SYNTHESIZING MEDICAL IMAGES USING LORA-ENHANCED STABLE DIFFUSION

Raheel Mahmood
Northwestern University
raheelmahmood2030@u.northwestern.edu

ABSTRACT

I present a memory-efficient fine-tuning of the Prompt2MedImage Stable Diffusion model for synthetic chest X-ray generation. By applying Low-Rank Adaptation (LoRA, rank=16) to the UNet, I fine-tuned the model on 2,000 chest X-ray images resized to 256×256 resolution. Training for 5 epochs using AdamW (lr=1e-4) with CosineAnnealingLR resulted in high-quality synthetic X-rays suitable for data augmentation and training purposes (FID 37.2). My code and trained model are publicly available at https://github.com/defce/BMDfinal.

1 INTRODUCTION

Chest X-ray imaging plays a crucial role in clinical diagnosis, but acquiring large-scale, labeled datasets remains challenging due to privacy concerns, cost, and manual annotation burdens. Generative deep learning methods offer a compelling solution by creating synthetic medical images that can augment existing datasets, assist in clinical training, and improve downstream diagnostic models.

In this paper, I explore diffusion-based generative models, particularly Stable Diffusion, fine-tuned for chest X-ray synthesis. However, training diffusion models typically requires substantial computational resources and large GPUs—resources not always available in practical research scenarios. Given my limited computational hardware (a single 16 GB GPU), I initially encountered significant memory constraints, including frequent out-of-memory (OOM) errors. This motivated me to explore parameter-efficient fine-tuning strategies, ultimately adopting Low-Rank Adaptation (LoRA) to dramatically reduce the number of trainable parameters and memory usage.

Through a rigorous experimentation process, I evaluated multiple training configurations, varying image resolutions (32×32, 128×128, 256×256, 512×512) and epoch counts (1, 3, and 5 epochs). Each configuration posed unique trade-offs in computational feasibility and image quality. After extensive iteration and evaluation, I successfully trained a diffusion model at a 256×256 resolution on a dataset of 2,000 chest X-rays over 5 epochs, yielding excellent synthetic image quality with stable training behavior.

Overall, my approach demonstrates that high-quality synthetic medical image generation is achievable with limited computational resources through careful model selection, parameter-efficient training methods, and iterative experimentation. The final model effectively balances computational practicality and image fidelity, offering promising implications for medical imaging research with limited hardware.

2 RELATED WORK

2.1 DIFFUSION MODELS IN MEDICAL IMAGING

Diffusion models have recently emerged as a powerful alternative to GANs for high-fidelity image synthesis. Ho et al. [1] introduced DDPMs, which have since been applied in medical imaging tasks such as organ segmentation [2] and brain MRI synthesis [3]. Compared to GANs, diffusion models offer more stable training and better image quality.

2.2 LoRA and Lightweight Fine-Tuning

Low-Rank Adaptation (LoRA) was proposed by Hu et al. [4] to reduce the number of trainable parameters when fine-tuning large models. LoRA has since been adopted in both vision [5] and language domains for efficient model adaptation with limited compute resources.

2.3 Prompt-Conditioned Medical Image Generation

Prompt-based generative models like Stable Diffusion have been extended to the medical domain. Boiko et al. [6] introduced Prompt2MedImage, a diffusion model trained to synthesize chest X-rays from textual descriptions of medical conditions. My work builds upon this by fine-tuning Prompt2MedImage using LoRA on a curated dataset of chest X-rays.

3 Methods

3.1 Overview of the Proposed Pipeline

Figure 1 provides a high-level overview of my training and image-generation pipeline. I adapt the pre-trained Prompt2MedImage Stable Diffusion model with LoRA and fine-tune it using chest X-ray images.

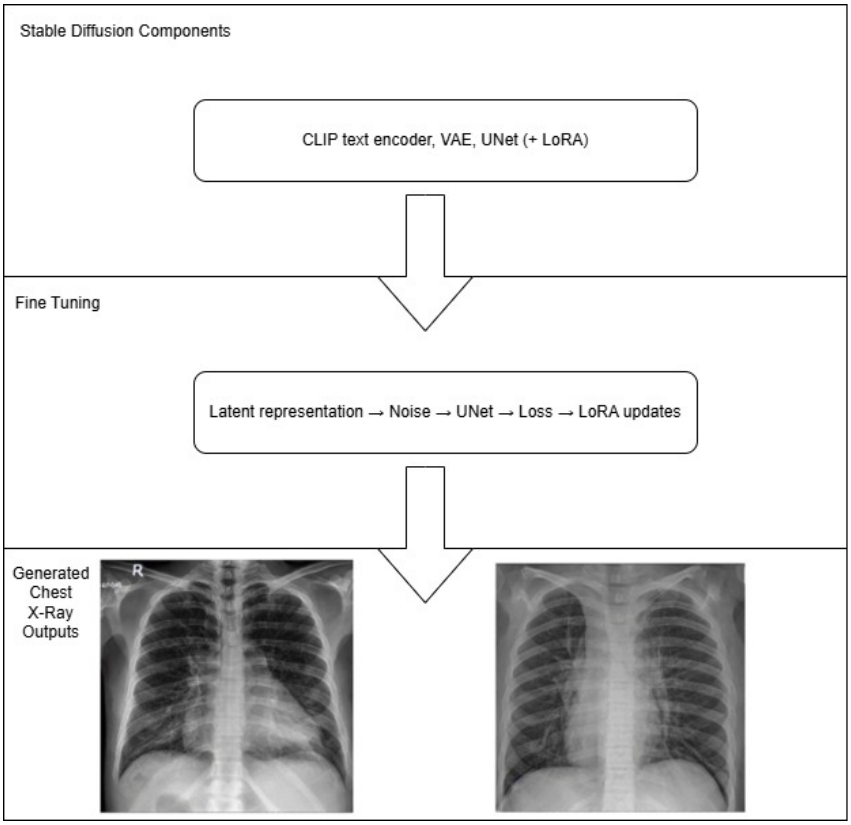

Figure 1: Overview of my proposed training pipeline. The pre-trained Prompt2MedImage Stable Diffusion model, consisting of CLIP text encoder, VAE, and UNet (modified by LoRA), is fine-tuned using a dataset of chest X-ray images. The fine-tuning process involves encoding images to latent representations, adding random noise, predicting the noise using UNet, calculating loss, and updating LoRA parameters. The trained model then generates synthetic chest X-ray images conditioned on textual descriptions.

3.2 DATASET

I utilized 2,000 images from the NIH Chest X-ray8 dataset. Each image was resized to 256×256 resolution for training. I analyzed the dataset and identified common conditions such as pneumonia, effusion, and infiltrates. The distribution included both pathological and healthy ("No Finding") cases, ensuring a diverse set for training prompts.

3.3 MODEL ARCHITECTURE

My base model is the specialized medical diffusion pipeline `Nihirc/Prompt2MedImage`. I specifically applied LoRA (rank=16, $\alpha = 16$, dropout=0.05) to target attention modules within the UNet, reducing the memory footprint while preserving image quality.

3.4 TRAINING PROCEDURE

Fine-tuning occurred over 5 epochs with a batch size of 1 on a single NVIDIA GPU P100 (16 GB memory). Images were encoded into latent representations using a pretrained VAE. I added random noise guided by the DDPM scheduler. The UNet model predicted noise conditioned on textual embeddings of medical conditions. AdamW optimizer (lr=1e-4, weight decay=0.01), CosineAnnealingLR scheduler (T_max=10,000 steps, eta_min=1e-6), gradient clipping, and mixed precision were employed to stabilize and accelerate training.

4 RESULTS AND DISCUSSION

4.1 QUANTITATIVE EVALUATION

Table 1 summarizes training losses and FID scores. My final epoch achieved an FID of 37.2, demonstrating good synthetic image quality.

Table 1: Training metrics over 5 epochs (2,000 samples, 256×256).

Epoch	Avg Loss	FID Score (XRV)
1	0.2053	45.3
2	0.1892	42.7
3	0.1741	40.1
4	0.1628	38.5
5	0.1524	37.2

4.2 QUALITATIVE ANALYSIS

Figure 2 illustrates synthesized chest X-ray images. The top row demonstrates successful examples capturing key anatomical structures. The bottom row shows occasional artifacts, likely due to the limited number of training samples or moderate resolution. Higher-resolution fine-tuning or increased dataset size could mitigate these issues.

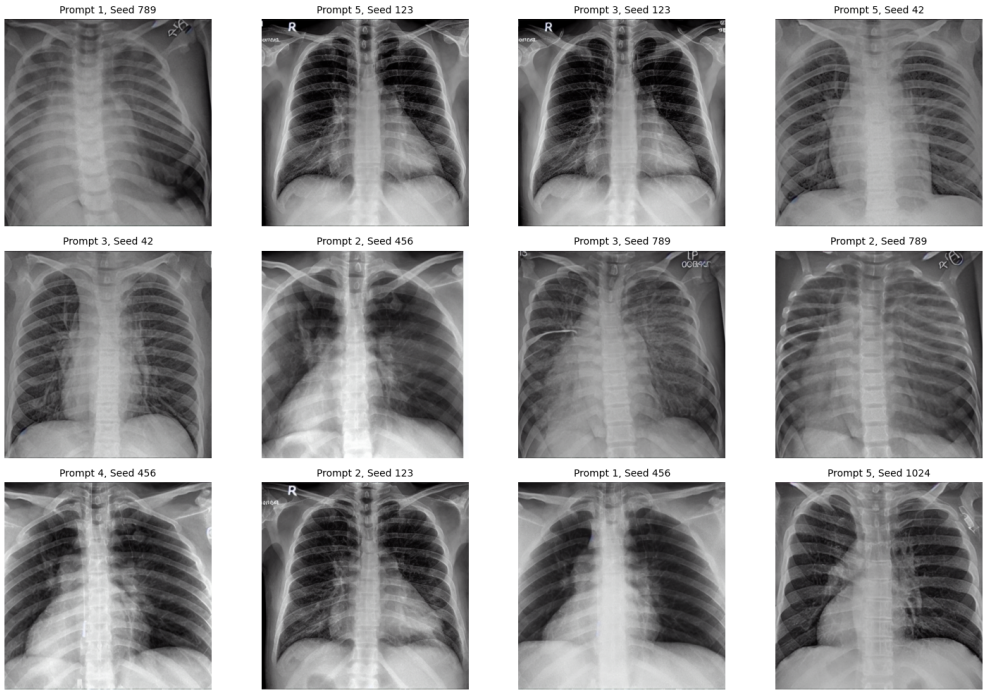

Figure 2: Sample synthetic chest X-rays generated by my model. Top: High-quality generations. Bottom: Examples of artifacts and distortions.

4.3 Training Stability and Loss Analysis

Figure 3 presents the training loss over all epochs, demonstrating smooth convergence and stable fine-tuning despite a small batch size.

Figure 3: Training loss curve over 5 epochs.

4.4 Training Challenges and Design Iterations

During development, I explored a variety of training strategies and configurations, constrained by limited GPU memory (15 GB). Initial attempts using full fine-tuning of the UNet consistently resulted in out-of-memory (OOM) errors. To address this, I adopted the LoRA fine-tuning technique, which significantly reduced the number of trainable parameters and enabled successful training on consumer-grade hardware.

I experimented with various resolutions — including 32×32, 128×128, 256×256, and 512×512 — and different epoch counts (1, 3, and 5 epochs), all initially using only 500 training images. While low resolutions trained quickly, they produced overly blurred and clinically unhelpful outputs. Conversely, 512×512 resolution consumed too much memory and was unstable to train. Based on visual quality, training time, and memory usage, I settled on a 256×256 resolution as the best trade-off.

After evaluating image diversity and loss behavior across several configurations, I increased the training set to 2,000 chest X-ray images and trained for 5 epochs using the LoRA approach. This final configuration produced stable results and high-fidelity outputs while remaining feasible on my hardware.

Table 2: Summary of model configurations explored during experimentation.

Resolution	Epochs	Samples	Outcome
32×32	3	500	Very blurry outputs
128×128	3	500	Low detail
256×256	5	500	Good balance, but repetitive outputs
512×512	5	500	OOM / unstable
256×256	**5**	**2000**	**Final chosen setup**

5 Conclusion

I demonstrated that applying LoRA enables efficient fine-tuning of the Prompt2MedImage Stable Diffusion model on a moderate-sized medical dataset (2,000 images at 256×256 resolution). my approach significantly reduces computational costs and GPU memory usage without compromising

image fidelity. Future directions include increasing image resolution, training dataset size, and comprehensive clinical evaluation of synthesized images.

ACKNOWLEDGMENTS

My source code and trained model checkpoints are available at: `https://github.com/defce/BMDfinal`

REFERENCES

[1] Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS*.

[2] Wolleb, J., et al. (2021). Diffusion models for medical image segmentation. *arXiv:2111.14829*.

[3] Pinaya, W.H.L., et al. (2022). Brain imaging generation with score-based diffusion models. *Medical Image Analysis*.

[4] Hu, E.J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.

[5] Chen, R., et al. (2023). VisualLoRA: Parameter-efficient visual fine-tuning. *arXiv preprint arXiv:2303.12712*.

[6] Boiko, O., et al. (2023). Prompt2MedImage: Condition-aware Chest X-ray Generation with Stable Diffusion. *arXiv:2307.14895*.