NANYANG TECHNOLOGICAL UNIVERSITY

Fine-tuning Stable Diffusion on Fashion with LoRA

Kozhimadam Shahin Shah

College of Computing and Data Science

2024

NANYANG TECHNOLOGICAL UNIVERSITY

MSAI Master Project MSAI/23/066

Fine-tuning Stable Diffusion on Fashion with LoRA

Submitted by:
Kozhimadam Shahin Shah
under the supervision of
Prof Zheng Jianmin

College of Computing and Data Science

2024

Statement of Originality

I hereby certify that the work embodied in this report is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

Kozhimadam Shahin Shah

Date: 29/10/2024

Supervisor Declaration Statement

I have reviewed the content and presentation style of this report and declare that it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and the writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Prof Zheng Jianmin

Date: 29/10/2024

Authorship Attribution Statement

This report **does not** contain any material from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

Kozhimadam Shahin Shah

Date: 29/10/2024

Contents

Abstract

This dissertation explores the use of Low-Rank Adaptation (LoRA) to fine-tune the Stable Diffusion XL model for generating high-quality fashion images from textual descriptions. The research is driven by the growing demand for precise and efficient text-to-image generation in the fashion industry, where creating realistic images of garments on models can be resource-intensive. Through a detailed review of foundational and advanced generative models, this study highlights the impact of diffusion models and LoRA fine-tuning. Comprehensive experiments were carried out to determine the optimal LoRA configuration that maximizes image quality while minimizing the number of trainable parameters. The findings reveal that applying LoRA to the UNet component of Stable Diffusion XL, specifically with a rank of 1, significantly enhances the model's capacity to generate highly detailed and semantically accurate fashion images. To demonstrate the effectiveness of the proposed approach, a practical application is presented where the optimal configuration is used to generate images of given garment being worn by models in a variety of poses, offering a cost-effective alternative to traditional photoshoots and showing its potential for real-world use in fashion marketing.

Keywords: Text-to-image generation, Stable Diffusion, Stable Diffusion XL, Low-Rank Adaptation (LoRA), fashion image synthesis, generative models, model fine-tuning, diffusion models, image quality, semantic accuracy.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. Zheng Jianmin, for his invaluable support and guidance throughout this research project. His expertise and insights were instrumental in shaping the direction of this work. I am particularly grateful to him for proposing the research topic, which has been both challenging and rewarding. His continuous review of my progress, constructive feedback, and unwavering encouragement have been crucial in guiding me towards the right direction, ensuring the successful completion of this dissertation.

I would also like to thank my family, friends, and colleagues for their constant support and understanding during this time. Their encouragement has been a source of motivation for me.

Finally, I would like to acknowledge everyone who has contributed to this research in one way or another. Your assistance and support have been greatly appreciated.

Acronyms

AI Artificial Intelligence

CLIP Contrastive Language–Image Pre-training

DDIM Denoising Diffusion Implicit Models

DDPM Denoising Diffusion Probabilistic Models

FID Fréchet Inception Distance

GAN Generative Adversarial Network

LoRA Low-Rank Adaptation

MS COCO Microsoft Common Objects in Context

PEFT Parameter-Efficient Fine-Tuning

SDXL Stable Diffusion XL

UNet U-shaped Convolutional Neural Network

VQGAN Vector Quantized Generative Adversarial Network

ViT Vision Transformer

Symbols

W_q	Query weight matrix in the attention mechanism
W_k	Key weight matrix in the attention mechanism
W_v	Value weight matrix in the attention mechanism
W_o	Output weight matrix in the attention mechanism
α	Scaling factor used in LoRA fine-tuning

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The field of text-to-image generation has seen remarkable advancements due to the evolution of deep learning and generative models. This technology enables the creation of images from textual descriptions, merging the domains of natural language processing and computer vision. Early models, such as alignDRAW [2] and GANs [3], laid the foundational groundwork by demonstrating the feasibility of generating images from text. Subsequent innovations, particularly diffusion models like Stable Diffusion, have significantly enhanced the quality and coherence of the generated images. More recently, techniques such as Low-Rank Adaptation (LoRA) [4] have been introduced to efficiently fine-tune these large models, further improving their performance for specific tasks. This research focuses on fine-tuning Stable Diffusion specifically for fashion applications using LoRA, aiming to generate high-quality images that accurately reflect detailed fashion descriptions.

1.2 Motivation

The motivation behind this research is to determine the optimal LoRA configuration for fine-tuning, with the goal of achieving high-quality image generation while minimizing the number of trainable parameters. Additionally, we explore an application where the model is fine-tuned using a single image of a garment, enabling the generation of images depicting models wearing that garment. This approach is particularly valuable for fashion marketing, where quickly producing specific, high-quality visual content is essential. Given that the fine-tuned LoRA layers are tailored to individual garments, it is crucial that the parameter count remains low without compromising the quality of the generated images.

1.3 Objectives and Specifications

The primary objectives of this dissertation are:

1. **To review the foundational models and advancements in text-to-image generation:** This includes a detailed review of early models, GAN-based approaches, and the latest diffusion models.

2. **To investigate the effectiveness of LoRA fine-tuning on Stable Diffusion XL for fashion applications:** This involves applying LoRA to the UNet and text encoder components of the model, testing different ranks, and assessing the impact on image quality and semantic accuracy.

3. **To develop and validate a methodology for efficient model fine-tuning:** The aim is to create a process that balances performance improvements with computational efficiency.
4. **To apply the fine-tuned model to practical fashion scenarios:** Specifically, generating images of specified garments worn by models in various poses, thereby demonstrating the practical applications of the improved model.

1.4 Organisation of the Dissertation

This dissertation is structured as follows:

- **Chapter Two: Literature Review** - This chapter provides a detailed review of the foundational models and key advancements in text-to-image generation, including early models, GANs, and diffusion models.
- **Chapter Three: Determining Optimal LoRA Configuration** - This chapter presents a comprehensive study aimed at identifying the optimal LoRA configuration for fine-tuning the Stable Diffusion XL model. It includes both the methodology and experiments conducted to evaluate the model's performance.
- **Chapter Four: Optimal LoRA Configuration in Practice** - This chapter explores a practical application using the identified optimal LoRA configuration, demonstrating its capabilities in generating images of a given garment being worn by models in various poses, providing an alternative to costly photoshoots.
- **Chapter Five: Conclusion and Future Work** - This chapter discusses the findings and key contributions of this research, and provides recommendations for future work.

Chapter 2

Literature Review

The evolution of text-to-image generation has been marked by significant advancements, driven by the continuous progress in deep learning and generative modeling techniques. This literature review provides a comprehensive overview of the foundational models, key advancements, and fine-tuning techniques that have shaped the current landscape of text-to-image synthesis. We begin by exploring the early models that laid the groundwork for this technology, followed by a detailed discussion on the progress made with diffusion models and subsequent innovations. Finally, we delve into the techniques used to fine-tune these models for specific tasks, highlighting both traditional methods and recent developments like Low-Rank Adaptation (LoRA) [4] and personalized image generation techniques such as Dreambooth [5], which allows for the customization of pre-trained models to capture user-specific visual concepts.

2.1 Foundation of Text-to-Image Models

This section reviews the foundational models that have laid the groundwork for current text-to-image generation techniques.

2.1.1 Early Models and alignDRAW

In 2015, Gregor et al. introduced the DRAW (Deep Recurrent Attentive Writer) architecture [6], which utilized a recurrent variational autoencoder with an attention mechanism to iteratively generate images. The alignDRAW model [2] extended this architecture by conditioning the image generation process on textual descriptions. This model represented one of the earliest attempts to integrate text and image modalities, allowing it to generate 32×32 pixel images from text sequences. Despite the low resolution and diversity of the images, alignDRAW demonstrated the potential of using deep learning for text-to-image generation by successfully generalizing to objects not present in the training data and handling novel prompts without simply memorizing the training set.

2.1.2 Generative Adversarial Networks for Text-to-Image

The next significant milestone came in 2016 when Reed et al. applied Generative Adversarial Networks (GANs) to the text-to-image task [3]. Their approach involved conditioning GANs on text descriptions to generate images of birds and flowers. The GAN architecture allowed for the creation of more visually plausible images compared to previous models. However, when applied to the more diverse COCO dataset, the generated images often lacked coherence in finer details, highlighting the challenge of generalizing across a broader range of objects and scenes.

2.1.3 Subsequent Models and Improvements

Following the initial successes of GAN-based models, researchers sought to improve the quality and coherence of generated images. VQGAN-CLIP [7] combined Vector Quantized Generative Adversarial Networks (VQGAN) [8] with Contrastive Language-Image Pre-training (CLIP) [9] to enhance the alignment between textual descriptions and generated images. This combination improved both the quality and fidelity of the outputs, addressing some of the limitations observed in earlier models. VQGAN-CLIP leverages CLIP's ability to understand and relate images and text, resulting in more accurate and contextually relevant image generation.

XMC-GAN [10] introduced cross-modal contrastive learning to further enhance text-to-image synthesis. By utilizing contrastive learning techniques, XMC-GAN improved the model's ability to generate coherent and contextually appropriate images from textual descriptions, making significant strides in the quality of generated images.

2.1.4 OpenAI's DALL-E

In January 2021, OpenAI unveiled DALL-E [11], a transformer-based model that captured widespread public attention. DALL-E demonstrated remarkable versatility and creativity in generating high-quality images from a diverse array of textual prompts. Leveraging the transformer architecture, which had shown great success in natural language processing tasks, DALL-E was able to interpret and generate detailed and contextually appropriate images from text. This model showcased the potential of large-scale transformer models in text-to-image generation, setting a new standard for the field.

2.2 Advances in Generative Models

Recent advancements in generative models have significantly improved the quality and efficiency of image synthesis. These developments have been driven by innovations in diffusion models and the introduction of advanced architectures like Stable Diffusion and Stable Diffusion XL. This section explores these key advancements, highlighting their contributions to the field of text-to-image generation.

2.2.1 Diffusion Models

Diffusion models have emerged as an effective class of generative models, especially in image synthesis, by employing an iterative denoising process to generate high-quality images. Denoising Diffusion Probabilistic Models (DDPMs) [12, 13] introduced the concept of formulating the generative process as a Markov chain of diffusion steps, where noise is gradually added to the data in the forward process, and the model learns to reverse this process to generate new data from noise.

In DDPMs, the objective is to approximate the data distribution $q(x_0)$ by learning a model distribution $p_\theta(x_0)$ that can be efficiently sampled. This objective is formalized as:

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T}, \quad \text{where } p_\theta(x_{0:T}) := p_\theta(x_T) \prod_{t=1}^{T} p_\theta^{(t)}(x_{t-1}|x_t)$$

In this formulation, $x_1, \ldots, x_T$ represent latent variables defined in the same space as x_0, and the distribution $p_\theta(x_{0:T})$ describes a reverse process from x_T back to x_0, parameterized by θ. Unlike traditional latent variable models, which often have a trainable inference process, DDPMs employ a fixed inference process, making them more computationally tractable.

In this framework, the forward diffusion process $q(x_{1:T}|x_0)$ progressively adds Gaussian noise to the data x_0 over a sequence of T steps. This process is represented as:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) := \mathcal{N}\left(x_t \mid \sqrt{\alpha_t/\alpha_{t-1}}\, x_{t-1}, (1-\alpha_t/\alpha_{t-1})\, I\right)$$

where $\alpha_{1:T}$ is a decreasing sequence in $(0,1]$. The forward process progressively adds noise to the observation x_0, resulting in a noisy sample x_T that approaches a standard Gaussian distribution as T grows.

One of the key properties of this process is that x_t can be expressed as a combination of x_0 and a noise variable $\varepsilon \sim \mathcal{N}(0,I)$:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\varepsilon$$

As α_T approaches zero, $q(x_T|x_0)$ converges to a standard Gaussian distribution, making it natural to define $p_\theta(x_T) := \mathcal{N}(0,I)$.

During training, DDPMs learn to predict the noise added at each step by minimizing the following simplified objective:

$$L(\varepsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0), \varepsilon_t \sim \mathcal{N}(0,I)} \left[\| \varepsilon_\theta^{(t)}(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\varepsilon_t) - \varepsilon_t \|_2^2 \right]$$

where $\varepsilon_\theta := \{\varepsilon_\theta^{(t)}\}_{t=1}^{T}$ is a set of functions with trainable parameters θ, each mapping x_t to a noise estimate.

A significant drawback of DDPMs is the large number of sampling steps T, often set to 1000 or more, which makes inference slow and computationally demanding. To mitigate this, Denoising Diffusion Implicit Models (DDIMs) [14] were introduced. DDIMs modify the forward process to allow for a non-Markovian, deterministic sampling process, reducing the number of steps required for high-quality image generation while preserving the sample quality of DDPMs.

In DDIMs, the forward process (inference) and reverse process (generation) are formulated similarly to DDPMs but assume a non-Markovian dependence in the inference distribution $q_\sigma(x_{0:T})$, with the marginal distribution $q_\sigma(x_t|x_0)$ remaining the same as in DDPMs. The DDIM generative process is defined as:

$$q_\sigma(x_{1:T}|x_0) := q_\sigma(x_T|x_0) \prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0)$$

with $q_\sigma(x_T|x_0) := \mathcal{N}(\sqrt{\alpha_T} x_0, (1-\alpha_T) I)$ and the transition kernel for $t > 1$ given by:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}} x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t} x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I \right).$$

This deterministic reverse process in DDIMs retains the DDPM marginals, maintaining the model's sampling quality while achieving higher computational efficiency.

2.2.2 Stable Diffusion

Stable Diffusion [1] is a latent diffusion model that operates within a learned latent space, rather than directly in the pixel space, as shown in Figure 2.1. This approach significantly reduces computational costs while preserving high image quality. The model first employs an encoder $\mathcal{E}$ to map input images x into a lower-dimensional latent space z, where the diffusion process occurs. After the iterative denoising process, a decoder $\mathcal{D}$ is used to transform the denoised latent representations back into the original pixel space.

In the latent space, Stable Diffusion leverages a U-Net architecture [15] for the core denoising tasks, enhanced with attention mechanisms for greater accuracy. As illustrated in the figure, the denoising U-Net ϵ_θ operates iteratively on the noisy latent variable z_t across multiple steps to recover the original latent representation z. The model also incorporates a CLIP (Contrastive Language–Image Pre-training) based text encoder [9] to condition image generation with textual information. These text embeddings are introduced into the U-Net through cross-attention layers, allowing for precise alignment of the generated images with the semantic content of the input text.

This combination of latent space processing, the U-Net architecture, and CLIP-based text conditioning, as demonstrated in Figure 2.1, enables Stable Diffusion to be both computationally efficient and highly flexible, supporting a wide range of applications in image synthesis.

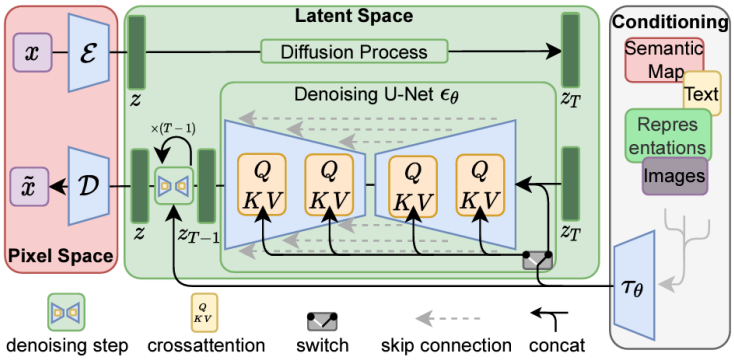


Figure 2.1: Architecture of Stable Diffusion. The model operates in latent space, utilizing a U-Net with cross-attention to integrate text-based conditioning from CLIP embeddings. Image taken from [1].

2.2.3 Stable Diffusion XL

Stable Diffusion XL (SDXL) [16] represents a significant advancement over its predecessors by incorporating several key improvements. SDXL uses a U-Net backbone three times larger than previous versions, with more attention blocks and a larger cross-attention context enabled by a second text encoder. This architecture enhances the model's ability to generate high-resolution images with fine details and accurate semantic alignment. The SDXL model features a heterogeneous distribution of transformer blocks within the U-Net, concentrating the bulk of transformer computations at lower levels of the network. This design choice improves efficiency and performance. The model employs OpenCLIP ViT-bigG [17] and CLIP ViT-L [9] as text encoders, concatenating their outputs to enhance text conditioning. SDXL introduces novel conditioning schemes, such as conditioning on the original image resolution and crop parameters, to better utilize training data and improve sample quality. Additionally, SDXL undergoes multi-aspect training, allowing it to handle images of varying resolutions and aspect ratios effectively.

A notable feature of SDXL is its refinement model, which further enhances the visual fidelity of generated images through a post-hoc image-to-image denoising process. This refinement stage helps produce highly detailed images, especially in complex scenes and human faces. Overall, the advancements in SDXL demonstrate significant improvements in both the architecture and training methodologies of diffusion models, setting a new standard for high-resolution image synthesis.

2.3 Fine-Tuning Techniques

Fine-tuning is a critical process in adapting pre-trained models to specific tasks or domains. This section explores two primary approaches to fine-tuning: full fine-tuning and Low-Rank Adaptation (LoRA), and their application in Stable Diffusion models.

2.3.1 Full Fine-Tuning

Full fine-tuning involves updating all the parameters of a pre-trained model on a task-specific dataset. This method is straightforward but computationally expensive, especially for large models like GPT-3, which has 175 billion parameters. The main advantage of full fine-tuning is that it can potentially yield the highest performance since all model parameters are optimized for the target task. However, the storage and computational requirements are significant. For instance, each fine-tuned model variant requires storing an additional 175 billion parameter set, making deployment and maintenance costly and inefficient.

2.3.2 Low-Rank Adaptation (LoRA)

To address the inefficiencies of full fine-tuning, Low-Rank Adaptation (LoRA) [4] was proposed. This technique freezes the original pre-trained model weights and injects trainable low-rank decomposition matrices into each layer of the Transformer architecture. By doing so, LoRA significantly reduces the number of trainable parameters needed for adaptation.

LoRA leverages the observation that the weight updates during fine-tuning have a low "intrinsic rank". Instead of updating the entire weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces two smaller matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, where $r \ll \min(d,k)$, such that the weight update can be represented as a low-rank product:

$$W_0 + \Delta W = W_0 + BA$$

During training, W_0 is kept frozen and does not receive gradient updates, while A and B contain trainable parameters. The product $\Delta W = BA$ is initialized to zero at the start of training, ensuring that the weight update ΔW is zero initially. The modified forward pass for an input x is therefore given by:

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

This reparameterization allows LoRA to achieve the following practical benefits:

- **Efficiency:** LoRA reduces memory and storage requirements significantly. For instance, fine-tuning large models like GPT-3 with LoRA requires only a fraction of the parameters compared to full fine-tuning.
- **Scalability:** LoRA enables efficient task switching by loading only the task-specific low-rank matrices while keeping the shared pre-trained model unchanged.
- **Performance:** Despite the reduction in trainable parameters, LoRA achieves performance comparable to or even surpassing full fine-tuning on several benchmarks.

LoRA's ability to efficiently fine-tune large language models with minimal computational overhead makes it an attractive alternative to traditional fine-tuning methods, especially when working with limited resources.

2.3.3 Fine-Tuning Stable Diffusion Models

Fine-tuning techniques are also applicable to Stable Diffusion models, which are pre-trained on large-scale datasets to generate high-quality images from textual prompts. The process involves adjusting the parameters of the model to better align with specific tasks or domains, such as generating images of specific objects or styles.

Full Fine-Tuning: In the context of Stable Diffusion models, full fine-tuning updates all parameters of the model, which can lead to significant improvements in performance. However, similar to language models, this approach is computationally expensive and requires substantial storage.

Low-Rank Adaptation (LoRA): Applying LoRA to Stable Diffusion models involves injecting trainable low-rank matrices into the model's architecture. This method is particularly beneficial for diffusion models due to their large size and complexity. By fine-tuning only a subset of parameters, LoRA makes it feasible to adapt Stable Diffusion models to new tasks without the heavy computational burden associated with full fine-tuning.

Overall, while full fine-tuning updates all model parameters, making it highly adaptable but resource-intensive, LoRA offers a more efficient alternative by adapting only a small subset of parameters through low-rank decomposition. This balance between efficiency and performance makes LoRA a powerful tool for fine-tuning large pre-trained models on diverse tasks, including Stable Diffusion models.

2.4 Personalized Image Generation Techniques

Personalized image generation techniques are designed to adapt large text-to-image models for generating images that reflect user-specific visual concepts using a limited number of reference images. One of the most prominent approaches in this domain is Dreambooth [5], which enables the fine-tuning of pre-trained models to accurately capture unique styles, objects, or individuals. Dreambooth introduces a unique identifier to associate the new visual concept with a placeholder token, allowing it to generate contextually varied images that feature the personalized concept without disrupting the model's ability to produce other images.

To achieve effective personalization, Dreambooth employs a special technique known as "prior preservation loss," which ensures that the model retains its generalization ability while learning new concepts. This loss function mitigates the risk of overfitting to the provided examples by maintaining the model's capacity to generate unrelated images when prompted with different contexts. As a result, the fine-tuned model can represent the new visual concept associated with the unique identifier in a wide variety of scenarios and prompts.

Overall, personalized image generation techniques like Dreambooth allow models to be adapted efficiently, even when only a few reference images are available. These methods have shown great promise in domains requiring customized visual outputs, such as art, design, and marketing.

Chapter 3

Determining Optimal LoRA Configuration

This chapter presents a comprehensive study aimed at identifying the optimal LoRA configuration for fine-tuning the Stable Diffusion XL model. The objective is to determine a configuration that balances high-quality image generation with minimal trainable parameters, making the model more efficient and easier to deploy. The chapter begins by outlining the methodology used for fine-tuning the model, followed by a detailed description of the experiments conducted to evaluate the impact of different LoRA configurations on model performance. By comparing various configurations, we aim to identify the optimal LoRA rank and component for fine-tuning, thus providing a clear guideline for effective model adaptation. The findings from this chapter inform subsequent applications and contribute to the development of efficient fine-tuning strategies for text-to-image generation models.

3.1 Methodology

This section outlines the methodologies employed to fine-tune and evaluate the Stable Diffusion XL model for text-to-image generation tasks. It provides a detailed description of the model selection, dataset preparation, training procedures, and evaluation methods used in this study. The methodologies are designed to identify the optimal LoRA configuration for fine-tuning, ensuring high performance while minimizing computational costs. Each step is meticulously described to facilitate transparency and reproducibility of the experimental results.

3.1.1 Model Selection

The choice of the base model is a critical factor in the success of any machine learning project, particularly in the field of generative models. For this research, we selected the Stable Diffusion XL model [18] as our base model. Below, we outline the rationale behind selecting Stable Diffusion XL and its relevance to our research objectives.

3.1.1.1 Rationale for Selecting Stable Diffusion XL

Stable Diffusion XL was chosen due to its superior performance in generating detailed and coherent images from textual prompts. This model's ability to operate efficiently in the latent space and produce high-fidelity images with intricate details makes it an ideal choice for our experiments focused on fine-tuning generative models. The model also features several cross-attention and self-attention layers, making it convenient to apply LoRA for fine-tuning. The availability of pre-trained weights allows for efficient fine-tuning, leveraging existing knowledge to accelerate the training process.

3.1.1.2 Key Features of Stable Diffusion XL

Several key features of Stable Diffusion XL include:

- **High Quality Image Generation:** The model is capable of producing images that are both visually appealing and semantically accurate to the given prompts.
- **Scalability:** Stable Diffusion XL can handle large datasets and complex image generation tasks, making it suitable for a wide range of applications.
- **Efficient Sampling:** The use of DDIM (Denoising Diffusion Implicit Models) sampling [14] accelerates the image generation process, making it more efficient.
- **Enhanced Quality:** The Stable Diffusion XL refiner [18] is used during sampling to improve the quality of generated images, ensuring they meet high standards of detail and realism.

3.1.1.3 Comparison with Stable Diffusion 2.0

In selecting Stable Diffusion XL, we considered Stable Diffusion 2.0 [1] as a potential base model. Stable Diffusion 2.0 is a prior version of the diffusion model, known for generating high-quality images but with certain limitations in terms of controllability and detail.

Key findings from our comparison include:

- **Image Quality:** Stable Diffusion XL produced samples with higher visual quality, exhibiting more intricate details and coherent structures.
- **Textual Control:** Stable Diffusion XL demonstrated better control with textual prompts, generating images that more accurately reflected the descriptions.

Based on this comparison, Stable Diffusion XL was identified as the most suitable base model for our research, offering superior quality and controllability.

3.1.2 Dataset

For this study, we utilized the DeepFashion-MultiModal dataset [19] for training and evaluation, and the MS COCO 2014 training dataset [20] for evaluating the models on general prompts.

3.1.2.1 DeepFashion-MultiModal Dataset

The DeepFashion-MultiModal dataset is a large-scale, high-quality human image dataset with rich multi-modal annotations. This dataset is particularly suitable for text-to-image generation tasks due to its detailed annotations and high-resolution images. The dataset includes captions describing clothing attributes and styles, which provide the necessary textual input for generating images. Key properties of the dataset include:

- **High-Resolution Images:** Contains 44,096 high-resolution human images, including 12,701 full-body human images.
- **Textual Descriptions:** Each image is provided with a textual description, which is essential for text-driven human image generation.

The DeepFashion-MultiModal dataset is proposed in the Text2Human project and can be applied to various tasks such as text-driven human image generation, text-guided human image manipulation, and human image captioning. Examples of images and captions from this dataset are shown in Figure 3.1.

Caption	”The shirt the gentleman wears has short sleeves and it is with cotton fabric and pure color patterns. The neckline of the shirt is crew.”	”The shirt this lady wears has short sleeves and its fabric is cotton. The pattern of it is pure color. It has a crew neckline.”	”This guy is wearing a short-sleeve T-shirt with solid color patterns. The T-shirt is with cotton fabric. The neckline of the T-shirt is lapel.”
Image			

Figure 3.1: Examples of images and captions from the DeepFashion-MultiModal dataset. The images include high-resolution human figures with detailed textual descriptions of clothing attributes and styles.

3.1.2.2 MS COCO 2014 Dataset

The MS COCO 2014 training dataset was used for additional evaluation to assess the model's performance on general prompts. This dataset features a wide variety of images including everyday scenes, objects, and activities, providing a comprehensive testbed for evaluating generative models. Key properties of the dataset include:

- **Large-Scale Dataset:** Contains 82,783 training images.
- **Caption Descriptions:** Each image is paired with five caption descriptions, offering rich textual annotations for evaluating the model's capability to generate images from varied and general prompts.

The MS COCO dataset is widely recognized for its extensive annotations and diversity, making it an ideal benchmark for assessing the generalization capabilities of generative models. Examples of images and captions from this dataset are shown in Figure 3.2.

3.1.3 Training

The training process for our experiments utilized the DeepFashion-MultiModal dataset. The training was conducted using the PyTorch framework, following a series of preprocessing steps and configurations to ensure optimal performance and efficiency.

3.1.3.1 Data Preprocessing

To prepare the dataset for training, the following preprocessing steps were applied:

- **Aspect Ratio Filtering:** Images with an aspect ratio greater than 0.69 were excluded to ensure compatibility with the downstream transformation scheme.

Caption	”A sandwich is sitting on a plate near some glasses.”	”Baby giraffe gets a drink from its Mama at the zoo.”
Image		

Figure 3.2: Examples of images and captions from the MS COCO 2014 dataset. The images feature a wide variety of everyday scenes, objects, and activities, each paired with caption descriptions.

- **Resizing:** The smaller dimension of each image was resized to 368 pixels, maintaining the aspect ratio.
- **Cropping:** A 512 x 360 pixel region was randomly cropped from the resized image to standardize input dimensions. This resolution was selected to both reduce memory usage and because Stable Diffusion XL has limited capability to generate realistic images at this resolution (see Section 3.2.3). Fine-tuning at this resolution allows us to evaluate the improvements from LoRA without interference from the model's pre-existing generation strengths.
- **Data Augmentation:** A random horizontal flip was applied to each image to increase the diversity of the training data and improve generalization.

The dataset was then split into training and test sets, with 95% of the data used for training and 5% for testing, to ensure no overfitting.

3.1.3.2 Model Configuration

The Stable Diffusion XL model was loaded from Stability AI's official repository on Hugging Face [16]. The model was loaded with bf16 precision, and xFormers memory-efficient attention was enabled to reduce the memory footprint.

3.1.3.3 LoRA Configuration

LoRA layers were applied to the UNet and/or text encoder with a specified rank, depending on the experiment. These layers were integrated into all four weight matrices in the attention modules: query (W_q), key (W_k), value (W_v), and output (W_o) matrices. This configuration was found to be optimal in the original LoRA paper [4]. A scaling factor alpha of 1 was consistently used across all experiments.

3.1.3.4 Training Procedure

The training procedure involved the following configurations:

- **Epochs:** The LoRA parameters were trained for 20 epochs to ensure convergence.
- **Batch Size:** A batch size of 16 was used for both training and evaluation to maintain a balance between computational efficiency and effective learning.
- **Optimizer:** The Adam optimizer was utilized with a learning rate of 1e-5, default betas of 0.9 and 0.999, and a weight decay of 1e-2 to prevent overfitting.
- **Learning Rate Schedule:** A cosine learning rate schedule was employed, with no warm-up steps, to gradually decrease the learning rate and enhance training stability.
- **Gradient Clipping:** Gradient clipping was applied with a maximum gradient norm of 1.0 to prevent exploding gradients and ensure stable training.
- **Memory Efficiency:** Training was performed with gradient checkpointing and bf16 precision to reduce memory usage and facilitate larger batch sizes.
- **Distributed Training:** The training process was distributed across 2 GPUs to leverage parallel processing and expedite the training time.

3.1.3.5 Libraries and Tools

The training process leveraged the Hugging Face ecosystem, utilizing the following libraries and tools:

- **Diffusers Library:** Utilized for loading and managing the Stable Diffusion model.
- **PEFT Library:** Used for implementing the LoRA configurations.
- **Accelerate Library:** Facilitated distributed training across multiple GPUs.

3.1.4 Evaluation

The evaluation of the models was conducted to comprehensively assess both their quantitative and qualitative performance. This section outlines the metrics used for evaluation, the datasets employed, the process of generating images, and the computational setup. The evaluation aimed to ensure the models performed well on the specific dataset used for fine-tuning and to verify that fine-tuning did not degrade the model's capability to handle general prompts.

3.1.4.1 Evaluation Metrics

We used the following quantitative metrics for evaluation:

- **Fréchet Inception Distance (FID)**: FID measures the distance between the feature distributions of generated images and real images, providing an indication of the quality and diversity of the generated images. A lower FID score indicates better quality and diversity. The method is described in detail in [21].
- **CLIP Score**: CLIP Score evaluates the similarity between text and image embeddings generated by the CLIP model. A higher CLIP Score indicates better alignment between the generated image and its corresponding textual description. This method is detailed in [9].

Qualitative Evaluation

In addition to quantitative metrics, qualitative evaluation was performed by visualizing samples generated for different configurations. This provided insights into the visual quality and semantic accuracy of the generated images.

3.1.4.2 Evaluation Procedure

Evaluation was performed on both the DeepFashion-MultiModal dataset and the MS COCO 2014 dataset to assess fine-tuning performance and to verify that the model's ability to generate images from general prompts was not compromised.

DeepFashion-MultiModal Dataset

To assess fine-tuning performance, we evaluated on 40,000 samples from the DeepFashion-MultiModal dataset, filtering out images with aspect ratios higher than 0.69. Each image was resized such that the smaller dimension was 360 pixels, followed by a center crop of 512x360 pixels to obtain the target images.

MS COCO 2014 Dataset

For general prompt evaluation, we used 40,000 samples from the MS COCO 2014 training dataset, filtering out images with aspect ratios smaller than 1.33. Each image was resized such that the smaller dimension was 480 pixels, followed by a center crop of 480x640 pixels to obtain the target images.

Image Generation Process

Images were generated based on the provided captions, with dimensions matching the target images:

- **Generation Parameters**: A fixed seed of 0 and a guidance scale of 5 were used to ensure consistent and high-quality generation.
- **Sampling Method**: DDIM sampling was employed for accelerated generation.
- **Quality Refinement**: The Stable Diffusion XL refiner was used to enhance image quality, with a refine fraction of 0.2.
- **Negative Prompt**: For generation from the base model, a negative prompt of "drawing, cartoon, painting, illustration, 3d, render, cgi" was used to steer the generation towards realistic images, addressing the tendency of the model to produce drawing-like images at the target resolutions.

Computational Setup

The evaluation was performed on two GPUs with a batch size of 8 to ensure efficient processing.

3.1.4.3 Evaluation on Training Data

Evaluating on the training data is a common practice in generative models to ensure that the model can reproduce the training data effectively, as noted in [13].

This comprehensive evaluation approach, combining both quantitative and qualitative methods, ensured a thorough assessment of the model's performance across different configurations and datasets.

3.2 Experiments

This section presents a series of experiments conducted to evaluate the performance of different LoRA configurations. The experiments focus on systematically assessing the impact of applying LoRA layers to various components of the Stable Diffusion XL model and testing different LoRA ranks to identify the optimal configuration. The results of each experiment are analyzed quantitatively using evaluation metrics like FID and CLIP Score, and qualitatively by visual inspection of generated samples. These experiments build on the methodology described previously and provide the basis for selecting the most effective fine-tuning strategy for the model.

3.2.1 Experiment 1: Effect of Applying LoRA to Different Components

3.2.1.1 Objectives

The primary objective of this experiment is to systematically evaluate the impact of applying Low-Rank Adaptation (LoRA) layers to different components of the Stable Diffusion XL model on the model's performance in generating high-quality images from textual prompts. Specifically, this experiment aims to achieve the following:

- **Identify the Optimal Component for LoRA Application:** Determine whether applying LoRA layers to the UNet component, the text encoder component, or both simultaneously yields the best performance in terms of image quality and semantic accuracy.
- **Inform Future Fine-Tuning Strategies:** Provide insights and recommendations for future fine-tuning strategies by identifying the most beneficial component(s) for LoRA application. This will help in optimizing the trade-off between model performance and computational efficiency in subsequent experiments.

By achieving these objectives, this experiment seeks to establish a clear understanding of the benefits and trade-offs associated with applying LoRA layers to different components of the Stable Diffusion XL model, thereby guiding the development of more efficient and effective fine-tuning techniques for text-to-image generation tasks.

3.2.1.2 Experimental Setup

In this experiment, we evaluated the impact of applying LoRA layers to different components of the Stable Diffusion XL model under three distinct configurations. To ensure a fair comparison, we set a trainable parameter budget of approximately 3 million for each configuration. The DeepFashion-MultiModal dataset was used for both fine-tuning and evaluation in all configurations. The configurations are as follows:

- **Configuration 1: UNet Only**
  - LoRA layers were applied only to the UNet component.
  - A LoRA rank of 2 was used for this configuration.
- **Configuration 2: Both UNet and Text Encoder**
  - LoRA layers were applied to both the UNet and the text encoder components.
  - A LoRA rank of 1 was used for the UNet and a rank of 4 for the text encoder.
- **Configuration 3: Text Encoder Only**
  - LoRA layers were applied only to the text encoder component.

– A LoRA rank of 8 was used for this configuration.

These configurations were chosen to explore the effectiveness of applying LoRA to different parts of the model and to identify the optimal component(s) for LoRA application. The approximate trainable parameter budget of 3 million ensures that the complexity of the model remains manageable while allowing us to observe significant differences in performance across the configurations.

3.2.1.3 Results

The results of Experiment 1, which evaluated the impact of applying LoRA layers to different components of the Stable Diffusion XL model, are summarized in Table 3.1. The table presents the trainable parameter count, Fréchet Inception Distance (FID), and CLIP Score for each configuration.

Table 3.1: Results for Experiment 1: Comparison of FID and CLIP Scores for Different Configurations

Configuration	**Trainable Parameter Count**	**FID**	**CLIP Score**
UNet Only	2,903,040	16.74	22.31
Both UNet & Text Encoder	3,057,152	17.96	21.41
Text Encoder Only	3,211,264	21.02	21.61

Figure 3.3 provides visual comparisons of images generated by the model under each configuration for three different prompts and seeds. This qualitative assessment complements the quantitative metrics by highlighting the visual and semantic accuracy of the generated images.

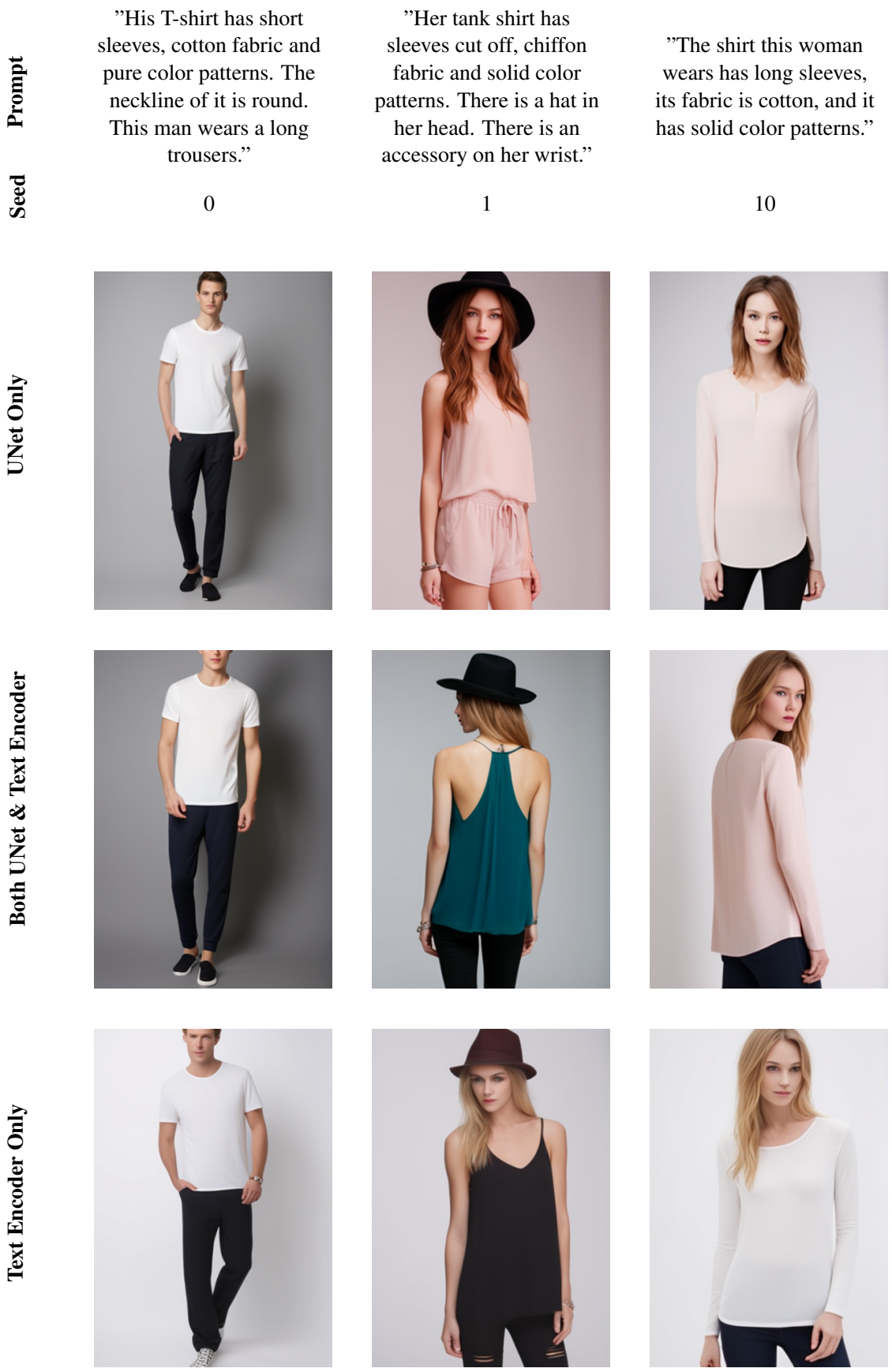


Figure 3.3: Generated images for Experiment 1: Performance comparison with three different prompts and seeds.

3.2.1.4 Discussion

The results from Experiment 1 provide several insightful observations regarding the application of LoRA layers to different components of the Stable Diffusion XL model.

General Insights

The Fréchet Inception Distance (FID) scores for all configurations are relatively low, indicating that the generated images exhibit high fidelity and diversity (see Table 3.1). This suggests that the model, regardless of the LoRA application, can produce realistic and varied images. However, the CLIP scores are also relatively low, which could be attributed to the fact that the CLIP model is trained on general image-text pairs, whereas the text and images in this experiment are domain-specific (fashion-related). This mismatch might have affected the CLIP model's ability to accurately evaluate the alignment between generated images and their corresponding textual descriptions.

Comparative Analysis

Upon comparing the different configurations, the best performance was observed when LoRA layers were applied only to the UNet component, yielding the lowest FID and the highest CLIP score (see Table 3.1). This configuration demonstrated superior image quality and better alignment with textual prompts. Conversely, the configuration with LoRA applied only to the text encoder resulted in the worst FID, significantly higher than the other two configurations. This disparity indicates that optimizing the UNet, which is responsible for generating pixel values from text embeddings, is more effective than optimizing the text encoder directly.

Qualitative Evaluation

The qualitative assessment further corroborates the quantitative findings. Samples generated with LoRA applied to the text encoder alone tend to be blurrier and less detailed compared to those generated by the other configurations. In contrast, samples from the UNet-only and combined UNet and text encoder configurations are sharper and more visually appealing (see Figure 3.3). Despite the differences in sharpness and detail, all configurations produced images that generally aligned well with the given prompts, as reflected by the similar CLIP scores across the board.

Interpretation

These results highlight the critical role of the UNet component in the Stable Diffusion XL model's performance in text-to-image generation tasks. The superior performance of the UNet-only configuration suggests that fine-tuning the component responsible for the actual image generation process (the UNet) is more beneficial than fine-tuning the component that handles text embeddings (the text encoder). This finding implies that optimizing the mapping of text embeddings to pixel values has a more pronounced impact on image quality than optimizing the text embeddings themselves.

Implications for Future Work

The insights gained from this experiment provide valuable guidance for future fine-tuning strategies. Focusing on the UNet component when applying LoRA layers appears to be the most effective approach for improving model performance. Additionally, the consistent performance across different configurations in terms of prompt alignment indicates that the model's

ability to understand and generate images based on textual descriptions is robust, even with domain-specific data.

3.2.2 Experiment 2: Effect of LoRA Rank

3.2.2.1 Objective

The primary objective of this experiment is to systematically evaluate the impact of different LoRA ranks on the performance of the Stable Diffusion XL model in generating high-quality images from textual prompts. Building on the findings from Experiment 1, where applying LoRA layers to the UNet component was identified as the most effective configuration, this experiment aims to:

- **Determine the Optimal LoRA Rank:** Assess the performance of the model when varying the LoRA rank applied to the UNet component, specifically testing ranks of 1, 2, 4, 8, and 64. The goal is to identify the lowest rank that still yields high-quality results, thus minimizing the number of additional parameters introduced during fine-tuning.
- **Evaluate the Trade-off Between Performance and Parameter Efficiency:** Understand how the trade-off between model performance and the number of trainable parameters varies with different LoRA ranks. This includes analyzing the balance between achieving good image fidelity and semantic alignment with textual prompts while keeping the parameter budget low.

By achieving these objectives, Experiment 2 seeks to deepen the understanding of how different LoRA ranks affect model performance, guiding the development of more efficient fine-tuning techniques for high-quality text-to-image generation.

3.2.2.2 Experimental Setup

In this experiment, we evaluated the impact of different LoRA ranks on the performance of the Stable Diffusion XL model. Based on the findings from Experiment 1, where applying LoRA layers to the UNet component yielded the best results, we focused exclusively on applying LoRA to the UNet for this experiment. The DeepFashion-MultiModal dataset was used for both fine-tuning and evaluation across all configurations. We considered five different ranks for the LoRA layers: 1, 2, 4, 8, and 64.

These configurations were designed to explore the effectiveness of different LoRA ranks and to identify the optimal rank that balances performance and parameter efficiency.

3.2.2.3 Results

The results of Experiment 2, which evaluated the impact of different LoRA ranks on the performance of the Stable Diffusion XL model, are summarized in Table 3.2. The table presents the trainable parameter count, Fréchet Inception Distance (FID), and CLIP Score for each LoRA rank configuration.

Figures 3.4 and 3.5 provide visual comparisons of images generated by the model under each configuration for three different prompts and seeds. This qualitative assessment complements the quantitative metrics by highlighting the visual and semantic accuracy of the generated images.

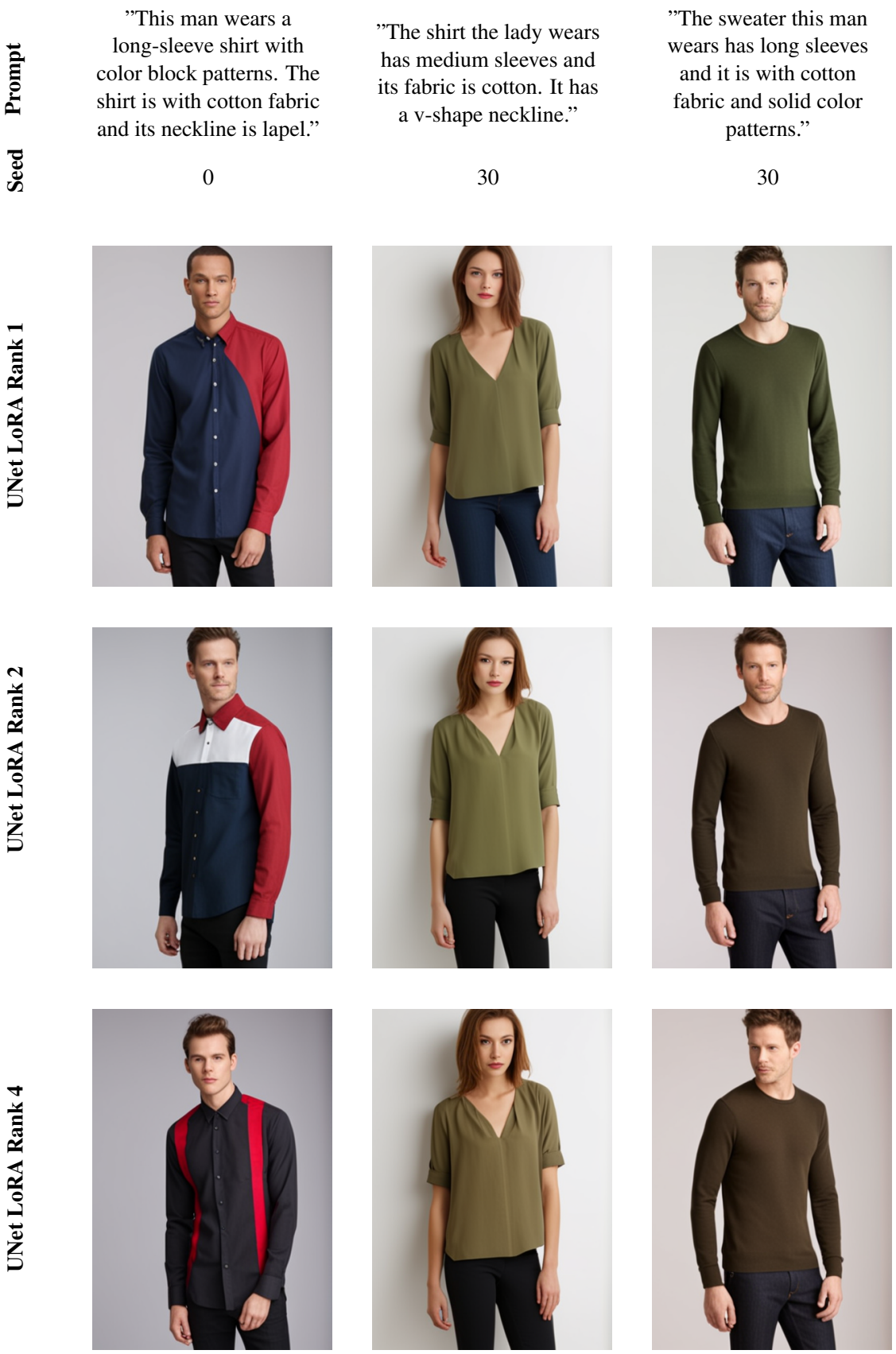


Figure 3.4: Generated images for Experiment 2: Performance comparison with three different prompts and seeds.

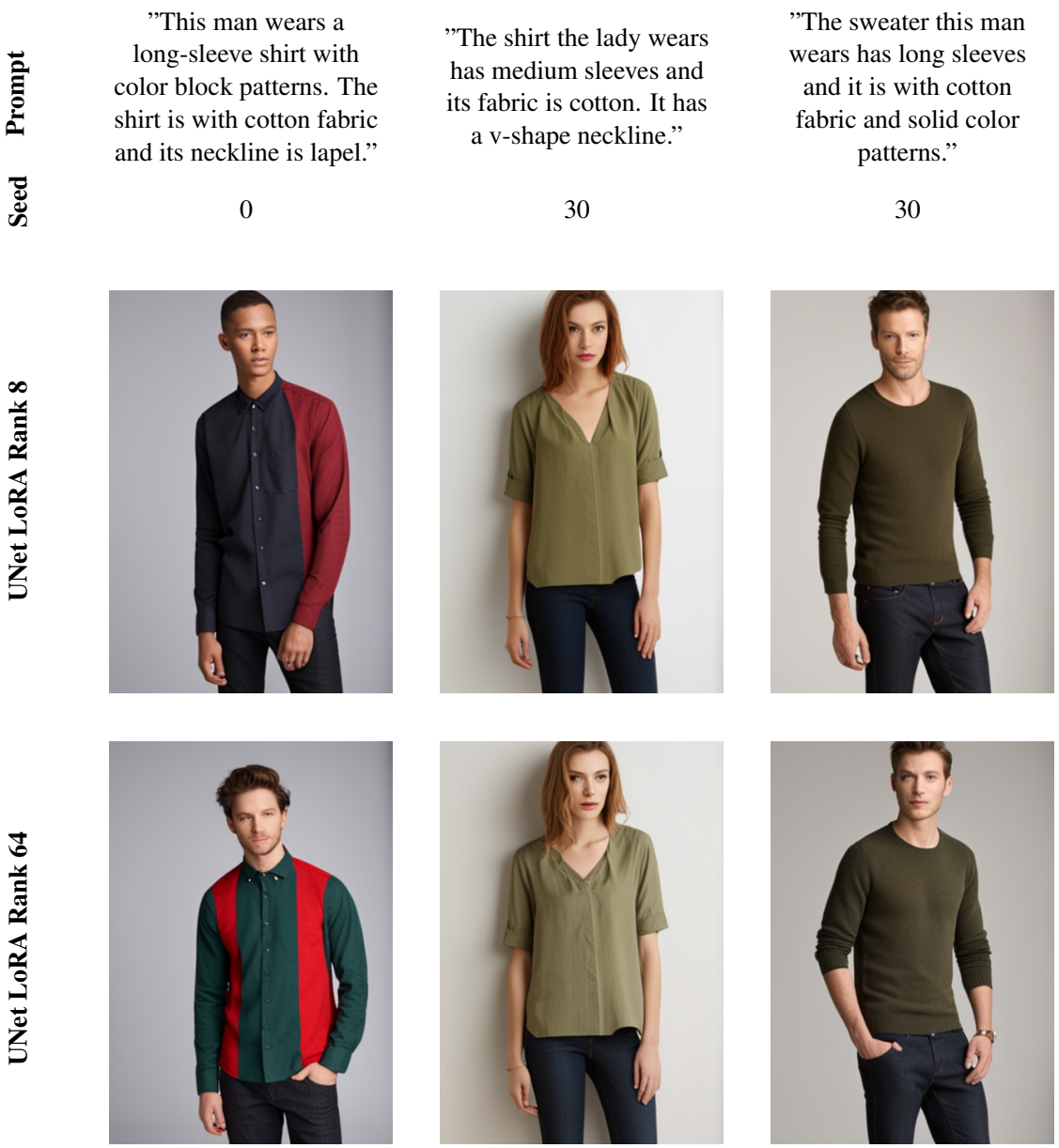


Figure 3.5: Generated images for Experiment 2: Performance comparison with three different prompts and seeds.

Table 3.2: Results for Experiment 2: Comparison of FID and CLIP Scores for Different LoRA ranks

Configuration	**Trainable Parameter Count**	**FID**	**CLIP Score**
UNet LoRA Rank 1	1,451,520	16.85	21.83
UNet LoRA Rank 2	2,903,040	16.74	22.31
UNet LoRA Rank 4	5,806,080	17.65	22.46
UNet LoRA Rank 8	11,612,160	16.72	22.80
UNet LoRA Rank 64	92,897,280	16.11	23.29

3.2.2.4 Discussion

The results from Experiment 2 provide several insightful observations regarding the impact of different LoRA ranks on the performance of the Stable Diffusion XL model.

General Insights

The Fréchet Inception Distance (FID) generally reduces and the CLIP Score increases with an increase in LoRA rank, as shown in Table 3.2. However, the improvements are not very significant across different ranks. This indicates that higher ranks do not substantially enhance the model's performance compared to lower ranks.

Comparative Analysis

Interestingly, the configuration with LoRA rank 64, which has 92.9 million trainable parameters, yields results comparable to the configuration with LoRA rank 1, which has only 1.45 million trainable parameters. This minimal difference in performance suggests that the additional parameters from higher ranks do not contribute significantly to the model's ability to generate high-quality images.

Qualitative Evaluation

The qualitative assessment supports the quantitative findings. Figures 3.4 and 3.5 illustrate that images generated with LoRA ranks 1, 2, 4, 8, and 64 exhibit similar quality and alignment with the prompts. This consistency across different ranks further reinforces the observation that higher LoRA ranks do not provide significant visual improvements.

Interpretation

The results suggest that the changes in weights during fine-tuning are of very low rank and can be effectively captured with a LoRA rank of just 1, particularly when applied to the UNet component. This finding implies that if we need different models for various applications, we only need to store approximately 1.45 million parameters (about 5.5 MB) on top of the base model for each application. This significantly reduces the storage requirements and enhances the efficiency of deploying fine-tuned models for specific tasks, especially when LoRA is applied to the UNet with a minimal rank.

Implications for Future Work

The findings from Experiment 2 highlight the efficiency of LoRA fine-tuning, especially when applied to the UNet with a rank of 1, and its potential for creating lightweight models tailored to specific applications without compromising performance. Future work could explore the efficiency of these lightweight models by assessing the improvements they achieve compared to the base model and also their performance degradation on general prompts.

3.2.3 Experiment 3: Evaluation of Fine-Tuned Model's Performance Compared to Base Model

3.2.3.1 Objective

The primary objective of this experiment is to systematically evaluate the improvement in the image generation capability of the Stable Diffusion XL model when fine-tuned with LoRA

applied to the UNet component using a rank of 1. Building on the findings from Experiments 1 and 2, this experiment aims to:

- **Quantify the Improvement Over the Base Model:** Assess the performance enhancements in image quality and alignment with textual prompts achieved by the fine-tuned model compared to the base model using quantitative metrics such as Fréchet Inception Distance (FID) and CLIP Score.

By achieving this objective, Experiment 3 seeks to demonstrate the effectiveness of fine-tuning the Stable Diffusion XL model with LoRA applied to the UNet component using a rank of 1.

3.2.3.2 Experimental Setup

In this experiment, we evaluated the fine-tuned model and the base model on the DeepFashion-MultiModal dataset to assess the improvements in image generation capability brought about by applying LoRA to the UNet component using a rank of 1. Based on the findings from Experiments 1 and 2, we focused exclusively on this configuration for the current experiment.

For generation from the base model, a negative prompt of "drawing, cartoon, painting, illustration, 3d, render, cgi" was used to steer the generation towards realistic images, addressing the tendency of the model to produce drawing-like images at the target resolutions.

This setup allows for a direct comparison of the base model and the fine-tuned model, highlighting the improvements in image generation capability achieved through LoRA fine-tuning.

3.2.3.3 Results

The results of Experiment 3, which evaluated the improvements in image generation capability of the Stable Diffusion XL model when fine-tuned with LoRA applied to the UNet component using a rank of 1, are summarized in Table 3.3. The table presents the Fréchet Inception Distance (FID) and CLIP Score for the fine-tuned model and the base model.

Table 3.3: Results for Experiment 3: Comparison of FID and CLIP Scores for UNet LoRA Rank 1 Configuration versus Base Model

Configuration	FID	CLIP Score
UNet LoRA Rank 1	16.85	21.83
Base Model	195.03	19.93

Figure 3.6 provides visual comparisons of images generated by the base model and the fine-tuned model under three different prompts and seeds. This qualitative assessment complements the quantitative metrics by highlighting the visual and semantic improvements achieved through fine-tuning.

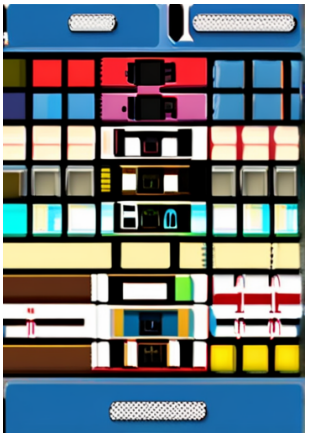
Prompt	”His tank top has no sleeves, cotton fabric and pure color patterns. The neckline of it is round.”	”This woman wears a medium skirt, with cotton fabric and color block patterns. This female wears a ring.”	”This lady is wearing a sleeveless tank top with graphic patterns and a three-point shorts. The tank top is with cotton fabric and its neckline is round.”
Seed	2	70	40
UNet LoRA Rank 1			
Base Model			

Figure 3.6: Generated images for Experiment 3: Performance comparison between the UNet LoRA Rank 1 configuration and the Base Model using three different prompts and seeds.

3.2.3.4 Discussion

The results from Experiment 3 provide several insightful observations regarding the improvements in image generation capability of the Stable Diffusion XL model when fine-tuned with LoRA applied to the UNet component using a rank of 1.

General Insights

The fine-tuned model performs well, as seen in the previous experiments. However, the performance of the base model is significantly worse, particularly with a very high Fréchet Inception Distance (FID), indicating that the generated images are very different from the target image

distribution. The CLIP Score is also relatively lower for the base model. These metrics suggest that fine-tuning the model significantly enhances its capabilities.

Comparative Analysis

Comparing the metrics, the fine-tuned model shows substantial improvements over the base model. The FID for the base model is 195.03, whereas the fine-tuned model achieves an FID of 16.85. Similarly, the CLIP Score for the base model is 19.93, compared to 21.83 for the fine-tuned model. These differences indicate a considerable enhancement in image quality and alignment with textual prompts due to fine-tuning.

Qualitative Evaluation

The qualitative assessment supports the quantitative findings. The images generated by the fine-tuned model are much closer to the target images in terms of realism and detail. In contrast, the images generated by the base model appear cartoonish or like drawings, even with the use of negative prompts intended to steer the generation towards realistic images. Figure 3.6 illustrate that the base model struggles to produce realistic images at the given resolution for these kinds of prompts, whereas the fine-tuned model generates images that are visually similar to the target images.

Interpretation

These results indicate that the base model lacks the capability to generate realistic images at the specified resolution for the given prompts. Fine-tuning the UNet component with a LoRA rank of 1 significantly improves the image generation capability. This suggests that the fine-tuning process enables the model to better capture and replicate the details and distributions present in the target dataset, thus enhancing overall performance.

3.2.4 Experiment 4: Evaluation of Model Performance on General Prompts

3.2.4.1 Objective

A common issue with fine-tuning is the potential for the model to become overly specialized to the fine-tuned dataset, which can lead to degraded performance on more general data. LoRA fine-tuning, particularly at lower ranks, can help mitigate this by reducing the complexity of the trainable parts of the model, thereby preventing overfitting.

The primary objective of this experiment is to assess the impact of fine-tuning the UNet component of the Stable Diffusion XL model with a LoRA rank of 1 on its performance with general prompts. Specifically, this experiment aims to:

- **Assess Performance Degradation:** Determine whether the fine-tuned model's ability to generate high-quality images from general, non-domain-specific prompts has been compromised compared to the base model.

By achieving this objective, Experiment 4 seeks to evaluate the robustness and generalization capabilities of the fine-tuned Stable Diffusion XL model, providing insights into any potential trade-offs in performance due to fine-tuning.

3.2.4.2 Experimental Setup

The experimental setup for Experiment 4 closely follows that of Experiment 3. We evaluated both the UNet component of the Stable Diffusion XL model fine-tuned with LoRA rank 1 and

the base model to assess their performance on general prompts. The key difference in this experiment is the use of the MS COCO 2014 dataset, which provides a diverse range of non-domain-specific images and textual descriptions. For the qualitative evaluation, images were generated at a resolution of 1024x1024, a resolution where the model has good capability and is commonly used.

3.2.4.3 Results

The results of Experiment 4, which evaluated the performance of the fine-tuned UNet component of the Stable Diffusion XL model with LoRA rank 1 compared to the base model on general prompts, are summarized in Table 3.4. The table presents the Fréchet Inception Distance (FID) and CLIP Score for both configurations.

Table 3.4: Results for Experiment 4: Comparison of FID and CLIP Scores for UNet LoRA Rank 1 Configuration versus Base Model on General Prompts

Configuration	**FID**	**CLIP Score**
UNet LoRA Rank 1	21.65	24.89
Base Model	29.29	24.09

Figure 3.7 provides visual comparisons of images generated by the base model and the fine-tuned model under two different prompts and seeds. This qualitative assessment complements the quantitative metrics by highlighting the visual and semantic differences between the two configurations.

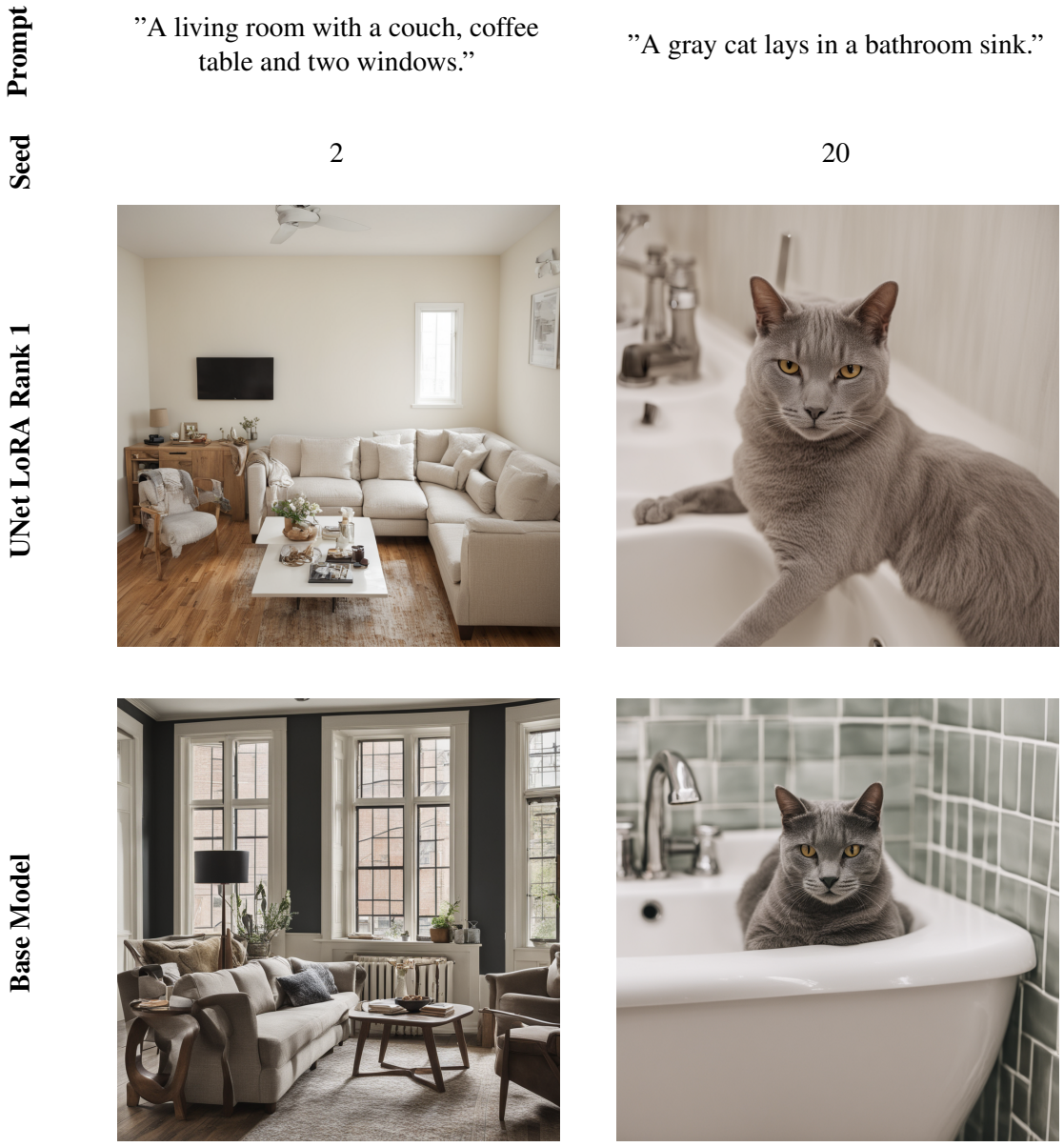


Figure 3.7: Generated images for Experiment 4: Comparison of UNet LoRA Rank 1 and Base Model performance with two different prompts and seeds.

3.2.4.4 Discussion

The results from Experiment 4 provide insights into the impact of fine-tuning the UNet component of the Stable Diffusion XL model with a LoRA rank of 1 on its performance with general prompts.

General Insights

The Fréchet Inception Distance (FID) scores for both the fine-tuned model and the base model are relatively low, indicating that both models generate high-quality samples. The CLIP scores are also higher compared to previous experiments, reflecting the general nature of the MS COCO 2014 dataset, which aligns well with the training data of the CLIP model.

Comparative Analysis

The fine-tuned model exhibits better performance compared to the base model in terms of both FID and CLIP Score. The FID for the fine-tuned model is 21.65, lower than the base model's FID of 29.29. Similarly, the CLIP Score for the fine-tuned model is 24.89, higher than the base model's score of 24.09. This improvement is likely due to the base model's limitations in producing realistic images at lower resolutions. Fine-tuning with realistic, domain-specific data appears to help the model generate realistic images even for general prompts.

Qualitative Evaluation

The qualitative evaluation supports these findings. As illustrated in Figure 3.7, the images generated by both the fine-tuned model and the base model are of higher quality when generated at a resolution of 1024x1024, which the models handle well. Both models produce images that align well with the given prompts.

Interpretation

The results suggest that LoRA fine-tuning does not degrade the model's capability to handle general prompts. On the contrary, it can enhance the model's ability to produce realistic images across a variety of prompts. The fine-tuning process injects the nature of realistic imagery into the model, improving its generalization capabilities without overfitting to the fine-tuning data.

Chapter 4

Optimal LoRA Configuration in Practice

In this chapter, we explore a practical application of the optimal LoRA configuration identified through our experiments to address a real-world problem in the fashion industry. The goal is to generate images of models wearing a specific garment provided as a single reference image, depicting the garment in various poses and contexts. This approach offers a cost-effective alternative to traditional photoshoots by eliminating the need for hiring models and arranging photoshoots for each garment variation. It allows designers and marketers to visualize their products on models in a range of styles and poses, significantly enhancing marketing, design workflows, and consumer engagement.

The optimal configuration, identified through our experiments, involves fine-tuning the UNet component of the Stable Diffusion XL model with a LoRA rank of 1. This configuration has been shown to effectively enhance the model's image generation capabilities while maintaining a balance between performance and memory efficiency.

For this application, each garment requires the training of a separate model. By using LoRA layers with the optimal configuration, we ensure that each fine-tuned model occupies only a fraction of the memory compared to full fine-tuning. This is crucial for practical deployment, as it allows for the efficient storage and use of multiple models without requiring extensive computational resources.

The methodology section will detail the process of training the model with a single image of a garment, generating images of models wearing the garment in different poses. The results section will present a qualitative evaluation of the generated images, showcasing the visual fidelity and realism of the model's outputs. Finally, the discussion section will analyze the effectiveness and limitations of this approach, providing insights for future improvements and applications.

4.1 Methodology

The methodology for this application is inspired by the Dreambooth approach [5], which enables the fine-tuning of generative models to learn specific visual concepts from a small number of reference images. However, our process focuses on fine-tuning the UNet component of the Stable Diffusion XL model with LoRA layers, using a single image of a garment as input. This image acts as a reference for the garment's unique characteristics such as fabric type, pattern, and color, which the model learns to replicate. Unlike the traditional Dreambooth method, which includes prior preservation loss to maintain the model's generalization ability, we omit this step as the model is tailored to a specific garment and context. The primary objective is to generate images that depict the garment in various poses on different models, accurately capturing its distinctive features.

This methodology is similar to that described in Section 3.1, with the following differences highlighted below:

4.1.1 Data Preprocessing

The training image is the single image of the garment. For the caption, a special token system is employed, where `[v]` (”sks”) represents the garment type and `[u]` (”ukj”) denotes the color/pattern. For example, the prompt ”Photo of a `[v]` sweater with `[u]` pattern” specifies the garment and its details. The single input image is resized to 736 pixels on the shorter side, and a random crop of 1024 x 720 pixels is taken. This higher image resolution leverages the model's prior generation capabilities, as the model performs well at higher resolutions..

4.1.2 LoRA Configuration

The LoRA configuration used involves applying LoRA layers to the UNet component with a rank of 1, as determined to be optimal in previous experiments.

4.1.3 Training

The training configuration includes the following specific details:

- **Training Procedure:**
  - **Epochs:** The training is conducted for 2000 epochs.
  - **Learning Rate:** A higher learning rate of 1e-4 is used.
  - **Learning Rate Schedule:** A constant learning rate schedule is employed.

4.1.4 Sampling and Generation

For generating the images, the following setup is used:

- **Resolution:** Images are generated at a resolution of 1024 x 720 pixels.
- **Guidance Scale:** A guidance scale of 5 is used to ensure high-quality image generation.
- **Prefix Prompt:** A prefix prompt ”A high-quality image of a professional photoshoot featuring” is added to each prompt to generate professional-looking photoshoot images.
- **Seed Selection:** Various seeds are tested to find the one that generates images closest to the original image in terms of color and pattern.
- **Sampling Method:** DDIM (Denoising Diffusion Implicit Models) sampling is employed for accelerated generation.
- **Quality Refinement:** The Stable Diffusion XL refiner is used to enhance image quality, with a refine fraction of 0.2.

This methodology ensures that the fine-tuned model generates high-quality, realistic images of the specified garments, leveraging the model's inherent capabilities and the effectiveness of LoRA fine-tuning.

4.2 Results

The results of this application demonstrate the capability of the fine-tuned model to generate realistic images of models wearing specific garments in various poses. The generated images were created using the prefix prompt ”A high-quality image of a professional photoshoot featuring” to ensure a consistent and professional appearance. Figure 4.1 showcases examples of the input images and the corresponding generated images based on different prompts.

4.3 Discussion

The results from our application showcase the remarkable capability of fine-tuning the UNet component of the Stable Diffusion XL model with a LoRA rank of 1 for generating realistic images of models wearing specific garments in various poses. The generated images, as illustrated in Figure 4.1, highlight the robustness and effectiveness of our methodology. The fine-tuned model consistently produced high-quality images that closely aligned with the provided prompts. The prefix prompt ”A high-quality image of a professional photoshoot featuring” was instrumental in ensuring that the generated images had a professional and polished appearance, enhancing their realism and visual appeal. This uniformity across different garments and poses indicates the model's ability to maintain a high standard of output quality.

Upon examining the generated images against the input images for different poses and prompts, it is evident that the fine-tuned model accurately preserved the color and pattern of the garments. The special token system, utilizing `[v]` for garment type and `[u]` for color/pattern, was effective in replicating the specific attributes of the garments. This was clearly demonstrated in the generated images for the shirtdress, sweater, and pants, where the model successfully captured the nuances of each garment type and style, indicating its adaptability and precision.

The findings suggest that the fine-tuning method using LoRA layers with a rank of 1 is highly efficient and practical for generating specific garment images. The model's capacity to produce high-quality outputs without prior preservation indicates that LoRA fine-tuning significantly enhances the model's functionality while being resource-efficient. This allows for the creation of multiple specialized models for different garments, each requiring only a fraction of the memory compared to traditional fine-tuning methods. The generated images' fidelity and detail reflect the method's efficacy in practical applications. This approach is particularly advantageous for the fashion industry, where such generated images can be utilized for various purposes including marketing and design. Future research could explore the scalability of this method to a broader range of garments and poses, as well as the inclusion of more complex and diverse prompts.

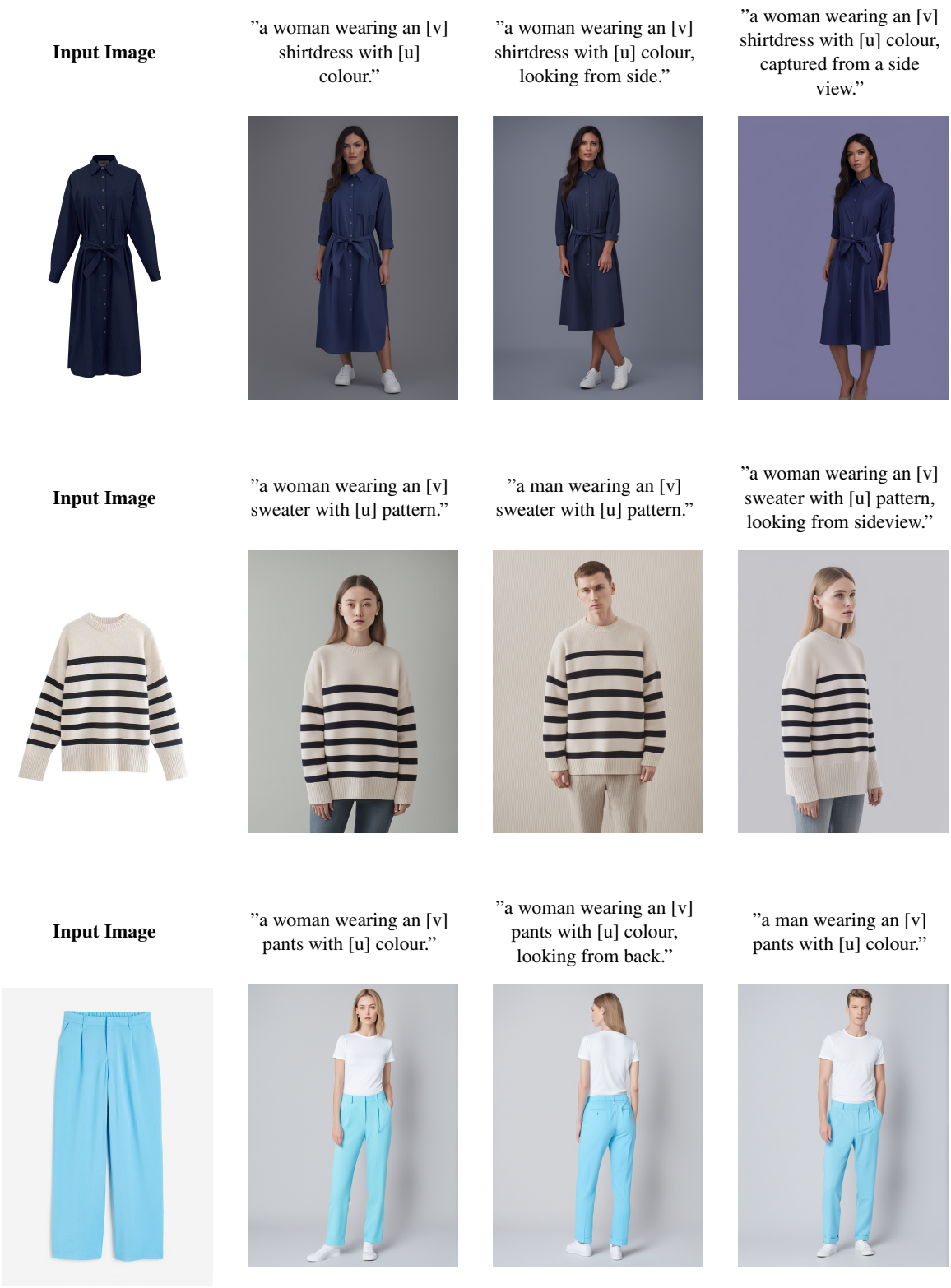


Figure 4.1: Generated images for different garments using the prefix prompt ”A high-quality image of a professional photoshoot featuring”. Each row shows the input image and the corresponding generated images for various prompts.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this dissertation, we explored the application of Low-Rank Adaptation (LoRA) for fine-tuning the Stable Diffusion XL model to generate high-quality images of fashion garments from textual descriptions. Our motivation stemmed from the need to develop efficient and scalable solutions for text-to-image generation in the fashion industry, where visually appealing and accurate outputs are essential for design and marketing. By leveraging LoRA's ability to reduce the number of trainable parameters, we sought to optimize the model's performance while minimizing computational resources.

The key contributions of this dissertation are summarized as follows:

1. **Comprehensive Literature Review:** We reviewed the evolution of text-to-image generation models, from early approaches such as alignDRAW and GANs to state-of-the-art diffusion models, including the advancements introduced in Stable Diffusion and Stable Diffusion XL.

2. **Systematic Evaluation of LoRA Configurations:** We investigated the impact of applying LoRA to different components of the Stable Diffusion XL model, focusing on the UNet and text encoder, and identified that applying LoRA to the UNet with a rank of 1 achieved the best trade-off between quality and efficiency.

3. **Development of an Efficient Fine-Tuning Methodology:** Using the optimal LoRA configuration, we developed a structured fine-tuning approach that enables effective model adaptation for generating high-quality fashion images with minimal computational overhead.

4. **Real-World Application:** We demonstrated the potential of our method by generating images of specific garments, provided as a single reference image, worn by models in various poses. This practical application serves as a cost-effective alternative to traditional photoshoots for visualizing fashion items.

Our research confirms that LoRA-based fine-tuning significantly enhances the quality and coherence of image generation in domain-specific tasks, making it a viable approach for resource-constrained environments. The optimized configuration preserves key visual attributes of garments while providing scalability and efficiency, making it suitable for various real-world scenarios.

5.2 Recommendations and Future Directions

Based on the findings and insights gained from this research, we propose the following recommendations and future research directions:

1. **Selective Application of LoRA:** Further reduce the number of parameters by applying LoRA only to a subset of UNet layers. This approach could streamline the model while maintaining or even enhancing its performance for specific tasks.

2. **Combining Fine-Tuning Techniques:** Explore the combination of prompt tuning [22] with LoRA, where prompt tuning adjusts the model's responses by learning soft prompt embeddings, while LoRA fine-tunes specific model layers for efficiency and performance.

3. **Utilization of More Recent Base Models:** Consider using more recent base models such as Stable Diffusion 3 [23]. This updated version may offer improvements in image generation quality and efficiency that could benefit the fine-tuning process.

4. **Memory-Efficient Fine-Tuning with QLoRA:** Implement QLoRA [24] to reduce the memory footprint during training by quantizing the base model and applying LoRA layers to the quantized model. This technique could make the fine-tuning process more accessible on hardware with limited resources, thereby broadening the practical applications of the model.

5.3 Closing Remarks

This dissertation has demonstrated the effectiveness of using LoRA to fine-tune Stable Diffusion models for generating high-quality fashion images. The approaches and findings presented here offer a foundation for extending LoRA applications to other generative models and domains, emphasizing its adaptability and potential for resource-efficient fine-tuning. We hope that this work will inspire further exploration and innovation in leveraging advanced fine-tuning strategies for both research and industry applications.

Bibliography

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[5] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[6] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015.

[7] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[10] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[12] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[17] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[18] Stability AI. Stable diffusion xl base 1.0. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0, 2023. Accessed: 2024-09-27.

[19] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[23] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[24] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms (2023). *arXiv preprint arXiv:2305.14314*, 52:3982–3992, 2023.