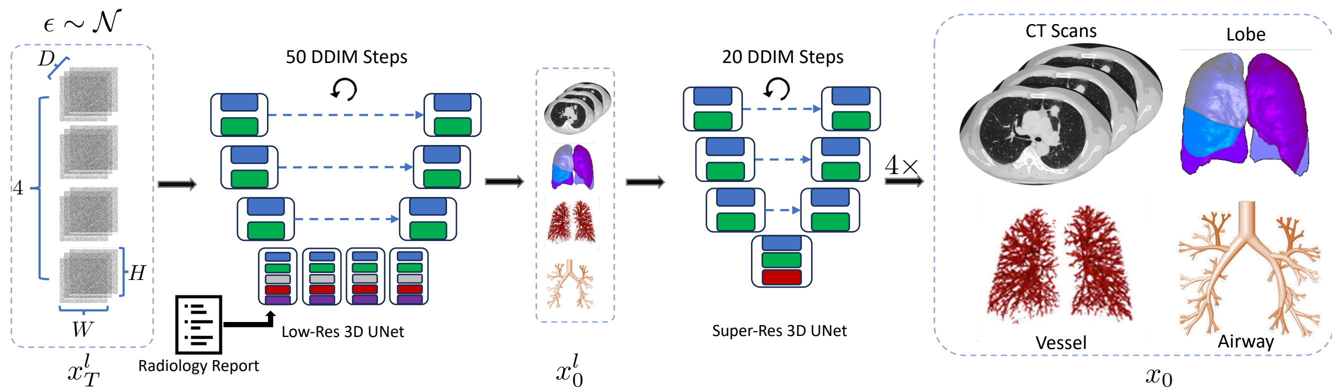

Fig. 1. Overview of our generative model, MedSyn. Using a hierarchical approach, we first generate a 64× 64 × 64 low-resolution volume, along with its anatomical components, conditioning on Gaussian noise ϵ and radiology report. The low-resolution volumes are then seamlessly upscaled to a detailed 256× 256 × 256 resolution.

a high-capacity denoiser to map the subtle and occasionally ambiguous pathologies mentioned in the reports to visual patterns. Enhancing the capacity of the UNet, further increases memory usage.

In 3D medical image synthesis, limitations often manifest as "hallucinations," leading to potential biases and inaccuracies—critically concerning in medical settings. Our paper concentrates on generating Computed Tomography (CT) images of the lung. Hallucinations may manifest as missing fissure lines separating lung lobes or the creation of implausible airway and vessel structures. This problem of hallucinations is less pronounced in X-ray image generation (e.g., RoentGen [5]) since the 2D X-ray projection obscures many fine details, like the lung's lobular structure, airway, and vessels. However, the issue becomes more pronounced in 3D image generation. Training on hundreds of thousands of 3D medical images isn't feasible, necessitating a strong prior to constraining the space of generative models.

To address these challenges, we propose MedSyn, a model tailored for high-resolution, text-guided, anatomy-aware volumetric generation. Leveraging a large dataset of approximately 9,000 3D chest CT scans paired with radiology reports, we employ a hierarchical training approach for efficient cross-modality generation. Given the input of tokenized radiology reports, our method initiates with a low-resolution synthesis ($64 \times 64 \times 64$), which is then fed to a super-resolution module to upscale the image to $256 \times 256 \times 256$. We have modified the UNet to bolster the network's capacity (i.e., the number of parameters) without significantly increasing memory requirements. MedSyn enhances controllability by harnessing textual priors from radiology reports to guide synthesis. We further regularize the generator by creating segmentation masks of the lung's airway, vessels, and lobular structure as additional output channels alongside synthetic CT images to conserve detailed anatomical features. The resultant model can condition not only on the text but also on single or paired anatomical structures. Our algorithmic experiments and blind evaluations conducted by 10 board-certified radiologists highlight the superior generative quality and efficiency of the proposed method compared to baseline techniques. To the best of our knowledge, this is the first work to evaluate a text-to-medical image model with radiologists empirically. We also delve into the significance of components within our proposed modules. Our code and pre-trained model is publicly available at https://github.com/batmanlab/MedSyn

II. Related Works

A. Image Synthesis Based on Text Prompt

Text-conditional image generation enables new applications and improves the diversity of generated images compared to models that are not conditioned on text. The most relevant studies in the domain of multi-modal generation include Imagen [3], Latent Diffusion (Stable Diffusion) [4], Video Diffusion [13], and RoentGen [14]. For natural images, both Imagen and Latent Diffusion pioneered the text-conditional diffusion models, producing high-fidelity 2D natural images. Dall·E 2 [15] uses pre-trained CLIP model to extract features, which guides text-to-image generation Video Diffusion introduces a method to generate videos using score-guided sequential diffusion models, accepting text prompts as conditional inputs. In medical imaging domain, RoentGen enhanced the Stable Diffusion prior to achieve an exceptional generative quality for Chest X-Ray scans. TauPETGen [16] proposes to utilize text description as a condition for generating 2D Tau PET images. However, these methodologies cannot be seamlessly integrated into multi-modal 3D medical volume generation because of their inherent 2D orientation or the challenges associated with efficient high-resolution 3D volume creation. In addition, most previous methods adopt language models pre-trained on generic text, then fine-tune on it using biomedical text. In contrast, we utilize a language encoder model [17] trained specifically with biomedical text data that learns domain-specific vocabulary and semantics.

B. Generative Models for 3D Medical Imaging

Generative models have emerged as a powerful tool in the field of medical imaging, offering a range of applications and benefits. Previous work [18], [19] leverages 3D GAN for volume generation. However, the generated images are limited to the small size of $128\times128\times128$ or below, due to insufficient memory during training. In order to address the challenges of producing efficient high-resolution medical volumes, HA-GAN [6] introduces an end-to-end GAN framework that operates on multiple-level feature maps. This can synthesize medical volumes of dimensions $256\times256\times256$ within a single model forward pass. Recently, diffusion models have evolved, and both cDPM [10] and SADM [11] aim to generate full 3D volumes. This sequential approach was adopted primarily because of the considerable memory demands of the 3D Attention UNet. In order to mitigate the discontinuity between slices due to sequential generation, MedGen3D [20] introduces a diffusion refiner that generates images from three different views and averages them. In order to address the memory constraint, Medical Diffusion [21] proposes to use VQ-GAN [22] to compress volume into latent space, then trains a diffusion model in the latent space. Yet, a compromise arises between memory efficiency and speed during inference, making it challenging to scale up for high-resolution and expansive cross-modality medical volume creation. Pinaya et al. [23] have previously used LDMs to generate 3D brain MRI. The aforementioned limitations of previous works underscore the unique value of our proposed method for efficient high-fidelity cross-modality medical volume generation, distinguishing it from prior techniques and applications.

III. Background

Diffusion models contain two processes: a forward diffusion process and a reversion diffusion process. The forward diffusion gradually generates corrupted data from $x_0 \sim q(x_0)$ via interpolating between the sampled data and the Gaussian noise as follows:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}),$$
$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}),$$

where T denotes the maximum time steps and $\beta_t \in (0,1]$ is the variance schedule. According to Sohl-Dickstein *et al.* [24], this can be reformed as a closed-form solution, from which we can sample x_t at any arbitrary time step t without needing to iterate through the entire Markov chain.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

in which $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$. The parameterized reversed diffusion can be formulated correspondingly:

$$p_\theta(x_{0:T}) := p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{2}$$
$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \sigma_t^2\mathbf{I}), \tag{3}$$

where we can parameterize $p_\theta(x_{t-1}|x_t)$ as Gaussian distribution when the noise addition between each adjacent step is sufficiently small, the denoised function μ_θ produces the mean value of the adjacent predictions, while the determination of the variance σ_t relies on β_t. The optimization objective can be written as follows:

$$\mathcal{L} = -\sum_{t>0}\mathbb{E}_{q(x_0)q(x_t|x_0)}\mathrm{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)), \tag{4}$$

which indirectly maximizes the ELBO of the likelihood $p_\theta(x_0)$. When x_0 is given, the posterior $q(x_{t-1}|x_t,x_0)$ is Gaussian. Thus, the above objective becomes ℓ_2 distance between the sampled x_{t-1} from the posterior and the predicted denoised data, conditioning on time step t. The practice of predicting x_0 is also used in previous work [3]. The final minimization objective for the diffusion model G can be simplified as follows:

$$\mathcal{L} = \mathbb{E}_{q(x_0)q(x_t|x_0)}\|G(x_t;\theta) - x_0\|_2^2 \tag{5}$$

IV. Method

Our radiology report conditional generated framework (MedSyn), as illustrated in Fig. 1, is a hierarchical model built upon the conditional diffusion models, which take the inputs of a random Gaussian noise ϵ and the text token features f_{text} to generate a $256\times256\times256$ medical volumes. The proposed model has three core components: 1. A pre-trained text-encoder (Medical BERT) for extracting language features from radiology reports. 2. A text-guided low-resolution 3D diffusion model that jointly synthesizes CT volume and its anatomical structure volumes. 3. A super-resolution 3D diffusion model for scaling up the low-resolution generated volumes and complementing the missing anatomical details. In the following session, we will introduce the details of our proposed method.

A. Learning Representations from Radiology Reports

To enable the text-guided medical image generation that utilize radiology reports as the guidance, we train a text encoder (Medical BERT [17]) that encodes natural language from radiology reports into high-dimensional vectors (f_{text}). The encoder schematic is shown in the lower part of Fig. 2. Specifically, we use the CT radiology reports in our collected data to fine-tune a BERT model [25] pre-trained on datasets of chest X-ray reports. We use the “Impression”, and the “Findings” sections of CT reports as the input for the model since they provide the most characteristic descriptions related to the disease symptoms and anatomy structures. The model is then fine-tuned using two tasks: `(a)` Mask language modelling; `(b)` Paired section prediction. In task (a), we randomly replace 15% of input tokens with `[MASK]`, and use the network to predict the masked token. In task (b), we randomly swap the “Impression” section of half of the samples with another subject’s “Impression” section. As a result, half of the samples in the dataset have unpaired “Impression” sections and “Findings” sections. In this way, we construct different `[CLS]` embedding for paired and unpaired samples. Therefore, we train a linear classifier that uses the `[CLS]` embedding to predict whether the input “Impression” and the “Findings” sections are paired or not.

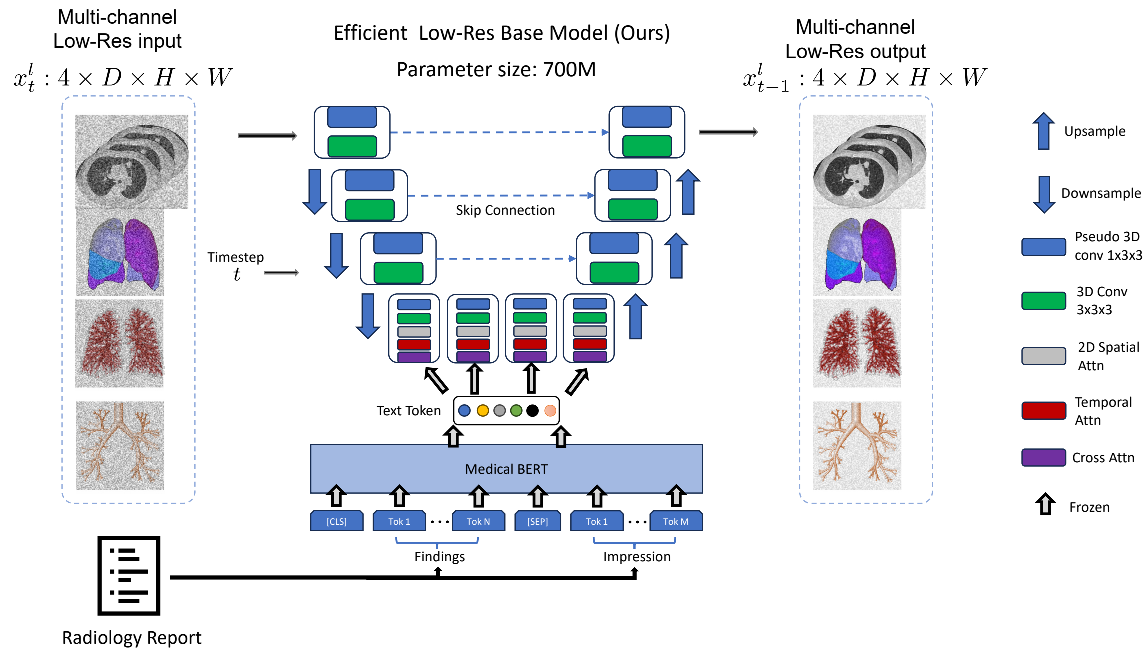

Fig. 2. This figure shows our efficient low-res generative model with the clinical tokens input. In this process, we train the denoising diffusion UNet and fix the pre-trained text feature extractor of Medical BERT. To be notified, our low-res base model has a large capacity of 700 million parameters.

B. Text-Conditioned Volume Generation

This section discusses how to incorporate the text embedding from the Medical BERT. Direct training of high-resolution diffusion models (such as $256 \times 256 \times 256$ for a given field of view) is highly memory-demanding and thus not feasible. We propose an efficient hierarchical model with a two-phase process. In the first phase, we generate a low-resolution volume ($64 \times 64 \times 64$) conditioned via the radiology report. In the second phase, the model outputs a high-resolution volume $256 \times 256 \times 256$ from a 3D super-solution module, which only takes low-resolution volume as input. The low-resolution image ensures the volumetric consistency of the final images. Using the radiology reports as a condition for low-res image generation is advantageous since we do not need to compromise between resolution and model capacity.

1) Low-Resolution Volume Generation Conditioned on Reports: To generate data conditioned via specific signals c, e.g., text information, we need to reformulate the unconditional DDPM objective in Equation 4 with the conditions,

$$\mathcal{L} = -\sum_{t>0} \mathbb{E}_{q(x_0)q(x_t|x_0,c)} \mathrm{KL}(q(x_{t-1}|x_t, x_0, c)||p_\theta(x_{t-1}|x_t, c)). \tag{6}$$

We first downsample the CT volume x_0 by $4\times$ to produce the low-resolution one x_0^l. Then, given the extracted features from the radiology report, we can finalize the training loss for the text conditional diffusion model as follows:

$$\mathcal{L}_{low} = \mathbb{E}_{q(x_0, f_{text})q(x_t^l|x_0^l)} ||G(x_t, f_{text}; \theta) - x_0^l||_2^2, \tag{7}$$

where G is the denoising model with cross-attention modules and θ is its parameters. We inject the text conditional information into the denoising network to reconstruct the original x_0^l via a cross-attention mechanism.

2) Super Resolution Model: Our model puts most of the learning capacity in the low-resolution module. We designed a lightweight diffusion UNet for the super-resolution module, which takes the low-resolution input and outputs the full high-resolution volume. For the super-resolution module, we design the loss to match the denoising distribution as follows:

$$\begin{aligned}\mathcal{L}_{sup} &= -\sum_{t>0} \mathbb{E}_{q(x_0, x_0^l)} q(x_t|x_0)[\\ &\qquad \mathrm{KL}(q(x_{t-1}|x_t, x_0^l)||p_\theta(x_{t-1}|x_t, x_0^l))]\\ \mathcal{L}_{sup}^{approximate} &= \mathbb{E}_{q(x_0, x_0^l)} q(x_t|x_0)[\\ &\qquad ||H(x_t, x_0^l; \phi) - x_0||_2^2],\end{aligned} \tag{8}$$

where H is the super-resolution denoising module with its parameters ϕ. To be notified, we do not contain the additional text information in our super-resolution module to save the computational cost.

C. Joint Generative Modeling of Volume and Anatomical Structures

Lung CT scans exhibit more intricate details compared to other organs. These include the bronchial tree formations, the delicate vascular network, and fissure lines delineating the lung lobes. Such complexity heightens the challenge of producing high-fidelity synthetic lung CT volumes. And, those subtle anatomical structures are easily ignored during the generative modeling. Moreover, anatomical structures are usually roughly described in the clinical report, which enlarges the difficulty of CT synthesis purely based on radiology reports. Therefore, we propose to model the volume generation jointly with the anatomical structure generation, all conditioned on the radiology reports. To this end, we extract the shape information for the core anatomical structures, i.e., the lung lobes, the airway and the vessels. We choose the commonly used pre-trained segmentation tools to provide stable shape information for these three structures. We use lungmask [26] to segment lobes from CT volumes and use TotalSegmentator [27] to segment vessels, and the NaviAirway [28] is utilized for airway segmentation. We denote the segmentation map as l, a, v for the lung, airway and vessels. To enable the previous model $G(\cdot;\theta)$ to jointly synthesize CT volume x along with its structures l, a, v, we simply add three more channels to the input and output layers of model G, while still conditioning on time step t. We then directly concatenate them in the channel dimension and construct a diffusion process on four-channel 3D volumes. To be notified, all the shape segmentations are paired with the volumes in the concatenation operation. We define the concatenation operation as $concat(\cdot)$ and the newly constructed volume as $x' = concat(x, l, a, v)$. Then, with minor modifications to Equations 7 and 8, we can write down our joint generation training objectives for the low-resolution phase and high-resolution phase as follows:

$$\mathcal{L}_{low} = \mathbb{E}_{q(x'_0,text)q(x'_t|x'_0)} \|G(x'_t, f_{text};\theta) - x'_0\|_2^2, \quad (9)$$

$$\mathcal{L}_{sup} = \mathbb{E}_{q(x'_0,x''_0)q(x'_t|x'_0)} \|H(x'_t, x''_0;\phi) - x'_0\|_2^2. \quad (10)$$

D. Anatomically Controllable Synthesis

Compared with the conditional modelling of ControlNet [29], our model is significantly different, where we model the volume generation with the structure information jointly $p(x, c)$ while ControlNet generates data conditioned on the structure $p(x|c)$. Ours is more flexible as MedSyn can still generate data when the predefined structures are unavailable, which is not feasible for ControlNet. Furthermore, if we marginalize one or multiple components in the joint denoising process, such as fixing the l given the predefined lobe structure inputs, we can achieve exactly what ControlNet can do. Further, if we fix the x, we can get the structure output from our model, such as segmentation maps. We will show these advantages of our model in the experimental section.

E. Efficient 3D Attention UNet

Although we propose an efficient two-phase text2volume generation, like other works in 3D generation [10], we still face the memory issue when generating such high-resolution volume. Sequential generation with conditional diffusion models [10], [11] works as one solution, but new issues will be easily introduced. Therefore, we design a new base neural architecture for much more efficient volume generation. Compared with the common 3D attention UNet [13] for video generation, we build an encoder-decoder with pure convolutions and move all the attention mechanisms into the bottom of the UNet. In this way, we propose a more efficient base model structure that drops all the computational burden from the attention mechanism while still benefiting it from the latent space where spatial resolution is much lower. This increases the parameter size via $10\times$ more but largely increases the computational efficiency. For the super-resolution network, we almost remove all the attention modules and keep one temporal attention at the bottom of the UNet, which makes it feasible for $256 \times 256 \times 256$ volume inputs.

F. Implementation Details

To pre-train our text encoder, we use 209,683 reports without paired images, and 7,728 reports with paired images as training set. We pre-train our text encoder for 5 epochs. Our training objective for the diffusion models is a simple reconstruction loss: the ℓ_2 pixel-level reconstruction between the ground truth and the prediction for different denoising time steps without any other terms, such as perceptual loss [30]. Following the common choice of training diffusion model, we use a continuous cosine time scheduler [1]. For the training of both the low-resolution diffusion model and the super-resolution model, the number of time steps is set as 1000. During inference, we use 50 DDIM steps for the low-resolution diffusion model, and 20 DDIM steps for the super-resolution model. For the optimizer, we apply AdamW [31] with the learning rate of 1×10^{-4} and $\beta = \{0.9, 0.999\}$ with clipping the gradient norm bigger than 1. We apply the mixed-precision for optimizing models to make the training more efficient. The gradient accumulation is applied during training to scale up the training batch size, as diffusion models are sensitive to the small training batch size [32]. Ultimately, we can train our 700M parameters low-resolution base model with a batch size of 64 and the super-resolution model with a batch size 32 on four NVIDIA A6000 GPUs. Both models have converged after undergoing 40k iterations of training.

V. Experiments

This section presents a comprehensive evaluation of the proposed generative model, MedSyn. We will first describe the dataset used in this experiment. Then, the MedSyn is compared with the state-of-the-art GANs and diffusion model, including WGAN [33], α-GAN [18], HA-GAN [6] and Medical Diffusion [34]. Extensive comparisons and analysis are finally given to evaluate the effectiveness of our method, qualitatively and quantitatively. For training of all baseline methods, we use the author's implementation. We made minimal modifications to the code to adapt to our dataset. All models are trained from scratch, for fair comparison with our method.

A. Dataset

We conduct experiments on a large-scale 3D dataset, which contains 3D thorax computerized tomography (CT) images and associated radiology reports from 8,752 subjects. The dataset also contains 209,683 reports without corresponding images. The images and reports are collected by the University of Pittsburgh Medical Center and have been de-identified. We randomly split our dataset of 8,752 subjects into a training set consisting of 7,728 subjects (88%), and a validation set of 1,024 subjects (12%). The images have been pre-aligned using affine registration and re-sampled to $1mm^3$. We resize the images to $256 \times 256 \times 256$. We use the nearest-neighbor downsampling to reduce the scans $4\times$ to train the

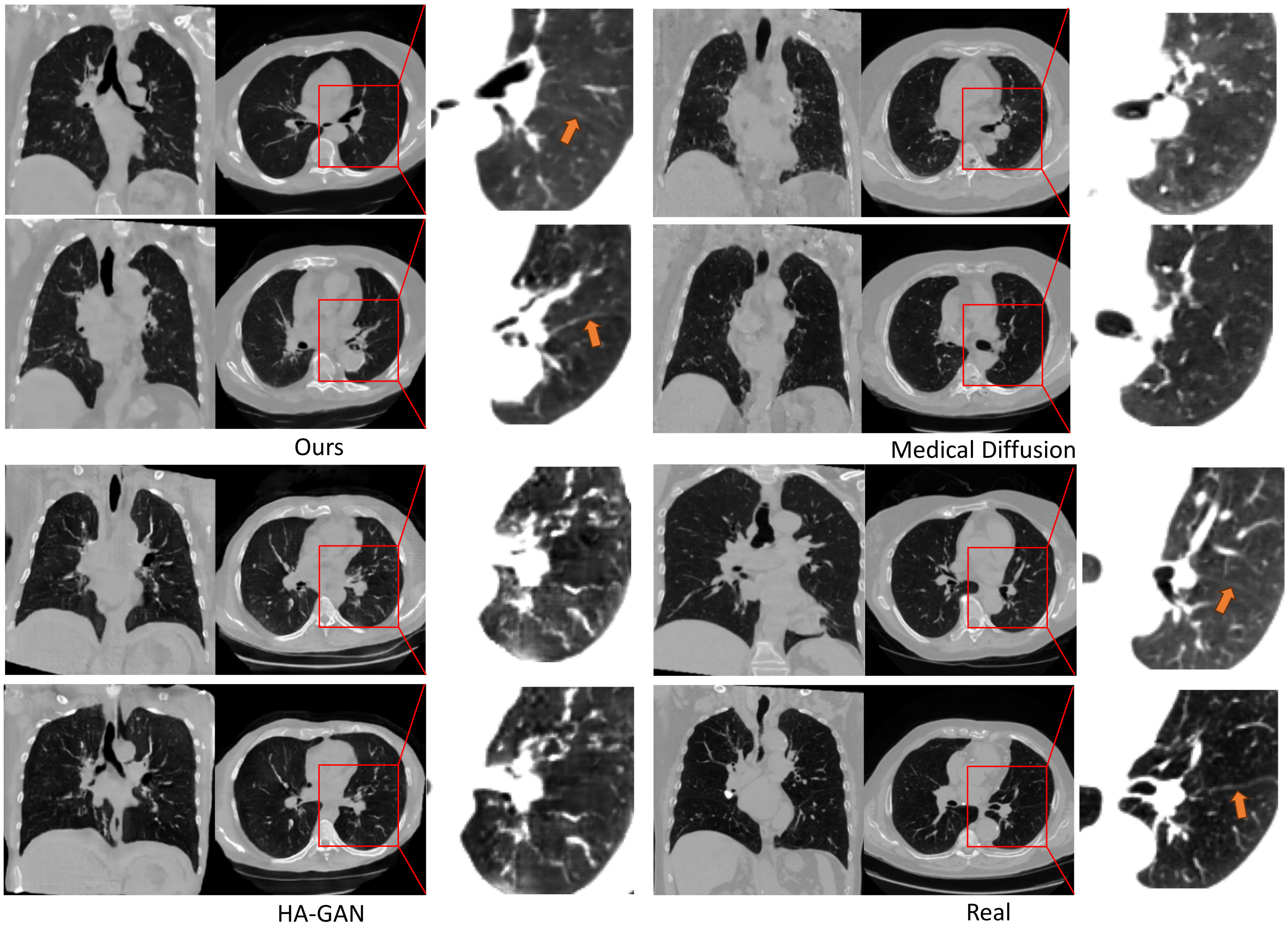

Fig. 3. Randomly generated images (from HA-GAN and Medical Diffusion) and the real images. The first two columns show axial and coronal slices, which use the HU range of [-1024, 600]. The last column shows the zoom-in region and uses HU range of [-1024, -250] to highlight the lung details. Our method is the only one that can preserve delicate anatomical details, including fissures, as indicated by the arrows.

low-resolution base model. The Hounsfield units of the CT images have been calibrated, and air density correction has been applied. The Hounsfield Units (HU) are mapped to the intensity window of $[-1024, 600]$ and normalized to $[-1, 1]$.

B. Evaluation for Synthesis Quality

TABLE I
QUANTITATIVE COMPARISON WITH DIFFERENT METHODS. OUR METHOD OUTPERFORMS BASELINE METHODS IN TERMS OF DISTANCE METRICS, AND PRESERVES AIRWAY BETTER

Method	FID↓	MMD↓	Airway ($\times 10^4 mm^3$)↑
WGAN	0.070	0.094	$1.07_{\pm 0.64}$
α-GAN	0.028	0.057	$1.14_{\pm 0.68}$
HA-GAN	0.023	0.054	$2.04_{\pm 0.73}$
Medical Diffusion	0.013	0.022	$1.77_{\pm 0.93}$
Ours	**0.009**	**0.019**	$\mathbf{3.34_{\pm 1.19}}$
Ours w/o shape	-	-	$1.99_{\pm 1.05}$
Real	-	-	$4.58_{\pm 1.45}$

1) Quantitative Evaluation: If the synthetic images are realistic, then their distribution should be indistinguishable from that of the real images. Therefore, we can quantitatively evaluate the quality of the synthetic images by measuring the distance to the real data, using Fréchet Inception Distance (FID) [35] and Maximum Mean Discrepancy (MMD) [36]. The lower the FID/MMD value is, the more similar the synthetic images are to the real ones. We use a sample size of 1,024 for computing the FID and MMD scores. For our method, we use randomly selected reports as the condition. To compute FID and MMD scores for 3D CT scans, like [6], we leverage a pre-trained 3D ResNet model on medical images [37] for feature extraction. As shown in Table I, our method achieves lowest FID and MMD than the baselines, which implies that our diffusion model generates more realistic images.

Quantitative Evaluation on Anatomical Details: While metrics like FID and MMD are widely used in literature and empirically work well for natural images, they highlight the semantic-wise similarly (distance) but may ignore subtle but important anatomical details in medical images, as implied by the small (FID/MMD) gap between different methods. Their real distances, as later shows in Fig. 3, could be much bigger when taking account into the anatomical details we are focused on. Therefore, we evaluate how well the generated images can preserve the anatomical details. Specifically, we

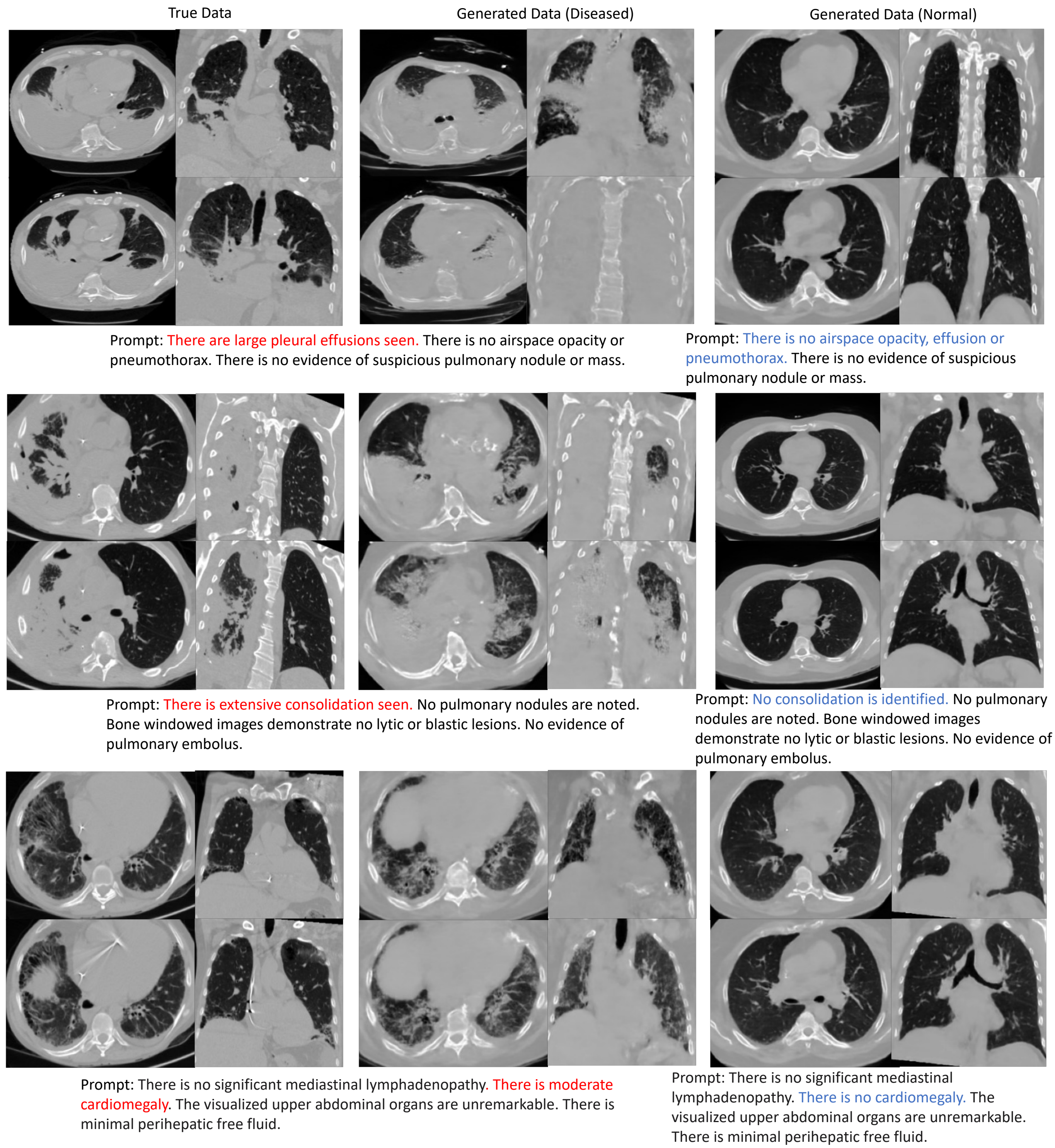

Fig. 4. Images conditionally generated with disease-related prompts. We show the real images in the first two columns. Then we extract disease-related mentions from their associated reports as conditions to generate images, which are shown in the third and fourth columns. We also show the synthesized samples by conditioning on prompts reversed of the disease in the last two columns. Four slices are shown for each subject. The generated images are conditioned on text only.

use Total Segmentor [38] to segment vessels and airways from generated images and real images, and measure the volume. The results are shown in Table. I. We also perform statistical tests (one-tailed two-sample t-test) on the evaluation results. At the significance level of $p < 0.05$, the results are significant for all three conditions, which further identify the effectiveness of our model on prompting generation.

2) Qualitative Evaluation: To qualitatively analyze the results, we show cases of synthetic images from current state-of-the-art GAN [6] and diffusion model [21]. As in Fig. 3, although synthetic images from different methods are all closed to the real ones in overall appearance, only our MedSyn consistently produces anatomically plausible CT scans upon closer inspection, showcasing its superiority.

C. Evaluation for Conditional Generation

TABLE II

EVALUATION FOR CONDITIONAL GENERATION OF PLEURAL EFFUSION. WE MEASURE THE SEGMENTED VOLUME OF PLEURAL EFFUSION FROM GENERATED IMAGES CONDITIONED ON DIFFERENT PROMPTS

Prompt type	Pleural effusion volume (L)
No effusion	$0.00_{\pm.00}$
Large effusion	$1.73_{\pm.22}$

TABLE III

EVALUATION FOR CONDITIONAL GENERATION OF BULLOUS EMPHYSEMA. THE RESULTS SHOW THAT THE BULLAE MENTIONING CAN INCREASE THE EMPHYSEMA VOLUME IN GENERATED VOLUMES

Prompt type	%LAA-950
No bullae	$0.019_{\pm.018}$
With bullae	$1.4_{\pm3.5}$

TABLE IV

EVALUATION FOR CONDITIONAL GENERATION OF CARDIOMEGALY. THE RESULTS SHOW THAT THE CARDIOMEGALY MENTIONING CAN INCREASE THE HEART SIZE IN GENERATED VOLUMES

Prompt type	CTR
No cardiomegaly	$0.48_{\pm.06}$
With cardiomegaly	$0.75_{\pm.24}$

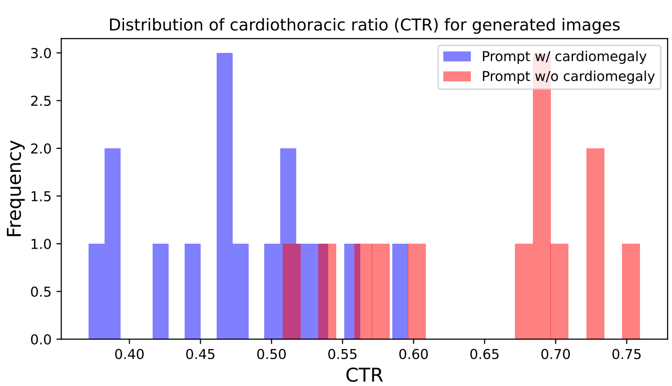

Fig. 5. Distribution of cardiothoracic ratio for images generated conditioning on different prompt types. The results show that when feeding prompt with cardiomegaly mentioning, the generated images will have higher CTR.

In this section, we perform experiments to study the relevance of generated images in response to specific input prompts, specifically on pleural effusion, bullous emphysema, and cardiomegaly. First, we build prompt pairs by selecting prompts from real reports that contain descriptions about certain pathology (e.g. *There are large pleural effusions seen*), and reversing its description (e.g. *There is no pleural effusion*) to build its prompt pair. The prompts used here can be found in Fig. 4. Then, we use our model to generate images conditioning on the original prompts and the modified prompts, respectively. Conditioned on each prompt, we generate 32 CT volumes and perform quantitative analysis to measure the alignment between the synthetic images and the abnormality condition specified in the prompt. For pleural effusion, we use Total Segmentor [38] to segment the effusion from generated images and measure the volume. For bullous emphysema, we measure the %LAA-950 (percentage of low attenuation areas less than a threshold of -950 Hounsfield units) for generated images. For cardiomegaly, we use Total Segmentor [38] to segment the heart and lung region from CT volume, and then we compute the cardiothoracic ratio (CTR) by measurement of the maximal cardiac width divided by the maximal thoracic width at the same axial scan level. The evaluation results for pleural effusion, bullous emphysema and cardiomegaly are shown in Table. VII, III and IV respectively. For pleural effusion, we found that when conditioning on the prompt with “large effusion,” the generated images show a greater volume of pleural effusion compared to images synthesized with a prompt containing “no effusion.” For bullous emphysema, we found that the generated images conditioning on prompt containing “no bullae” have higher %LAA-950 values, which suggests more severe emphysema. For cardiomegaly, we found that when conditioning on the prompt with “There is cardiomegaly,” the generated images have higher CTR, which suggests a greater degree of cardiomegaly. We also provide the distribution of CTR in Fig. 5.

For the qualitative examples generated from our model, we chose those three distinct prompts paired with negative prompts to show the prompting effect on synthetic images. In Fig. 4, we show the volumes from the real and synthetic data with the text description and the negative prompting synthetic data. Our model shows the ability to generate unseen data and control the generative process through prompting.

D. Controllable Synthesis via Conditioning on Anatomical Structures

In this section, we explore the application of conditional generation. In this study, we aim to generate data when the anatomical structures are available, such as we can simulate the structures of lobes, airways or vessels. Conditioned on those priors, we are able to provide volumetric CT scans through our model. In Fig 6, we fix the lobes, vessels or the airway in the input channel, respectively, which are segmented from the real data. Then, we generate the lungs with those anatomical structures, which shows great consistency with those anatomical priors and variance for different seeds.

To be notified, if we marginalize the input channel of the volumes, we can get the segmentation from the rest of the channels, which means our model can perform the segmentation task as well. However, the generative quality is mostly

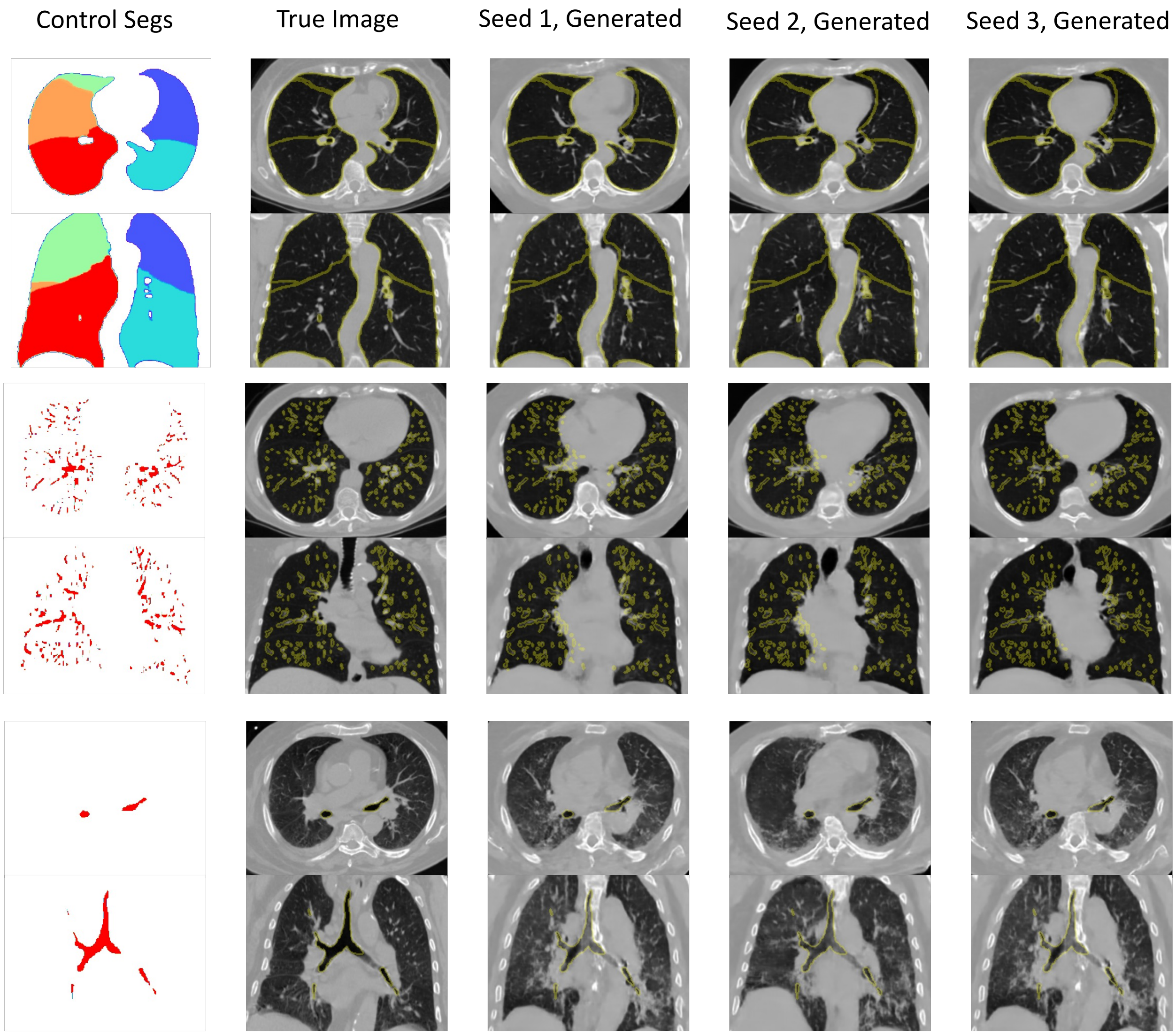

Fig. 6. Controlled volume synthesis via the anatomical priors. The first column shows the anatomical mask used as the condition. The second column shows the corresponding real images. The remaining columns show samples of conditionally generated images. The results show that the generated images can preserve the conditioning anatomical structures.

dominated by the low-res module, thus, the segmentation cannot be compared with the state of the art. To show the potential of future multimodal segmentation via prompt, we compare the segmentation quality of the model with prompt or without the prompt (by setting zero values for the conditional text embedding). In our experiments, we use the embedding extracted from the associated radiology report as the condition. The evaluation results are shown in Table. V. We use Dice score as the evaluation metric for segmentation. We note that the baseline methods can't use segmentation as conditioning, so we don't include the results here. We find that our method achieves decent segmentation performance. With additional information from reports, our method performs even better. In this sense, we show that our generative model has the potential for prompt-guided volumetric segmentation.

TABLE V
EVALUATION FOR SEGMENTATION. WITH ADDITIONAL INFORMATION FROM REPORTS, THE SEGMENTATION IS FURTHER IMPROVED

Dice score↑	Airway	Lobe
No text prompt	$0.70_{\pm.14}$	$0.69_{\pm.12}$
With text prompt	$0.75_{\pm.10}$	$0.77_{\pm.12}$

E. Data Augmentation for Supervised Learning

In this experiment, we used the synthesized samples from MedSyn to augment the training dataset for a supervised classification task. Previous work [39] has shown that synthetic samples improve the diversity of the training dataset, resulting in a better discriminative performance of the classifier. Motivated by their results, we conduct new experiments of using our MedSyn for data augmentation. First, we train a classifier for predicting lung opacity and pleural effusion from CT scans in RADChest dataset [40]. Second, we use the

SARLE labeler [40] to parse the reports in the validation set of our UPMC dataset to derive the labels for lung opacity and pleural effusion. Next, we randomly sampled 100 reports parsed for positive and negative for the two diseases mentioned above. Then, we use the 200 reports as the condition, and feed them into our MedSyn model to generate additional samples. Finally, we use the synthesized samples as extra samples to train the classification models. We perform 5-fold cross-validation. The results are shown in Table. VI. We found that when using augmented samples from our MedSyn, the performance of the classifier can be improved.

TABLE VI

EVALUATION FOR DATA AUGMENTATION. BASELINE MODEL IS TRAINED ONLY WITH REAL RADCHEST DATA. WE ALSO AUGMENT THE TRAINING SET WITH 200 MEDSYN-GENERATED SAMPLES, AND REPORT THE ACCURACY AND F1 SCORE.

Method	Pleural effusion		Lung opacity	
Metric	Accuracy%↑	F1↑	Accuracy%↑	F1↑
Baseline	$90.7_{\pm 3.2}$	$0.79_{\pm .05}$	$61.0_{\pm 2.2}$	$0.72_{\pm .03}$
Augmented w/ MedSyn	$\mathbf{94.0}_{\pm .2}$	$\mathbf{0.84}_{\pm .01}$	$\mathbf{62.0}_{\pm 1.5}$	$\mathbf{0.75}_{\pm .00}$

F. Evaluation by Radiologists

To complement the evaluations of our method, we designed a blind evaluation survey that elicits board-certified radiologists' opinions on the anatomical feasibility of structures generated by our approach in comparison with those generated by existing methods, including Medical Diffusion [34] and HA-GAN [6], and real CT scans. The survey also exposes how accurately radiologists can recognize pathologies in CT scans generated by pathology prompts using our method. To achieve a blind evaluation of the generated CT scans, we intentionally did not mention to the radiologists that some of the CT scans they were about to interpret were generated by AI to avoid potentially biased results due to potentially negative perceptions towards AI.

Overall, our findings from the survey with 10 radiologists with varying years of experience (4 – 23 years) reveal that radiologists can correctly recognize the pathologies defined by the prompt in the CT scans generated by our method with high accuracy. Additionally, our findings reveal that our method generates CT scans with fissure lines and lobe structures that are significantly more anatomically feasible than in CT scans generated by the Medical Diffusion [34] and HA-GAN [6] methods. Our findings also reveal that our methods generate CT scans with airway structures that are significantly indistinguishable from real CT scans and more anatomically feasible than those in CT scans generated by the Medical Diffusion method. We expand upon the experiment design that led to these findings and additional statistical analyses below.

1) Experiment Design: We designed an online survey to elicit radiologists' opinions on how accurately our method represents different diseases and how our method compares to generated CT images from related works and real CT images. To explore this, our survey consisted of pathology recognition and anatomical feasibility evaluation tasks. We designed an interface for both tasks to simulate how radiologists view CT scans by showing them the axial, coronal, and sagittal views. Our interface, built on a medical research image viewer called Papaya [1], is embedded into a Qualtrics survey. We chose to use Papaya due to the limitations of Qualtrics, and professional viewers such as ITKSnap or 3DSlicer cannot be embedded into an anonymized survey platform like Qualtrics. Hence, we have limitations for the contrast window that radiologists use when viewing vessels. Therefore, the evaluation of the vessels has limitations in terms of visualization. Our interface did allow radiologists to adjust the contrast of the images and swap the main image with one of the other two views.

All radiologists were provided an instruction video to show them how to interact with the interface before they started the survey. During the instruction video, all radiologists were told that they would review CT scans "that may belong to different patients and were acquired based on different image acquisition devices and image reconstruction methods". Additionally, one board-certified radiologist and one of the authors were available while the participant took the survey to help address any confusion or technical issues.

We recruited 10 board-certified radiologists with varying years of experience to participate in our study. All radiologists were recruited through our professional network.

a) Pathology Recognition Task: The pathology recognition task in the survey asks radiologists to identify the most prominent finding in the CT scan from a selection of options: consolidation, pleural effusion, cardiomegaly, no abnormalities, and other abnormalities not listed. They could leave a note in an optional text field if they feel multiple prominent or other findings are present. Radiologists are shown six CT scans in total: five generated CT scans by our method and one real CT scan. We generated five CT scans, including one for cardiomegaly, one for consolidation, two for pleural effusion, and one for no abnormalities. The real CT scan shown is randomly selected to present one of those conditions. The six CT scans are shown to the radiologist in a random order.

b) Anatomical Feasibility Task: The anatomical feasibility task specifically asks radiologists to rank the four CT scans based on how well they preserve the given anatomical structure, considering which most looks like it is of a real image where 1 is best preserved and 4 is least preserved. This task evaluates the realism of three categories of anatomical structures: the lobe structure along with the fissure lines, the vessel structures, and the airway structures. For each category, we show the radiologists four different sets of four CT scans to rank. The four sets and the four CT scans for a given set are randomized in the order they appear. Lastly, the anatomical structure to be identified is shown in a randomized order.

2) Analysis: For the pathology recognition task, we calculate the total number of radiologists that correctly identify the pathology used in the prompt to generate that CT scan. We cross-check the radiologists' open-ended responses for correctness with a radiologist within our professional network.

[1] https://github.com/rii-mango/Papaya

For the anatomical feasibility task, we collect four data points from each participant for the rank of each method for the three categories (lobe structure along with lung fissures, vessel structure, and airway structure). With this ranked data, we calculate the frequency of participants that ranked each method for each rank and calculate the mean rank. We use a non-parametric Friedman test to determine if the mean rank for each method is significantly different from that of the other. We conduct a pairwise analysis to determine which mean ranks are significantly different from each other and use Bonferonni Correction to adjust the p-values due to multiple comparisons. We set our significance threshold as $p <= 0.05$. We calculate Kendall's Coefficient of Concordance to determine how consistent the rankings are across the radiologists.

3) Findings: The radiologists who completed the survey included four senior-level radiologists with more than 15 years of experience in radiology and six junior-level radiologists with four to eight years of experience in radiology. One of the radiologists is currently in their residency, and another is a faculty member. The survey took the radiologists 37.47 minutes on average to complete ($std = 12.06$ minutes). Below are additional statistics on each task.

a) Pathology Recognition Task: Figure 7 shows that every radiologist's interpretation was consistent with the pathology of the real CT scan and the prompted pathology in the generated CT scan representing consolidation. Nine of ten radiologists correctly interpreted the pathology for the CT scans generated with a pathology prompt for cardiomegaly, pericardial effusion, and normal conditions. Eight out of 10 radiologists correctly recognized the other pericardial effusion case. One radiologist interpreted this CT scan to present consolidation, while another mentioned another abnormality that was not listed.

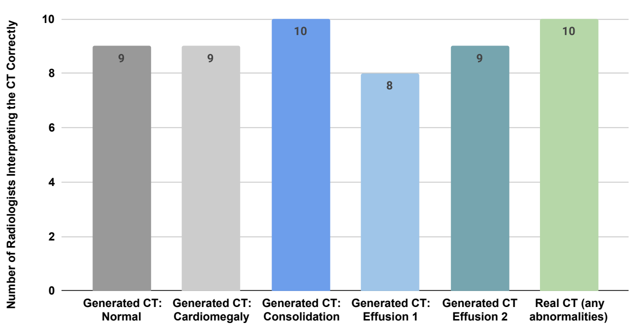

Fig. 7. The number of radiologists that correctly interpreted the CT scans generated by our method for different conditions and real CT scans for different conditions.

b) Anatomical Feasibility Task: We found that the CT scans generated by our method present lobe structures and lung fissures that are significantly more anatomically feasible than those generated by the Medical Diffusion ($p < .05$) and weakly significantly better than CT scans generated by the HA-GAN ($p = .056$). As seen in Figure 8, the real CT scans are significantly better than all three of these methods, as expected. Research images achieve higher quality than clinical images because of limited data and limited exposure patients can have with the imaging machines. The W from Kendall's Coefficient Test is 0.528, indicating moderate agreement among the rankings provided by the radiologists for this structure ($p < .001$), validating the effect of these findings.

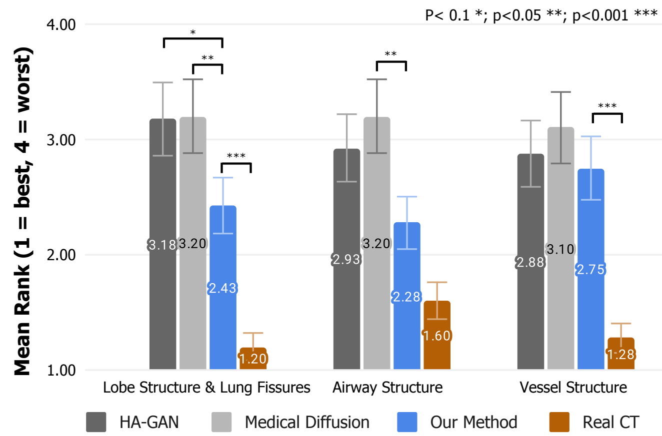

Fig. 8. The mean ranks for each method and each structure. We show the p-values for our method compared to the other methods and the real CT images. P-values for comparisons that do not include our method are not included.

For the airway structures, CT scans generated by our method produce airway structures that are significantly more anatomically feasible than those generated by the Medical Diffusion method ($p < .05$). Interestingly, our method is able to generate the airway structures in CT scans to be significantly indistinguishable from those in real CT scans ($p = 0.1$). The radiologists, on average, ranked the anatomical feasibility of the airway structures in the real CT images as 1.6 and 2.28 for our method (see Figure 8).

For the vessel structure, the real CT scans are significantly better than all three methods ($p = 0.0$), and our method is indistinguishable from the Medical Diffusion and HA-GAN ($p = 1.0$); (Kendall's Coefficient $W = 0.413, p =< .001$). We hypothesize that we do not see any significance for the vessel structure because we train our MedSyn with masks predicted by pre-trained segmentation models, and those segmentation methods are more reliable for the airway (NaviAirway [28]) than the vessel structures (TotalSegmentator [38]). Additionally, the contrast of the CT scan was not pre-determined for the radiologists for this structure, and some chose not to adjust the contrast.

G. Ablation study on the language model

In our MedSyn model, we leverage a language model pre-trained on biomedical text to extract conditioning features for the diffusion model. In this section, we conducted new ablation experiments to demonstrate the effectiveness of our biomedical language model. Specifically, we adopt BERT-Base [25] as pre-trained language model, which is pre-trained on general text corpora, including English Wikipedia and BookCorpus dataset [41]. Next, we fine-tuned the model with our radiology report dataset. Finally, we re-train our diffusion model with the standard language model, and perform the evaluation. The results are shown in Table. VII. It's evident that our choice of using a language model specifically pre-trained on biomedical text improves the sensitivity of the diffusion model to text prompts.

TABLE VII
EVALUATION FOR CONDITIONAL GENERATION OF PLEURAL EFFUSION. OUR BIOMEDICAL LANGUAGE MODEL IS MORE SENSITIVE TO THE EFFUSION MENTIONING IN THE PROMPT

Measured effusion volume (mL)	Generic pre-trained LM	Biomedical pre-trained LM
Prompting w/o effusion	$0.6_{\pm 1.4}$	$0.0_{\pm 0.0}$
Prompting w/ effusion	$2.4_{\pm 5.9}$	$1725.7_{\pm 219.5}$

VI. DISCUSSION

In this study, we achieve high-fidelity, anatomy-aware synthesis of volumetric lung CT scans using guidance from radiology reports. Nonetheless, our model has certain limitations. First, the anatomical structures of the lobe, airway, and vessel are derived from pre-trained segmentation networks. This method may not always align with the ground truth, particularly concerning detailed airway and vessel structures. Second, the radiology reports only provide a condition on the low-resolution images. Therefore, if the text condition mentions subtle changes that require high-resolution, the model will likely be unable to generate that. In other words, our current model is good at generating global and large-scale changes. While our approach demonstrates remarkable adaptability in generating volumetric lung CT scans from radiology reports, evaluating more intricate lung diseases remains challenging due to the complexities presented in the reports. Such evaluations might best be deferred to radiology experts. Therefore, we conduct a blind user study by radiologists in Section. V-F, which verified the quality and fidelity of the generated images. During the inference time, our models support both generation with only one conditioning type, and generation with both text and shape conditioning. If the user thinks there might be a conflict between the text prompt and anatomical shape conditioning, he/she can choose to use one conditioning.

The existing conditional generative models utilized in medical imaging have limited capabilities. They can either accommodate discrete conditions (such as the presence and absence of a disease) or are limited to only 2D images (i.e., X-ray images) when conditioned on text. Utilizing free-style text as a condition can yield substantial enhancements in the diversity of the generated samples. One can control the pathology and the anatomical location, severity, size, and many other aspects of the pathology. This technique could mitigate the longstanding issue of releasing extensive medical imaging datasets. For example, collecting datasets for rare pathology is challenging. Synthetic samples from a well-trained generative model with a radiology report can be viewed as the second-best replacement in such a scenario. While our approach cannot entirely substitute the release of the real data, collaborative efforts within the medical imaging community to refine this model on diverse datasets and share it can significantly mitigate this issue in the foreseeable future.

Our goal is to enhance the memory efficiency of the diffusion model without compromising its fidelity. Memory demand poses a significant bottleneck for generating high-resolution 3D medical images. To address this challenge, we introduce a two-stage stacked diffusion model. Unlike latent diffusion models (LDM), our hierarchical approach conducts denoising operations directly on image pixels, offering finer control over image generation and preserving higher fidelity details. Our hierarchical diffusion model incurs higher computational costs when compared to LDMs. There exists a natural trade-off between fidelity and computational efficiency. However, in our scenario, the advantages of heightened fidelity surpass the disadvantages of slower generation times. Recently, there have been advancements in methods [42], [43] that markedly decrease the necessary denoising steps, sometimes even to just one step. We anticipate that integrating these methods could narrow the disparity between our approach and LDMs, and we leave this as a prospect for our future work.

The proposed model has several fascinating applications that could be pursued in the future. Data augmentation is the most obvious application of the conditional generative models [44], [45]. One can use the generative model as a building block for model explanation, as suggested in [46]–[48]. Generated samples can be used to audit the uncertainty of pre-trained Deep Learning models by conditioning on the pathology, changing various aspects of the anatomy, and assessing the DL model's output distribution. One can deploy such an approach to evaluate and improve the out-of-sample distribution of the DL model for various tasks such as classification and segmentation [49]–[52]. Since our model can condition anatomical segmentation and generate a consistent volumetric image, one can use synthetic data to train a data-free and robust segmentation method similar to [53], [54].

VII. CONCLUSION

Our research takes a significant step forward by synthesizing high-resolution 3D CT lung scans guided by detailed radiological and anatomical information. While GANs and cDPMs have set benchmarks, they come with inherent limitations, particularly when generating intricate chest CT scan details. Our proposed MedSyn model addresses these challenges using a comprehensive dataset and a hierarchical training approach. Innovative architectural designs not only overcome previous constraints but also pioneer anatomy-conscious volumetric generation. Future work can leverage our model to enhance clinical applications.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[5] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, "Roentgen: Vision-language foundation model for chest x-ray generation," *arXiv preprint arXiv:2211.12737*, 2022.

[6] L. Sun, J. Chen, Y. Xu, M. Gong, K. Yu, and K. Batmanghelich, "Hierarchical amortized gan for 3d high resolution medical image synthesis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3966–3975, 2022.

[7] H. Shan, Y. Zhang, Q. Yang, U. Kruger, M. K. Kalra, L. Sun, W. Cong, and G. Wang, "3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1522–1534, 2018.

[8] W. Jin, M. Fatehi, K. Abhishek, M. Mallya, B. Toyota, and G. Hamarneh, "Applying artificial intelligence to glioma imaging: Advances and challenges," *arXiv preprint arXiv:1911.12886*, 2019.

[9] M. D. Cirillo, D. Abramian, and A. Eklund, "Vox2vox: 3d-gan for brain tumour segmentation," *arXiv preprint arXiv:2003.13653*, 2020.

[10] W. Peng, E. Adeli, T. Bosschieter, S. H. Park, Q. Zhao, and K. M. Pohl, "Generating realistic brain mris via a conditional diffusion probabilistic model," 2023.

[11] J. S. Yoon, C. Zhang, H.-I. Suk, J. Guo, and X. Li, "SADM: Sequence-aware diffusion model for longitudinal medical image generation," in *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2023, pp. 388–400.

[12] W. H. Pinaya, M. S. Graham, E. Kerfoot, P.-D. Tudosiu, J. Dafflon, V. Fernandez, P. Sanchez, J. Wolleb, P. F. da Costa, A. Patel *et al.*, "Generative ai for medical imaging: extending the monai framework," *arXiv preprint arXiv:2307.15208*, 2023.

[13] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *arXiv:2204.03458*, 2022.

[14] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, "Roentgen: vision-language foundation model for chest x-ray generation," *arXiv preprint arXiv:2211.12737*, 2022.

[15] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[16] S.-I. Jang, C. Lois, E. Thibault, J. A. Becker, Y. Dong, M. D. Normandin, J. C. Price, K. A. Johnson, G. E. Fakhri, and K. Gong, "Taupetgen: Text-conditional tau pet image synthesis based on latent diffusion models," *arXiv preprint arXiv:2306.11984*, 2023.

[17] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[18] G. Kwon, C. Han, and D.-s. Kim, "Generation of 3d brain mri using auto-encoding generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 118–126.

[19] S. Xing, H. Sinha, and S. J. Hwang, "Cycle consistent embedding of 3d brains with auto-encoding generative adversarial networks," in *Medical Imaging with Deep Learning*, 2021.

[20] K. Han, Y. Xiong, C. You, P. Khosravi, S. Sun, X. Yan, J. Duncan, and X. Xie, "Medgen3d: A deep generative framework for paired 3d image and mask generation," 2023.

[21] F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han *et al.*, "Denoising diffusion probabilistic models for 3d medical image generation," *Scientific Reports*, vol. 13, no. 7303, 2023.

[22] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[23] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*. Springer, 2022, pp. 117–126.

[24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020.

[27] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. Boll, J. Cyriac, S. Yang *et al.*, "Totalsegmentator: robust segmentation of 104 anatomical structures in ct images," *arXiv preprint arXiv:2208.05868*, 2022.

[28] A. Wang, T. C. C. Tam, H. M. Poon, K.-C. Yu, and W.-N. Lee, "Naviairway: a bronchiole-sensitive deep learning-based airway segmentation pipeline," *arXiv preprint arXiv:2203.04294*, 2022.

[29] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[30] S. Lin and X. Yang, "Diffusion model with perceptual loss," *arXiv preprint arXiv:2401.00110*, 2023.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[32] S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle, "Differentially private diffusion models generate useful synthetic images," *arXiv preprint arXiv:2302.13861*, 2023.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[34] F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, "Medical diffusion - denoising diffusion probabilistic models for 3d medical image generation," 2022. [Online]. Available: https://arxiv.org/abs/2211.03364

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017.

[36] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[37] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[38] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[39] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[40] R. L. Draelos, D. Dov, M. A. Mazurowski, J. Y. Lo, R. Henao, G. D. Rubin, and L. Carin, "Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes," *Medical image analysis*, vol. 67, p. 101857, 2021.

[41] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[42] X. Liu, X. Zhang, J. Ma, J. Peng, and Q. Liu, "Instaflow: One step is enough for high-quality diffusion-based text-to-image generation," *arXiv preprint arXiv:2309.06380*, 2023.

[43] Y. Xu, M. Gong, S. Xie, W. Wei, M. Grundmann, T. Hou *et al.*, "Semi-implicit denoising diffusion models (siddms)," *arXiv preprint arXiv:2306.12511*, 2023.

[44] X. Chen, Y. Li, L. Yao, E. Adeli, and Y. Zhang, "Generative adversarial u-net for domain-free medical image augmentation," *arXiv preprint arXiv:2101.04793*, 2021.

[45] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*. Springer, 2018, pp. 1–11.

[46] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, "Explaining the black-box smoothly—a counterfactual approach," *Medical Image Analysis*, vol. 84, p. 102721, 2023.

[47] H. Montenegro, W. Silva, and J. S. Cardoso, "Privacy-preserving generative adversarial network for case-based explainability in medical image analysis," *IEEE Access*, vol. 9, pp. 148 037–148 047, 2021.

[48] C. Mauri, S. Cerri, O. Puonti, M. Mühlau, and K. Van Leemput, "Accurate and explainable image-based prediction using a lightweight generative model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 448–458.

[49] D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes, "Interpretability-driven sample selection using self supervised learning for disease classification and segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2548–2562, 2021.

[50] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8300–8311.

[51] W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao, "The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 623–631.

[52] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, and Y. Zheng, "Mi 2 gan: Generative adversarial network for medical image domain adaptation using mutual information constraint," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 516–525.

[53] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: domain randomisation for segmentation of brain scans of any contrast and resolution," *arXiv preprint arXiv:2107.09559*, 2021.

[54] B. Billot, D. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca, "A learning strategy for contrast-agnostic mri segmentation," *arXiv preprint arXiv:2003.01995*, 2020.