

MedSynV2: Flexible Control of 3D CT Generation via Text and Semantically-Defined Segmentation Prompts

Weicheng Dai¹[0009–0009–2811–2198], Chenyu Wang¹[0009–0008–4468–6434],
Shantanu Ghosh¹[0000–0003–4085–541X], and Kayhan
Batmanghelich¹[0000–0001–9893–9136] ^{*}

Boston University School of Engineering, Boston, MA, USA
{wd2119, chyuwang, shawn24, batman}@bu.edu

Abstract. Generative models for volumetric medical images have found many applications in medical imaging, ranging from data augmentation to serving as priors for inverse problems. For these applications, generating high-resolution 3D images with strong controllability is essential but remains highly challenging. Existing approaches typically control generation either through radiology reports used as text prompts or through full image segmentation. While text-based prompting is flexible, it provides limited spatial control over the location, shape, and boundary of abnormalities. In contrast, segmentation-based methods receive precise spatial guidance but are restrictive in requiring full-organ annotations. In this work, we propose MedSynV2, a flexible multimodal framework for controllable volumetric image generation that supports input from radiology reports and segmentation prompts (both optional). Our approach allows users to provide segmentation of a specific anatomy or abnormality without requiring full-organ annotations. The semantic meaning of the segmentation mask is specified through an accompanying text description, resulting in a highly flexible and scalable conditioning mechanism. We develop a memory-efficient architecture based on a modified diffusion transformer that jointly processes image and segmentation tokens. The model further incorporates gated attention to effectively attend to long radiology reports. Experiments demonstrate that our method achieves state-of-the-art perceptual and semantic scores (*e.g.*, $\sim 24\%$ relative improvement in mean FID), generates high-resolution anatomically consistent CT volumes, and improves data efficiency when used for data augmentation. Radiologists’ evaluation further confirms strong alignment between generated and real medical images. We also explore the ability of the model to generalize towards concepts beyond training data. *Code will be released at* <https://github.com/batmanlab/MedSynV2>.

1 Introduction

Generative models play an important role in medical imaging. They enable privacy-preserving data augmentation [12], serve as building blocks for model

^{*} Corresponding author.

explanation [17], provide strong priors for solving inverse problems [6], and support the training of radiologist residents [26]. For these applications, controlling the generation process is essential, as the semantic properties of the synthetic image must be specified explicitly. Most existing approaches control generation using radiology reports as textual prompts [1, 9, 12, 21, 25, 37, 42]. However, radiology reports provide limited control over spatial attributes, *e.g.* they cannot precisely specify the shape of a nodule or the precise size of opacity. Such problems can be mitigated by segmentation masks, which clearly define boundaries of pathologies and anatomies. Thus other methods use masks for synthesis [7, 10, 13, 15, 16, 28, 36, 44, 46], yet they require full annotations on all organs which is expensive. This paper proposes a flexible and scalable multi-modal framework that controls high-resolution volumetric CT synthesis using either radiology reports or segmentation masks, or a combination of both, where the user specifies semantic meaning of the segmentation.

Currently, there are two main ways to control volumetric radiology image generation: using radiology reports as prompts [1, 9, 12, 21, 25, 37, 42] or using segmentation masks as spatial inputs [7, 10, 13, 15, 16, 28, 36, 44, 46]. Radiology reports contain rich clinical information describing observed abnormalities. However, these descriptions are always written in approximate terms. Text itself cannot precisely specify the location, shape, or boundary of an abnormality, although such details may carry important clinical value. Segmentation masks, on the other hand, alleviate the burden by defining clear spatial descriptions, helpful in defining sizes of abnormalities. Methods that condition generation on segmentation masks address this limitation by providing explicit spatial information, where anatomical structures and lesion boundaries are defined precisely. However, current state-of-the-art approaches typically require full segmentation covering all organs and abnormalities in the volume (*e.g.*, 117 classes in [44]). As a result, they are often limited to a small set of abnormality labels defined during training. Moreover, these approaches impose a heavy burden at inference time, as full-organ segmentation is required before generation. They also reduce the diversity of the generated images, since all organs and abnormalities must be predefined. Importantly, segmentation and radiology reports provide complementary information and should be ideally used together. Our goal therefore is to develop a method that combines the strengths of both modalities, enabling precise control while remaining flexible and scalable.

A desirable method should be *flexible*, allowing users to provide segmentation of one specific abnormality or anatomy without requiring full-organ annotations. In addition, the semantic meaning of the segmentation prompt should be defined explicitly and flexibly, since a mask may represent different concepts such as a lung nodule, heart boundaries, or mask of pleural effusion, etc. This design increases the diversity of the generated images, as the model retains more degrees of freedom compared to approaches that rely on full segmentation where all boundaries are predefined. Our architecture builds on a modified diffusion transformer (DiT) [29]. We tokenize both the image and the segmentation mask, when available, with a fine-tuned video tokenization method. This design is memory

efficient for high-resolution CT generation, as image and segmentation tokens are processed jointly without increasing the number of tokens, incurring negligible additional memory cost. To better handle long radiology reports, we incorporate gated attention [32] to ensure the model attends to all text tokens. The semantic meaning of the segmentation prompt is provided as optional text appended to the radiology report, enabling *scalable* conditioning where new abnormalities can be introduced through fine-tuning or even in a zero-shot setting.

This work makes three **contributions**. (1) We develop MedSynV2, a framework where users can optionally provide a radiology report or a segmentation of an abnormality, or both to control image generation, resulting in a highly versatile generative model. (2) We propose an architecture that attends to long text prompts and segmentation while generating high-resolution radiology images on modest GPU resources. (3) Our method achieves state-of-the-art perceptual and semantic scores. We further show that the generated images exhibit high anatomical fidelity, improve data efficiency when used for data augmentation, and align well with real medical images according to evaluation by two radiologists.

2 Related Work

Text-to-Image Lung CT Generation. Generating high-resolution 3D medical images is computationally expensive due to the large spatial dimensions of volumetric data. CTFlow [37] mitigates this challenge by leveraging 2D backbones to generate slices and auto-regressively assemble them into volumes. While this design reduces training memory requirements, it often leads to long inference times and anatomical inconsistencies across orthogonal planes. GenerateCT [12] and MedSyn [42] address memory limitations using multi-resolution generation strategies. However, these approaches synthesize images directly in the voxel domain, which becomes computationally expensive when extending to high-resolution volumetric images. More recent methods, including Text2CT [25], T2CT [9], and Report2CT [1], adopt latent diffusion models (LDM) [33], which are more computationally scalable. Despite this improvement, these models rely solely on textual prompts and therefore inherit the limitations of text-only control, such as limited ability to specify the precise size, shape, or spatial extent of anatomical structures and abnormalities. Such drawbacks can be mitigated via segmentation masks as they absorb more abundant spatial information.

Segmentation-to-Image Lung CT Generation. Using semantic segmentation masks as prompts provides an alternative approach for controlling image generation. SemGen [46], mDDPM [16], and MedGen3D [13] generate CT images conditioned on segmentation masks using 2D backbones. However, similar to text-based methods built on 2D architectures, these approaches inherit limitations when generating consistent volumetric structures. LungDDPM [15] and LAND [28] instead adopt 3D backbones, but they focus on inpainting local abnormalities such as lung nodules in existing scans rather than synthesizing full CT volumes. Med-DDPM [7] generates full volumes conditioned on paired images and semantic masks in the form of image-to-image style transfer. However,

requiring a full volumetric mask as input limits the flexibility and applicability of the method for downstream tasks, since the model cannot generate complete images without training with corresponding masks. MAISIs [10, 44] and 3D Med-Diffusion [36] attempt to alleviate this limitation by first training unconditional volumetric generators and then fine-tuning them using ControlNet [43] to accept segmentation masks as conditions. Nevertheless, these methods still require full segmentation masks covering all organs in the field of view. It reduces the diversity of the generated images and limits the model accepting segmentation labels only defined during fine-tuning, requiring retraining when new mask types are introduced. Joint mask and image generation approaches (SimGen [2], CoSimGen [4]) fail to accept single mask condition, still pre-define full segmentation masks, therefore fail to address the flexible control problem. Importantly, none of these segmentation-based approaches incorporates radiology reports as prompts, even though textual reports provide complementary clinical information beyond spatial annotations. Recent joint-modality approaches (DGM [41], Trace [35]) attempt to combine text and segmentation prompts, but similar to MAISIs they require full segmentation masks defining all organs, which limits flexibility in practical clinical scenarios.

3 Method

Overview. Our goal is to develop a controllable framework for 3D CT synthesis that uses both long-form radiology reports and segmentation masks as prompts. The model should have the following capabilities: (1) A user should be able to control the image generation with either a radiology report, or a segmentation mask, or both. (2) The user can *define* the semantic meaning of a segmentation mask with extra text appended to the radiology prompt, where both prompts are optional. (3) The model jointly attends to spatial and semantic conditions while preserving *high-fidelity anatomical details* in *high-resolution* volumes. We address these requirements via a rectified flow formulation [20] operating on latent tokens, where a diffusion transformer (DiT) [29] predicts the final clean latent state (i.e., $\mathbf{x}_0$ prediction) from noisy inputs. This setting introduces several challenges: (a) generating volumetric images is highly memory demanding, (b) radiology reports are long and standard architectures often focus on early tokens [32], and (c) segmentation masks do not have fixed semantics and the user should be able to define the meaning of the mask prompt. To address these issues, we introduce a compound prompting strategy that explicitly encodes mask semantics as a textual description together with the radiology report. We further address (a) through a unified tokenization scheme adapted from video tokenization (OSP [19]), which treats images and segmentations in the same latent space. We further address (b) by adopting gated attention [32] to improve reasoning over long and complex text prompts (Ablation in Supp. 7.5). Finally, for (c) we modify the DiT architecture to jointly reason over segmentation and report conditioning while improving spatial consistency and anatomical fidelity in generated volumes. Fig. 1 shows the schematic of our model architecture.

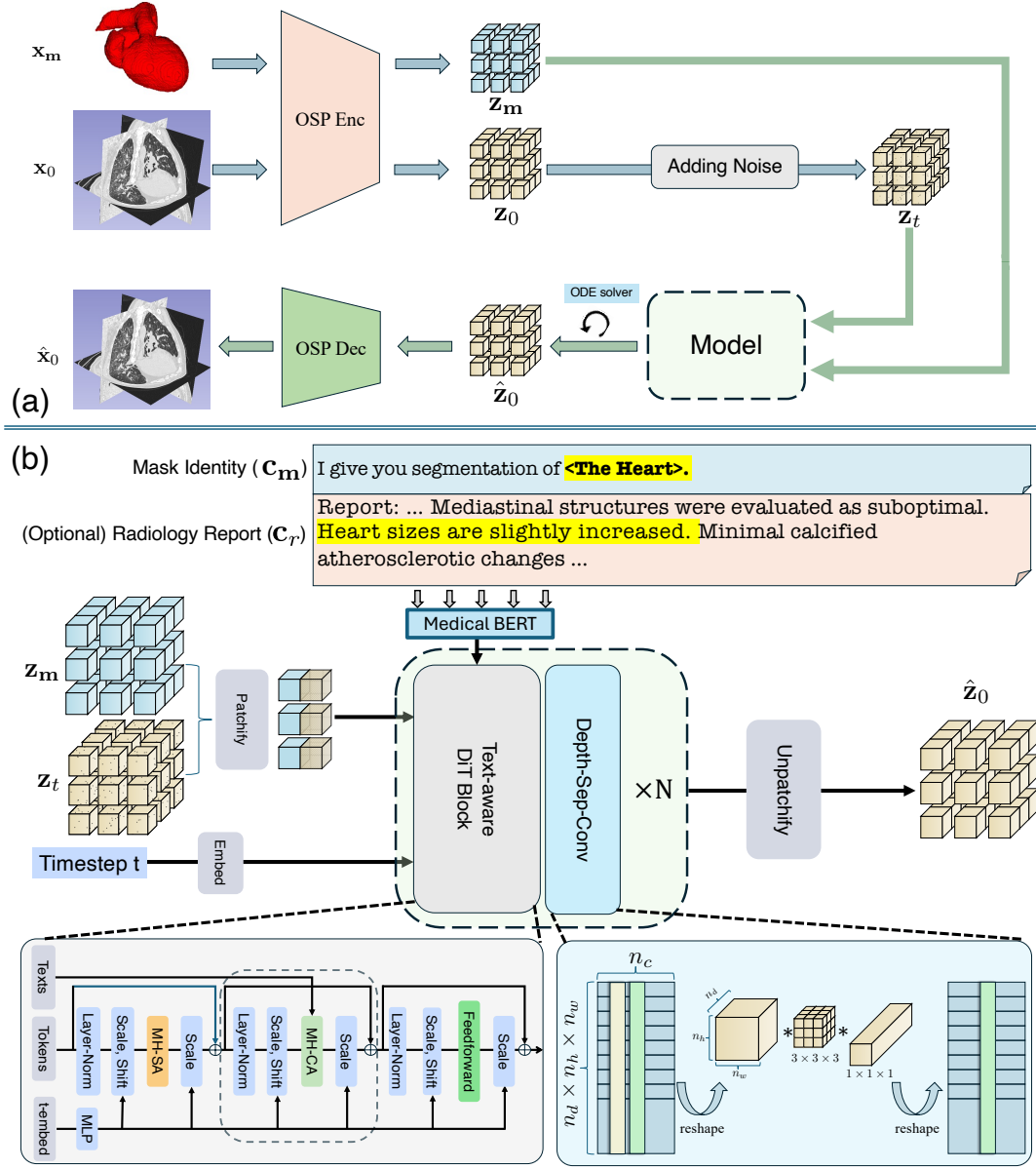


Fig. 1: Overview of MedSynV2. (a.) Our Encoder-Decoder utilizes OSP trained on CT-Rate images, while we inference on both x_0 and x_m . (b.) We inject compound text embeddings (c_m , optional c_r) into text-aware DiT Block with multi-head cross-attention. The paired segmentation latent z_m and noisy image latent z_t are patchified together in early stages to learn a clean image latent z_0 . A light-weighted Depth separable convolution layer (Depth-Sep-Conv) is appended after each Block for maintaining inter-token feature consistency.

3.1 Data Preparation and Image Tokenization

Our framework follows a latent diffusion formulation and therefore requires a tokenization strategy that preserves anatomical fidelity while efficiently handling volumetric data. Let $\text{Enc}(\cdot)$ denote the encoder used to tokenize volumetric inputs into latent representations. We denote a CT image volume by $\mathbf{x}$ and its corresponding segmentation mask by $\mathbf{m}$. Prior to tokenization, all CT volumes are affinely registered to a common template such that the initial and terminal slices approximately correspond to the same anatomical regions. This alignment reduces inter-subject spatial variance and stabilizes token learning. Given the aligned volumes, we interpret each 3D CT scan as videos alongside axial axis [9], and adopt a video tokenization paradigm. Among several alternatives, we fine-tune the Open-Sora [30, 45] tokenizer on CT images to obtain compact yet high-fidelity latent representations (reconstruction analysis in Supp. 7.3).

The image tokens are defined as $\mathbf{z}_0 = \text{Enc}(\mathbf{x}) \in \mathbb{R}^{(n_d \times n_h \times n_w) \times n_c}$, where n_d , n_h , and n_w denote the number of tokens along depth, height, and width, respectively, and n_c is the latent channel dimension. Following prior work [38], the tokenizer is fine-tuned using images only and is not trained on segmentation masks. During inference, however, the same encoder is applied to segmentation masks to obtain $\mathbf{z}_m = \text{Enc}(\mathbf{m}) \in \mathbb{R}^{(n_d \times n_h \times n_w) \times n_c}$. This unified tokenization strategy compresses both images and segmentations within a shared latent space, allowing the diffusion model to process spatial and anatomical conditions in a consistent manner.

3.2 Proposed Architecture

Patchification and Compound Prompting. Given a noisy version of the tokenized image $\mathbf{z}_t$ at time t and a clean segmentation representation $\mathbf{z}_m$, we perform early fusion by concatenating them along the channel dimension. The fused latent representation is converted into transformer tokens through the patchification operator in Fig. 1 illustrating:

$$\mathbf{h}_t = \text{Patchify}([\mathbf{z}_t; \mathbf{z}_m]), \quad (1)$$

where $[\cdot; \cdot]$ denotes channel concatenation and $\mathbf{h}_t$ denotes the sequence of latent tokens processed by the DiT throughout the remainder of the paper.

Segmentation masks may correspond to different anatomical or pathological concepts. To explicitly specify mask semantics, we introduce a compound text prompt noted $\mathbf{c}_{\text{text}}$. A semantic pre-prompt describing the mask identity $\mathbf{c}_m$ is concatenated before the optional radiology report $\mathbf{c}_r$. When segmentation is absent, the pre-prompt is replaced with a sentence indicating an empty condition (“no image segmentation is provided”). This design enables optional conditioning while preserving consistent text structure.

Modified DiT Block. Our architecture builds upon the diffusion transformer (DiT) [29] but introduces several modifications to satisfy controllability and

anatomical fidelity requirements. First, to incorporate token-level textual conditioning, we introduce a multi-head cross-attention (MHCA) module to DiT that attends image tokens to compound text tokens. Adaptive layer normalization (AdaLN) is applied before and after MHCA to modulate features using timestep embeddings. Second, to stabilize learning and improve anatomical fidelity, we introduce a depthwise separable convolution after each transformer block, enforcing local spatial consistency between neighboring tokens. Third, to improve reasoning over long radiology reports and compound text, gated attention [32] is incorporated in every layer. Ablation studies in Supp. [7.5] further prove the efficacy of each component in the aforementioned modifications.

Illustrated in Fig. [1] let $\mathbf{h}^j$ denote tokens at layer j and $\mathbf{c}_{\text{text}}$ denote compound text tokens, our modified DiT block is summarized as:

$$\begin{aligned}
\mathbf{Q}_h &= \mathbf{h}^j \mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{c}_{\text{text}} \mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{c}_{\text{text}} \mathbf{W}_h^V, \quad \mathbf{G}_h = \mathbf{h}^j \mathbf{W}_h^G, \\
\tilde{\mathbf{A}}_h &= \sigma(\mathbf{G}_h) \odot \text{Softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_h}}\right) \mathbf{V}_h, \\
\mathbf{u}^j &= \text{AdaLN}(\text{MHCA}(\mathbf{h}^j, \mathbf{c}_{\text{text}})) + \mathbf{h}^j, \\
\mathbf{v}^j &= \text{AdaLN}(\text{FFN}(\mathbf{u}^j)) + \mathbf{u}^j, \\
\mathbf{h}^{j+1} &= \text{Conv}(\mathbf{h}^j) + \mathbf{v}^j,
\end{aligned} \tag{2}$$

where the $\text{Conv}(\cdot)$ in the last line involves a spatial and channel-wise convolution to enhance inter-patch consistency as shown in Fig. [1] (ablation in Supp. [7.5]). For intra-patch learning, we adopt gated attention to multi-head self-attention (MHSA) block as well. After all blocks are finished, We perform the unpatchify operation *i.e.*, $\hat{\mathbf{z}}_0 = \text{UnPatchify}(\mathbf{h}^J)$, on which we apply our loss function.

3.3 Training and Inference Procedure

We train the model using conditional rectified flow [20][22] in the latent token space following the x -prediction formulation of [18]. Let $\mathbf{z}_0$ denote clean latent tokens obtained from the tokenizer and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denote Gaussian noise. A noisy sample is constructed using linear interpolation:

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\epsilon, \quad t \in [0, 1]. \tag{3}$$

Our network directly predicts the clean latent tokens $\mathbf{z}_\theta = f_\theta(\mathbf{z}_t, t, \mathbf{c})$, where $\mathbf{c}$ includes $\mathbf{c}_m, (\mathbf{c}_r), \mathbf{z}_m$. Thus velocity prediction and conditional flow matching objective are computed as:

$$v_\theta(\mathbf{z}_t, t, \mathbf{c}) := \frac{\mathbf{z}_\theta - \mathbf{z}_t}{0 - t}; \quad \mathcal{L} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon} \left[\left\| v_\theta(\mathbf{z}_t, t, \mathbf{c}) - \frac{\mathbf{z}_0 - \mathbf{z}_t}{0 - t} \right\|_2^2 \right], \tag{4}$$

which corresponds to x -prediction with v -loss as described in [18]. Directly predicting clean latent tokens encourages learning on the data manifold and improves anatomical fidelity.

Conditioning Randomization. During training, conditioning modalities are randomized within each batch to encourage robustness. Samples may include (i) radiology report only (i.e., $\mathbf{c}_m = \emptyset$), (ii) segmentation mask with semantic description only (i.e., $\mathbf{c}_r = \emptyset$), (iii) both prompts, or (iv) neither prompt. This strategy enables the model to operate under arbitrary prompt availability during inference. We set segmentation tokens to zeros ($\mathbf{z}_m = \mathbf{0}$) when segmentation masks are absent.

Inference. Given the learned velocity field v_θ , generation is performed by solving the ordinary differential equation (ODE):

$$\frac{d\mathbf{z}_t}{dt} = v_\theta(\mathbf{z}_t, t, \mathbf{c}), \quad (5)$$

starting from $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and integrating to $t = 0$. We use a Dormand–Prince solver with a fixed 50 steps to solve the ODE. After the last step of denoising, we pass the denoised $\mathbf{z}_0$ to our decoder to obtain the image, *i.e.*, $\hat{\mathbf{x}} = \text{Dec}(\hat{\mathbf{z}}_0)$.

4 Experiments

We conduct experiments to show **flexibility**, **controllability**, **scalability** and diversity of our model. We compare our method with recent open-source SOTA methods with publicly available weights, including GenerateCT [12], MedSyn [42], and Text2CT [25] (top candidate in VLM3D Challenge [11]). We compare results in text-only, mask-only, and joint-modality settings to show flexibility and controllability, different random seeds to show diversity, and two downstream tasks to show scalability. Moreover, we provide radiologists’ assessment to present close alignment of our results with real images. (Ablation study in Supp. 7.4)

4.1 Dataset

We perform our training and validation on the CT-Rate dataset [11], a large public chest CT dataset consisting of 25,692 pairs of CT images and radiology reports, with the corresponding 18 abnormality labels. We follow their official train-validation setting. We follow MedSyn [42] and affine-register all CT images, with a fixed pixel spacing $0.7 \times 0.7 \times 0.7$ followed by padding or cropping to a fixed size $448 \times 448 \times 448$. Hounsfield Unit (HU) intensities are clipped to $[-1024, 1024]$ and normalized to $[-1, 1]$ for stabilized training [25]. CT-Rate does not contain corresponding segmentation masks, thus we leverage multiple state-of-the-art models for anatomical and pathological masks. We include four anatomy and five abnormality masks: lung lobes (lungmask [14]), airways (TotalSegmentator [39]), vessels ([31]), heart (TotalSegmentator [39]); consolidation (MedPSeg [5]), ground glass opacity (MedPSeg [5]), pericardial effusion ([40]), pleural effusion ([34]) and lung nodules (TotalSegmentator [39]).

Table 1: FID↓ on different CT windows with reports only. We use different intensity windows used by radiologists to increase contrast for different organs: ([Level, Width]= Lung: [−600, 1500], Vessel: [100, 700], Soft Tissue: [50, 350], Bone: [300, 2000]) when comparing FID scores. (best **bold**, second underlined) Our method performs best over all windows on almost all axes. (XY: Axial, YZ: Sagittal, ZX: Coronal)

Model	Lung				Vessel				Soft Tissue				Bone			
	XY↓	YZ↓	ZX↓	Mean↓	XY↓	YZ↓	ZX↓	Mean↓	XY↓	YZ↓	ZX↓	Mean↓	XY↓	YZ↓	ZX↓	Mean↓
GenerateCT [12]	7.03	<u>14.07</u>	21.71	14.27	4.78	11.77	14.57	10.37	6.49	13.89	17.25	12.54	4.57	9.20	12.35	8.71
MedSyn [42]	<u>9.37</u>	14.27	<u>14.84</u>	12.83	13.59	16.33	17.24	15.72	13.96	18.59	20.00	17.52	<u>5.86</u>	8.23	8.22	7.44
Text2CT [25]	11.51	<u>8.55</u>	17.69	<u>12.58</u>	8.91	<u>9.76</u>	<u>9.37</u>	<u>9.35</u>	9.12	<u>13.18</u>	<u>13.58</u>	<u>11.96</u>	8.67	<u>7.04</u>	<u>6.37</u>	<u>7.36</u>
MedSynV2(Ours)	10.96	6.40	11.40	9.59	<u>7.97</u>	5.17	6.24	6.46	<u>8.97</u>	7.45	8.99	8.47	7.13	3.43	3.19	4.58

4.2 Experimental details

Textual Conditions. We encode the compound pre-prompt $\mathbf{c}_m$ and the radiology reports $\mathbf{c}_r$ separately to reduce storage space. Both the radiology reports (including both Findings and Impressions) and the compound pre-prompt are processed with a pretrained CXR-BERT [3] model from MedSyn [42]. For training data variance, we augment $\mathbf{c}_m$ using different formats shown in Supp. [7.2]. We set the $\mathbf{c}_m$ max token length as 32, describing identity of abnormalities. For the optional reports, we concatenate the findings and impressions with a <SEP> token in between, and set token length 512. Thus, we acquire textual condition shaped (544, 768) concatenating them. (More details in Supp. [7.1])

Mask Conditions and Tokens. We normalize all masks to $[-0.5, 0.5]$ for matching our image intensity and avoiding outliers. For all images and segmentation masks shaped $448 \times 448 \times 448$, our encoder processes them into latents $\mathbf{z}$ of a unified shape $(n_d \times n_h \times n_w) \times n_c = (112, 56, 56) \times 4$. We z-norm all latents with their individual channels for training with the loss function in Eq. [4] and reverse the z-norm when applying our decoder. (More details in Supp. [7.1])

4.3 Quantitative Results

Modified DiT. We choose patch size $(2 \times 2 \times 2)$ to balance between memory issues and image quality, yielding 43,904 patches (latent tokens). We use 16 layers and feature dimension 1152 [29]. All the experiments are run on a single A100 GPU (PyTorch 2.6.0+cu126) with AdamW optimizer, learning rate 10^{-4} and batch size 1 with 4 accumulated forward steps. (More details in Supp. [7.1])

Baselines and Metrics. We compare on various metrics against current state-of-the-art text-to-image generative models, namely GenerateCT [12], MedSyn [42], Text2CT [25]. For perceptual evaluation, we follow radiology practice and define contrast intensity windows for four tissue types: lung, vessel, soft tissue, and bone (See Supp. [7.1]). We compute FID for each window across the three orthogonal planes. For semantic perspectives, we leverage two frequently used CLIP models: CT-CLIP [11] and Pillar-0 [27] to compute CLIP-Score on both text-to-image (T2I) and image-to-image (I2I) scopes. Note Pillar-0 applies eleven windows

Table 2: CLIP-Score (%) using reports only with CT-CLIP (1 window) and Pillar-0 (11 windows). Our result achieves good results on these four metrics with both CLIP models.

Source	CT-CLIP (1 window)		Pillar-0 (11 windows)	
	T2I ↑	I2I ↑	T2I ↑	I2I ↑
Real	16.46	–	26.36	–
GenerateCT [12]	14.70	59.03	22.10	32.11
MedSyn [42]	12.70	81.61	22.75	44.59
Text2CT [25]	20.17	72.37	25.80	49.38
MedSynV2(Ours)	14.74	83.00	27.61	46.79

Table 4: Expert evaluation of images matching report and mask.

	Match	Partial Match	Not Match
Radiologist 1 (4 years experience)			
Real	49	10	16
Synthetic	40	18	17
Radiologist 2 (17 years experience)			
Real	45	25	5
Synthetic	36	30	9

Table 3: Classification results (%) using reports only. We use zero-shot CT-CLIP as classifier, showing our method performs on par with existing method.

	Accuracy ↑	Precision ↑	AUROC ↑	F1 ↑
Real	67.81	32.00	67.34	58.95
GenerateCT [12]	28.24	19.77	50.57	26.33
MedSyn [42]	79.65	35.65	51.32	40.12
Text2CT [25]	45.06	23.76	50.96	37.43
MedSynV2(Ours)	63.55	24.90	52.24	48.40

Table 5: Results of incorporating synthesized data for Opacity segmentation on SemiSeg test set. We report results with the real images, our synthesized images, and their combination. (**best, second best**)

Training Data	Dice(%) ↑	Jaccard(%) ↑	ASD (pixel) ↓	HD95 (pixel) ↓
Real (50 images)	58.78	43.49	21.65	65.33
Synthesized (416 images)	48.82	32.30	56.32	10.61
Combined (466 images)	68.13	54.27	43.32	4.69
Inf-Net [8]	62.41	45.35	32.67	20.63
MedPSeg [5]	65.20	48.13	28.45	5.51

in their image encoder. For factual correctness, we leverage the zero-shot CT-CLIP [11] to predict the 18 pathological labels based on the generated images. We report accuracy, precision, AUROC and F1 score based on the ground truth.

5 Results

5.1 Quantitative Results

Comparing Baselines. Table 1 demonstrates that our method outperforms all baselines in the mean FID scores among all contrast windows. We achieve **24%** improvement compared to TextCT on the lung window (ours **9.59** *vs.* TextCT **12.58**). The good performances result from our gated attention, which captures most details in long CT reports. GenerateCT performs good on axial (XY) view, due to its design of slice-wise super resolution on axial views. It however loses continuity and suffer on the other two axes. MedSyn suffers from soft tissue generation due to its training on another dataset, while Text2CT uses unregistered images for training, yielding to inferior results. Likewise, tables 2, 3 also substantiate that our method performs well on CLIP T2I, I2I, factual correctness AUROC and F1 scores (table 2 Pillar-0 T2I ours **27.6%** *vs.* Text2CT **25.8%**). This results from our modified DiT-based method, as it enhances report understanding by token level cross attention, with inter-patch consistency.

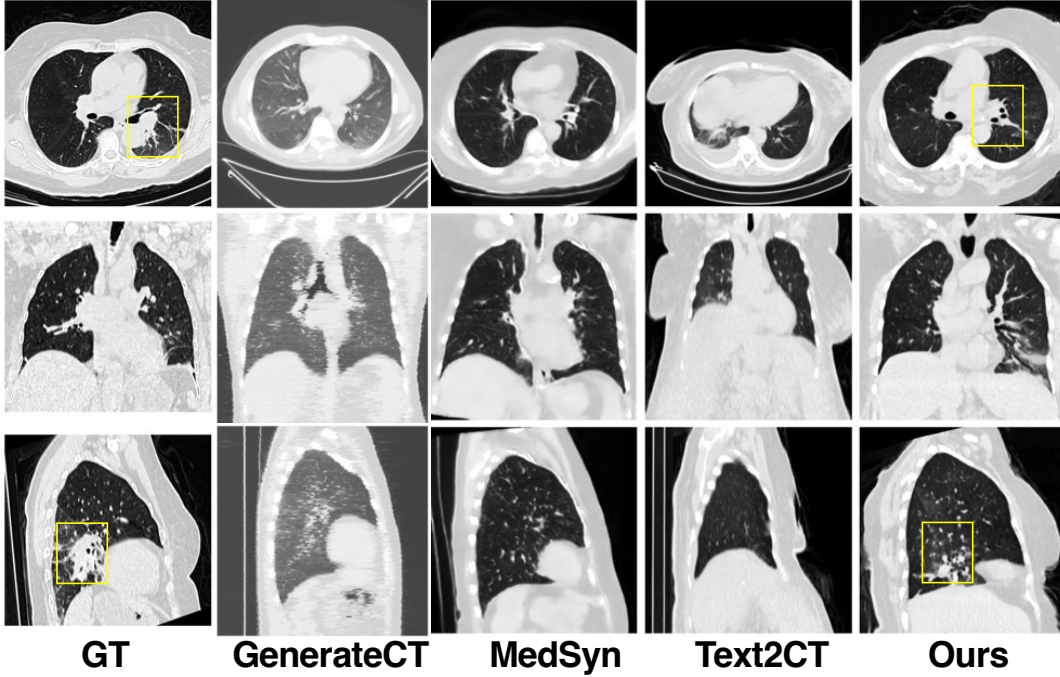


Fig. 2: Comparisons using text guidance only. All images are shown in lung contrast window with pixel spacing $(0.7 \times 0.7 \times 0.7)$. Corresponding report mentions ‘volume loss in the left lung and widespread atelectatic changes’. Our method correctly synthesizes this in bounding box, showing **controllability**.

Radiologist Evaluation. We randomly sample 15 pairs each abnormality, and synthesize another 75 cases with reports and masks, forming a dataset totaling 150 cases. A double-blind test with two radiologists to evaluate whether the images match the report and mask is conduct. Table 4 shows the assessment of synthetic ones and real ones. Our synthetic images show majority in the ‘Match’ and ‘Partial Match’ ones, with a distribution similar to the real ones. Moreover, both radiologists show agreement on low ‘Not Match’ counts, further implying the same pattern. The radiologist evaluation shows an overall high fidelity result of our image towards the input condition. (More details in Supp. 7.4)

5.2 Qualitative Results

Comparing Baselines. In Fig. 2 although GenerateCT shows good on axial view (first row), it suffers on the other two, as it performs super resolution on axial slices. MedSyn generates a rather healthy image, lacking the focus on the abnormality region. Text2CT shows abdomen which leads to good soft tissues yet it produces blurry image compared to ours, since it adopts un-registered image and therefore loses detail. In contrast, our method generates high quality images, while it clearly synthesizes the volume loss and atelectasis (yellow boxes) noted in the report, attributing to our encoding of word level tokens and gated attention module, allowing for more details preserved. In Fig. 3 our

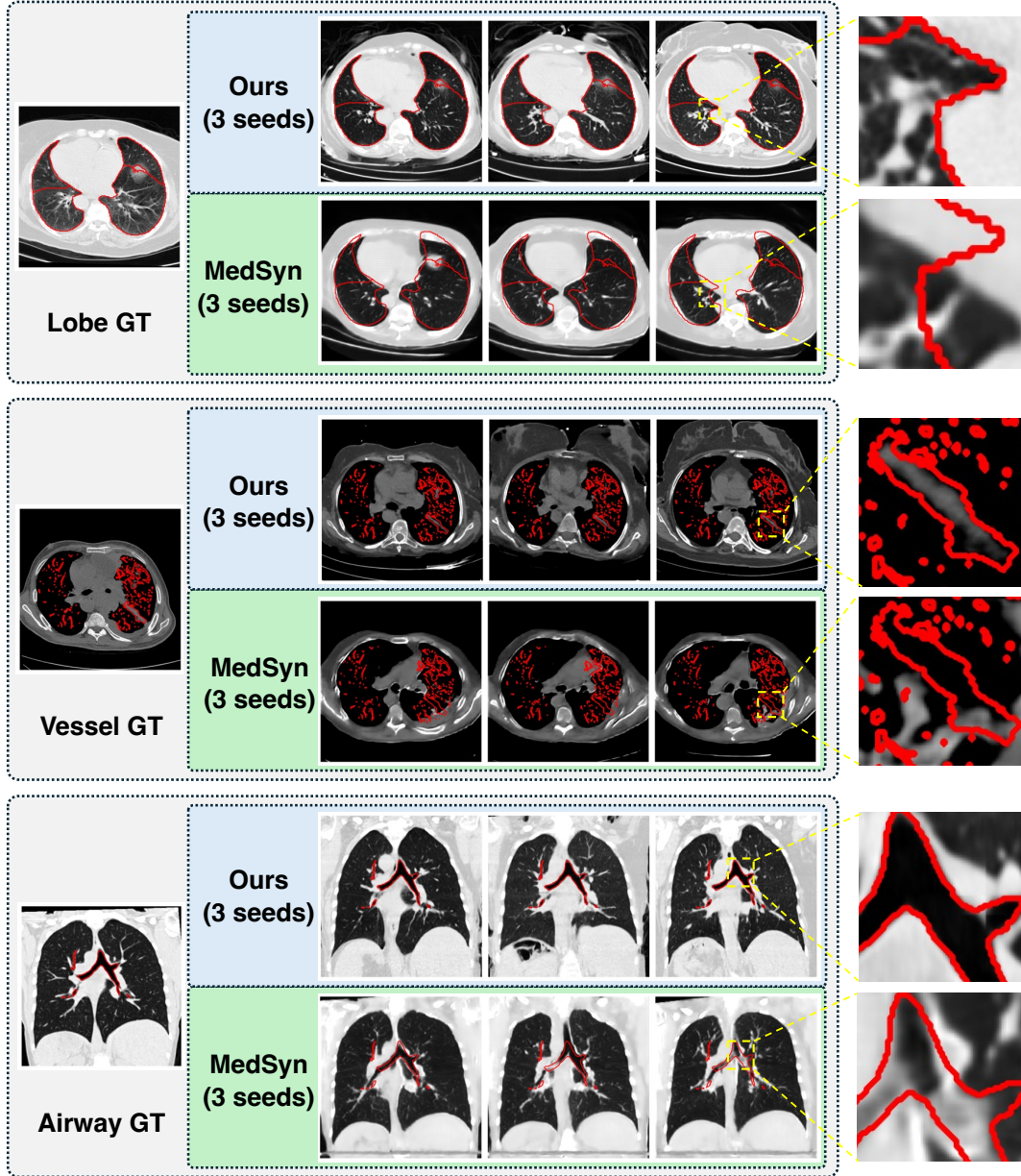


Fig. 3: Comparison of anatomically conditional generation using lobe, airway and vessel masks (without report). The second block is shown in vessel contrast window, while the other two in lung contrast window. We show ground truth images, ours and MedSyn’s with 3 different random seeds. The zoomed-in views on the right show MedSyn fails to follow anatomical masks. In contrast, our results closely respect the conditional masks, exhibiting both **controllability** and **diversity**.

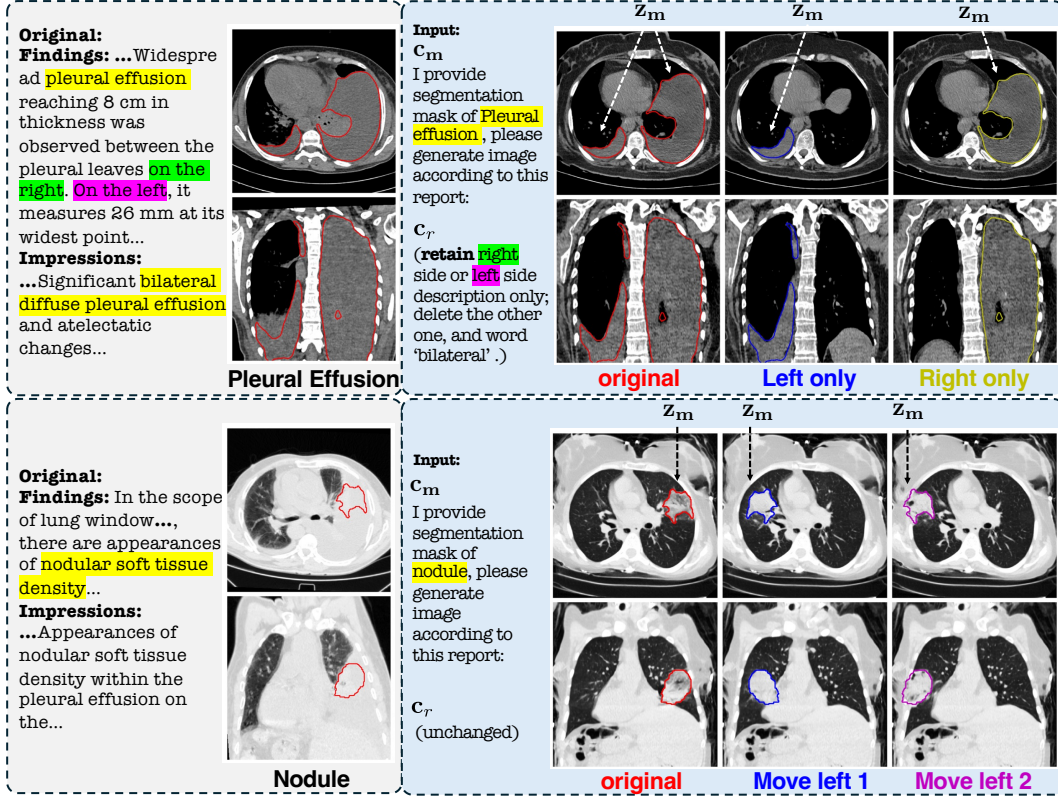


Fig. 4: Results of modifying pathology masks and reports. For pleural effusion, we **retain** either **left** or **right** mask, and modify the corresponding report by deleting the abnormality description on the other side. For nodule, since no location was described in report, we spatially **move** the mask to left side of body by either **flipping** or **translating** without changing the report. Our results closely follow the conditional masks, showing both **controllability** and **flexibility**.

method shows closer alignment towards the segmentation mask. As noted by the zoomed-in views (right most), MedSyn generates organs falling out of the given mask, which results from the loose incorporation of segmentation masks. Ours on the contrary, directly incorporates the noisy latent and mask together, enforces better inter-patch consistency with our convolution layer, therefore yielding better alignment. Results in Fig. 4 show great controllability over masks of different appearance. Whether we retain one side of the abnormality, or completely flip/translate the mask, the generated results always show strong correspondence with the condition. All these results indicate our controllability and flexibility. (More results in Supp. 7.1)

5.3 Scalability

We investigate the scalability with two downstream tasks: 1) augmenting datasets assisting training segmentation model. 2) Multiple conditional synthesis simulating gradually growing abnormalities. (More results in Supp. 7.4)

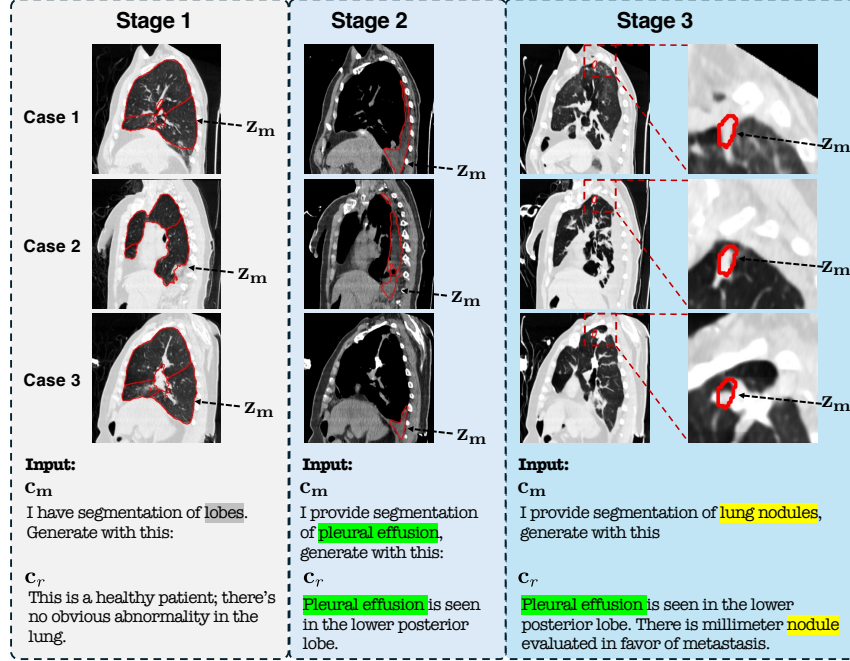


Fig. 5: Results of progressively adding conditions. We start by generating healthy images with lobes only. Progressively, we add a pleural effusion mask and a nodule mask, describing them in the radiology report. In each row, following stages keep the characteristic from previous stages, while exhibiting new abnormalities.

Augmenting dataset. We augment the SemiSeg dataset [8] segmenting Ground Glass Opacity (GGO), whose original training set contains only 50 2D images with segmentation labels. We sample 416 2D masks from 3D ones (details in Supp. [7.1]) and compare three settings: 1) real images only (50 images), 2) synthesized only (416 images), and 3) combination of both (466 images). We train vanilla UNet models and show quantitative results in table [5]. Training with the combined dataset achieves the highest Dice score (68.13%), Jaccard (54.27%), and the lowest HD95 (4.69), even surpassing the SOTA models (MedPSeg Dice 65.2%), showing great efficacy with our augmented images. In addition, training solely on synthetic images achieves a fair Dice score (48.82%) and HD95 (10.61). In conclusion, 1) our method effectively assists training segmentation models by augmenting training data. 2) our model receives only segmentation masks and preserves data and patient privacy. (More results in Supp. [7.4])

Multiple Conditional Synthesis. Fig. [5] shows the results where we progressively add more details to a case, synthesizing growing multiple diseases. We start with lobe mask and healthy report (pair one) to generate stage one images, then fix the same random seed and add noise again to apply additional pairs of conditions following SDEdit [24]. Images in column two all preserve the original layout of the first column, with only pleural effusions added and marked in boundaries. Similarly in column three, while lobe positions and pleural effusions are preserved, the zoomed-in views show delicate nodules, indicating our

scalability of multiple conditions. In conclusion, 1) scalability of our model indicates further details can be added given single or more segmentation masks with our low-cost controlling, 2) the results form procedures with arbitrary abnormalities and simple reports, educating residents to various combinations of diseases on the same CT.

6 Conclusion

This paper proposes the first joint-modality to lung CT generation model enabling single object of anatomy or pathology as condition, enforcing scalability and flexibility via compound prompts (pre-prompt and radiology reports). We enforce the scalability of various binary mask by introducing a compound pre-prompt, optionally paired with radiology reports. Extensive experiments including radiologist evaluation and downstream dataset augmentation prove the controllability, flexibility and scalability of our method. Evaluations by radiologists, downstream data augmentation utility and practical clinical scenario simulation of developing diseases all indicate the potential of our method.

Ethical Statement and Data Usage The retrospective data used in this study was obtained from the publicly available CT-Rate repository [11]. All data utilization was performed in compliance with the repository’s data use agreement and institutional ethical guidelines.

Acknowledgements This work was supported by NIH grant 5R01HL141813 and NSF CAREER grant 2443167. We thank Binxu Li (Stanford), Afroz Zandifar (UPMC), Christina LeBedis (BU) for their valuable assistance.

Disclosure of Interests The authors declare no competing interests in the paper.

References

1. Amirrajab, S., Salahuddin, Z., Kuang, S., Woodruff, H.C., Lambin, P.: Radiology report conditional 3d ct generation with multi encoder latent diffusion model (2025), <https://arxiv.org/abs/2509.14780>
2. Bhat, A., Bose, R., Nwoye, C.I., Padoy, N.: Simgen: A diffusion-based framework for simultaneous surgical image and segmentation mask generation (2025), <https://arxiv.org/abs/2501.09008>
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision-language processing (2022). <https://doi.org/10.48550/ARXIV.2204.09817> <https://arxiv.org/abs/2204.09817>
4. Bose, R., Nwoye, C.I., Bhat, A., Padoy, N.: Cosingen: Controllable diffusion model for simultaneous image and mask generation (2025), <https://arxiv.org/abs/2503.19661>
5. Carmo, D.S., Ribeiro, J.A., Comellas, A.P., Reinhardt, J.M., Gerard, S.E., Rittner, L., Lotufo, R.A.: Medpseg: Hierarchical polymorphic multitask learning for the segmentation of ground-glass opacities, consolidation, and pulmonary structures on computed tomography (2024)
6. Chung, H., Ryu, D., McCann, M.T., Klasky, M.L., Ye, J.C.: Solving 3d inverse problems using pre-trained 2d diffusion models (2023)
7. Dorjsembe, Z., Pao, H.K., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics* **28**(7), 4084–4093 (2024). <https://doi.org/10.1109/JBHI.2024.3385504>
8. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging* **39**(8), 2626–2637 (2020). <https://doi.org/10.1109/TMI.2020.2996645>
9. Guo, P., Zhao, C., Yang, D., He, Y., Nath, V., Xu, Z., Bassi, P., Zhou, Z., Simon, B., Harmon, S., Turkbey, B., Xu, D.: Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model (05 2025). <https://doi.org/10.48550/arXiv.2505.04522>
10. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: Maisi: Medical ai for synthetic imaging (09 2024). <https://doi.org/10.48550/arXiv.2409.11169>
11. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., İrem Hatice Doğan, Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A.K., Lafci, B., Ozdemir, M.K., Menze, B.H.: Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature biomedical engineering* (2024), <https://api.semanticscholar.org/CorpusID:268691507>
12. Hamamci, I.E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Doğan, I., Dasdelen, M.F., Prabhakar, C., Reynaud, H., Pati, S., Bluethgen, C., Ozdemir, M.K., Menze, B.: Generatect: Text-conditional generation of 3d chest ct volumes. In: *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*. p. 126–143. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72986-7_8, https://doi.org/10.1007/978-3-031-72986-7_8

13. Han, K., Xiong, Y., You, C., Khosravi, P., Sun, S., Yan, X., Duncan, J.S., Xie, X.: Medgen3d: A deep generative framework for paired 3d image and mask generation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. pp. 759–769. Springer Nature Switzerland, Cham (2023)
14. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental* **4**, 50 (08 2020). <https://doi.org/10.1186/s41747-020-00173-2>
15. Jiang, Y., Lemaréchal, Y., Plante, S., Bafaro, J., Abi-Rjeile, J., Joubert, P., Després, P., Manem, V.: Lung-ddpm: Semantic layout-guided diffusion models for thoracic ct image synthesis. *IEEE Transactions on Biomedical Engineering* **73**(3), 1134–1145 (2026). <https://doi.org/10.1109/TBME.2025.3599011>
16. Krishna, A., Wang, G., Mueller, K.: Guided synthesis of annotated lung ct images with pathologies using a multi-conditioned denoising diffusion probabilistic model (mddpm). *Physics in Medicine & Biology* **70**(6), 065007 (mar 2025). <https://doi.org/10.1088/1361-6560/adb9b3> <https://doi.org/10.1088/1361-6560/adb9b3>
17. Kumar, A., Kriz, A., Havaei, M., Arbel, T.: PRISM: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. In: *Medical Imaging with Deep Learning* (2025), <https://openreview.net/forum?id=UpJMA1ZNuo>
18. Li, T., He, K.: Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720* (2025)
19. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., Jia, T., Zhang, J., Tang, Z., Pang, Y., She, B., Yan, C., Hu, Z., Dong, X., Chen, L., Pan, Z., Zhou, X., Dong, S., Tian, Y., Yuan, L.: Open-sora plan: Open-source large video generation model (2024), <https://arxiv.org/abs/2412.00131>
20. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: *The Eleventh International Conference on Learning Representations* (2023), <https://openreview.net/forum?id=PqvMRDCJT9t>
21. Liu, C., Yuan, X., Yu, Z., Wang, Y.: Texdc: Text-driven disease-aware 4d cardiac cine mri images generation. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. pp. 3005–3021 (December 2024)
22. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: *The Eleventh International Conference on Learning Representations* (2023), <https://openreview.net/forum?id=XVjTT1nw5z>
23. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence* **0**(ja), e210315 (0). <https://doi.org/10.1148/ryai.210315> <https://doi.org/10.1148/ryai.210315>
24. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: *International Conference on Learning Representations* (2022), https://openreview.net/forum?id=aBsCjcPu_tE
25. Molino, D., Caruso, C.M., Ruffini, F., Soda, P., Guarrasi, V.: Text-to-ct generation via 3d latent diffusion model with contrastive vision-language pretraining (2025), <https://arxiv.org/abs/2506.00633>

26. Morrison, K., Mathur, A., Bradshaw, A., Wartmann, T., Lundi, S., Zandifar, A., Dai, W., Batmanghelich, K., Eslami, M., Perer, A.: A human-centered approach to identifying promises, risks, & challenges of text-to-image generative ai in radiology (07 2025). <https://doi.org/10.48550/arXiv.2507.16207>
27. Nercessian, M., Agrawal, K., Liu, L., Lian, L., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L., Fintelmann, F., Darrell, T., Bai, Y., Chung, M., Yala, A.: Pillar-0: A new frontier for radiology foundation models (11 2025). <https://doi.org/10.21203/rs.3.rs-8196619/v1>
28. Oliveras, A., Marí, R., Redondo, R., Guardiola-Olivella, O., Tost, A., Nagarajan, B., Migliorelli, C., Ribas, V., Radeva, P.: LAND: Lung and nodule diffusion for 3d chest CT synthesis with anatomical guidance. In: Submitted to Medical Imaging meets EurIPS: MedEurIPS 2025 (2025), <https://openreview.net/forum?id=VTQwLZLq0a>, under review
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
30. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., Wang, Y., Ye, A., Ren, G., Ma, Q., Liang, W., Lian, X., Wu, X., Zhong, Y., Li, Z., Gong, C., Lei, G., Cheng, L., Zhang, L., Li, M., Zhang, R., Hu, S., Huang, S., Wang, X., Zhao, Y., Wang, Y., Wei, Z., You, Y.: Open-sora 2.0: Training a commercial-level video generation model in 200k. arXiv preprint arXiv:2503.09642 (2025)
31. Poletti, J., Bach, M., Yang, S., Sexauer, R., Stieltjes, B., Rotzinger, D.C., Bremerich, J., Walter Sauter, A., Weikert, T.: Automated lung vessel segmentation reveals blood vessel volume redistribution in viral pneumonia. *European Journal of Radiology* **150**, 110259 (2022). <https://doi.org/10.1016/j.ejrad.2022.110259>
32. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., Lin, J.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), <https://openreview.net/forum?id=1b7wh04SfY>
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
34. Sexauer, R., Yang, S., Weikert, T., Poletti, J., Bremerich, J., Roth, J., Sauter, A., Anastasopoulos, C.: Automated detection, segmentation, and classification of pleural effusion from computed tomography scans using machine learning. *Investigative Radiology* **Publish Ahead of Print** (03 2022). <https://doi.org/10.1097/RLI.0000000000000869>
35. Shao, M., Miao, X., Duan, H., Wang, Z., Chen, J., Huang, Y., Wu, X., Deng, J., Long, Y., Zheng, Y.: Trace: Temporally reliable anatomically-conditioned 3d ct generation with enhanced efficiency. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. pp. 627–637. Springer Nature Switzerland, Cham (2026)
36. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging* **44**, 4960–4972 (2024), <https://api.semanticscholar.org/CorpusID:274789446>
37. Wang, J., Reynaud, H., Erick, F.X., Kainz, B.: Ctflow: Video-inspired latent flow matching for 3d ct synthesis (2025), <https://arxiv.org/abs/2508.12900>

38. Wang, Z., Xia, X., Chen, R., Yu, D., Wang, C., Gong, M., Liu, T.: Lavin-dit: Large vision diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20060–20070 (2025)
39. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5**(5), e230024 (2023). <https://doi.org/10.1148/ryai.230024>
<https://doi.org/10.1148/ryai.230024>
40. Wilder-Smith, A.J., Yang, S., Weikert, T., Bremerich, J., Haaf, P., Segeroth, M., Ebert, L.C., Sauter, A., Sexauer, R.: Automated detection, segmentation, and classification of pericardial effusions on chest ct using a deep convolutional neural network. *Diagnostics* **12**(5) (2022). <https://doi.org/10.3390/diagnostics12051045>, <https://www.mdpi.com/2075-4418/12/5/1045>
41. Xing, X., Ning, J., Nan, Y., Yang, G.: Deep generative models unveil patterns in medical images through vision-language conditioning. *arXiv preprint arXiv:2410.13823* (2024)
42. Xu, Y., Sun, L., Peng, W., Jia, S., Morrison, K., Perer, A., Zandifar, A., Visweswaran, S., Eslami, M., Batmanghelich, K.: Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. *IEEE Transactions on Medical Imaging* (2024). <https://doi.org/10.1109/TMI.2024.3415032>
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
44. Zhao, C., Guo, P., Yang, D., Tang, Y., He, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss (2025), <https://arxiv.org/abs/2508.05772>
45. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404* (2024)
46. Zhuang, Y., Hou, B., Mathai, T.S., Mukherjee, P., Kim, B., Summers, R.M.: Semantic image synthesis for abdominal ct. In: Deep Generative Models: Third MICCAI Workshop, DGM4MICCAI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings. p. 214–224. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-53767-7_21
https://doi.org/10.1007/978-3-031-53767-7_21