frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 11 November 2025
DOI 10.3389/fphy.2025.1661146



OPEN ACCESS

EDITED BY
Lipeng Ning,
Brigham and Women's Hospital and Harvard Medical School, United States

REVIEWED BY
Ziqi Yu,
Fudan University, China
Jun Lyu,
Yantai University, China

*CORRESPONDENCE
XiaoLing Luo,
xlluo@szu.edu.cn

RECEIVED 09 July 2025
REVISED 13 August 2025
ACCEPTED 29 August 2025
PUBLISHED 11 November 2025

CITATION
Tang G, Cai S, Meng X, Huo S, Wang M, Lu Z, Chen Z and Luo X (2025) High-fidelity medical image generation: controllable synthesis of high-resolution medical images via hierarchical fusion in vector-quantized generative networks.
*Front. Phys.* 13:1661146.
doi: 10.3389/fphy.2025.1661146



# High-fidelity medical image generation: controllable synthesis of high-resolution medical images via hierarchical fusion in vector-quantized generative networks

Guangfa Tang[1,2,3], Shanshan Cai[4], Xiangjun Meng[5], SiYan Huo[6], Mengbo Wang[1], Zichen Lu[3], Zhuokang Chen[5] and XiaoLing Luo[2]*

[1]School of Information and Intelligent Engineering, Guangzhou Xinhua University, Dongguan, China, [2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [3]Department of Health Industry Research, Dongguan Zhongke Institute of Cloud Computing, Dongguan, China, [4]Division of Biomedical and Life Sciences, Faculty of Health and Medicine, Lancaster University, Lancaster, United Kingdom, [5]Department of Urology, Dongguan People's Hospital, Dongguan, China, [6]Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI, United States

**Objective:** High-resolution medical images are scarce, and existing image generation methods perform poorly at high resolutions, struggling with the representation of small lesions, loss of detailed information, distortion of anatomical structure, high computational cost, and mode collapse. This study aims to develop a novel generative framework to address the challenges of high-resolution medical image generation.

**Methods:** Clinical X-ray data from 255 patients and a public dataset containing 1,657 lung CT images with lung nodules were collected. We propose a pioneering medical image generation method that employs a two-route synthesis strategy: a foreground generation route that utilizes a generative model from a single lesion image (SinGAN) to create new lesion configurations and structures while preserving the original patch distribution and a background generation route that utilizes a high-fidelity medical image generation model, high-resolution medical image (HiResMed) Vector-Quantized Generative Adversarial Network (VQGAN), which incorporates a hierarchical dual-path fusion block (HDFB) and integrates it into a VQGAN, trained on the collected data. The HDFB module combines a dual-path learning strategy: a residual path with skip connections to capture hierarchical dependencies and multi-scale textures and a multi-scale convolutional feedforward feature extraction module (MSConvFE) that preserves low-level anatomical features through localized detail enhancement. Finally, based on the location of lesions in historical data as prior knowledge to guide the fusion position of the synthesized lesions in the background image, a high-resolution synthetic medical image with small lesions is obtained. We compared our method with denoising diffusion model (DDM), StyleSwin, VQGAN, and SinGAN using Frechet Inception Distance (FID), learned perceptual image

patch similarity (LPIPS), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM). Two urologists participated in a visual Turing test to assess perceptual fidelity.

**Results:** The experimental results demonstrate that the proposed method achieves state-of-the-art performance, reducing FID by 43.3% (145.64 vs. 256.11) and LPIPS by 5% (0.48 vs. 0.51), enhancing the PSNR by 4% (59.03 vs. 56.54) and SSIM by 6% (0.67 vs. 0.63), and accelerating training convergence by 83% compared to baseline VQGAN. Clinicians misclassified 55% of synthetic images as real, validating their anatomical fidelity.

**Conclusion:** This study proposes a method for generating high-resolution medical images of small lesions. It not only ensures high-quality lesion generation but also allows controls over the number and location of lesions. Moreover, the innovative architecture enhances the detailed quality of anatomical structures and improves computational efficiency during training.

KEYWORDS

**controllable synthesis, two-route synthesis strategy, high-resolution medical image generation, hierarchical dual-path learning, detail preservation, high fidelity**

## 1 Introduction

High-resolution imaging is essential for numerous medical applications, including surgical navigation systems, high-precision diagnostic technologies, and early disease screening. Preoperative path planning for percutaneous nephrolithotomy for kidney stones requires comprehensive X-ray and CT imaging of the entire upper torso [1, 2]. Such applications require computation and processing of high-resolution images to provide detailed anatomical information, which is essential for accurate diagnosis and effective surgical planning. However, the limited maturity of high-resolution imaging pipelines, data silos across hospitals, and strict privacy/ethics constraints make case collection and annotation difficult, time-consuming, and expensive [3–6]. Despite the existence of synthetic data generation methods, existing methods mainly focus on low-resolution medical images of 128 pixels × 128 pixels or 256 pixels × 256 pixels and rarely exceed 512 pixels × 512 pixels [7], or the generated effects still lack high definition and anatomical fidelity [8]. Zhao et al. [9] and Cao et al. [10] explored transformer-based improvements for high-resolution synthesis, but these methods have not been validated on medical images. The increasing demand for large-scale imaging in various medical fields has gradually exposed the limitations of the existing methods, including high consumption of computing resources, loss of detailed information, and distortion of anatomical structures, making it hard to achieve clinical-grade detail under limited data [11, 12].

Generative adversarial networks (GANs) and their variants provide advanced medical image synthesis [13–17], yet the adversarial setup often prioritizes global fidelity for fooling the discriminator, which conflicts with the high-dimensional, sparse, and strongly constrained nature of medical images and can lead to mode collapse. Another significant issue is that the GANs often lose detailed information due to some convolution operations, such as downsampling, and their GANs to focus on global distributions. These problems are magnified in medical images where detailed information is particularly important and the resolution is high, resulting in severe distortion of anatomical structures in the reconstructed images and the inability to generate detailed information, such as small lesions and their texture features [18, 19]. In 2020, the denoising diffusion model (DDM) [20] achieved improved fidelity but required prohibitive computational resources, had long training and generation times, and could not easily meet the demands of immediate diagnosis. Moreover, high-frequency information is prone to over-smoothing during the denoising step and relies on large-scale, high-quality datasets to accurately learn data distribution, but medical images are usually limited in sample size, multimodal, and exhibit strong domain specificity. More recently, transformer-enhanced GANs, such as StyleSwin [21], have introduced attention mechanisms to better preserve structural details, but processing high-resolution images produces very long token sequences and incurs high self-attention costs [22]; the enlarged parameter space also complicates optimization and can yield divergent attention weights between the generator and discriminator, producing structural noise. Additionally, traditional evaluation metrics such as the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [23], and other pixel-level indicators cannot evaluate anatomical rationality, and distribution similarity metrics such as Frechet Inception Distance (FID) and IS ignore medical specificity. Therefore, clinical experts are also required to evaluate the diagnostic value of the generated images, but this process is complex and expensive. Consequently, the ability to generate high-resolution medical images with high fidelity has become a crucial research objective [24–26].

Recent studies have explicitly embedded anatomical or hierarchical priors to improve high-resolution medical image synthesis. Kang [27] proposed a method that explicitly introduces anatomical structure preservation loss, which significantly improves the consistency of organ contours during cross-domain migration. However, it is still limited to 256 × 256 resolution, and small lesion details are easily lost. Yu [28] proposed a HiFi-Syn, which includes multi-scale discriminators with layered supervision to achieve high-fidelity 512 × 512 MRI synthesis with superior structural fidelity to traditional GANs. However, the cascaded network doubles the number of parameters, placing heavy demands

on hardware computing resources, and its generalization to non-brain medical image synthesis requires further research. Yu [29] focused on cross-granular comparative representation of unsupervised lesion segmentation in medical images, and although it has unique explorations in lesion segmentation tasks, it does not involve the medical image generation link, has poor adaptability in multi-modal medical image data fusion scenarios, and cannot be directly applied to high-fidelity medical image synthesis tasks. Efficient-Vector-Quantized Generative Adversarial Network (VQGAN) [10] introduces a hierarchical transformer module that captures the global anatomical structure and local details through self-attention at different scales. However, the transformer's high computational complexity makes it difficult to process high-resolution images, such as 1,024 × 1,024 resolution, and it does not optimize feature weights for the sparsity of medical images, such as small lesions. Although these works demonstrate the value of medical priors, none address the dual challenge of sub-millimeter lesion fidelity and computational tractability at 1,024 pixel resolutions. Our high-resolution medical image (HiResMed)-VQGAN addresses this challenge through parameterized hierarchical fusion, explicitly preserving macro-anatomical structures via residual skip connections and micro-textures via MSConvFE, while reducing computational cost. This approach enables adaptive integration of macroscopic structures, such as spinal morphology, and microscopic lesions, such as pulmonary nodules. Decoupled foreground synthesis enables precise manipulation of lesion characteristics such as size and location, which is impossible in diffusion and transformer frameworks. It achieves the collaborative optimization of "high fidelity–high efficiency–controllability," thus providing a new paradigm for the synthesis of small-sample, high-resolution medical images in clinical practice.

VQGAN [30] is an advanced generative model proposed at the 2021 IEEE International Conference on Computer Vision and Pattern Recognition, which has demonstrated excellent performance in various applications such as high-resolution image generation, texture synthesis, and video generation, and it provides a partial solution [31, 32]. The advantage of the network's codebook [33, 34] discrete calculation mechanism is that it improves computational efficiency, but its disadvantage is that it fails to coordinate multi-scale feature learning, resulting in the inability to simultaneously preserve the macroscopic information of the anatomical structure and the microscopic structural information of tiny lesions. Therefore, the application of VQGAN to high-resolution medical image generation remains underexplored [33] [35–38].

In the context of high-resolution medical imaging, which is above 512 × 512 pixel images, the challenge is further compounded by the scarcity of cases, especially for small lesions that are critical for early detection of diseases such as kidney stones, early-stage tumors, and nodules. Traditional data augmentation approaches, such as downsampling, have been shown to result in the loss of critical details about small lesions, thereby compromising the quality of synthetic data. This loss of information can lead to suboptimal performance of AI models in detecting and diagnosing diseases at their earliest stages. The need for a novel approach that can generate high-resolution small-lesion medical images while preserving lesion details and maintaining data diversity is, therefore, imperative.

This study introduces a pioneering method that harnesses the power of the single-image generative adversarial network (SinGAN) [39] model for lesion generation as a foreground synthesis, complemented by an improved VQGAN model for background synthesis. We propose a novel approach to enhance the performance of the VQGAN by introducing a residual convolutional feedforward network module. This module is integrated into the encoder and decoder of a VQGAN framework. Unlike prior works, the hierarchical dual-path fusion block (HDFB) employs a dual-path learning strategy. An MSConvFE path preserves low-level anatomical structures. A residual path utilizes depth-wise convolutions and channel scaling to capture multi-scale textures. This integration accelerates the model's convergence, reducing training time and enhancing the detailed information in the generated high-resolution medical images. This work aims to fill the gap in the current literature and provide a robust solution for high-resolution medical image generation. Our contributions are summarized as follows:

1. Controllable two-route synthesis: We decouple training into a foreground lesion route and a background route and then compose them at inference with explicit control over the lesion size and location. This enables flexible recombination and substantially expands data diversity, which is particularly valuable for rare cases.
2. HDFB for high-fidelity, efficient background generation: We introduce a dual-path block that combines residual connections for multi-scale texture modeling with an MSConvFE path for low-level anatomical preservation, addressing the fidelity–efficiency trade-off at high resolution.
3. Architectural innovation: To the best of our knowledge, this is the first integration of a hybrid HDFB into a VQGAN encoder–decoder for high-resolution medical imaging, improving feature extraction, gradient propagation, computational efficiency, and training speed.
4. Strong potential for clinical application: Clinicians misjudged 55% of the synthetic images as real images, which proved that the synthetic images had high anatomical fidelity, which strongly verified the feasibility and effectiveness of the framework in clinical application and provided strong support for the application in actual medical scenarios.

# 2 Materials and methods

## 2.1 Datasets

Our study utilizes a public and a proprietary dataset. The public dataset is LIDC-IDRI, one of the most popular benchmarks in deep learning research, containing 1,657 lung CT images with lung nodules of 512 × 512 × 3 resolution. The proprietary dataset DGPH-KUB comprises 255 high-resolution kidney–ureter–bladder (KUB) X-ray images at 3,292 × 3,141 resolution collected from the Urology Department of Dongguan People's Hospital. In particular, this study has been authorized by the Ethics Committee of Dongguan People's Hospital (No.: KYKT2022-040). In order to eliminate the influence of other factors on our reported results, image processing software was used to adjust the resolution of the original image, and the

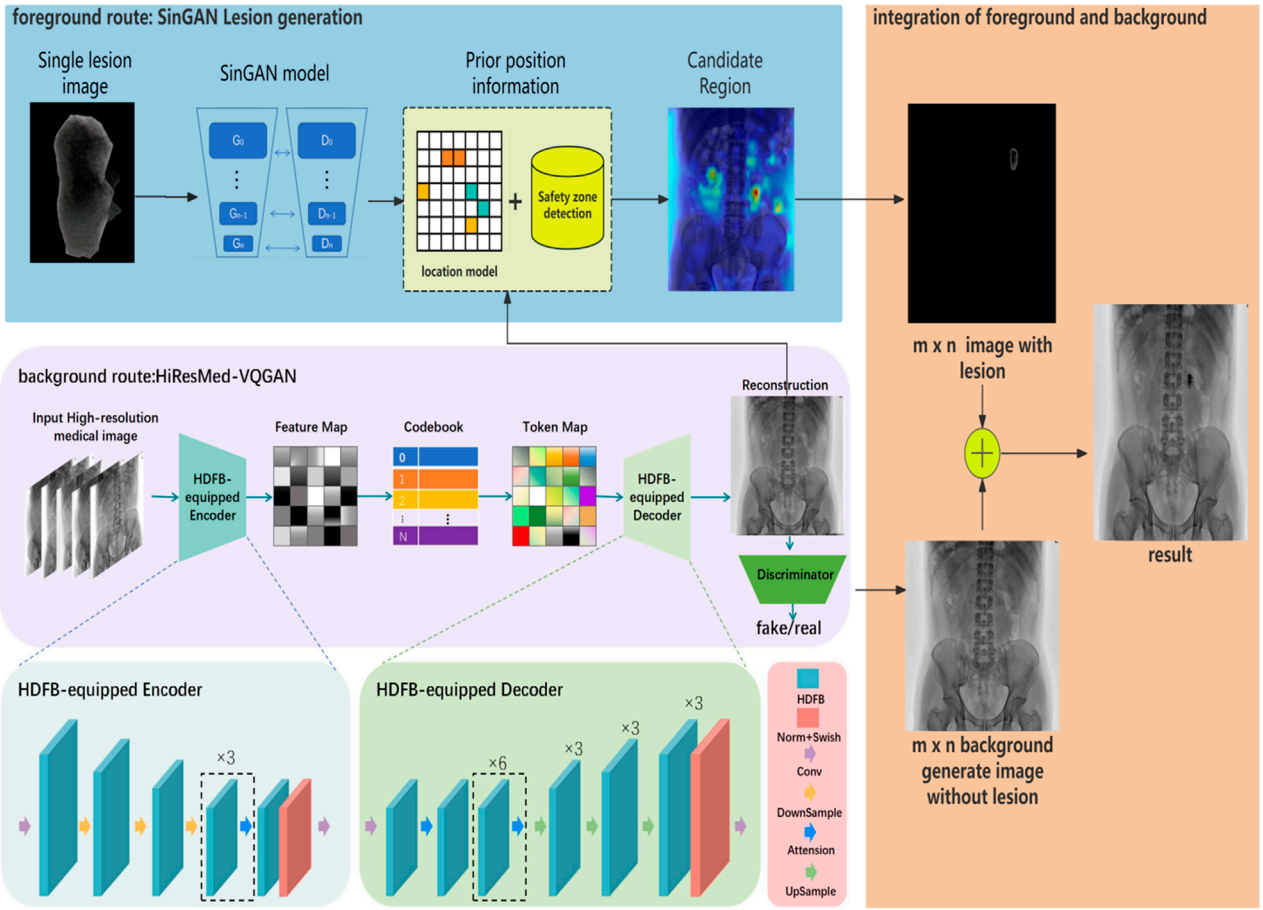


FIGURE 1
Overview of the controllable high-resolution medical image synthesis network. The proposed framework consists of two generation routes: the foreground and background. The foreground uses the SinGAN model to generate lesion images, and the background image is generated using the HiResMed-VQGAN model. Finally, the foreground and background are fused to obtain the result.

images were uniformly changed to a resolution of 1,024 × 1,024 × 3. This dataset is unique in its focus on high-resolution X-ray images and is particularly valuable for research on kidney stone diagnosis and surgical navigation systems.

## 2.2 High-resolution medical image VQGAN network

The high-fidelity, high-resolution medical image VQGAN network is proposed as a novel architecture that integrates HDFB into the encoder and decoder of VQGAN. The HDFB proposed in this paper is inserted into the encoder and decoder of VQGAN, and the specific construction method is shown in Figure 1.

The generation process is as follows. First, the real high-resolution medical image I is input into the HDFB-equipped encoder. The purpose of this process is to perform multi-scale feature extraction and latent space mapping. This process is mainly divided into two stages. The first stage is layered convolutional downsampling, which uses convolutional blocks with residual connections to perform downsampling three times. This process gradually compresses the spatial resolution from 1,024 × 1,024 to 16 × 16 while increasing the number of channels from 3 to 512, layer by layer, forming a feature pyramid that contains contextual information at different scales. The second stage is the processing of the multi-scale convolutional feedforward feature extraction. Through the parallel structure of depth-wise separable convolution and residual convolution, multi-scale features from the local texture to the global structure are captured, and the features of different branches are fused across scales by element-by-element addition.

The feature map M processed by the HDFB-equipped encoder is compressed into a continuous latent space representation through a 1 × 1 convolution and then mapped to a discrete codebook space through a vector quantization layer. The codebook is a set of predefined vectors that maps the continuous latent space to the discrete codebook space [16]. Let $B = \{b_n \in R^D\}_{n=1}^N$ denote a codebook containing N entries, with each entry $b_i$ being a D-dimensional trainable embedding with random initialization. These vectors are continuously updated during the training process so that the model can learn a discrete representation to better represent the features of the input image. Subsequently, the quantizer in the codebook maps M to a token map $M^t$, where each token is an entry in B based on the cosine distance between M and B.

Then, the HDFB-equipped decoder reconstructs the original image from the token map $M^t$. This process is divided into two stages. The first stage is layered deconvolution upsampling. Deconvolution blocks with skip connections are used for upsampling three times to gradually restore the spatial resolution from 16 × 16 to 1,024 × 1,024, and the number of channels is compressed from 256 to 3 layer by layer. After each upsampling, a residual convolution feedforward network is inserted, and the weights of features of different scales are dynamically adjusted through the channel attention mechanism. The second stage is the detail enhancement convolution calculation. After upsampling the last layer, the high-frequency texture of the reconstruction is captured through the parallel structure of the depth-separable convolution and the residual convolution, and a high-quality reconstructed image is generated.

Finally, the discriminator, composed of two convolutional layers, a normalization operation, and an activation function, calculates the authenticity probability of the real image and the reconstructed image, and it distinguishes the authenticity of local details, medium-scale structures, and global layout of the generated image, respectively. The ultimate goal is to continuously optimize the generator based on the feedback from the discriminator, enabling the generator to produce reconstructed images capable of deceiving the discriminator.

The entire network is optimized using a combination of losses, which is expressed as follows (Equation 1):

$$L = \|\hat{I} - I\|^2 + \alpha\|sg(M^t) - M\| + \gamma\|sg(M) - M^t\| + L_p + L_{GAN}, \quad (1)$$

where sg(·) denotes the stop-gradient operation. $\|\hat{I} - I\|^2$, $\alpha\|sg(M^t) - M\| + \gamma\|sg(M) - M^t\|$, $L_p$, and $L_{GAN}$ represent the reconstruction loss, quantization loss, VGG-based perceptual loss [27], and GAN loss [27], respectively. The hyper-parameters $\alpha$ and $\gamma$ are, respectively, set to 1.0 and 0.33 by default.

## 2.3 Hierarchical dual-path fusion block

The HDFB is designed to optimize feature representation and gradient propagation in high-resolution medical image synthesis. The structure is shown in Figure 2. By integrating sequential normalization, activation, and multi-sale feature learning with skip connections, the HDFB ensures both anatomical fidelity and computational efficiency. In the HDFB, the input data tensor, which represents the height, width, and number of channels, is first passed through a GroupNorm–SiLU pair (Equations 2, 3):

$$X_{norm1} = \text{GroupNorm}(X), \quad (2)$$

$$X_{act1} = \text{SiLU}(X_{norm1}). \quad (3)$$

We use a smoothly gated non-linear activation defined as follows (Equation 4):

$$\text{SiLU}(x) = \frac{x}{1 + e^{-x}}. \quad (4)$$

Here, x is the input, and the function is derivable over the whole real number field, which makes the gradient smoother and continuous during the backpropagation process, so there is no problem of gradient disappearance, and it helps improve the stability and convergence speed of training. The non-monotonicity property can, therefore, switch between positive and negative values, providing richer information-processing ability. It can better capture the detailed information of the anatomical structure. SiLU [40] preserves gradient information better than ReLU, especially for subtle features. After applying 2D convolutional layers to extract local spatial features, we repeat normalization and activation (Equations 5, 6), thus amplifying discriminative features while suppressing noise.

$$X_{norm2} = GroupNorm(X_{act1}), \quad (5)$$

$$X_{act2} = SiLU(X_{norm2}). \quad (6)$$

The final output is fed into the MSConvFE block (Equation 7) to enhance multi-scale feature learning:

$$F_{convffn} = ConvFFN(X_{act2}). \quad (7)$$

In order to mitigate vanishing gradients and preserve low-frequency anatomical structures, we introduce a skip connection (Equation 8):

$$Y = X + F_{convffn}. \quad (8)$$

The residual block, previously used in both the encoder and decoder, was replaced by the HDFB module, resulting in several significant improvements. First, the convolutional feedforward network further analyzes and processes these details by retaining low-level details in residual blocks. In particular, it contains multiple convolutional layers and fully connected layers, and its complex structure can capture finer-grained patterns and relationships in the data. When processing medical images, the convolutional feedforward network can perform an in-depth analysis of details, such as texture and density changes in organs, soft tissues, and bone regions, thereby extracting more subtle features. Through in-depth analysis, this detailed information enables the decoder to reconstruct high-resolution medical images with greater accuracy, thereby enhancing overall network performance in terms of reconstruction quality and generation fidelity. The enhanced feature extraction in the encoder and the improved detail-handling in the decoder result in more accurate reconstructions and higher-quality generated outputs. Second, the HDFB-equipped VQGAN is more robust in terms of noise and input variations. The skip connections in HDFB and its non-linear transformation capabilities help the network to better adapt to different input conditions, which is beneficial in real-world applications where the input data may be corrupted or have diverse characteristics. Third, the combination of HDFB and VQGAN can lead to more efficient training. The HDFB blocks' ability to mitigate the vanishing gradient problem and their effective feature processing can accelerate the convergence of the network during training, thus reducing the overall training time and computational resources required.

To enhance the multi-scale feature learning and preserve fine-grained details simultaneously, we design a hybrid architecture for the multi-scale convolutional feedforward feature extraction module, addressing the dual challenges of anatomical coherence and texture fidelity in high-resolution medical image generation. While

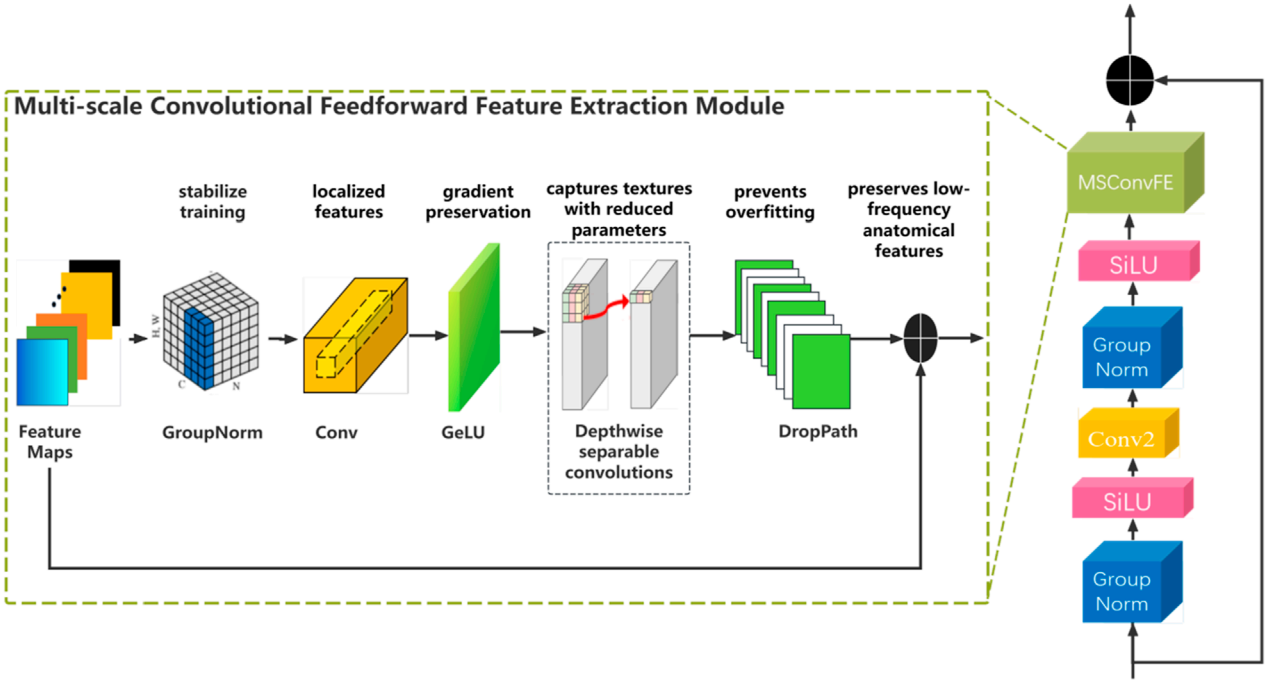


FIGURE 2
Structural diagram of HDFB. A residual path utilizes skip connections to capture hierarchical dependencies and multi-scale textures. The multi-scale convolutional feedforward feature extraction module preserves low-level anatomical features through localized detail enhancement.

TABLE 1 Training parameters utilized in the HiResMed-VQGAN model.

| Parameter item | Value |
|---|---|
| Batch size | 8 |
| Epochs | 3,500 |
| Loss function | Reconstruction loss + adversarial loss + perceptual loss |
| Learning rate | 2.25e-05 |
| Number codebook vectors | 1,024 |
| Optimizer | Adam (eps = 1e-08, betas = (0.5, 0.9)) |

classical feedforward modules focus on global context aggregation, our MSConvFE uniquely integrates localized detail enhancement, hierarchical multi-scale modeling, and improved computational efficiency through a dual-path structure. The architectural components are shown in Figure 2. Moreover, the 2D convolutional layer can extract local spatial features. Group normalization [41] divides the channels into several groups, calculates the mean and variance within each group for normalization (Equation 9), and calculates the formula as follows:

$$GN(x) = \frac{x - \mu_G}{\sqrt{\sigma_G^2 + \varepsilon}} \cdot \gamma + \beta. \quad (9)$$

Here, G represents the number of groups; $\mu_G$ and $\sigma_G^2$ represent the mean and variance of channels in each group, respectively; and γ and β represent the parameters of learnable scaling and translation, respectively. This method does not depend on the batch size and helps stabilize the training process. In order to prevent the gradient explosion problem and improve the convergence speed of the model, we introduce a batch normalization operation after SiLU activation function processing. The GeLU activation function introduces nonlinearity through the probability of a Gaussian distribution, and its calculation formula is as follows (Equation 10):

$$GeLU(x) = x \cdot \Phi(x). \quad (10)$$

Here, x is the input, and Φ(x) is the cumulative distribution function of the normal distribution. Due to its high computational complexity, an approximate expression is often used to simplify the calculation [42].

$$GeLU(x) \approx 0.5 \cdot x \cdot \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right). \quad (11)$$

The nonlinear nature of the GeLU activation function (Equation 11) can enhance the model's ability to fit complex data, and the activation degree of the GeLU function is proportional to the size of the input value, which is helpful for the learning and generalization of the model. In particular, this paper introduces a depth-wise separable convolution layer, which is composed of a depth-wise convolution and a point-wise convolution, where the depth-wise convolution is a convolution operation performed on each channel of the input feature map. Specifically, for

TABLE 2 Quantitative results (mean ± SD) compared with state-of-the-art methods on two datasets. Lower FID/LPIPS and higher PSNR/SSIM indicate better performance. The best result is shown in bold, and the second-best result is underlined; significance testing is based on a paired t-test. Results are averaged over five independent samplings per background.

| Method | Dataset | FID↓ (mean ± SD) | LPIPS↓ (mean ± SD) | PSNR↑ (mean ± SD) | SSIM↑ (mean ± SD) |
|---|---|---|---|---|---|
| DDM | DGPH-KUB | 178.31 ± 8.24 | **0.46 ± 0.03** | **63.10 ± 1.27** | **0.69 ± 0.03** |
| | LIDC-IDRI | 171.53 ± 7.91 | **0.46 ± 0.02** | 63.57 ± 1.19 | 0.55 ± 0.02 |
| StyleSwin | DGPH-KUB | 243.12 ± 10.56 | 0.49 ± 0.04 | 62.63 ± 1.32 | 0.64 ± 0.04 |
| | LIDC-IDRI | 205.92 ± 9.83 | 0.47 ± 0.03 | 64.36 ± 1.08 | 0.60 ± 0.06 |
| VQGAN | DGPH-KUB | 256.11 ± 12.37 | 0.51 ± 0.05 | 56.54 ± 1.45 | 0.63 ± 0.02 |
| | LIDC-IDRI | 280.23 ± 11.72 | 0.57 ± 0.06 | 61.33 ± 1.21 | 0.52 ± 0.03 |
| SinGAN | DGPH-KUB | 277.11 ± 13.15 | 0.48 ± 0.04 | 58.58 ± 1.52 | 0.65 ± 0.05 |
| | LIDC-IDRI | 268.07 ± 12.89 | 0.47 ± 0.03 | 64.00 ± 1.15 | 0.59 ± 0.06 |
| Ours | DGPH-KUB | **145.64 ± 5.23★★★** | 0.48 ± 0.02★★ | 59.03 ± 0.95★★ | 0.67 ± 0.03★ |
| | LIDC-IDRI | **180.29 ± 6.87★★★** | 0.47 ± 0.02★★ | **64.46 ± 0.84★** | 0.59 ± 0.03★ |

Significance test: "★★★" represents $p < 0.05$; "★★" represents $p < 0.01$; "★" represents $p < 0.001$ (vs. baselines).

an RGB three-channel image, the depth-wise convolution uses three single-channel convolution kernels to convolve the three input channels, respectively, and output the feature maps of the three channels. In this way, the convolution kernel of each channel only needs to process the data of one channel, which greatly reduces the number of parameters and the amount of calculation. Point convolution is a 1 × 1 convolution operation applied to the output of the depth-wise convolution to merge the features of different channels. Specifically, the point-wise convolution uses a 1 × 1 convolution kernel to convolve the output of the depth-wise convolution, fuses the features of different channels, and generates the final output feature map. Therefore, this combination can not only significantly improve the performance of the model but also optimize the computing resources.

For an input feature map $X \in R^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels, respectively, the MSConvFE processes the feature as follows (Equation 12):

$$\begin{cases} X_{norm} = GroupNorm(X) \\ F_{conv} = Conv2D_{3x3}(X_{norm}) \\ F_{act} = GeLU(F_{conv}) \\ F_{DW} = DWConv_{5x5}(F_{act}) \\ F_{channel} = Conv2D_{1x1}(F_{DW}) \\ F_{drop} = DropPath(F_{channel}) \\ Y = X + F_{drop} \end{cases} . \quad (12)$$

The skip connection retains raw anatomical features, while the processed branch drop refines high-frequency details. This modification ensures that both high-frequency details and low-frequency features are preserved, which is critical for high-resolution medical image synthesis.

## 2.4 Lesion synthesis

In this study, the preprocessed lesion images obtained from the previous step serve as the input to the lesion generation model. We employ the SinGAN model for generating synthetic lesion images. SinGAN is a single-image GAN that is particularly well-suited for medical image synthesis tasks where data scarcity is a common challenge. Unlike traditional GANs that require large datasets for effective training, SinGAN can achieve convergence with only a single training image, making it an ideal choice for generating lesion images in scenarios with limited data availability. Therefore, the problem of poor generation quality due to insufficient data volume can be avoided. Moreover, data scarcity and data silos have always been common problems in medical data. The SinGAN model is based on a pyramid of fully convolutional GANs, where each level of the pyramid learns to capture the statistical properties of the input image at different scales. This hierarchical structure enables the model to generate high-quality synthetic images that preserve the fine-grained details of the original lesion. The key advantage of SinGAN lies in its ability to learn from a single image, which is particularly beneficial for medical imaging applications where annotated datasets are often limited. Given a preprocessed lesion image $I_{lesion}$, the SinGAN model generates synthetic lesion images by learning the distribution of the input image across multiple scales. The generation process can be formally described as follows (Equation 13):

$$I_{syn\_lesion} = SinGAN(I_{lesion}), \quad (13)$$

where $SinGAN(\cdot)$ represents the SinGAN generator network.

During the training stage, the SinGAN model is trained on a single preprocessed real lesion image, learning a multi-scale representation of its texture and structure. During inference, no real lesion image is fed into the network. Instead, new lesions are

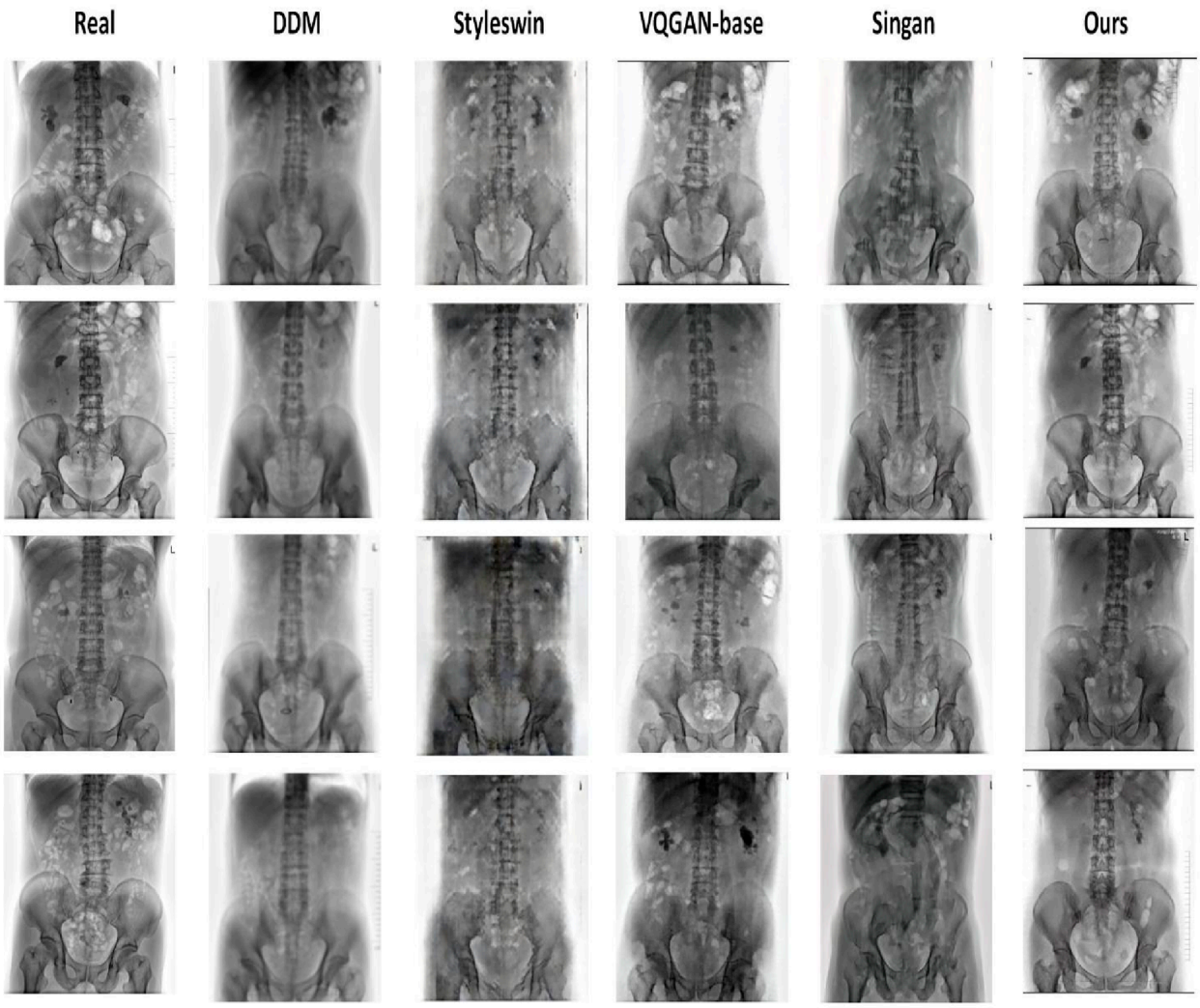


FIGURE 3
Generation performance of our method compared with the state-of-the-art on DGPH-KUB X-ray images.

synthesized by sampling random noise at the coarsest scale and progressively refining it through the trained scales. This process allows lesion generation to be conditioned solely on the learned internal distribution of the training exemplar without reusing the original image.

To generate high-resolution medical images containing small lesions, the synthetic lesion images $I_{syn\_lesion}$ are placed into a high-resolution background image $I_{canvas}$. The background image $I_{canvas}$ is initialized as a zero matrix with the same dimensions as the background high-resolution image $I_{background}$, which is the output of the background route. The placement of the synthetic lesions is guided by prior knowledge of lesion locations derived from historical patient data.

To ensure anatomically plausible placement of synthetic lesions within the background image $I_{canvas}$, we introduce a dual-model prior position information framework. This framework combines the location model and danger zone detection model. The location model is a YOLOv11-based detector trained on historical lesion annotations to probabilistically predict likely lesion locations. Formally, for the background image $I_{background}$, the model outputs the following set of candidate coordinates (Equation 14):

$$C_{loc} = (x_i, y_i)|i = 1, ..., k), \tag{14}$$

where each coordinate represents a high-probability lesion occurrence region learned from historical distributions.

The danger zone detection model is a U-Net segmentation network that is trained to identify anatomically implausible regions (e.g., bones, major vessels, and spinal column in KUB X-rays and pleural surfaces in lung CTs). The model generates a binary mask $M_{danger}$, where (Equation 15)

$$M_{danger}(x,y) = \begin{cases} 0\ (\textit{forbidden zones: spine, pelvis, etc.}) \\ 1\ (\textit{permissible regions}) \end{cases}. \tag{15}$$

We design a candidate region filtering $C_{final}$ for the final candidate region calculation (Equation 16). $C_{final}$ is derived by

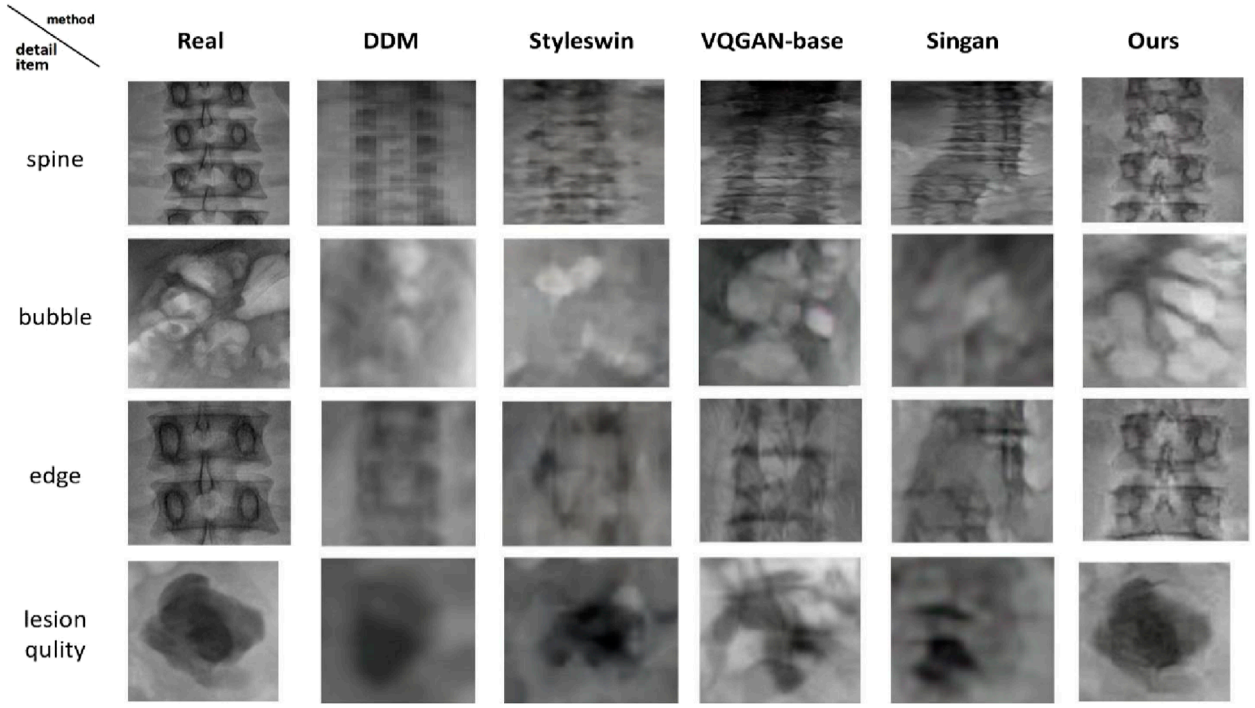


FIGURE 4
Details of the performance of our method compared with the state-of-the-art on DGPH-KUB X-ray images.

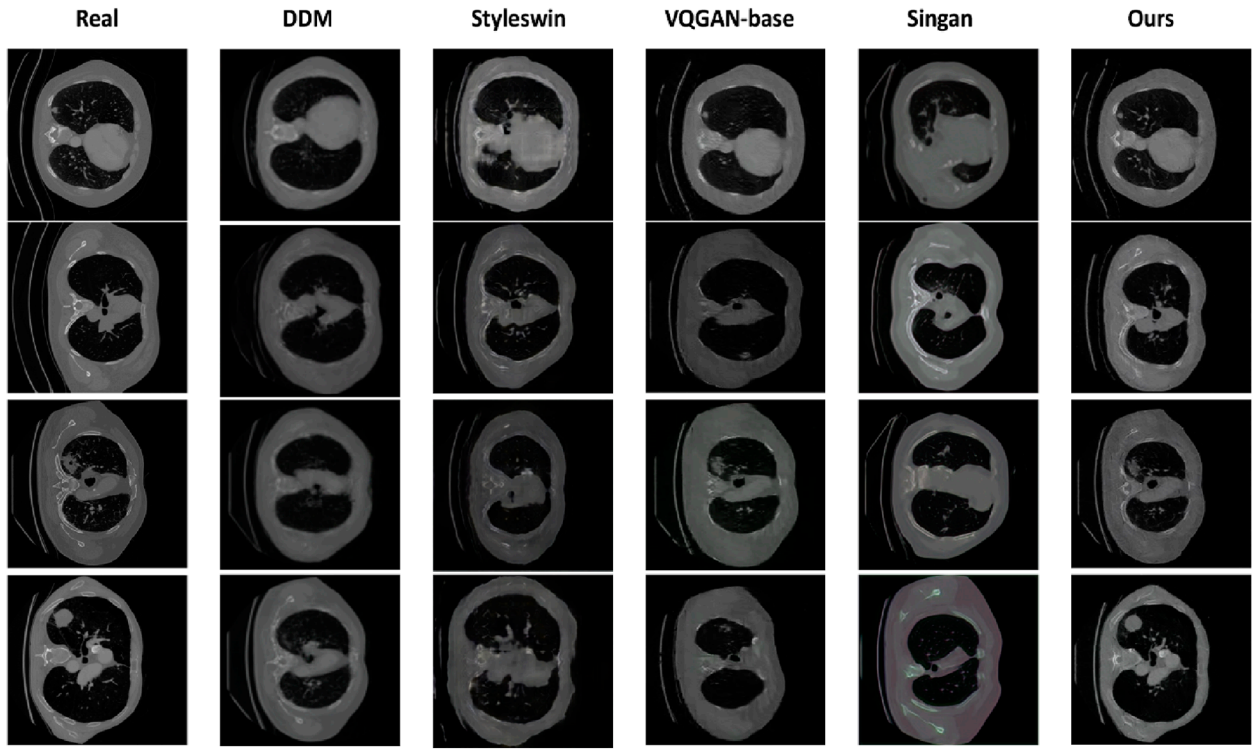


FIGURE 5
Generation performance of our method compared with the state-of-the-art on LIDC-IDRI data.

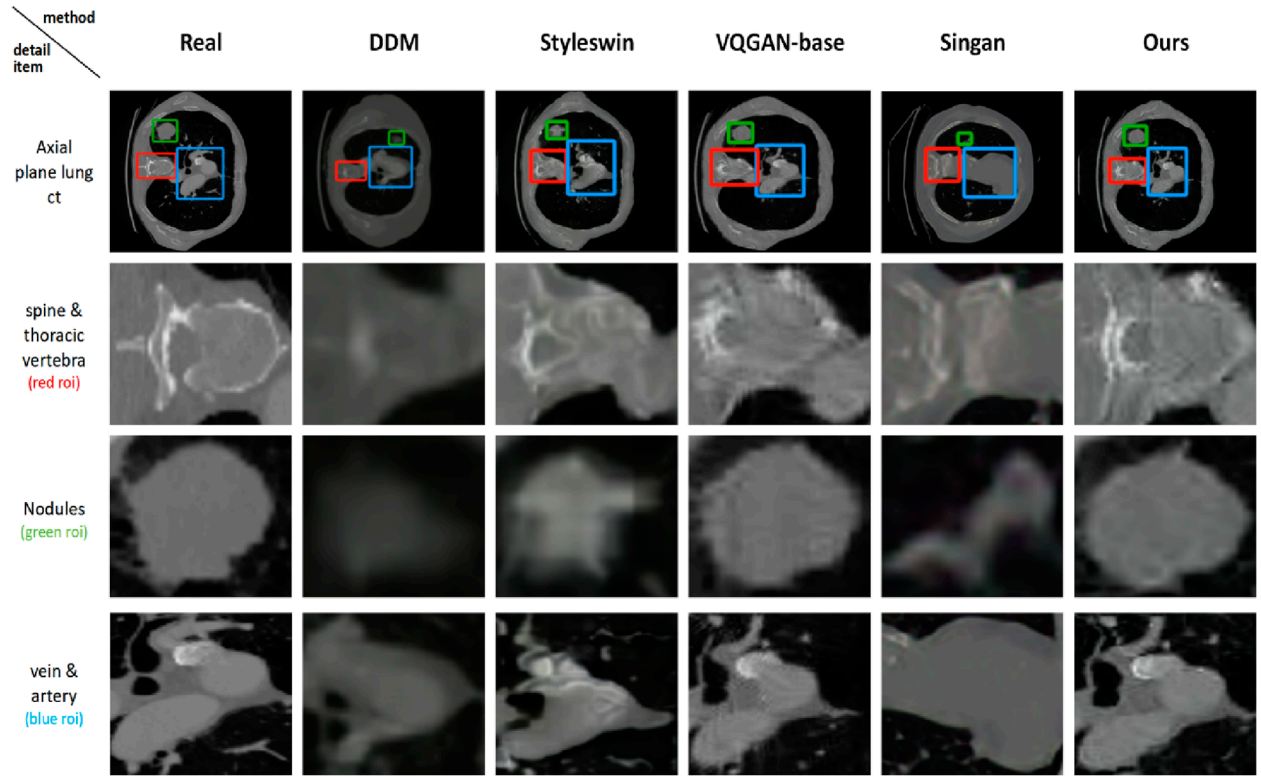


FIGURE 6
Details of the performance of our method compared with the state-of-the-art on LIDC-IDRI data.

TABLE 3 Confusion matrices of the senior and intermediate urologists in the visual Turing test. The urologists completed the authenticity judgment on 200 KUB X-ray images, which included both real and synthetic images. "Synt" denotes synthesized images, and P and N indicate positive and negative classes.

| 2 urologists prediction / Actual | Senior professional | | Intermediate professional | |
|---|---|---|---|---|
| | Real(P) | Synt(N) | Real(P) | Synt(N) |
| Truth Real(P) | 85 | 15 | 65 | 35 |
| Truth Synt(N) | 55 | 45 | 55 | 45 |

TABLE 4 Results of the visual Turing test (n = 200). Analysis of the sensitivity, specificity, accuracy, and consistency of the senior and intermediate urologists in the authenticity judgment of synthetic KUB X-ray images.

| Metric | Senior professional | Intermediate professional |
|---|---|---|
| Sensitivity (TPR, %) | 85.0 | 65.0 |
| Specificity (TNR, %) | 45.0 | 45.0 |
| Accuracy (%) | 65.0 | 55.0 |
| Kappa (agreement) | 0.3 | 0.1 |
| *p*-value (vs. 50%) | <0.001 | 0.046 |

imposing anatomical constraints as follows:

$$C_{final} = \{(x_i, y_i) \in C_{loc} | M_{danger}(x_i, y_i) = 1\}. \quad (16)$$

Users may either manually select the coordinates of interest from these safe regions or allow random sampling to determine the final lesion insertion regions, $R_{lesion}$. Finally, the pixel values of $I_{syn\_lesion}$ are filled into $R_{lesion}$. $I_{canvas}$ contains $I_{syn\_lesion}$ information, and we may also define $I_{canvas}$ as $I_{foreground}$.

Specifically, this process ensures that the generated lesions are anatomically plausible and consistent with real-world medical imaging scenarios. The final high-resolution image $I_{output}$ containing the synthetic lesion is obtained by combining $I_{foreground}$ and $I_{I_{background}}$ using a pixel-wise addition operation, which is as follows (Equation 17):

$$I_{output}(x, y) = I_{background}(x, y) + I_{syn\_lesion}(x, y). \quad (17)$$

In particular, the pixel values in the $R_{lesion}$ regions of $I_{background}$ need to be set to 0.

Since this merging method is pixel-level, there will inevitably be excessive seams between the edge of the lesion and the background.

TABLE 5 Variability analysis under fixed lesion/varying background and fixed background/varying lesion conditions. The metrics (mean ± SD) show minimal fluctuations, confirming robustness against sampling stochasticity.

| Metric | Fixed lesion–varying background (mean ± SD) | Fixed background–varying lesion (mean ± SD) |
|---|---|---|
| FID | 147.21 ± 4.56 | 145.64 ± 3.82 |
| LPIPS | 0.49 ± 0.03 | 0.48 ± 0.02 |
| PSNR | 58.76 ± 1.12 | 59.03 ± 0.87 |
| SSIM | 0.69 ± 0.03 | 0.68 ± 0.03 |

We leverage Sobel edge detection and Gaussian blurring to achieve natural pixel-level continuity.

First, the contour of the synthetic lesion is extracted using the Sobel operator, which computes gradient magnitudes along both the horizontal and vertical directions to identify edge pixels. This step isolates the boundary between the lesion and its surrounding area, ensuring precise targeting of the transition region. Subsequently, a Gaussian blur (with a kernel size of 3 × 3 and standard deviation σ = 1.0, which is empirically optimized for medical image textures) is applied to the detected edge. This blurring operation creates a gradual intensity transition between the lesion and the background: edge pixels are weighted by a Gaussian distribution, with values smoothly decreasing from the periphery of the lesion to the background.

This approach minimizes abrupt intensity changes at the lesion boundary, thus enhancing the visual coherence of the integrated image without introducing excessive computational overhead.

# 3 Results

## 3.1 Implementation details

Weights were initialized using torch.nn.init (mean 0 and standard deviation 0.02), and training was conducted for up to 3,500 epochs. The codebook dimension for vector quantization is selected as 256 to align with the feature dimension of the encoder output. More training hyperparameters are summarized in Table 1. All experiments were conducted on a single NVIDIA V100 GPU with 32 GB of memory. We synthesize images at 1,024 × 1,024 resolutions for both datasets. Owing to computational constraints, DDM was trained to generate 128 × 128 × 3 images, which were subsequently upsampled to 1,024 × 1,024 × 3 for comparison.

## 3.2 Quantitative evaluation

We used several quantitative metrics to assess the quality of the generated high-resolution medical images, which are detailed as follows.

1. Frechet Inception Distance (FID)

The FID [43] calculates indicators of the quality and diversity of the generated image by comparing the distribution of the generated image with the real image in a specific space. The definition is as follows (Equation 18):

$$FID = \left\|\mu_r - \mu_g\right\|^2 + Tr\left(\sum_r + \sum_g - 2\left(\sum_r \sum_g\right)^{1/2}\right), \tag{18}$$

where $\mu_r$ and $\sum_r$ are the mean and covariance matrix of real image features, respectively, and $\mu_g$ and $\sum_g$ are the mean and covariance matrix of the generated image features, respectively. Tr is the trace of a matrix.

2. Learned perceptual image patch similarity (LPIPS)

The LPIPS [44] is a perceptual similarity measure based on deep learning, which is used to measure the perceptual difference between two images. Its definition is formulated as follows (Equation 19):

$$LPIPS = \sum_L \frac{1}{H_l W_l} \sum_{h,w} \left\|\omega_l \otimes \left(y_h^l - y_o^l\right)\right\|_2^2, \tag{19}$$

where $y^l$ is the lth feature maps. It is normalized with respect to the initial feature map $y_o^l$ in the channel dimension using unit normalization, and the number of activated channels is scaled using $\omega_l$; the L2 distance value is then calculated. Here, ⊗ is the dot product operation.

3. Peak signal-to-noise ratio (PSNR).

The PSNR measures the pixel-wise similarity between the generated images and the ground truth. The definition is as follows (Equation 20):

$$PSNR = 10 \times \log_{10}\left(\frac{(2^n - 1)^2}{MSE}\right), \tag{20}$$

where n is the number of sampling points. In this study, we process the RGB images, so n = 24. MSE stands for the mean squared error, which is defined as follows (Equation 21):

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X(i,j) - Y(i,j))^2, \tag{21}$$

where H × W is the number of pixels in the image, H and W are the length and width of the image, X is the enhanced image, and Y is the real clear image.

4. Structural similarity (SSIM)

The similarity between two images is measured from three dimensions: brightness, contrast, and structure. The value range is [0, 1], and the closer the value is to 1, the more similar it is. The calculation formula is as follows (Equation 22):

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{22}$$

Here, $\mu, \sigma^2, \sigma\{xy\}$ represent the local mean, variance, and covariance of the image, respectively.

To rigorously evaluate the stochasticity induced by lesion sampling, we generated five independent samples per test background (using different random seeds) and report the mean ± standard deviation (SD) in Table 2. Three key findings emerged.

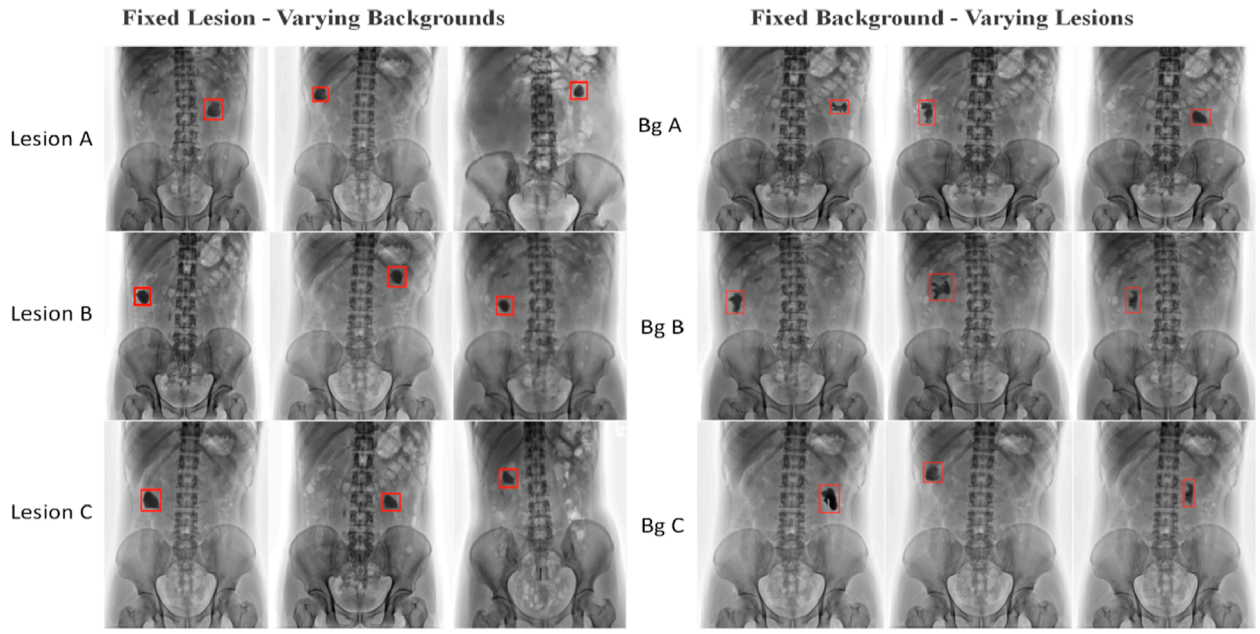


FIGURE 7
Visual examples of generated images under a fixed lesion with varying backgrounds (left), where each row represents the same lesion image, and a fixed background with varying lesions (right), where each row represents the same background image (Bg is short for background, and the red boxes indicate the lesion images).

TABLE 6 SSIM comparison for computing outside the lesion mask.

| Method | Dataset | SSIM (no mask) | SSIM(mask lesion) |
|---|---|---|---|
| VQGAN | DGPH-KUB | 0.63 | 0.62 |
| OUR | DGPH-KUB | 0.67 | 0.66 |

First, in terms of lesion generation fidelity, on KUB X-ray data, the FID decreased to 145.64 ± 5.23, which is a significant 43.3% reduction compared to the baseline VQGAN ($p < 0.001$), and the standard deviation (SD = 5.23) was the lowest among all the methods, demonstrating optimal generation stability. Although DDM performed well in terms of LPIPS (0.46 ± 0.03) and PSNR (63.10 ± 1.27), it failed to generate visible lesions (Figures 3, 4). Our method, on the other hand, successfully synthesized small lesions while preserving the anatomical structure (LPIPS = 0.48 ± 0.02, PSNR = 59.03 ± 0.95). Second, in terms of cross-modal generalization ability, in the CT dataset (LIDC-IDRI), the FID (180.29 ± 6.87) of this method significantly outperformed all baselines ($p < 0.001$), and the PSNR (64.46 ± 0.84) was the best ($p < 0.05$). The variability caused by the prior lesion (LPIPS fluctuation SD ≤ 0.02) is far below the human eye perception threshold (LPIPS > 0.05 can be perceived [44]), proving the clinical reliability of the synthesized results. Third, statistical significance was verified by paired t-tests (Bonferroni correction, $\alpha = 0.05$). The FID improvement of this method was significant for all baselines ($p < 0.001$), and PSNR was significantly better than DDM on CT data ($p < 0.05$). Due to computational resource limitations, DDM can generate images only at 128 × 128 resolution, which must then be upsampled to 1,024 × 1,024. This results in high PSNR values while failing to capture true high-resolution details (Figure 5).

## 3.3 Qualitative comparison with state-of-the-art approaches

Figures 3, 4 present KUB X-ray results. Our method synthesizes target images with accurate anatomical structures and fine lesion details. The details of the synthetic KUB X-ray images are displayed in Figure 4. DDM produces structurally reasonable yet overall blurry images and often fails to generate lesion signals. The StyleSwin model produces inferior quality results, and the structure of the spinal cord is unreasonable and unclear. In this comparative experiment, the target map generated by the VQGAN model demonstrates better overall quality but lacks sharp bone edges and clear lesion depiction. The generation effect of SinGAN is not satisfactory, and additionally, the spine is broken, indicating that the model fails to learn global anatomical logic. Overall, our results are visually closest to real images, providing clearer cortical bone boundaries, a more realistic lesion appearance, and more natural representation of intrabody bubbles. Comparison of generation details (Figure 4) shows that our method most closely resembles real images in the synthetic quality of the spine and intrabody bubbles, whereas the results of other methods deviate substantially from realism. A crucial point is that the texture, edge, and clarity of kidney stone lesions generated by the proposed method are superior.

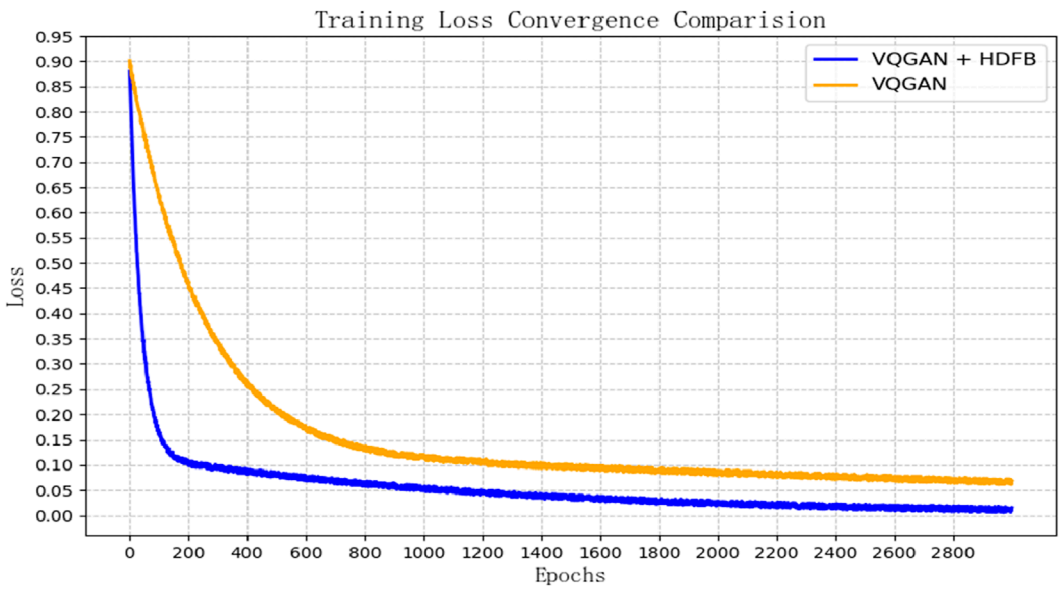


FIGURE 8
Training loss convergence comparison on the KUB images dataset.

Figures 5, 6 show the visual comparison of the generated CT medical images and their details using our proposed method and other state-of-the-art techniques. Similarly, the generation effect of DDM is still vague, and the information on pulmonary blood vessels is not generated. The generation result of StyleSwin is only at a normal level for pulmonary blood vessels but severely distorted for other tissue structures. The texture features generated by the VQGAN model are better than those mentioned above, but the pulmonary vascular information is almost missing. SinGAN generates higher-quality bone and blood vessel information, but it introduces severe distortions in the morphology of other tissues. In our method, both the overall morphology and local texture features, including the vascular features of both lungs and the information about the spine, are very close to those of the real image. Representative synthetic detail information is displayed in Figure 6. The spine generation quality of all the comparison methods is poor and does not reach the level of clinical application.

Comparatively, the generation quality of this method and StyleSwin is acceptable and can roughly show the shape of the spinal bone cross-section. When comparing the generation quality of pulmonary nodules, the results of the proposed method and the VQGAN method are closest to real images, whereas other methods produce ground glass-like nodules that lack clarity and suggest a risk of malignant transformation. The last row shows the comparison of the imaging quality of arteries. It is easily observable that the proposed method can clearly generate contours and textures similar to those of real images, while the other methods cannot even present the contours of arteries.

The above results fully confirmed the feasibility of the proposed method in generating X-ray and CT images with high resolution. In particular, compared with the baseline VQGAN effect, the overall quality and details are significantly improved, and the effectiveness of the proposed structural optimization of HDFB is confirmed.

## 3.4 Clinical validation using visual Turing test

To evaluate the perceptual fidelity of synthetic images, we conducted a visual Turing test on 200 KUB X-ray images, where 100 X-rays are real and taken from the DGPH-KUB dataset and the other 100 are images synthesized using our method. Two urologists (senior: 20 years of experience; intermediate: 10 years of experience) take part in this test to complete the authenticity judgment of the 200 KUB X-ray images. As shown in Table 3, the true positive of the senior urologist is 85, reflecting familiarity with authentic anatomical features. Nevertheless, the true negative is 45, which means that 55% of the synthetic images were misclassified as real. Our method can mimic clinical data. The false positives of the intermediate urologist are 72, indicating that 72% of the synthetic images were mistaken as real, which validates our method's perceptual fidelity. A further analysis of the clinical implications is shown in Table 4. Sensitivity (the true positive rate for real images) and specificity (the true negative rate for synthetic images) were calculated, and statistical significance was assessed using McNemar's [45] test against random guessing (50%). Inter-rater agreement was quantified using Cohen's kappa (κ) [46]. Senior physicians demonstrated significantly higher sensitivity (85.0% vs. 65.0%, $p < 0.001$), reflecting their expertise in familiarity with the characteristics of real KUB X-ray images. However, both groups exhibited critically low specificity (senior: 45.0%; intermediate: 45.0%, $p < 0.01$ vs. 50% random guessing), with 55% of the synthetic images being misclassified as real. The low kappa values (0.3 for senior, 0.1 for intermediate) suggest variability in individual judgment criteria, yet the consistent 55% misclassification rate is sufficient to support the validity of the model's ability to generate clinically plausible images.

## 3.5 Ablation studies

### 3.5.1 Effect of lesion–background variability on synthesis stability

To assess the effect of lesion-sampling variability, we analyzed two cases: (1) varying lesion texture, scale, and placement on a fixed background and (2) placing a fixed lesion across varying backgrounds. All metrics were computed per the synthesized image, and the reported values are the mean ± SD across B = 10 backgrounds × L = 5 lesion samples per background (N = 50). As shown in Table 5, both scenarios exhibited low metric fluctuations (FID: 145.64–147.21; LPIPS: 0.48–0.49; PSNR: 58.76–59.03; SSIM: 0.68–0.69), indicating that the learned prior introduces controlled diversity without compromising visual realism. These variations are below clinical perceptibility thresholds, confirming the method's stability. Some visual examples are shown in Figure 7.

### 3.5.2 Background fidelity assessment via lesion masking

To evaluate the impact of controlled lesion synthesis on background anatomy fidelity, we conducted a specialized analysis on 100 background samples from the DGPH-KUB test set. The experiments computed two variants: SSIM computed only on pixels outside the lesion mask (background), and SSIM with no mask computed the global pixels. As shown in Table 6, the difference between the mask SSIM 0.66 and the global SSIM 0.67 of our proposed method was only 0.01, which is comparable to the difference observed in the baseline VQGAN, demonstrating that controlled lesion insertion did not disrupt the background anatomy. Furthermore, the global SSIM of our proposed method was significantly higher than that of the VQGAN, validating the enhanced background fidelity achieved by the HDFB module. This conclusion demonstrates that the innovative approach in this paper achieves flexible integration of pathological features while maintaining the integrity of the background.

### 3.5.3 Accelerated convergence via HDFB integration

For the ablation experiment, we compared the quantitative results with the benchmark model VQGAN in Section 3.2 and the visual results in Section 3.3, all of which prove the effectiveness of the proposed HDFB module and have a significant effect on the performance improvement of the VQGAN model. In addition, we compared the training loss convergence of our proposed method with that of VQGAN. As shown in Figure 8, integrating the HDFB module into the VQGAN framework leads to faster convergence and more stable training. The training loss of our method decreases more rapidly and reaches a lower value than VQGAN. In addition, the loss of the model with the HDFB block was reduced to 0.1 at approximately the 200th epoch, while the baseline model needed to reach 0.1 at approximately the 1,200th epoch. A total of 10,000 epochs were run in this experiment. Compared with the baseline model, the final loss value of the model with the HDFB block was 0.023, which was 0.057 less than 0.08 of the VQGAN model. This indicates that the HDFB module helps mitigate the vanishing gradient problem and accelerates the training process. This results in reduced training time and improved computational efficiency.

# 4 Discussion

In this study, we proposed a controllable lesion synthesis framework that integrates a SinGAN-based lesion generator with anatomically guided placement and a high-fidelity background synthesis network (HiResMed-VQGAN). The experimental results demonstrate the superiority of the proposed method in generating high-resolution medical images with small lesions. The two-route synthesis strategy addresses a critical bottleneck in medical AI: the scarcity of rare-lesion data. By decoupling SinGAN's lesion generation from background synthesis through HiResMed-VQGAN, our framework achieves flexible lesion control while ensuring high-quality anatomical structures in the generated images and establishes a new benchmark for high-resolution medical image generation, with transformative potential in surgical planning and early-disease detection.

Although the proposed method demonstrates promising results in high-resolution medical image generation, some limitations need to be addressed. In this study, the inter-rater agreement was low (kappa = 0.3 for senior, kappa = 0.1 for intermediate), corresponding to a fair and slight agreement by Landis and Koch's criteria. This may be due to the intrinsic difficulty of the "real vs. synthetic" visual judgment task, especially in the absence of standardized evaluation criteria. The results suggest that while sensitivity was relatively high, low specificity and low agreement limit the reliability of purely visual assessments, warranting methodological refinements in future work. Although the evaluation metrics indicate superior perceptual quality, the absence of task-specific evaluation, such as lesion detection or segmentation limits claims, regarding diagnostic fidelity. This is partly mitigated by clinical validation, which shows a high misclassification rate of 55%, indicating that the synthetic lesions are anatomically plausible, and we report SSIM values computed specifically within the lesion mask. Our method achieves SSIM = 0.68 ± 0.03 for lesions, which is significantly higher than that of VQGAN. This objectively confirms that synthetic lesions retain structural similarity to real lesions. Furthermore, low LPIPS variance (SD ≤ 0.02) implies perceptual consistency below human-discernible thresholds.

To bridge this gap, the future work will train lesion detectors and segmenters on hybrid datasets to quantify diagnostic utility. Lesion morphological control will be extended by enhancing SinGAN to generate diverse lesion shapes and textures, enabling the synthesis of atypical pathologies. The findings will be validated across modalities, and generalizability will be tested to MRI/PET, where structural constraints differ.

# Data availability statement

The LIDC-IDRI dataset presented in this article are available at https://www.kaggle.com/datasets/jokerak/lidcidri. The DGPH-KUB datasets are not readily available because the data are part of an ongoing study. Requests to access the datasets kindly directed to guangfa_tang@163.com.

## Ethics statement

The studies involving humans were approved by the Ethics Committee of Dongguan People's Hospital (number: KYKT2022-040). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

## Author contributions

GT: Writing – review and editing, Formal analysis, Writing – original draft, Methodology, Conceptualization. SC: Writing – review and editing, Methodology, Formal analysis, Resources. XM: Formal analysis, Data curation, Validation, Supervision, Writing – review and editing, Funding acquisition. SH: Writing – review and editing, Formal analysis, Resources, Validation. MW: Visualization, Writing – review and editing, Software. ZL: Investigation, Funding acquisition, Writing – review and editing. ZC: Formal analysis, Data curation, Writing – review and editing, Validation. XL: Formal analysis, Funding acquisition, Writing – original draft, Methodology, Writing – review and editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the Guangdong Basic and Applied Basic Research Foundation under grant no. 2021B1515140038, the National Natural Science Foundation of China (NSFC) under grant no. 82274413, the Key Discipline Research Capacity Enhancement Project of Guangdong Province in 2024 under grant no. 2024ZDJS130, the National Natural Science Foundation of China under Grant 62502320, the Natural Science Foundation of Guangdong Province under Grant 2025A1515010184, the project of Shenzhen Science and Technology Innovation Committee under Grant JCYJ20240813141424032 and JCYJ20240813112420027, Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515220079, the Foundation for Young innovative talents in ordinary universities of Guangdong under grant no. 2024KQNCX042, and the Young Innovative Talents Project for Ordinary Universities in Guangdong Province in 2023 under grant no. 2023KQNCX123.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note



## References

1. Wakabayashi T, Cacciaguerra AB, Abe Y, Dalla Bona E, Nicolini D, Mocchegiani F, et al. Indocyanine green fluorescence navigation in liver surgery: a systematic review on dose and timing of administration. *Ann Surg* (2022) 275(6):1025–34. doi:10.1097/SLA.0000000000005406

2. Giri M, Dai H, Puri A, Liao J, Guo S. Advancements in navigational bronchoscopy for peripheral pulmonary lesions: a review with special focus on virtual bronchoscopic navigation. *Front Med* (2022) 9:989184. doi:10.3389/fmed.2022.989184

3. Currie GM, Hawk KE, Rohren EM. Generative artificial intelligence biases, limitations and risks in nuclear medicine: an argument for appropriate use framework and recommendations. *Semin Nucl Med* 55 (2024). p. 423–36. doi:10.1053/j.semnuclmed.2024.05.005

4. Hölscher D, Reich C, Gut F, Knahl M, Clarke N. Exploring the efficacy and limitations of histogram-based fake image detection. *Proced Computer Sci* (2024) 246:2882–91. doi:10.1016/j.procs.2023.102288

5. Kumar A, Soni S, Chauhan S, Kaur S, Sharma R. Navigating the realm of generative models: GANs, diffusion, limitations, and future prospects—A review. In: *International conference on cognitive computing and cyber physical systems*. Singapore: Springer Nature Singapore (2023).

6. Dwivedi DN, Dwivedi VN. Critiquing the limitations' challenges in detecting GAN-generated images with computer vision. In: *International conference on communication and intelligent systems*. Singapore: Springer Nature Singapore (2023). doi:10.1007/978-981-97-2053-8_7

7. Uzunova H, Ehrhardt J, Jacob F, Frydrychowicz A, Handels H. Multi-scale gans for memory-efficient generation of high resolution medical images. In: *The 22nd international conference on medical image computing and computer assisted intervention (MICCAI)*. Shenzhen, China: Springer International Publishing (2019). doi:10.1007/978-3-030-32226-7_6

8. Khatun A, Yeter-Aydeniz K, Weinstein YS, Usman M. Quantum generative learning for high-resolution medical image generation. *Machine Learn Sci Technology* (2025) 6(2):025032. doi:10.1088/2632-2153/add1a9

9. Zhao L, Zhang Z, Chen T, Metaxas D, Zhang H. Improved transformer for high-resolution gans. *Adv Neural Inf Process Syst* (2021) 34:18367–80. doi:10.48550/arXiv.2104.11233

10. Cao S, Yin Y, Huang L, Liu Y, Zhao X, Zhao D, et al. Efficient-vqgan: towards high-resolution image generation with efficient vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Paris, France: IEEE (2023). p. 7368–77. doi:10.1109/ICCV42023.2023.00249

11. Hu T, Ge W, Zhao Y, Lee GH. X-ray: a sequential 3d representation for generation. *Adv Neural Inf Process Syst* (2024) 37:136193–219. doi:10.48550/arXiv.2401.13619

12. Zhao M, Cong Y, Carin L. On leveraging pretrained gans for generation with limited data. In: Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria. JMLR.org (2020):11340–11351. doi:10.48550/arXiv.2002.07781

13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* (2020) 63(11):139–44. doi:10.1145/3422622

14. Zhou T, Li Q, Lu H, Cheng Q, Zhang X. GAN review: models and medical image fusion applications. *Inf Fusion* (2023) 91:134–48. doi:10.1016/j.inffus.2022.10.017

15. Xia W, Zhang Y, Yang Y, Xue JH, Zhou B, Yang MH. Gan inversion: a survey. *IEEE Trans pattern Anal machine intelligence* (2022) 45(3):3121–38. doi:10.1109/tpami.2022.3181070

16. Dash A, Ye J, Wang G. A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: from medical to remote sensing. *IEEE Access* (2023) 12:18330–57. doi:10.1109/access.2023.3346273

17. Wu AN, Stouffs R, Biljecki F. Generative adversarial networks in the built environment: a comprehensive review of the application of GANs across data types and scales. *Building Environ* (2022) 223:109477. doi:10.1016/j.buildenv.2022.109477

18. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). doi:10.1109/CVPR46437.2022.00574

19. Li W, Li J, Polson J, Wang Z, Speier W, Arnold C. High resolution histopathology image generation and segmentation through adversarial training. *Med Image Anal* (2022) 75:102251. doi:10.1016/j.media.2021.102251

20. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* (2020) 33:6840–51. doi:10.48550/arXiv.2006.11239

21. Zhang B, Gu S, Zhang B, Bao J, Chen D. Styleswin: transformer-Based gan for high-resolution image generation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). doi:10.1109/CVPR.2022.01614

22. Aboutalebi H, Pavlova M, Gunraj H, Shafiee MJ, Sabri A, Alaref A, et al. MEDUSA: multi-scale encoder-decoder self-attention deep neural network architecture for medical image analysis. *Front Med* (2022) 8:821120. doi:10.3389/fmed.2021.821120

23. Martini MG. Measuring objective image and video quality: on the relationship between SSIM and PSNR for DCT-based compressed images. *IEEE Trans Instrumentation Meas* (2025) 74:1–13. doi:10.1109/tim.2025.3529045

24. Zhang L, Dai H, Sang Y. Med-SRNet: GAN-Based Medical Image Super-Resolution via High-Resolution Representation Learning. *Comput. Intell. Neurosci.* (2022). 1744969. doi:10.1155/2022/1744969

25. Wang B, Liao X, Ni Y, Zhang L, Liang J, Wang J, et al. High-resolution medical image reconstruction based on residual neural network for diagnosis of cerebral aneurysm. *Front Cardiovasc Med* (2022) 9:1013031. doi:10.3389/fcvm.2022.1013031

26. Kim J, Li Y, Shin BS. 3D-DGGAN: a data-guided generative adversarial network for high fidelity in medical image generation. *IEEE J Biomed Health Inform* (2024) 28(5):2904–15. doi:10.1109/JBHI.2024.3367375

27. Kang M, Chikontwe P, Won D, Luna M, Park SH. Structure-preserving image translation for multi-source medical image domain adaptation. *Pattern Recognition* (2023) 144:109840. doi:10.1016/j.patcog.2023.109840

28. Yu Z, Zhao B, Zhang S, Chen X, Yan F, Feng J, et al. HiFi-Syn: hierarchical granularity discrimination for high-fidelity synthesis of MR images with structure preservation. *Med Image Anal* (2025) 100:103390. doi:10.1016/j.media.2024.103390

29. Yu Z, Zhao B, Zhang Y, Zhang S, Chen X. Cross-grained contrastive representation for unsupervised lesion segmentation in medical images. In: *InProceedings of the IEEE/CVF international conference on computer vision 2023*. Paris, France: IEEE (2023). p. 2347–54.

30. Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. Nashville, TN, USA: IEEE (2021). doi:10.1109/CVPR46437.2021.00376

31. Huang M. Towards accurate image coding: improved autoregressive image generation with dynamic vector quantization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Vancouver, BC, Canada: IEEE (2023). p. 12873–83. doi:10.1109/CVPR.2023.01893

32. Gu Y, Wang X, Ge Y, Shan Y, Shou MZ. Rethinking the objectives of vector-quantized tokenizers for image synthesis. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Seattle, USA: IEEE (2024). p. 7631–40.

33. Verma L, Mohan V. Vector quantization loss analysis in VQGANs: a single-GPU ablation study for image-to-image synthesis. *arXiv preprint arXiv:2308.05242* (2023). doi:10.48550/arXiv.2308.05242

34. Zhu L, Wei F, Lu Y, Chen D. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99. *arXiv preprint arXiv:2406.11837* (2024). doi:10.48550/arXiv.2406.11837

35. Zhan F, Yu Y, Wu R, Zhang J, Lu S. Multimodal image synthesis and editing: a survey. *arXiv preprint arXiv:2112* (2022):13592. doi:10.48550/arXiv.2112.13592

36. Zhang J, Zhan F, Theobalt C, Lu S. Regularized vector quantization for tokenized image synthesis. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Vancouver, BC, Canada: IEEE (2023). p. 18467–76.

37. Podell D, English Z, Lacey K, Blattmann A, Dockhorn T. SDXL: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023). doi:10.48550/arXiv.2307.01952

38. Chen T, Zhang D, Qian Y. Enhancing VQGAN performance through integration of multiple vision transformers. In: *2024 IEEE 8th international conference on vision, image and signal processing (ICVISP)*. Malaysia: IEEE: Kuala Lumpur (2024). p. 1–8. doi:10.1109/ICVISP64524.2024.10959589

39. Shaham TR, Dekel T, Michaeli T. Singan: learning a generative model from a single natural image. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Seoul, Korea (South): IEEE (2019). p. 4570–80. doi:10.1109/ICCV.2019.00123

40. Pappas C, Kovaios S, Moralis-Pegios M, Tsakyridis A, Giamougiannis G, Kirtas M, et al. Programmable tanh-elu-sigmoid-and sin-based nonlinear activation functions for neuromorphic photonics. *IEEE J Selected Top Quan Electronics* (2023) 29(6):1–10. doi:10.1109/jstqe.2023.3277118

41. Gunawan A, Yin X, Zhang K. Understanding and improving group normalization. *arXiv preprint arXiv:2207.01972* (2022). doi:10.48550/arXiv.2207.01972

42. Hendrycks D, Gimpel K. Gaussian error linear units (gelus). *arXiv* (2016). arXiv:1606.08415. doi:10.48550/arXiv.1606.08415

43. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *The thirty-first annual conference on neural information processing systems (NIPS)*. Long Beach, California: Curran Associates Inc (2017). p. 6629–40.

44. Ghildyal A, Liu F. Shift-tolerant perceptual similarity metric. In: *The 17th European conference on computer vision*. Tel Aviv, Israel: Springer (2022). p. 91–107. doi:10.1007/978-3-031-19836-6_6

45. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* (1947) 12(2):153–7. doi:10.1007/bf02295996

46. Altman DG. *Practical statistics for medical research*. Chapman and Hall/CRC (1990). doi:10.1201/9780429258589