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

Anna Oliveras[1,2] Roger Marí[1] Rafael Redondo[1] Oriol Guardià[1] Ana Tost[1]
Bhalaji Nagarajan[3] Carolina Migliorelli[1] Vicent Ribas[1] Petia Radeva[2]

[1]Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain
[2]Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
[3]Barcelona Supercomputing Center (BSC), Spain

Contributions

- Novel latent diffusion model for high-quality 3D chest CT volume synthesis ($256 \times 256 \times 256$ outputs at 1 mm resolution), optimized for memory efficiency and scalability, capable of running on a single 20 GB GPU.
- Anatomically-conditioned nodule synthesis: 3D semantic masks with lung and nodule areas are used to guide the generative process and improve the spatial and morphological consistency of the output samples. In addition, the texture of the nodules can be explicitly controlled.

Method

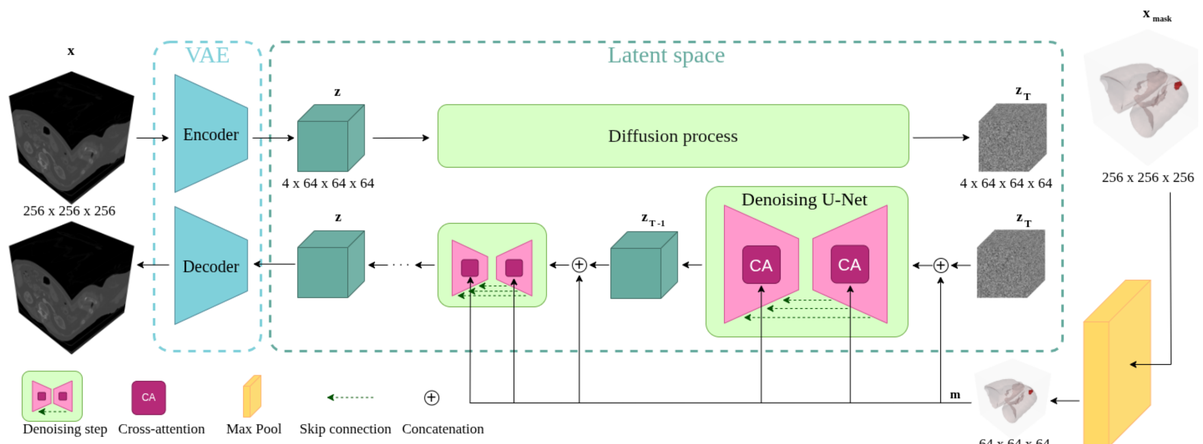

Figure 1: LAND architecture.

- LAND is a 3D latent diffusion model combining a 3D VAE for latent compression and a 3D U-Net for denoising.

3D VAE (Latent Compression)

The VAE encodes each 256^3 CT volume into a $64^3 \times 4$ latent representation ($4\times$ spatial downsampling, $4\times$ channel expansion). We adopt a lightweight MAISI-inspired architecture with 3 resolution levels and one residual block per level. **VAE Objective.** The VAE is trained using the following loss:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{MAE}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{LPIPS}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{ADV}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{KL}}(\mathcal{E}(\mathbf{x})),$$

where $\hat{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$. MAE and LPIPS enforce numerical and perceptual fidelity, $\mathcal{L}_{\text{ADV}}$ discourages artifacts, and $\mathcal{L}_{\text{KL}}$ regularizes the latent space.

3D U-Net (Latent Diffusion)

The denoiser is a 5-level 3D U-Net with two residual blocks per level and additive skip connections to reduce memory. Cross-attention layers re-inject conditioning information at multiple resolutions. **Velocity Prediction Training.** A linear noise schedule is used, and the network predicts the velocity:

$$\mathcal{L}_{\text{min-SNR}} = \gamma(\text{SNR}_t) \left\| \hat{\mathbf{v}}_t(\mathbf{z}_t, \mathbf{m}) - \mathbf{v}_t \right\|^2,$$

where $\mathbf{z}_t$ is the noisy latent, $\mathbf{m}$ the conditioning mask, $\mathbf{v}_t$ and $\hat{\mathbf{v}}_t$ the target and predicted velocities, and $\gamma(\cdot)$ the Min-SNR weight.

Volumetric Conditioning

To ensure 3D anatomical plausibility, LAND is conditioned on volumetric masks:

- **Lungs** encoded with value 0.5
- **Nodules** encoded with values 1–5 (non-solid $\rightarrow$ solid)

Masks are normalized to $[0, 1]$, downsampled $4\times$ via 3D max pooling or encoded with a dedicated VAE (LAND-LatentMask variant), concatenated with $\mathbf{z}_t$, and injected into U-Net cross-attention layers.

Datasets

- **LIDC-IDRI**: Public dataset containing 1,010 CT volumes with lung nodule annotations from four radiologists (Lung nodule segmentation masks and texture scores 1–5: Non-Solid $\rightarrow$ Solid). Used for **training**.
- **NLST** subset: Public dataset containing 881 CT volumes with nodule locations. Nodule segmentation masks are produced via an ad-hoc U-Net. Used as unseen conditioning **masks** for **inference**.

Results

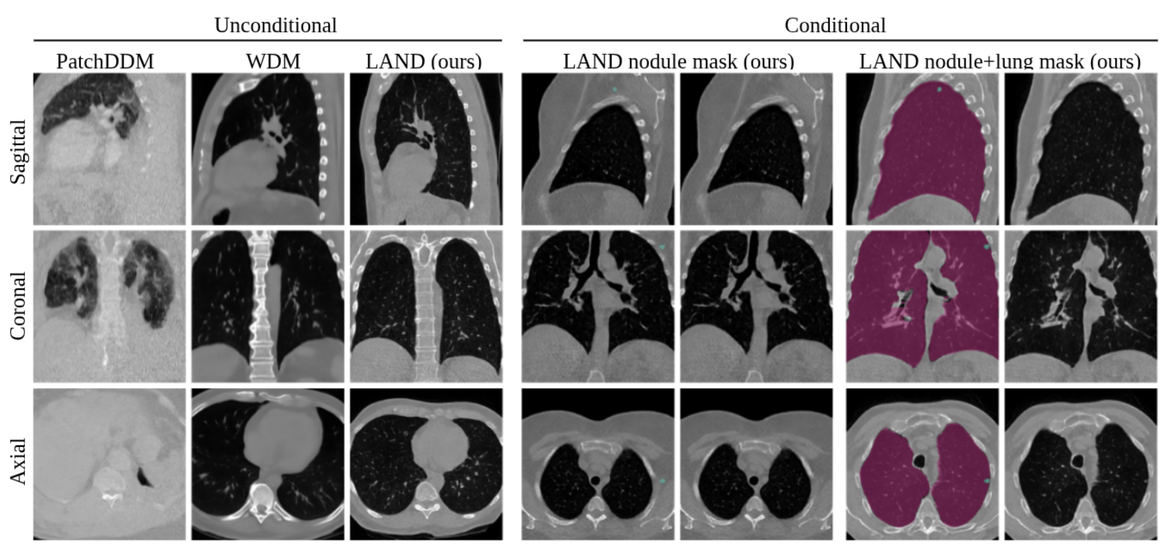

Figure 2: Comparison of unconditional (left) and conditional (right) CT generation using LAND and baseline methods PatchDDM and WDM.

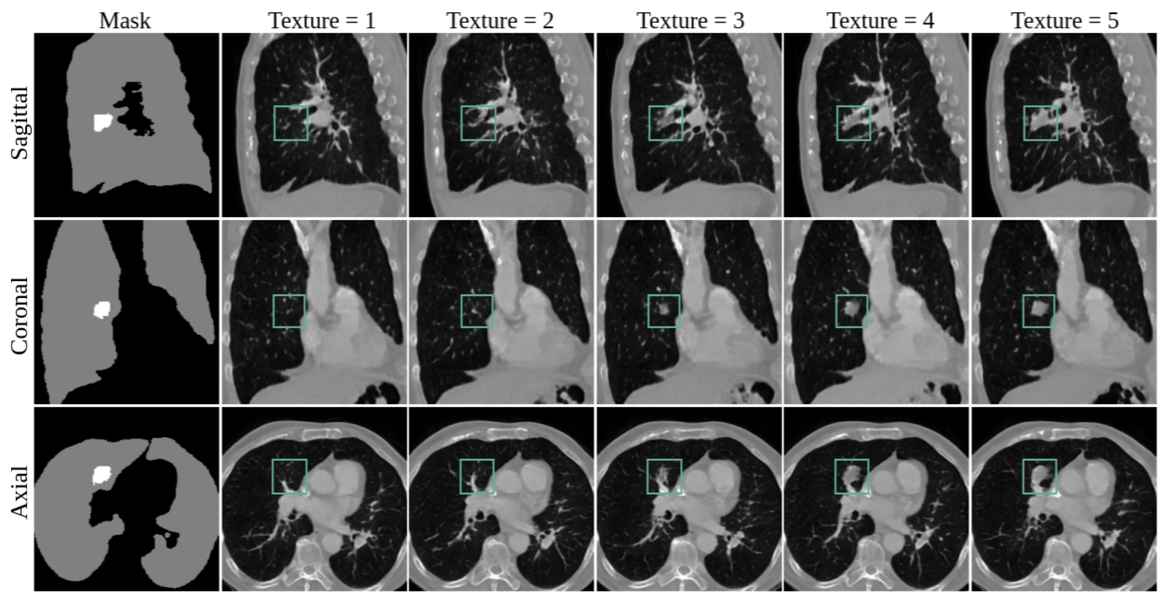

Figure 3: LAND samples conditioned on nodule+lung+texture masks.

Table 1: Comparison of LAND (ours) with SOTA methods. FID values are multiplied by 10^3. "Mem" indicates peak GPU memory during inference.

Unconditional Method	FID↓ (LIDC)	FID↓ (NLST)	MS-SSIM↓	Mem↓ (GB)	Conditional Method	FID↓ (LIDC)	FID↓ (NLST)	MS-SSIM↓	Mem↓ (GB)
PatchDDM	167.90	183.54	0.39	19.61	LAND-Mask	3.75	1.78	0.29	7.52
WDM	12.83	15.78	**0.27**	**7.27**	LAND-LatentMask	**3.74**	**1.77**	**0.29**	**7.52**
LAND	**2.66**	**2.271**	0.29	7.38	LAND-LatentMask+	3.84	1.76	**0.29**	**7.52**

- **Conditional generation:** Including lung areas in the conditioning masks improves FID and ensures realistic nodule placement.
- **Anatomical fidelity:** Generated samples are sharp and anatomically consistent with the provided lung and nodule masks, improved with LAND-LatentMask.
- **Texture control:** LAND+ allows precise control over nodule solidity.

Take-Home Message

- LAND generates high-quality, anatomically realistic 3D CT volumes.
- The model is memory-efficient: full training fits on a single 20GB GPU.
- Conditioning masks provide intuitive control over nodule placement and texture.
- **Future work:** improve fidelity to conditioning masks for small nodules and evaluate generated data on downstream tasks.

Acknowledgements

This research is part of the PHASE IV AI project, which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101095384. The authors thankfully acknowledges the computer resources at Finis Terrae III (CESGA) and the technical support provided by Barcelona Supercomputing Center (RES-BCV-2025-2-0043). Anna Oliveras is a fellow of Eurecat's "Vicente López" PhD grant program. Bhalaji Nagarajan acknowledges AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.