LAND: Lung and Nodule Diffusion for 3D Chest CT Synthesis with Anatomical Guidance

Anonymous Author(s)
Affiliation
Address
email

Abstract

This work introduces a new latent diffusion model to generate high-quality 3D chest CT scans conditioned on 3D anatomical masks. The method synthesizes volumetric images of size $256 \times 256 \times 256$ at 1 mm isotropic resolution using a single mid-range GPU, significantly lowering the computational cost compared to existing approaches. The conditioning masks delineate lung and nodule regions, enabling precise control over the output anatomical features. Experimental results demonstrate that conditioning solely on nodule masks leads to anatomically incorrect outputs, highlighting the importance of incorporating global lung structure for accurate conditional synthesis. The proposed approach supports the generation of diverse CT volumes with and without lung nodules of varying attributes, providing a valuable tool for training AI models or healthcare professionals. Code for LAND is available at: https://github.com/anonymous/LAND-3DCT.

1 Introduction

Deep learning in medical imaging is hindered by the scarcity of large, diverse datasets, constrained by privacy concerns, costs, and the need for expert labeling. Synthetic data offers a promising solution, with potential impact in critical areas such as lung cancer, the leading cause of cancer-related deaths [3]. Diffusion models [11] have emerged as the most powerful generative framework, surpassing VAEs [16] and GANs [7] in realism and stability [5, 19]. However, scaling diffusion models to large synthetic volumes such as CT scans remains challenging due to extreme computational demands [14]. Recent methods have explored efficiency trade-offs. Previous Latent Diffusion Models (LDMs) [22] for 3D synthesis use autoencoders for data compression, but are often limited in resolution [21, 15]. PatchDDM [2] and WDM [6] bypass autoencoders with subvolume or wavelet representations but still require large GPU memory. NVIDIA's LDM MAISI [8] attains the highest resolution to date ($512 \times 512 \times 768$), but demands 49.7GB GPU memory, unaffordable for most users.

We introduce LAND (Lung-And-Nodule-Diffusion), a memory-efficient latent diffusion model for 3D chest CT synthesis. It generates 256^3 volumes at 1mm resolution on a single 20GB GPU, uses lung and nodule masks for anatomical conditioning, and controls nodule texture for realistic pathological diversity. LAND combines computational efficiency with fine-grained anatomical control to achieve state-of-the-art (SOTA) high-resolution volume synthesis with practical hardware requirements.

2 Method

LAND is a latent diffusion model comprising a 3D U-Net and a 3D VAE architecture (Fig. 1).

3D VAE A 3D VAE encodes input CT images into latent representations compressing $4\times$ the spatial resolution and expanding $4\times$ the feature dimensionality: each $256 \times 256 \times 256$ volume

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

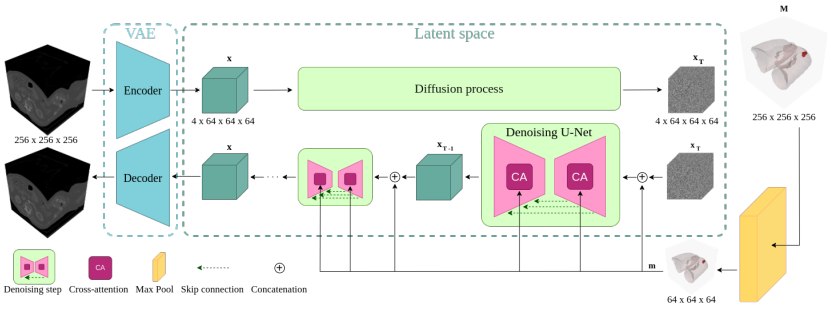

Figure 1: LAND uses a 3D VAE to encode CT volumes into a latent space, where a 3D U-Net performs diffusion on the latent samples $\mathbf{x}$, optionally conditioned on anatomical masks $\mathbf{m}$ (lungs/nodules).

is encoded as a $64 \times 64 \times 64 \times 4$ latent sample. We adopt a lightweight variant of the MAISI architecture [8], using 3 resolution levels with one residual block per level and the same number of channels in both encoder and decoder. The VAE is trained using a combination of an L_1 loss $\mathcal{L}_{\text{MAE}}$, a perceptual similarity loss $\mathcal{L}_{\text{LPIPS}}$, an adversarial loss $\mathcal{L}_{\text{ADV}}$ and a Kullback-Leibler term $\mathcal{L}_{\text{KL}}$: $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{MAE}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{LPIPS}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{ADV}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{KL}}(\mathcal{E}(\mathbf{x}))$, where $\hat{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$ is the reconstructed encoded-decoded volume. $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{LPIPS}}$ enforce numerical and perceptual fidelity [25], while $\mathcal{L}_{\text{KL}}$ regularizes the latent space [16] and $\mathcal{L}_{\text{ADV}}$ prevents unrealistic artifacts [8, 7].

3D U-Net The denoising network is a 3D U-Net with 5 resolution levels and two residual blocks per level. Additive skip connections [2] reduce memory load while preserving spatial information. To enhance conditional generation, cross-attention modules [22, 17] re-inject conditioning masks (if any) at multiple resolution levels. Training uses velocity prediction [23], enabling the U-Net to learn denoising by estimating a linear combination of clean latent and added noise, which stabilizes training and improves high-resolution synthesis [12, 23]. A linear noise schedule is applied, and training follows a Min-SNR-γ loss weighting [9] to balance timestep contributions by signal-to-noise ratio: $\mathcal{L}_{\text{min-SNR}} = \gamma(\text{SNR}_t)|\hat{\mathbf{v}}_t(\mathbf{z}_t, \mathbf{m}) - \mathbf{v}_t|^2$, where $\mathbf{z}_t$ is the noisy latent, $\mathbf{m}$ the conditioning mask, $\gamma(\cdot)$ the Min-SNR weight, and $\mathbf{v}_t, \hat{\mathbf{v}}_t$ the target and predicted velocities. To ensure anatomical plausibility in 3D, LAND can be conditioned on masks $\mathbf{m}$ covering lungs and nodules. Unlike prior 2D work [17], where nodule-only masks led to implausible nodule placements, our volumetric setting proposes richer conditioning. Spatial and textural cues are encoded by assigning lungs a value of 0.5 and nodules 1–5 (non-solid to solid). Masks are normalized to [0,1], downsampled four times via 3D max pooling, concatenated with the noisy latent, and injected into U-Net cross-attention layers [22, 17].

3 Experimental Results

Datasets and Evaluation Two publicly available datasets were used. LIDC-IDRI [1] includes 1,010 CT volumes with nodule masks and attribute ratings from four radiologists; for this study, nodule textures scored 1–5 (1: Non-Solid, 2: Non-Solid/Mixed, 3: Part-Solid, 4: Solid/Mixed, 5: Solid) were considered. From NLST [20], we selected 881 CT volumes with at least one nodule annotation [18] and generated the nodule masks using an ad-hoc U-Net. Lung regions in both datasets were segmented with a pre-trained open-source U-Net [13]. All scans were preprocessed as in [6]. LIDC-IDRI was used for training, while the NLST subset provided unseen anatomical masks for inference. Evaluation follows the protocol of previous SOTA models [6], using Fréchet Inception Distance (FID) [10] for synthesis quality and MS-SSIM [24] for sample diversity. FID is computed on 881 real and synthetic CT scans using a ResNet-50 pretrained on 23 medical imaging datasets [4]; lower FID indicates closer distributional alignment between real and synthetic samples. MS-SSIM is computed on 10k synthetic pairs, with lower scores indicating higher diversity.

Implementation Details Training was performed on a single Nvidia Grid A100-20C (20 GB) GPU. The 3D VAE was trained independently for 100 epochs with AdamW (learning rate 1×10^{-4}, batch size 1). The 3D U-Net was trained for 500k steps with AdamW (learning rate 1×10^{-5}, batch size 1). The diffusion process used $T = 1000$ timesteps with a linear noise schedule from $\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 0.02$. Inference uses the same number of steps to prioritize sample quality.

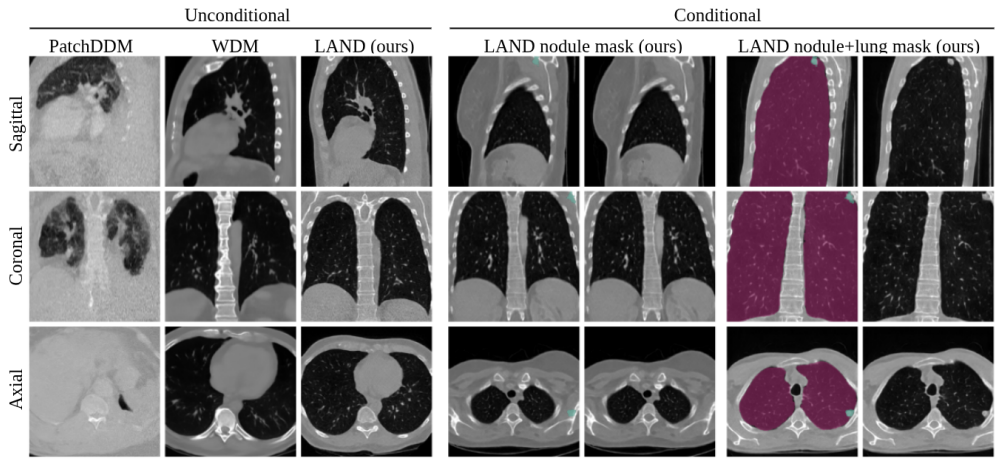

Figure 2: Comparison of unconditional (left) and conditional (right) CT generation using LAND and baseline methods PatchDDM [2] and WDM [6]. Mask overlays are shown where applicable.

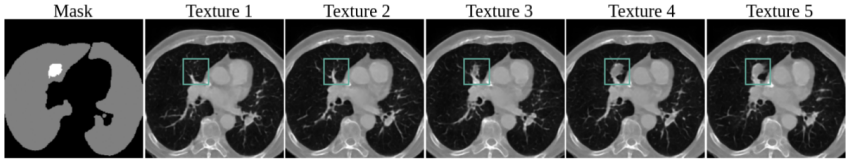

Figure 3: LAND samples conditioned on nodule+lung+texture masks, with increasing texture scores.

Discussion We evaluate the unconditional LAND pipeline against WDM [6] and PatchDDM [2], with WDM using the official pre-trained weights and PatchDDM retrained for our task. Quantitatively, LAND achieves the lowest FID (Table 1-Left), reflecting higher fidelity and semantic alignment through its VAE latent space; WDM shows slightly higher MS-SSIM, indicating greater diversity but potentially related to some unrealistic samples, while PatchDDM performs worst, likely due to the inability to handle unregistered volumes. Qualitatively, LAND produces sharp, anatomically consistent samples, WDM tends to appear slightly blurred, and PatchDDM exhibits higher levels of noise and structural variability (Fig.2-Left). Note that LAND and WDM have similar inference memory requirements, but only LAND can be trained on a single 20GB GPU, whereas WDM requires double the memory. Differences in WDM performance compared to [6] likely stem from the differing sample count used for the FID computation, as FID can be sensitive to sample count. Conditional LAND experiments with masks—(1) nodules, (2) nodule+lung, and (3) nodule+lung+texture—show improved FID when lungs are included, highlighting the importance of global context, while MS-SSIM remains similar (Table 1-Right). Only nodule+lung masks (Fig.2-Right) ensure realistic nodule placements. Nodule masks (without lung areas) may incorrectly lead to nodules outside the lungs. Nodule+lung+texture masks further allow control over the synthetic nodule solidity (Fig.3).

Table 1: Comparison of LAND (ours) with SOTA methods. FID values are multiplied by 10^3.

Unconditional Method	**FID↓ (LIDC)**	**FID↓ (NLST)**	**MS-↓ SSIM**	**Mem↓ (GB)**	**Conditional Method**	**FID↓ (LIDC)**	**FID↓ (NLST)**	**MS-↓ SSIM**	**Mem↓ (GB)**
PatchDDM [2]	317.53	376.4	0.39	19.61	LAND nodule	4.52	5.82	0.3	7.52
WDM [6]	15.24	32.66	**0.27**	**7.27**	LAND nodule+lung	**4.48**	**3.37**	0.29	7.52
LAND	**5.062**	**4.76**	0.29	7.38	LAND nodule+lung+texture	4.60	3.87	0.29	7.52

4 Conclusion

This paper presents LAND, a latent diffusion model that generates high-quality chest CT volumes from anatomical masks. The method enables precise control of lung and nodule characteristics while remaining efficient on a single mid-range GPU. Future work includes testing LAND synthetic samples for tasks such as nodule classification and segmentation, extending the model with additional clinically relevant features, and adding a mask generation module to enhance anatomical diversity.

5 Potential Negative Societal Impact

While the proposed approach offers useful tools for medical research and education, it also presents potential risks that should be acknowledged. High-quality synthetic 3D chest CT scans could be mistaken for real clinical data if not clearly labeled, which might lead to confusion or reduce trust in medical imaging workflows. There is also a possibility that biases present in the training data could be reflected or amplified in the generated outputs.

The proposed method enhances image quality and lowers computational cost without introducing new misuse risks beyond those already known in generative modeling. Careful data governance, clear labeling of synthetic content, and responsible use are important to minimize unintended negative consequences.

References

[1] Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, et al. Data from LIDC-IDRI, 2015. URL `https://www.cancerimagingarchive.net/collection/lidc-idri/`.

[2] Florentin Bieder, Julia Wolleb, Alicia Durrer, Robin Sandkuehler, and Philippe C Cattin. Memory-efficient 3D denoising diffusion models for medical image processing. In *Medical Imaging with Deep Learning*, pages 552–567, 2024.

[3] Freddie Bray, Mathieu Laversanne, Hyuna Sung, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024.

[4] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[6] Paul Friedrich, Julia Wolleb, Florentin Bieder, Alicia Durrer, and Philippe C Cattin. WDM: 3D wavelet diffusion models for high-resolution medical image synthesis. In *MICCAI Workshops*, pages 11–21, 2024.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[8] Pengfei Guo, Can Zhao, Dong Yang, et al. MAISI: Medical AI for synthetic imaging. In *WACV*, pages 4430–4441, 2025.

[9] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *ICCV*, pages 7441–7451, 2023.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[12] Jonathan Ho, William Chan, Chitwan Saharia, et al. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[13] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4:1–13, 2020.

[14] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, et al. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.

[15] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1), 2023. ISSN 2045-2322.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Roger Marí Molas, Paula Subías-Beltrán, Carla Pitarch Abaigar, Mar Galofré Cardo, and Rafael Redondo Tejedor. Characterization of synthetic lung nodules in conditional latent diffusion of chest CT scans. In *Artificial Intelligence Research and Development*, pages 44–51. 2024.

[18] Peter G Mikhael, Jeremy Wohlwend, Adam Yala, et al. Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, pages JCO–22, 2023.

[19] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

[20] National Lung Screening Trial Research Team. Data from the National Lung Screening Trial (NLST), 2013. URL `https://www.cancerimagingarchive.net/collection/nlst/`.

[21] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshops*, pages 117–126, 2022.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[23] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[24] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, volume 2, pages 1398–1402, 2003.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.