Adapting Pretrained Stable Diffusion models to generate Chest X-ray images

Alaqian Zafar, Chinmay Tompe, Akshay Gowda and Haoyang Pei
Department of Electrical and Computer Engineering
New York University, Tandon School of Engineering
aa7z118@nyu.edu, cst9314@nyu.edu, ah6147@nyu.edu, hp2173@nyu.edu

***Abstract*—Deep learning-based image synthesis methods have shown promising advancements in generating diverse, high-quality medical images in recent years. We present an approach for generating synthetic chest X-rays using latent diffusion models (LDMs), that achieve convincing results in image synthesis. LDMs are typically trained on powerful pre-trained autoencoders in the latent space, significantly reducing the computational resources required while retaining quality and flexibility. This paper uses Stable Diffusion, a popular open-source diffusion model. We explore two different finetuning techniques to teach the model the concepts of a chest X-ray. We propose adopting a low-rank adaptation (LoRA) method with Dreambooth for fine-tuning large pre-trained language models, reducing the number of trainable parameters, for downstream tasks while maintaining model quality. Our approach can help generate synthetic chest X-rays and improve the availability of healthcare datasets. These images can potentially be used for training machine learning models, data augmentation, and clinical applications.**

I. Overview

One of the biggest challenges in medical imaging is the scarcity of data available for training deep-learning models, especially for rare diseases. For instance, the lack of dermatology data for darker skin tones is a problem that results in reduced model performance for underrepresented groups [1]. Acquiring images through special techniques such as Diffusion MRI can be time-consuming and costly, leading to a shortage of data. The confidentiality of this data and the costs associated with labelling the data using medically trained professionals, stifles scientific development [2].

Stable Diffusion is a promising new generative Artificial Intelligence technique that could help to overcome this limitation by generating synthetic versions of medical images. This can be especially beneficial for startups that seek to use medical images but face limitations in accessing sensitive and private information. Synthetic images can help to improve the performance of deep-learning models by providing them with a more diverse and comprehensive dataset to train on. This is especially important for models that are used to diagnose diseases or make other medical decisions, where accuracy is critical. Synthetic images can be generated much faster and cheaper than real images. which can help to accelerate the development of new medical imaging applications.

II. Related Work

Recently, researchers at Stanford used Stable Diffusion to generate chest X-rays and found that training on a combination of natural and synthetic images can increase classification performance [3][4]. The model was also able to generate images of specific diseases, such as pneumothorax. The authors use image quality metrics and human domain expert evaluations to evaluate the model's performance. They find that the resulting model, called RoentGen, can create visually convincing and diverse synthetic chest X-ray images and that the output can be controlled using radiology-specific language in the text prompts. The authors demonstrate the usefulness of RoentGen for data augmentation and disease representation. They show that finetuning the model on a fixed training set can improve the performance of a classifier trained jointly on synthetic and real images and that the model can improve its representation capabilities of certain diseases like pneumothorax by 25%.

The Stable Diffusion pipeline used in the study consists of a variational autoencoder (VAE), a conditional denoising U-Net, and a conditioning mechanism using a CLIP text encoder, as shown in Figure 1. The variational autoencoder (VAE) is a neural network that learns to represent images as a latent vector [5]. This vector is a lower-dimensional representation of the image that contains all of the essential information about the image. The conditional denoising U-Net is a neural network that learns to denoise images [6]. It is called a "conditional" denoising U-Net because it is trained on a dataset of images that are paired with their corresponding latent vectors. This allows the network to learn to denoise images while also preserving the information that is encoded in the latent vector. The conditioning mechanism using a CLIP text encoder allows the model to be controlled using radiology-specific language in the text prompts. The CLIP text encoder is a neural network that learns to map text to images. This allows the model to learn to associate certain words and phrases with specific medical conditions.

Other applications of Stable Diffusion in the domain of generating medical images include a study where the authors used latent diffusion models to generate 3D MRI brain images conditioned on several covariates [7]. Researchers have also shown that these models could generate real microscopy image data in 2D and 3D based on simulated sketches of cellular structures [8]. These synthetic images were used to train a segmentation model that accurately segmented cells in an actual image dataset. These examples demonstrate the potential of Stable Diffusion for augmenting limited medical imaging datasets and improving deep-learning model performance.

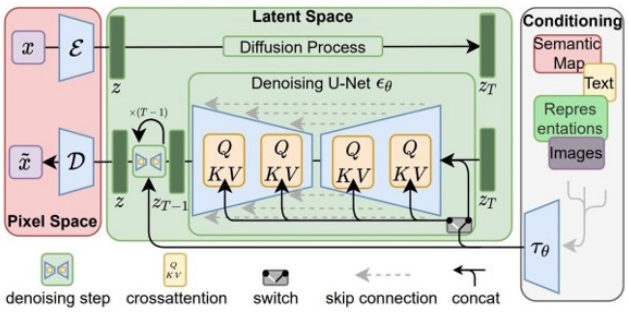

Fig. 1. The architecture of an implemented Standard Diffusion model[9]

III. Theoretical Principles

A. Stable Diffusion

The paper "High-Resolution Image Synthesis with Latent Diffusion Models" proposes a new method for generating high-resolution images using latent diffusion models (LDMs) [9]. The authors argue that current image synthesis methods need assistance to create high-resolution images with fine details, especially for large-scale datasets. They propose to use LDMs, a probabilistic model that iteratively refines a latent code to create an image.

The proposed method, called "Diffusion Probabilistic Models" (DPMs), uses a series of diffusion steps to refine the latent code. The latent code is perturbed by noise during each stage and gradually refined through an iterative diffusion process. The authors introduced several techniques to improve the training and inference efficiency of the model, including a dynamic down-sampling approach that allows for efficient computation of high-resolution images.

The approach described focuses on training diffusion models for high-resolution image synthesis. The first step involves analyzing already trained diffusion models in pixel space to determine the rate-distortion trade-off. The learning process is divided into two stages: a perceptual compression stage, which removes high-frequency details, and a semantic compression stage, which learns the semantic and conceptual components of the data.

To make the training process more computationally efficient, the authors aimed to find a perceptually equivalent but computationally more suitable space to train the diffusion models. They accomplished this by introducing an auto-encoder to provide a lower-dimensional representational space that is perceptually equivalent to the data space. This reduced complexity enabled efficient image generation from the latent space with a single network pass.

This approach has the notable advantage of requiring the universal auto-encoding stage to be trained only once, allowing it to be reused for multiple diffusion model training or exploring different tasks. The authors also design an architecture connecting transformers to the latent diffusion model's U-Net backbone to enable arbitrary token-based conditioning mechanisms for text-to-image tasks [10]. This approach thus allows for the efficient exploration of many diffusion models for various image-to-image and text-to-image tasks.

B. Dreambooth

Dreambooth is a novel approach for personalizing text-to-image diffusion models to generate user-specific images [11]. The method expands the language-vision dictionary of the model by binding new words with specific subjects that the user wants to generate. The approach represents a given subject with rare token identifiers and fine-tunes a pre-trained diffusion-based text-to-image framework. Autogenous, class-specific prior preservation loss is proposed to prevent language drift, which can cause the model to associate the class name with the specific instance.

The technique is applied to various text-based image generation applications, including subject recontextualization, property modification, and original art renditions. Ablation studies and user studies are conducted to evaluate subject and prompt fidelity in synthesized images. This is the first technique to tackle the problem of subject-driven generation, allowing users to synthesize novel renditions of a subject in different contexts while maintaining its distinctive features.

Rather than writing detailed image descriptions for each image, the images are labeled with `"a {identifier} {class noun}"`, where the identifier is a unique identifier linked to the subject. The class noun is a coarse class descriptor of the subject (e.g., "cat," "dog," "watch," etc.). The class descriptor is used to tether the prior of the class to the unique subject, allowing the model to leverage the visual prior of the specific class and entangle it with the embedding of the subject's unique identifier. This enables the generation of new poses and articulations of the subject in different contexts.

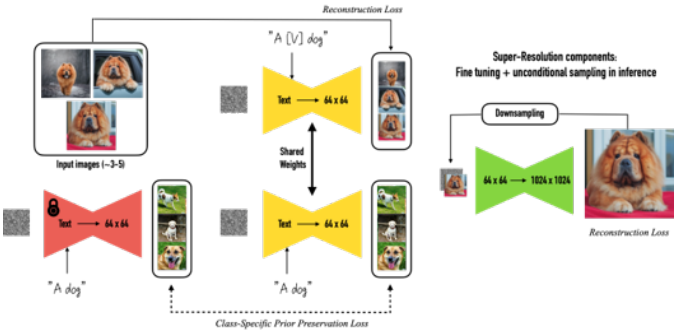

Fig. 2. Architecture of implemented Dreambooth model [11]

.

The authors propose using rare-token identifiers to label images in their diffusion model. They argue that using existing English words can be suboptimal as the model has to disentangle them from their original meaning. Instead, they look for rare tokens in the vocabulary and invert them into text space to minimize the probability of the identifier having a strong prior. They perform a rare-token lookup in the vocabulary and obtain a sequence of rare token identifiers using a tokenizer

function. The sequence can be of variable length, and they find that short sequences of 1 to 3 tokens work well. They then invert the vocabulary using the detokenizer on the token sequence to obtain a sequence of characters that define the unique identifier.

To achieve this, we first generate many images of the same class as the target subject using the diffusion model with random text embeddings. We then use these images to compute the class-specific prior by computing the mean and covariance of their diffusion model latent. During fine-tuning, we encourage the diffusion model to generate images with similar diffusion latent as those of the class-specific prior by minimizing the distance between the target subject's diffusion latent and the class-specific prior diffusion latent. This loss is added to the existing losses used to train the diffusion model.

The autogenous, class-specific prior preservation loss encourages the model to retain knowledge about the class-specific prior and thereby counteract language drift. It also encourages diversity in the generated images by allowing the model to generate images in novel viewpoints, poses, and articulations.

C. Low-Rank Adaptation (LoRA) of Large Language Models

Fine-tuning large pre-trained language models for multiple downstream applications can be challenging as there are billions of trainable parameters [12]. While fine-tuning can be effective, updating all the pre-trained parameters requires a new model with as many parameters as the original model. This can be a critical deployment challenge for larger models, as it may pose a trade-off between efficiency and model quality.

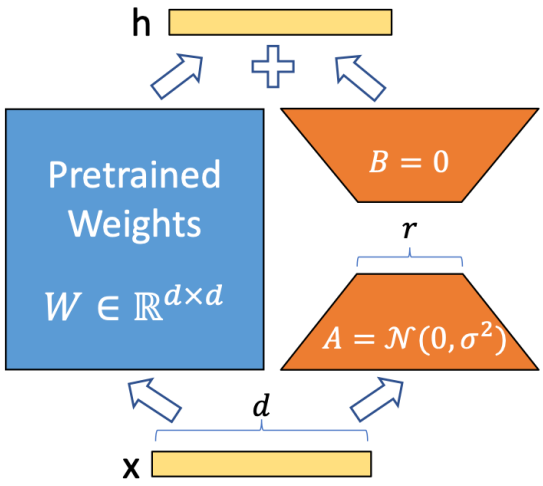

Fig. 3. The structure of LoRA. The Pre-trained Weights are frozen. A and B will be trained on downstream tasks [13].

The author proposes LoRA, a method for training over-parameterized models that is inspired by a previous work showing that these models reside on a low intrinsic dimension [13]. LoRA allows indirect training of some dense layers by optimizing rank decomposition matrices of the dense layers' change during adaptation while keeping the pre-trained weights frozen. LoRA is shown to be storage- and compute-efficient and allows a pre-trained model to be shared and used to build many small LoRA modules for different tasks.

The author also highlights that LoRA makes training more efficient and lowers the hardware barrier to entry by up to 3 times when using adaptive optimizers since only the injected, much smaller low-rank matrices need to be fine-tuned. Finally, LoRA is shown to be orthogonal to many prior methods and can be combined with them, such as prefix-tuning.

IV. Methodology

A. Project Objective

This project is largely inspired by and builds upon work that was done by researchers at Stanford University that successfully demonstrated that a Stable Diffusion model can be fine-tuned to generate a diverse set of high-fidelity chest x-ray images [3][4]. They used simple prompts, consisting of natural and medical language, such as "A photo of a lung xray" or "A photo of a lung xray with a visible pleural effusion", showing that these models can be used to generate images of specific medical conditions. They concluded that the best performance was observed by jointly fine-tuning both the pre-trained model and text encoder.

We wanted to expand on their work and investigate whether we could use the same approach to generate images of specific medical conditions, but using more complex prompts and concept, such as the specific location of the condition in the image, or having more than one condition at different locations in an image. In addition to this, we wanted to incorporate the ability to condition the generative process on variables like age, sex, projection and modality similar to the work involving the generation of synthetic brain MRI images [7]. In doing so, we waned to develop a comprehensive tool that could be used to generate synthetic images of specific medical conditions. This would serve as a proof of concept and could be used to generate synthetic medical images in other domains.

Furthermore, we also wanted to explore the use of an alternative fine-tuning technique that incorporates Dreambooth with the LoRA method to fine-tune the pre-trained model. This would allow us to train the model on a smaller dataset, which would be more accessible. LoRA layers can be stacked on top of each other, which would allow us to train the model on different X-ray images, each with a different level of specificity to the medical domain.

Finally, we wanted to test some of the more recent open-source Stable Diffusion models (v1.5 and v2) that have a larger number of parameters and are trained on a larger dataset making them capable of generating more realistic and detailed image.

B. Dataset

Stanford used the MMIC-CXR dataset that has 14 different labels for chest x-rays assigned by radiologist based on the presence or absence of specific findings [14]. In order to train a more extensive model, we decided to use the PadChest dataset

that includes annotations for 174 disease labels [15]. PadChest is a publicly available chest X-ray dataset that contains over 160,000 chest X-ray images with corresponding radiology reports. The annotations were created using a combination of automatic and manual labeling techniques.

The PadChest dataset is used in machine learning research for image classification and disease diagnosis tasks. It's important to note that the dataset has some limitations, such as the fact that the distribution of disease labels may not represent real-world populations.

- Image size: 1024 x 1024 pixels
- Image format: PNG
- Number of patients: 67,864
- Number of unique reports: 376,737
- Number of disease labels: 174
- Number of images: 160,868

C. Prompt Formulation

For fine-tuning the model, we used prompts of varying levels of complexity with the simplest prompt being "chest x-ray". For more complex prompts, we used the the following labels as input text prompts:

- `StudyDate_DICOM`
- `PatientBirth`
- `PatientSex_DICOM`
- `ViewPosition_DICOM`
- `Projection`
- `Modality_DICOM`
- `LabelsLocalizationsBySentence`

The highest complexity prompt would be structured in this format "chest x-ray, age {StudyDate_DICOM - PatientBirth} {sex}, view {ViewPosition_DICOM}, projection {view_projection}, modality {Modality_DICOM}, diagnosis {LabelsLocalizationsBySentence}". For instance, "chest x-ray, age 47 male, view PA, projection PA, modality DX, diagnosis calcified granuloma, loc right upper lobe". These prompts, along with their x-ray images, formed a image-prompt pair.

D. Data Preprocessing

The data used to fine-tune the model were manually inspected. Images with bad cropping, poor contrast, low quality and mislabelled images were discarded.

The original x-ray images were saved in 1-channel 32-bit unsigned integer grayscale format which were converted to 3-chanel 8-bit unsigned integer RGB images, which is the format used to train Stable Diffusion models.

E. Training details

Experiments were conducted using the full fine-tuning technique and the combination of Dreambooth and LoRA method. Three different devices were used in order to train the models; an Nvidia T4 with 16 GB Vram usning Google Colab, an Nvida V100 with 16 GB VRAM on a Google Cloud Platform virtual machine compute engine and an Nvida RTX 3070ti with 8 GB VRAM on a local machine. The limited amount of VRAM available as well limited time an experiment could be run on Colab or GCP served as a significant bottleneck. For full-finetuning, we were limited to using around 500 images and a batch size of 2, which was significantly smaller in comparison to the models trained by Stanford that were fine-tuned on a subset of 1.1 - 5.5K images.

Model weights were obtained from the Hugging Face platform. For the experiments, we tested Stable Diffusion v1.4, v1.5 and v2.0. Initial testing showed the best results fine-tuning on a smaller dataset were observed using Stable Diffusion v1.5 and thus they were used in all of the subsequent experiments.

F. Evaluation

In recent years, there have been many methods proposed to evaluate synthetic medical images [16]. Some of these include creating feature maps and evaluating the quality of noise in the generated image. While these methods are not reliable entirely, they provide some idea about how to approach this task. For our implementation, we use Fréchet Inception Distance (FID) to calculate the similarity score between images from the original set of images and generated images. The distance between the distribution of the synthetic images and real images is calculated to get the FID score. Lower score indicates better quality images.

V. Results

A. Full Fine-tuning

This section discusses two full fine-tuning experiments that were conducted. The first experiment consisted of a dataset consisting of 8 AP/PA radiography images, whereas the second consisted of a dataset of 325 images. Both these runs used a simple prompt of "Chest X-ray" and used normal x-rays i.e. images where there was no finding. This was in order to teach the model the concept of a "chest x-ray" to replicate the methods used by Stanford and to serve as a control to compare the other method with.

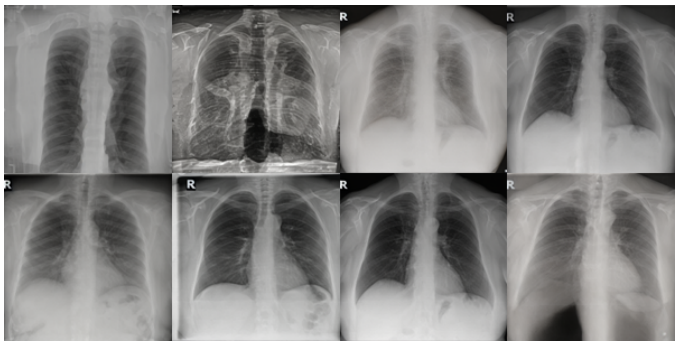

Fig. 4. Samples generated while full fine-tuning using a dataset of 8 normal AP/PA radiography images with train prompt "Chest X-ray" with the following step count (left to right, top to bottom): 100, 200, 600, 1000, 2000, 3800, 5700 and 6000.

Figure 4 and Figure 5 show samples taken during training. It did not take long for both models converge to a loss of around 0.1 in about 1000 steps. The model does not take long to start producing visually convincing xray images. The images have a similar look to the training images.

In the first run, the diversity of the images is low, and they look quite similar to each other and to the training set. The consistency of the results was also poor, with a lot of images produced with obvious errors. This could be explained by the small dataset as well as over-fitting. Both consistency and diversity of the images improved as the dataset size was increased in the second run.

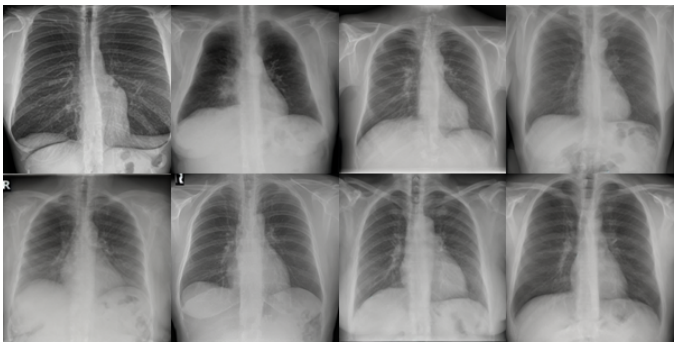

Fig. 5. Samples generated while full fine-tuning using a dataset of 325 normal AP/PA radiography images the prompt "Chest X-ray" with the following step count (left to right, top to bottom): 100, 200, 800, 2400, 4000, 5000, 7400 and 9100.

B. Dreambooth and LoRA

For the first run, a dataset of 960 AP/PA radiography images were used with normal as well as combinations of different morbidities. The highest complexity prompts were used to train these images. Figure 6 shows samples generated during the training of this model. The first one or two samples show the model generating images very close to what the pre-trained model would produce. They seem to resemble more closely to 3D illustrations of the chest anatomy than xray images. As the training progresses, the samples become black and white and begin resembling xray images. The texture of these images look glossy, which is similar to the results from Stanford's first paper [3]. In comparison, the full fine-tuning produced more realistic xray images. This can be explained by the fact that we are only training a few layers and preserving the original network. Rather than trying to recreate the training images, it is transforming its original results to match the training data and minimize the loss between the images.

When testing different prompts, the model was able to correctly generate x-rays of male and female patients. In another run, where different projections were used, the model was also able to learn the difference between an AP/PA projection with a lateral projection. However, the model did not learn how to display different morbidities in the image, such as scoliosis or pacemaker. Moreover, there was no apparent different between digital and classical x-ray modalities. This is due the difference between images of different morbidities being way too minor for the model to learn. Additionally, the text encoder might be assigning a higher weightage to the sex labels, but a lower weightage to the labels concerning the diagnosis. This is done using a technique called attention, which allows the model to learn which tokens are most important for generating the

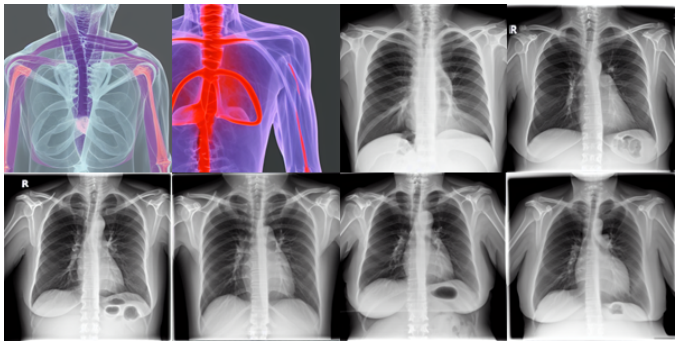

Fig. 6. Samples generated while using Dreambooth and LoRA using a dataset of 960 mixed AP/PA radiography images using a complex prompt at the following epochs (left to right, top to bottom): 1, 2, 10, 50, 80, 130, 155 and 180 steps.

desired image [6]. The weights are then used to control the probability of each token being included in the generated image. Using text encoders specifically designed for medical or radiography language, such as RadBERT [17] as opposed to the default encoder is likely to yield better results.

For the second run, a dataset of 325 AP/PA radiography images were used with normal diagnosis. Figure 7 shows the samples taken during training of this run. The trend in training is quite similar to that of the previous run where the initial images are colored 3D illustrations which gradually transform to match the color and features of the training set.

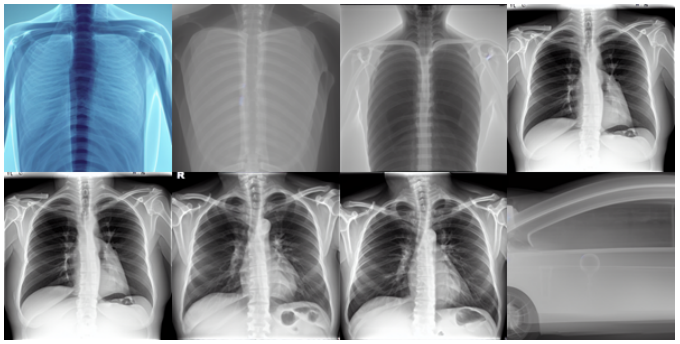

Fig. 7. Samples generated while using Dreambooth and LoRA using a dataset of 325 normal AP/PA radiography images using a complex prompt with the following epochs (left to right, top to bottom): 1, 2, 3, 10, 25, 35, 40 steps and "Car X-ray prompt".

When testing the model with the prompt "An X-ray of a car" it was able to produce a result that looks like a mix of an xray and a car showing that the model has learnt the concept of an xray while retaining the prior concepts. This is further illustrated in Figure 8 that shows the results produced by the two full fine-tuned and Dreambooth/LoRA models when given the prompt "A bird sitting on a tree". The fully fine-tuned has completely forgotten its prior concepts while the Dreambooth/LoRA model has preserved its prior concepts.

This might serve as a baseline model to train subsequent LoRA layers in future experiments with additional concepts of morbidities built over the new layers.

Fig. 8. Images generated using the prompt, "A bird sitting on a tree", from left to right, top to bottom: Full fine-tuning with 8, and 325 images followed by Dreambooth/LoRA using 960 and 325 images.

TABLE I
TASKS AND SUB-TASKS OF THE PROJECT

	Akshay	Alaqian	Chinmay
Literature Survey			
Stable Diffusion	✓	✓	✓
U-Net	✓		✓
CLIP			✓
Dreambooth	✓		
LoRA	✓	✓	✓
FID			✓
Medical imaging applications	✓	✓	✓
Code Implementations	✓	✓	✓
Datasets		✓	✓
Midterm Report			
Abstract			✓
Overview	✓	✓	
Remaining work & schedule		✓	
References		✓	
Experiments			
Data acquisition		✓	✓
Image preprocessing	✓	✓	
Prompt/label generation	✓	✓	
Data organization		✓	
Setting up GCP		✓	✓
Design of experiments	✓	✓	
Running experiments/training	✓	✓	
FID score calculation			✓
GitHub		✓	
Presentation			
Introduction		✓	✓
Theoretical principles	✓	✓	✓
Related work	✓	✓	
Methodology	✓	✓	✓
Results		✓	✓
Conclusions	✓		
Final Report			
Abstract	✓		✓
Related work	✓-	✓	
Theoretical principles	✓		
Methodology	✓	✓	✓
Results	✓	✓	✓
Conclusions	✓		
References	✓	✓	✓

VI. CONCLUSION

A. Project Accomplisments

We were able to learn from the literature survey about the principles of stable diffusion and implemented different fine-tuning techniques using less demanding processes to replicate similar results. Able to reflect on mistakes and correct them to Trained a model with the concept of chest X-ray with prior preservation. Although these models have promising results, there is a challenge of ethics and free distribution owing to the terms dataset curators, as the reproduced results learned from the original source

Our literature survey provided us with profound insights into the principles governing stable diffusion. By examining research papers and studies, we have been able to delve into this details and underlying mechanisms of stable diffusion. The comprehensive analysis of existing literature has equipped us with a deep understanding of the fundamental principles governing various systems' stability and diffusion. This extensive exploration has broadened our knowledge and enabled us to identify critical factors and variables that contribute to stable diffusion.

By implementing various fine-tuning techniques, we replicated the results of full fine-tuning while employing less demanding processes. One such technique we explored is DreamBooth, which selectively adapts parameters and incorporates external modules for specific tasks. Additionally, we integrated LoRA (Low-Rank Adaptation) into the DreamBooth framework, leveraging the low intrinsic dimensionality of over-parameterized models. By optimizing rank decomposition matrices during adaptation and keeping the pre-trained weights frozen, LoRA significantly reduced the number of parameters and storage requirements.

Table I displays what member of the group completed each sub-task.

B. Ethical concerns

Although such models seem promising and their resurgence has open up new avenues for the medical field, there is an ever bigger need for efforts towards developing guidelines that prohibit misuse and unethical practices. The goal of generating synthetic data is help train unbiased and accurate models that can help human medical professionals to oversee more cases.

The use of models to generate chest X-rays has shown promise in their results. However, one of the challenges is the ethical implications and the issue of freely distributing these generated images. This challenge stems from the terms and

conditions set by the dataset curators, which may restrict the reproduction of results learned from the original data source.

Therefore, when working with pre-trained models or generating chest X-rays, it is essential to respect these ethical considerations and adhere to the terms and conditions set by the dataset curators. This may include seeking permissions, obtaining necessary licenses, or adhering to specific usage restrictions.

C. Carbon Emissions Related to Experiments

Experiments were conducted using Google Cloud Platform in region us-west1, which has a carbon efficiency of 0.3 $kgCO_2eq/kWh$. A cumulative of 40 hours of computation was performed on hardware of type Tesla V100-PCIE-16GB (TDP of 300W).

Total emissions are estimated to be 3.6 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider.

Estimations were conducted using the MachineLearning Impact calculator presented in [18].

D. Future Work

For future work, we can test specialized VAEs and vary different parameters of training, such as learning rate, optimizers or newer iterations of the Stable Diffusion models. We can add more LoRA layers to the original model to teach it increasingly complex concepts. We need to do experiments with larger datasets which demand more computational resources. Finally, this approach might be adapted to other domains in the medical imaging field.

References

[1] L. W. Sagers, J. A. Diao, M. Groh, P. Rajpurkar, A. S. Adamson, and A. K. Manrai, "Improving dermatology classifiers across populations using images generated by large diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2211.13352

[2] V. Fernandez, W. H. L. Pinaya, P. Borges, P.-D. Tudosiu, M. S. Graham, T. Vercauteren, and M. J. Cardoso, "Can segmentation models be trained with fully synthetically generated data?" in *Simulation and Synthesis in Medical Imaging*. Cham: Springer International Publishing, 2022, pp. 79–90. [Online]. Available: https://arxiv.org/abs/2209.08256

[3] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," 2022. [Online]. Available: https://arxiv.org/abs/2210.04133

[4] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, "Roentgen: Vision-language foundation model for chest x-ray generation," 2022. [Online]. Available: https://arxiv.org/abs/2211.12737

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022. [Online]. Available: https://arxiv.org/abs/2204.06125

[7] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2209.07162

[8] D. Eschweiler and J. Stegmaier, "Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image data sets," 2023. [Online]. Available: https://arxiv.org/abs/2301.10227

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://arxiv.org/abs/1505.04597

[11] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023. [Online]. Available: https://arxiv.org/abs/2208.12242

[12] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, pp. 1–16, 2023.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[14] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," 2019. [Online]. Available: https://arxiv.org/abs/1901.07042

[15] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101797, dec 2020. [Online]. Available: https://doi.org/10.10162Fj.media.2020.101797

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," pp. arXiv–1706, 2018. [Online]. Available: https://arxiv.org/abs/1706.08500

[17] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics

to improve biomedical vision–language processing," in *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2022, pp. 1–21. [Online]. Available: https://doi.org/10.10072F978-3-031-20059-5_1

[18] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv preprint arXiv:1910.09700*, 2019.

APPENDIX

The code for our work can be found at this GitHub Repository and sample images and models can be found on this Google Drive folder.