Medical domain-adapted stable diffusion model for the generation of realistic chest X-ray images

Zhaowei Zhang*
Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92092
zhz180@ucsd.edu

Abstract

Generating realistic medical images from textual descriptions remains a significant challenge due to the domain gap between natural and clinical data. In this work, we explore the adaptation of a pre-trained Stable Diffusion model to the medical domain, specifically for synthesizing chest X-ray (CXR) images from radiology report impressions. We employ parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) to update only the cross-attention layers of the model, enabling efficient training on limited computational resources.

Using a publicly available chest X-ray dataset containing paired image-report samples, we fine-tune the model and evaluate its performance both qualitatively and quantitatively. Our experiments demonstrate that the LoRA-finetuned model significantly improves over the base model in terms of Fréchet Inception Distance (FID), reducing the score from 363.91 to 166.55, while maintaining comparable Inception Score (IS). Visual comparisons further highlight the improved anatomical realism and alignment with textual prompts, particularly at higher resolutions (512×512).

These results suggest that domain-adapted latent diffusion models can serve as powerful tools for medical image synthesis, offering scalable and privacy-preserving alternatives to real data in clinical AI development.

1 Background

In recent years, latent diffusion models (LDMs) have emerged as powerful generative frameworks capable of producing high-fidelity images across a range of domains. Their ability to learn rich data distributions and synthesize novel samples makes them especially valuable in fields where acquiring large, high-quality datasets is challenging. One such field is medical imaging, where access to well-annotated and diverse datasets is often restricted due to several factors, including the need for expert radiological interpretation, stringent privacy regulations, and data siloing across healthcare institutions.

Chest X-ray imaging, a fundamental diagnostic tool in medicine, exemplifies the broader challenges in the medical domain. Developing robust machine learning models in this area is hampered by limited data availability, particularly for rare pathologies or nuanced clinical presentations. This scarcity poses a significant barrier to training and evaluating reliable AI systems for clinical deployment.

Generative models such as LDMs offer a promising solution by enabling the synthesis of realistic medical images that can supplement existing datasets. These models can help alleviate data imbalance,

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

facilitate data sharing without compromising patient privacy, and provide a basis for pretraining or augmenting downstream diagnostic models. However, off-the-shelf generative models, particularly large vision-language models trained primarily on natural image datasets, often struggle to capture the unique visual and semantic characteristics of medical imagery. The vocabulary used in medical contexts is highly specialized, and the distribution of features in medical images significantly differs from that found in natural images.

This mismatch highlights the need for domain-adapted generative approaches. By fine-tuning diffusion models within the medical context—specifically using domain-relevant data and medical language supervision—it may be possible to generate chest X-ray images that are not only realistic in appearance but also faithful in representing a wide range of medical conditions and anatomical variations. Such advances could play a critical role in enhancing AI applications in radiology, from education and algorithm development to testing underrepresented diagnostic categories.

2 Objective

The primary objective of this project is to adapt a pre-trained, general-domain stable diffusion (SD) model to the medical imaging domain, with a specific focus on generating realistic and diverse chest X-ray (CXR) images. By leveraging publicly available datasets consisting of chest X-rays and their associated radiology reports, this project aims to fine-tune a text-conditional latent diffusion model (LDM) such that it can synthesize high-fidelity medical images controlled by free-form, domain-specific text prompts.

This domain adaptation strategy is intended to overcome the distributional gap between natural images—on which current large vision–language diffusion models are typically trained—and medical images, which exhibit distinct visual characteristics and semantic structures. The adapted model will utilize the rich vision–language alignment inherent in the diffusion framework to generate synthetic CXRs that are both visually plausible and clinically meaningful.

Specifically, the goals of the project are:

To fine-tune a stable diffusion model using chest X-ray image–report pairs to better align with medical domain semantics.

To evaluate the quality, diversity, and realism of the generated synthetic images.

To enable controlled image generation based on medically relevant text prompts, such as pathology names, anatomical references, or clinical descriptors.

To explore the potential of the adapted model as a data augmentation tool for downstream machine learning tasks in radiology.

By achieving these objectives, this project aims to demonstrate the feasibility and value of domain-adapted generative models for addressing data scarcity and enhancing AI development in medical imaging.

3 Significance

The ability to generate high-quality, text-conditioned synthetic medical images represents a significant advancement in addressing critical data challenges in the medical imaging domain. The proposed domain-adaptation strategy for stable diffusion models has the potential to transform how medical datasets are augmented and shared, particularly in scenarios where access to real patient data is restricted due to privacy, regulatory, or logistical constraints.

By enabling the generation of realistic chest X-ray images from free-form medical text prompts, this project offers a scalable and privacy-preserving solution to augment existing datasets. The synthetic images produced can enhance the diversity and representation of training data, especially for rare conditions or underrepresented patient populations. This can lead to more robust and generalizable diagnostic models.

Moreover, the ability to generate clinically meaningful images without direct access to real patient data opens new possibilities for collaborative AI research across institutions, eliminating many of the legal and ethical barriers associated with data sharing. The adapted model can serve as a

valuable resource for training, validation, and benchmarking of machine learning systems in radiology, ultimately contributing to improved diagnostic performance and broader access to medical AI tools.

In summary, this project addresses a pressing bottleneck in medical AI development and offers a practical, innovative approach to overcoming data limitations in a safe, efficient, and scalable manner.

4 Related Work

4.1 Latent diffusion models (LDMs)

Latent Diffusion Models (LDMs) are a class of generative models that address the high computational cost of traditional diffusion models by operating in a compressed latent space rather than directly on high-dimensional pixel data[1]. Instead of denoising images in raw pixel space—which is both memory- and compute-intensive—LDMs first use a pretrained autoencoder to map images into a lower-dimensional, perceptually meaningful latent space. Diffusion processes are then applied within this space, significantly reducing the complexity of training and inference while preserving high visual fidelity.

A key innovation of LDMs is their ability to maintain fine detail and semantic richness during generation, achieving state-of-the-art performance on tasks like text-to-image synthesis, inpainting, and super-resolution. By introducing cross-attention mechanisms, LDMs can condition image generation on arbitrary inputs such as text or segmentation maps, making them highly flexible and efficient tools for high-resolution image synthesis.

4.2 Cross-attention

Cross-attention is a mechanism that enables models to integrate and align information from different modalities, such as text and images, during the generative process. In the context of latent diffusion models (LDMs), cross-attention layers allow the model to condition image generation on textual descriptions by linking latent image representations with text embeddings derived from language models like CLIP or T5. This mechanism enhances the model's ability to produce semantically consistent and visually coherent images based on user-provided prompts[1][2].

By leveraging cross-attention, LDMs and other diffusion-based architectures can interpret complex natural language inputs and accurately translate them into high-resolution images. This has been a key driver behind recent breakthroughs in text-to-image synthesis, enabling fine-grained control over generated content without retraining the entire model[3].

4.3 Stable Diffusion (SD)

Stable Diffusion is a cutting-edge text-to-image synthesis model that leverages the power of latent diffusion and large-scale vision-language datasets to generate high-fidelity images from textual prompts. Central to its development is the use of the LAION-5B dataset, an open-access corpus containing over five billion image-text pairs, which provides a diverse and comprehensive foundation for training multimodal generative models[4].

As a foundation model, Stable Diffusion exemplifies the paradigm shift toward models trained on vast, general-purpose data that can be adapted for a wide range of downstream tasks with minimal fine-tuning. While this brings significant opportunities in accessibility and creative applications, it also introduces challenges regarding bias, control, and responsible deployment—issues that are increasingly central in the discourse around foundation models[5].

5 Methods

5.1 Dataset

Dataset We used the publicly available *Chest X-ray dataset from Indiana University*, sourced via Kaggle, which contains **7,466 chest X-ray images** across different views (Frontal and Lateral) along with corresponding radiology reports from **3,851 patients**. Due to memory limitations, all images were resized to a resolution of **128×128 pixels**. For the text prompts, we extracted the *"Impression"*

section of each report—a concise summary highlighting the key findings and interpretations. To ensure compatibility with the CLIP tokenizer, we excluded empty entries and those exceeding the **77-token limit**, resulting in a final dataset of **7,281 unique image-text pairs** for the fine-tuning experiments. The dataset was then split into **80% training (5,825 pairs)** and **20% testing (1,456 pairs)**.

5.2 Model Architecture

This project utilizes the Latent Diffusion Model (LDM) Stable Diffusion 1.4 (SD1.4) as the foundational generative architecture. The SD pipeline consists of three main components:

- **Variational Autoencoder (VAE):** A Kullback–Leibler-regularized VAE trained using a combination of perceptual loss and a patch-based adversarial loss. It compresses high-dimensional image inputs into a lower-dimensional latent space and reconstructs images from these representations during decoding.
- **Denoising U-Net Backbone:** The core denoising model is a U-Net structure, enhanced by a cross-attention mechanism. This allows the model to condition its denoising process on auxiliary inputs by attending to embeddings from a text encoder.
- **Text Conditioning via CLIP:** The conditioning mechanism employs a pre-trained CLIP ViT-L/14 text encoder, which encodes input prompts into a 768-dimensional latent space. Each text input is limited to a maximum of 77 tokens, in accordance with the CLIP tokenizer's constraints.

To generate an image, the process begins with a latent vector sampled from Gaussian noise. This vector is denoised through multiple iterations using the U-Net, where cross-attention integrates semantic information from the CLIP text encoder at various layers. The final denoised latent vector is then passed through the VAE decoder to reconstruct a high-resolution image in pixel space.

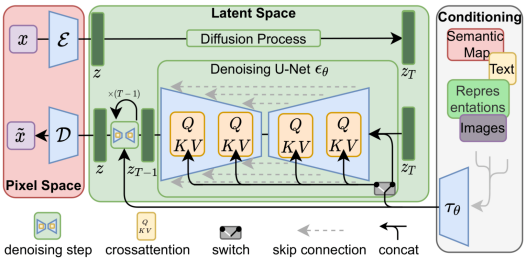

Figure 1: A conditional denoising U-Net iteratively denoises a latent vector sampled from a Gaussian distribution over t timesteps. The process is conditioned via cross-attention (QKV for query, key, value of the attention process) through embeddings created from short medical free-text inputs processed by a text encoder ET. The decoder D of the VAE maps the denoised latent vector to pixel space, resulting in a high-fidelity CXR image showing imaging features corresponding to the initial text prompt[1].

5.3 Fine-Tuning Procedure

We fine-tuned the Stable Diffusion model on CXR-report pairs by training it to predict the noise component added to latent image representations. For each image-text pair $(x_{\text{text}}, y_{\text{pixel}}) \in \mathcal{D}$, and for a randomly chosen timestep T, Gaussian noise $N_{i,t} \sim \mathcal{N}(0_{h\times w}, I_{(h\times w)^2})$ was added to the latent encoding of y_{pixel} using the VAE encoder E_V. The U-Net ψ, conditioned on the text embedding $E_T(x_{\text{text}})$, was trained to predict the last added noise:

$$\hat{N}_{i,T} = \psi\left(E_T(x_{\text{text}}), E_V(y_{\text{pixel}}) \oplus_T N_{i,t}, T\right), \tag{1}$$

where $\oplus_T$ denotes iterative noise addition over T steps. Training was supervised using the mean squared error between the predicted and true noise:

$$\mathcal{L} = \frac{1}{h \times w} \sum_{k=1}^{h} \sum_{l=1}^{w} (\hat{N}_{i,T,k,l} - N_{i,T,k,l})^2. \tag{2}$$

Only the unfrozen components of the SD pipeline were updated during backpropagation.

5.4 Low-Rank Adaptation (LoRA)

To enable parameter-efficient fine-tuning of the Stable Diffusion model, we adopt Low-Rank Adaptation (LoRA), which injects trainable low-rank matrices into attention layers while keeping the original model weights frozen. This approach significantly reduces computational cost and memory usage.

In transformer-based architectures like the U-Net backbone of Stable Diffusion, attention layers use linear projections:

$$\mathbf{y} = W\mathbf{x}, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input and $W \in \mathbb{R}^{d \times d}$ is a learned weight matrix. Instead of updating W, LoRA introduces a trainable residual term:

$$\mathbf{y} = W\mathbf{x} + \Delta W\mathbf{x}, \quad \text{with} \quad \Delta W = BA, \tag{4}$$

where $A \in \mathbb{R}^{r \times d}$, $B \in \mathbb{R}^{d \times r}$, and $r \ll d$. Only A and B are updated during training.

We apply LoRA to the query, key, value, and output projections in the U-Net's cross-attention layers. This enables domain-specific adaptation (e.g., chest X-ray generation) while preserving the general capabilities of the pretrained model. The resulting LoRA modules are lightweight and modular, allowing efficient and flexible deployment at inference time.

5.5 Reconstruction Quality Assessment

To evaluate the fidelity of the generated chest X-ray images, we will compute standard quantitative metrics that assess both the quality and diversity of the synthesized samples. Specifically, we will use Fréchet Inception Distance (FID) and Inception Score (IS). FID measures the distance between the feature distributions of real and generated images, providing insight into both realism and variety. IS evaluates the confidence and diversity of predicted labels from an image classifier, capturing the clarity and variability of generated outputs. Together, these metrics offer a comprehensive assessment of generative performance in terms of both visual realism and sample diversity.

6 Experiments

6.1 Fine-tuning of a pre-trained LDM

The fine-tuning experiment was performed on a pretrained latent diffusion model, specifically *Stable Diffusion v1.5*, using an NVIDIA P100 GPU provided by Kaggle. LoRA-based adaptation was applied exclusively to the U-Net component of the model, with all other modules frozen during training. The training data consisted of chest X-ray (CXR) images resized to 128×128 pixels, and the model was trained with a batch size of 4. The pretrained weights were sourced from the Hugging Face repository `stable-diffusion-v1-5/stable-diffusion-v1-5`. Implementation was carried out using the `diffusers` library (version 0.32.2) along with Hugging Face's official LoRA training framework.

Fine-tuning on chest X-ray (CXR) and report pairs was carried out by training the model to predict the noise component added to the latent image representations. Given a dataset of paired samples $\mathcal{D} = \{(x_{\text{text}}, y_{\text{pixel}})_i\}_{i=1}^n$, each image y_{pixel} was encoded by the VAE encoder E_V into a latent representation, and each text prompt x_{text} was encoded by a text encoder E_T. For each pair and timestep t, Gaussian noise was sampled in latent space of dimensions (h, w):

$$N_{i,t} \sim \mathcal{N}(0_{h \times w}, I_{(h \times w)^2})$$

This noise $N_{i,t}$ was added to the latent image embedding $E_V(y_{\text{pixel}})$ for T timesteps using a scheduler, simulating the forward diffusion process. The noisy latent image and the encoded text prompt were then passed to the U-Net ψ, which was trained to predict the last added noise component $\hat{N}_{i,T}$:

$$\hat{N}_{i,T} = \psi\left(E_T(x_{\text{text}}),\ E_V(y_{\text{pixel}}) \oplus_T N_{i,t},\ T\right)$$

where $\oplus_T$ denotes iterative noise addition for T timesteps. The training objective was to minimize the mean squared error (MSE) between the predicted noise $\hat{N}_{i,T}$ and the true noise $N_{i,T}$ over all spatial positions (k, l):

$$\mathcal{L} = \frac{1}{h \times w} \sum_{k=1}^{h} \sum_{l=1}^{w} \left(\hat{N}_{i,T,k,l} - N_{i,T,k,l}\right)^2$$

Backpropagation using this loss allowed updating only the unfrozen components of the Stable Diffusion pipeline.

7 Results

7.1 Hyperparameters

Hyperparameters

The fine-tuning process was conducted using a resolution of **128×128 pixels** and a **batch size of 4**. The model was trained for **20 epochs** using the **Adam optimizer** with a **learning rate of** 5×10^{-5}. To enable parameter-efficient fine-tuning, **Low-Rank Adaptation (LoRA)** was employed on the attention layers of the U-Net. The LoRA configuration used a **rank of 8** and set `lora_alpha = 8`, with weights initialized from a **Gaussian distribution**. Adaptation was applied specifically to the attention projection modules: `to_q`, `to_k`, `to_v`, and `to_out.0`.

7.2 Training Loss Curve

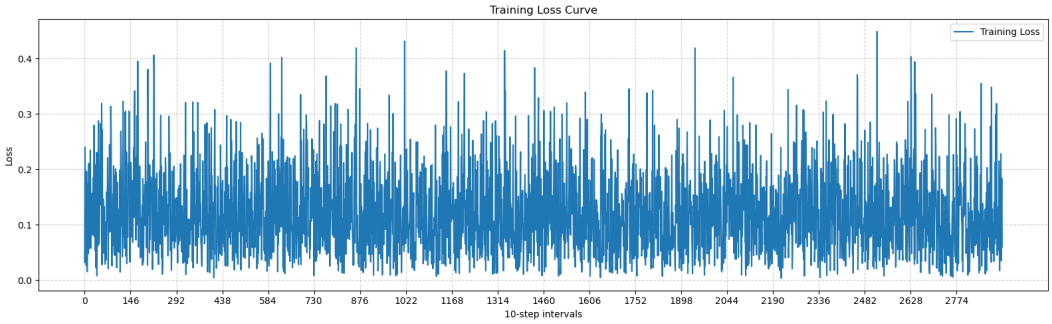

Figure 2: Training loss curve during LoRA fine-tuning of the Stable Diffusion model. The x-axis represents 10-step intervals, and the y-axis denotes the loss values.

Figure 2 illustrates the training loss curve observed during LoRA-based fine-tuning of the Stable Diffusion model on chest X-ray and report text pairs. The loss was logged every 10 training steps, covering approximately 2800 intervals.

The curve exhibits noticeable oscillations, which are typical for diffusion model training, particularly under noisy latent-space supervision and small batch sizes. Despite the high variance, the training loss remains bounded within a reasonable range (approximately 0.0 to 0.4), and no signs of instability or divergence are observed throughout the training process.

Such fluctuations are common in denoising diffusion models, where the loss is computed between predicted and sampled noise components at randomly selected timesteps. Additionally, the use of

LoRA introduces lightweight updates to select layers (e.g., attention projections in U-Net), which may lead to slower but more stable convergence patterns.

Overall, the training dynamics suggest that the model successfully learned from the paired CXR-report data, while maintaining training stability under the low-rank adaptation constraints.

7.3 Image Comparison

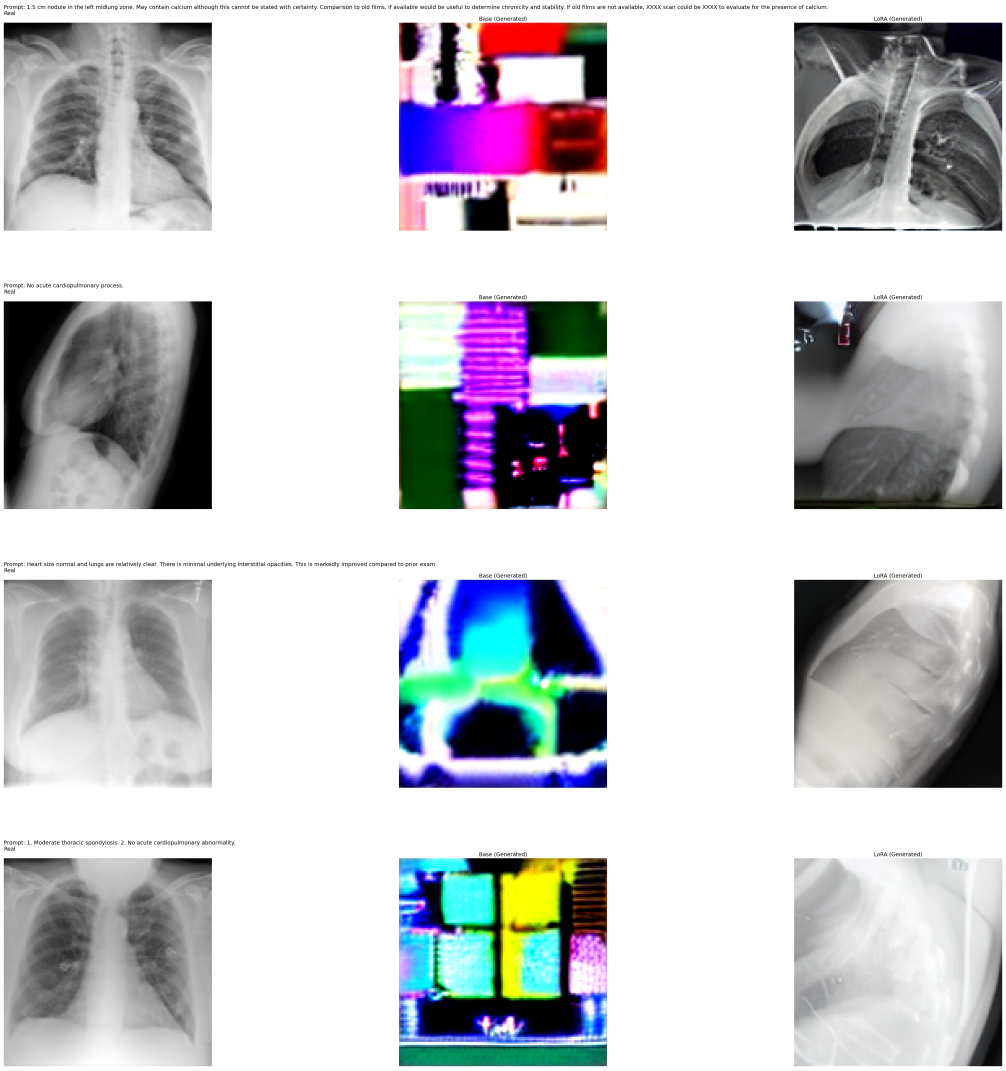

Figure 3: Generated image samples at resolution 128×128. Left: Ground-truth CXR image; Center: Base Stable Diffusion model output; Right: LoRA-finetuned model output.

Figures 3 and 4 provide qualitative comparisons between real chest X-ray images, base Stable Diffusion outputs, and images generated by the LoRA-finetuned model. In each row, the original input image and radiology report impression are used to condition both the base model and the LoRA-enhanced model.

At a resolution of 128×128 (Figure 3), the generated images suffer from significant degradation and hallucinations. Both the base and LoRA models struggle to maintain structural and anatomical realism, and the outputs often lack meaningful spatial organization. This can be attributed to the limited representational capacity at low resolutions, as well as the challenge of modeling fine-grained anatomical details using such compressed spatial dimensions.

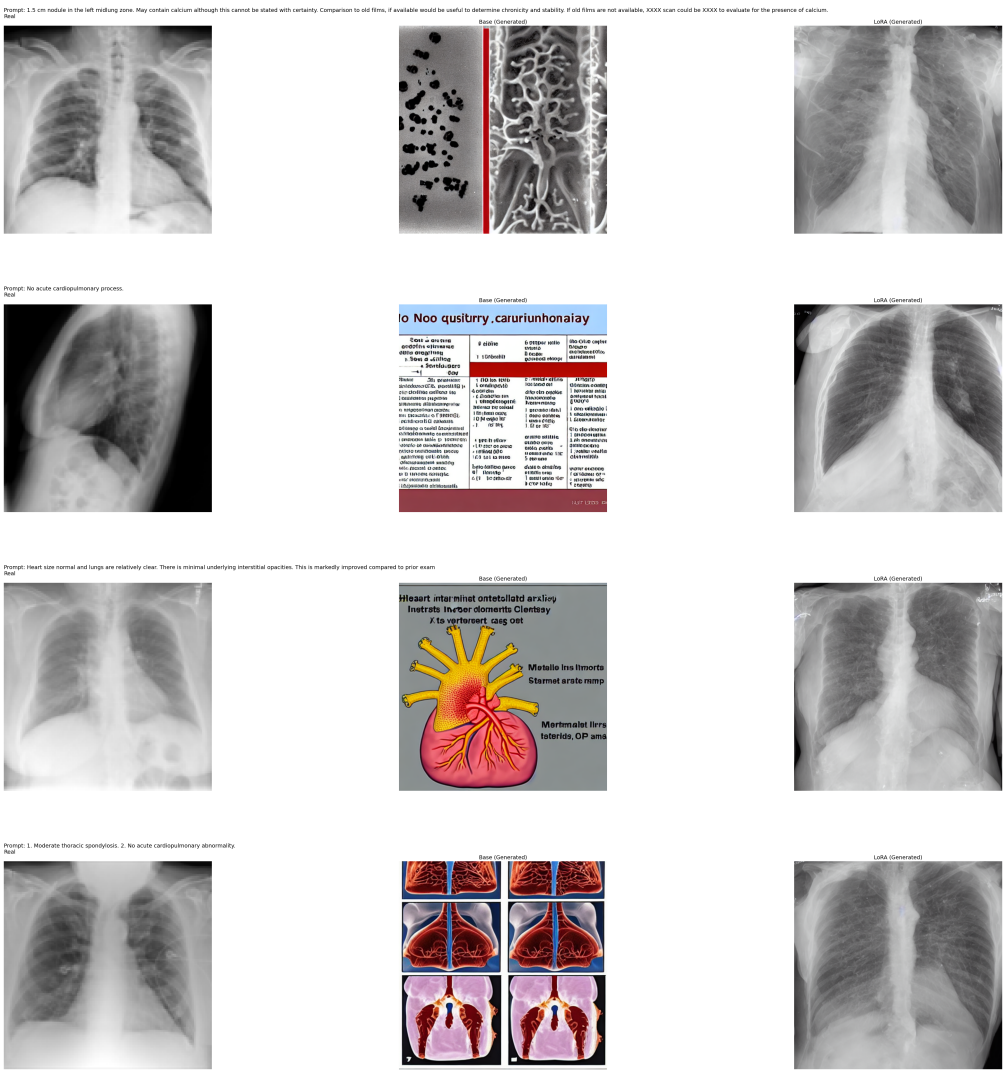

Figure 4: Generated image samples at resolution 512×512. Left: Ground-truth CXR image; Center: Base Stable Diffusion model output; Right: LoRA-finetuned model output.

In contrast, the 512×512 results (Figure 4) exhibit a dramatic improvement in visual fidelity. The LoRA-finetuned model generates outputs that closely resemble real CXRs in texture, anatomy, and orientation. Compared to the base model at this resolution, the LoRA-enhanced images show better alignment with the semantic content of the input reports and avoid the artifacts and incoherent patterns that frequently appear in the base model's outputs.

These results strongly suggest that fine-tuning with LoRA is effective at enhancing the generative capabilities of Stable Diffusion when working at clinically relevant resolutions. The combination of higher resolution and low-rank adaptation appears crucial for preserving critical radiographic features in medical image synthesis tasks.

7.4 Evaluation Metrics

To quantitatively assess the quality and realism of the generated chest X-ray images, we computed two widely-used generative evaluation metrics: Fréchet Inception Distance (FID) and Inception Score (IS). These metrics provide complementary insights—FID measures the distance between the real and generated image distributions, while IS evaluates both the confidence and diversity of generated samples.

- **FID (Base Model)**: 363.91
- **FID (LoRA Model)**: 166.55
- **Inception Score (Base Model)**: 4.30 ± 0.41
- **Inception Score (LoRA Model)**: 3.92 ± 0.37

The LoRA-finetuned Stable Diffusion model significantly outperformed the base model in terms of FID, showing a substantial reduction from 363.91 to 166.55. This improvement indicates that the LoRA-enhanced model produces images that are much closer to the distribution of real chest X-rays in the embedding space of the Inception network. This result aligns with visual inspections, which revealed more anatomically consistent and realistic images at higher resolutions.

Interestingly, while the Inception Score of the LoRA model (3.92) is slightly lower than that of the base model (4.30), the difference remains within the standard deviation range. This suggests that although the LoRA model may produce slightly less diversity or classifier confidence as measured by IS, its image realism—as captured by FID—is notably better. The lower IS may also reflect the domain mismatch between chest X-ray images and the ImageNet-trained Inception model used for evaluation.

Overall, these metrics validate that LoRA fine-tuning meaningfully improves generative quality in a medically relevant context.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685. New Orleans, LA: IEEE.

[2] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/abs/2204.06125v1 (2022).

[3] Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. 35, 36479–36494 (2022).

[4] Schuhmann, C. et al. LAION-5B: an open large-scale dataset for training next generation imagetext models. In Advances in Neural Information Processing Systems Vol. 35 (eds Koyejo, S. et al.) 25278–25294 (Curran Associates, Inc., 2022)

[5] Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258v3 (2022).