Artificial Intelligence Research and Development
T. Alsinet et al. (Eds.)

doi:10.3233/FAIA240408

Characterization of Synthetic Lung Nodules in Conditional Latent Diffusion of Chest CT Scans

Roger MARÍ MOLAS [a,1], Paula SUBÍAS-BELTRÁN [b], Carla PITARCH ABAIGAR [b], Mar GALOFRÉ CARDO [b] and Rafael REDONDO TEJEDOR [a]

[a] *Eurecat, Centre Tecnològic de Catalunya, Multimedia Technologies, Barcelona, Spain*
[b] *Eurecat, Centre Tecnològic de Catalunya, Digital Health, Barcelona, Spain*

Abstract. This study delves into the characterization of synthetic lung nodules using latent diffusion models applied to chest CT scans. Our experiments involve guiding the diffusion process by means of a binary mask for localization and various nodule attributes. In particular, the mask indicates the approximate position of the nodule in the shape of a bounding box, while the other scalar attributes are encoded in an embedding vector. The diffusion model operates in 2D, producing a single synthetic CT slice during inference. The architecture comprises a VQ-VAE encoder to convert between the image and latent spaces, and a U-Net responsible for the denoising process. Our primary objective is to assess the quality of synthesized images as a function of the conditional attributes. We discuss possible biases and whether the model adequately positions and characterizes synthetic nodules. Our findings on the capabilities and limitations of the proposed approach may be of interest for downstream tasks involving limited datasets with non-uniform observations, as it is often the case for medical imaging.

Keywords. diffusion models, generative AI, CT scan, lung cancer, lung nodules

1. Introduction

Deep learning generative models have revolutionized the landscape of image-based medical applications, providing novel methodologies for data synthesis, augmentation, and interpretation [1, 2]. These models are capable of learning complex data distributions, which allows them to generate high-quality, diverse images that closely mimic real data. This capability is crucial in medical fields where data scarcity and privacy concerns limit the availability of large datasets [2]. However, generative models may face limitations in fine-tuning specific image attributes without altering the overall data integrity.

This work explores the synthesis of lung nodules in chest computed tomography (CT) scans. In particular, the synthesis of lung nodules involves generating images with controlled variations in nodule characteristics, such as size, shape and location. Having control over the characterization of synthetic nodules is a fundamental step to generate well-balanced datasets. Preventing over- or under-represented attributes in synthetic data is essential for robust training of artificial intelligence (AI) systems in downstream tasks.

[1] Corresponding Author: Roger Marí Molas, roger.mari@eurecat.org.

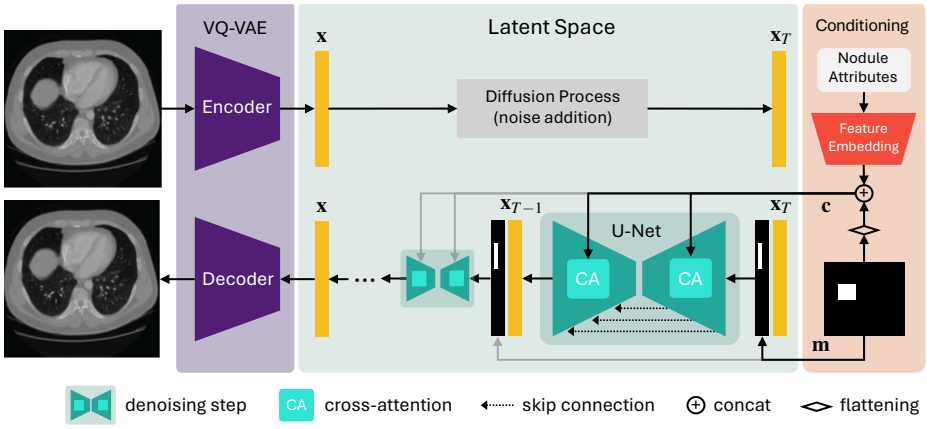

Figure 1. Diagram of the conditional diffusion pipeline. The input CT image is converted by the VQ-VAE encoder into a latent space of lower dimensionality, where the diffusion process takes place. The U-Net is trained to predict the noise in the latent samples (denoising step). A localization mask of the nodule, **m**, is attached to the noisy latent vector $\mathbf{x}_t$ to condition each denoising step t. In addition, the mask **m** and an embedding of the nodule's attributes, **c**, can also be injected into the cross-attention layers of the U-Net to further guide the generative process towards specific visual attributes. Synthetic images are generated by T denoising steps.

In-context synthesis has been previously studied [3–5], showing promising results, especially to artificially insert malignant nodules into healthy lung scans. However, techniques based on adversarial architectures have shown limitations in terms of fidelity and diversity of synthetic data [6].

Today, diffusion models constitute the state of the art in image generation, also in medical imaging problems [7], such as data augmentation [8], super-resolution [2], anomaly detection [9], data repair or alteration [10], and image reconstruction from partial measurements [11, 12]. This work explores different techniques to gain fine-grained control and fidelity of synthetic nodule characterization in chest CT scans based on conditional latent diffusion models.

2. Methodology

The objective of this work is to study how the conditional information of nodule attributes and their location can be injected into diffusion models to guide the synthesis process. Diffusion models using U-Net architectures can be effectively conditioned by incorporating conditional data as input as well as cross-attention layers [13]. The following sections describe the architecture of the method (Figure 1) and the different conditioning mechanisms, as well as the preparation of the training data.

2.1. Data preparation

LIDC-IDRI is a well-known commonly used public dataset [14, 15]. It contains chest CT scans from 1018 patients annotated by 4 experienced thoracic radiologists. The annotations consist of segmentation masks of lung nodules, as well as various nodule attributes scored from 1 to 5. For this study, only annotations related to objective attributes were considered. The rating of each attribute is summarized in Table 1 and illustrated in Figure 2.

Score	Sphericity	Margin	Lobulation	Spiculation	Texture
1	Linear	Poorly defined	None	None	Non-Solid/GGO
2	Ovoid/Linear	Near poorly defined	Nearly none	Nearly none	Non-Solid/Mixed
3	Ovoid	Medium	Medium	Medium	Part-Solid/Mixed
4	Ovoid/Round	Near sharp	Nearly marked	Nearly marked	Solid/Mixed
5	Round	Sharp	Marked	Marked	Solid

Table 1. Attribute annotations of lung nodules in the LIDC-IDRI dataset considered for this study. Sphericity: shape of the nodule or roundness. Margin: sharpness of the nodule contour. Lobulation: degree of lobulation present. Spiculation: degree of spiculation present. Texture: radiographic solidity.

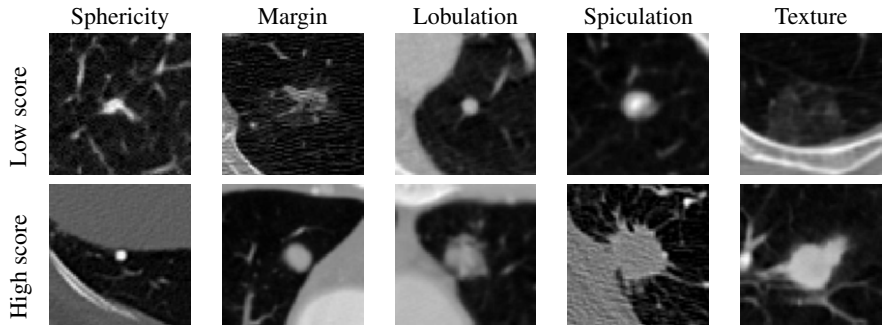

Figure 2. Examples of nodules (2D) in the LIDC-IDRI dataset with high and low attribute scores.

CT scans were resampled to have the same isotropic resolution ($0.7 \times 0.7 \times 0.7$ mm) with pixel intensities clipped between [-1000, 2500] Hounsfield Units and finally normalized between $[0, 1]$. This pre-processing separates air from tissue and ensures no saturated nodules [16]. A total of 1587 CT images with at least one nodule were used for training.

2.2. *Latent diffusion pipeline*

The method consists of a latent diffusion pipeline to generate synthetic 2D images individually at each forward step. Note that in the generative process, synthetic chest CT scans always exhibit at least one pulmonary nodule. To guide the synthesis process, the diffusion model is conditioned on particular data, i.e., a binary mask to indicate the approximate nodule location in the shape of a square bounding box, and the nodule attributes to describe the appearance of the nodule (Table 1). The latent diffusion pipeline consists of three main components: a VQ-VAE, a U-Net and an embedding module, as illustrated in Figure 1.

VQ-VAE. A variational autoencoder performing the conversion between the image space and latent space, with a $\times 4$ spatial compression factor. The encoder incorporates a vector quantization step to map the continuous-valued output to a discrete set of codes. This code book narrows the representation learning problem, improving quality and efficiency [17]. Performing the diffusion process on a latent space instead of the image domain requires less steps and makes it possible to synthesize high-resolution images more efficiently [13].

U-Net. A convolutional neural network performing the diffusion process in latent space. More specifically, the *forward* diffusion process gradually adds noise to the input samples, and the *reverse* process generates new samples by performing the inverse operation [18, 19]. The efficiency of this architecture and the preservation of the resolution of input and output make the U-Net a common choice in diffusion models [1]. Instead of generating a denoised image, the U-Net actually learns to predict the noise present in the noisy input sample. The predicted noise is then subtracted from the input sample. In the reverse process, this step is performed iteratively a fixed number of time steps until the output converges to a realistic synthetic sample. Additionally, it is quite common to insert some self-attention layers in the U-Net to capture long-range dependencies and global context, enhancing the quality and coherence of generated images [13].

Embedding module. This network encodes the nodule attributes into a feature vector that is used in the activations of the U-Net. The scalar value of each attribute is used as input to a dictionary of learnable parameters, which ultimately embeds each value into a vector of dimensionality D. After concatenation, the embedded vectors feed two fully connected layers of size 128, each one followed by non-linear ReLU activation functions, and finally merged into a single vector of length D'. Hereafter, $D = 10$ and $D' = 10$.

Conditioning and cross-attention layers. To guide the spatial information more precisely, a basic and common approach consists of attaching the binary mask to the input of the U-Net —previously downsampled to accommodate to the latent space dimensionality. The square bounding box in the binary mask indicates the position and size of a single nodule at a time. Additionally, cross-attention layers [6, 13] at different levels of the U-Net can dynamically attend to relevant parts of the input and activations, aligning the semantic information more effectively. Both the binary mask and the embedding of nodule attributes can be injected into the cross-attention layers by flattening and concatenating the vectors. These conditioning mechanisms encourage the model to focus on the conditional data and to generate outputs consistent with it.

The diffusion model is optimized to minimize the square of the L^2-norm between the actual ε and predicted noise $\hat{\varepsilon}$, formulated as $\|\varepsilon - \hat{\varepsilon}(\mathbf{x}_t, \mathbf{m}, \mathbf{c}, t)\|_2^2$. Thus, at time t the prediction of the diffusion process relies on —depending on the conditioning mechanism— a noisy latent $\mathbf{x}_t$, the resized binary mask $\mathbf{m}$, and optionally a conditional vector $\mathbf{c}$ containing the embedded nodule attributes in addition to the flattened mask. The larger the time step t is, the noisier the input sample $\mathbf{x}_t$, which by the end of the process follows a standard normal distribution $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A total of 1000 time steps are used for both training and inference. Both the weights of the U-Net and the feature embedding are optimized during training. The VQ-VAE is instead pre-trained [13] and frozen from the open-source Diffusers library [20]. For feasibility reasons, slices are resized to 256×256 pixels, so latent samples $\mathbf{x}_t$ are shrunken to 64×64 spatial dimensions.

3. Experiments

In this section the synthesis quality is evaluated globally (image-level) and locally (nodule-level) across 3 incremental conditioning configurations: (1) the conditional binary mask of the nodule is attached to the sample input, (2) the mask is additionally feed-forwarded through the cross-attention layers, and (3) cross-attention layers additionally access the embedded attributes of the nodule.

FID↓	Global	Nodules
(1) Input nodule mask	31.3069	49.6125
(2) + CA mask	29.6736	46.3218
(3) + CA mask and attributes	29.9101	47.0072

Table 2. Comparison of global and local synthesis quality on the Fréchet Inception Distance (FID) across 3 incremental configurations: (1) binary nodule mask attached to the input, (2) additionally injecting the mask into the cross-attention (CA) layers, and (3) adding the embedding vector of nodule attributes to the CA layers. All FID values were computed using 2048 synthetic samples.

3.1. Quantitative results

The Fréchet Inception Distance (FID) [21] is used to quantify the synthesis quality with respect to real samples. At the global level, the FID is computed using the entire image with dimensions of 256×256. At the nodule level, crops of 32×32 pixels centered around the nodules are used instead. The local and global impact of the localization mask in the synthetic images is also visually inspected.

According to Table 2, configuration 2 demonstrated superior performance in terms of FID score at both the global and nodule levels. This indicates that configuration 2 produces images that are statistically closer to the real images in terms of overall distribution, suggesting higher synthesis quality. Configuration 1, on the other hand, achieved the worst FID values, highlighting the importance of injecting the spatial mask into the cross-attention (CA) layers at different levels of the U-Net.

Configuration 3 uses the most conditional data in the CA layers, including the embedded nodule attributes. However, its FID is worse compared to configuration 2. We attribute this result to several factors. The resolution of the input images, downsampled to 256×256 for feasibility reasons, is likely insufficient to fully capture the nodule visual attributes, especially in the smallest nodules of the training data, corresponding to areas of 1-2 pixels after downsampling. Additionally, the noise in the four radiologist annotations of the LIDC-IDRI dataset may contribute to misleading results, as discrepancies of 1 or 2 scores on the scale of Table 1 are common (during training one of the four annotations is randomly selected to characterize each nodule). Another possible factor is that the embedding vectors may not be encoded in the optimal manner —other studies have explored encoding text captions instead of scalar values [22, 23].

Another key aspect shown in Table 2 is that all approaches achieved better FID at image-level compared to nodule-level. This is probably related to the input and the latent space resolution. Previous work on diffusion models for medical image synthesis has indicated that over-compressing the spatial resolution leads to a loss of small anatomic details [24].

Note that the relative differences between configurations are likely more significant than the absolute values in Table 2. This is because the image features used in the FID computation are extracted using the Inception model, which was trained on ImageNet images and not medical images [21]. Previous studies have pointed out this issue and explored alternative networks for the feature extraction step [25].

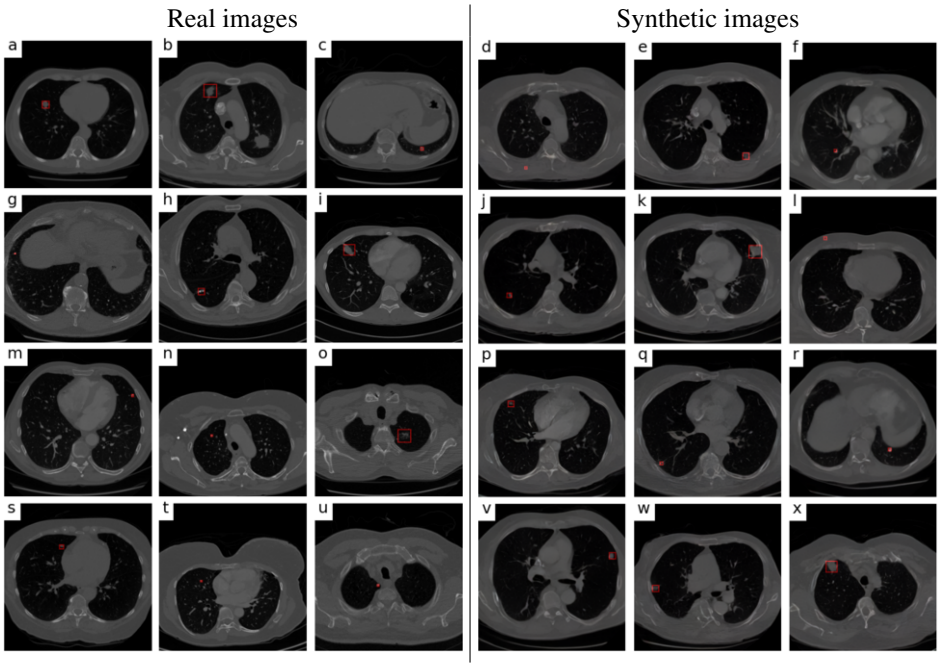

Figure 3. Examples of real and synthetic 2D CT slices containing a nodule. The location and size of the nodule are represented by a red bounding box corresponding to the binary mask that conditions the synthesis. Synthetic samples were generated using the latent diffusion pipeline configuration 2 listed in Table 2.

3.2. Qualitative results

Visual inspection of the synthetic images reveals additional points of interest. Notably, all configurations, including configuration 1, generate a nodule within the bounding box of the localization mask, even configuration 1. Figure 3 displays sets of real and synthetic images generated using the top-performing configuration of the experiments (configuration 2). The synthetic images exhibit similar anatomical diversity compared to the real images. However, there are some failure cases, such as images (d) and (l), where the binary mask points to an unrealistic area of the CT slice. For example, a nodule cannot be inside muscle tissue, in a bone, or outside the body. We attribute these failure cases to the choice of the mask $\mathbf{m}$ given a latent vector $\mathbf{x}_i$. Ideally, $\mathbf{m}$ defines the size and location of the nodule, while $\mathbf{x}_i$ defines the rest of the body structure (i.e., the CT slice depth, lung shape, tissue textures, etc.). However, the process we use to generate the synthetic mask $\mathbf{m}$ during inference is independent of $\mathbf{x}_i$. In particular, a probability of nodule presence is computed for each pixel of the CT slice based on the training masks, from which the centroid of the bounding box is drawn. The size of the bounding box is decided next by sampling from a second probability distribution over the square size of the training bounding boxes. Lastly, the synthetic nodule attributes are sampled from uniform distributions in a range of 1 to 5. Note that this procedure does not consider correlation between all these distributions modeling the nodule attributes as well as location and size, which might neglect relevant morphological and pathological factors.

Conclusions and future work

Synthetic medical images play a crucial role in ensuring privacy compliance and anonymization in medical data handling. By generating realistic, yet artificial, medical images, researchers and developers can access rich datasets without risking the exposure of sensitive patient information.

Our work experimented with pipelines of diffusion models to customize specific visual attributes in synthetic medical images. Different conditioning techniques were compared to address the position, size and appearance of nodules in 2D chest CT slices. The results are promising and show the advantages of using spatial masks and cross-attention layers to better exploit conditional information.

Future work involves exploring strategies to better capture the visual attributes of nodules, such as reducing spatial compression to preserve more image details [24], or encoding nodule attributes in alternative ways, like using text prompts [22, 23].

A final noteworthy point is that recent diffusion models operating directly in 3D have also shown promising results in medical imaging, at the expense of sometimes unaffordable computational resources. The described approach could be extended to a 3D data by means of 3D layers [24] or by adding loss terms to reinforce coherence between adjacent CT slices [12].

Acknowledgments

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.

References

[1] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, p. 102846, 2023.

[3] C. Han, Y. Kitamura, A. Kudo, A. Ichinose, L. Rundo, Y. Furukawa, K. Umemoto, Y. Li, and H. Nakayama, "Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection," in *2019 International Conference on 3D Vision (3DV)*, pp. 729–737, 2019.

[4] M. Nishio, C. Muramatsu, S. Noguchi, H. Nakai, K. Fujimoto, R. Sakamoto, and H. Fujita, "Attribute-guided image generation of three-dimensional computed tomography images of lung nodules using a generative adversarial network," *Computers in Biology and Medicine*, vol. 126, p. 104032, 2020.

[5] Q. Jin, H. Cui, C. Sun, Z. Meng, and R. Su, "Free-form tumor synthesis in computed tomography images via richer generative adversarial network," *Knowledge-Based Systems*, vol. 218, p. 106753, 2021.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[7] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, *et al.*, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, no. 1, p. 12098, 2023.

[8] M. Akrout, B. Gyepesi, P. Holló, A. Poór, B. Kincső, S. Solis, K. Cirone, J. Kawahara, D. Slade, L. Abid, *et al.*, "Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 99–109, 2023.

[9] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 35–45, 2022.

[10] A. L. Y. Hung, K. Zhao, H. Zheng, R. Yan, S. S. Raman, D. Terzopoulos, and K. Sung, "Med-cDiff: Conditional medical image generation with diffusion models," *Bioengineering*, vol. 10, no. 11, p. 1258, 2023.

[11] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *International Conference on Learning Representations*, 2022.

[12] H. Chung, D. Ryu, M. T. McCann, M. L. Klasky, and J. C. Ye, "Solving 3D inverse problems using pre-trained 2D diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22542–22551, 2023.

[13] 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*, pp. 10684–10695, 2022.

[14] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, p. 915–931, 2011.

[15] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "Data from LIDC-IDRI," 2015. The Cancer Imaging Archive.

[16] X. Rafael-Palou, A. Aubanell, I. Bonavita, M. Ceresa, G. Piella, V. Ribas, and M. A. G. Ballester, "Re-identification and growth detection of pulmonary nodules without image registration using 3d siamese neural networks," *Medical image analysis*, vol. 67, p. 101823, 2021.

[17] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

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

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[20] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf, "Diffusers: State-of-the-art diffusion models." https://github.com/huggingface/diffusers, 2022.

[21] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[22] 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.

[23] R. Montoya-del Angel, K. Sam-Millan, J. C. Vilanova, and R. Martí, "MAM-E: Mammographic synthetic image generation with diffusion models," *Sensors*, vol. 24, no. 7, p. 2076, 2024.

[24] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, *et al.*, "Denoising diffusion probabilistic models for 3d medical image generation," *Scientific Reports*, vol. 13, no. 1, 2023.

[25] 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*, pp. 117–126, Springer, 2022.