國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

透過 CLIP 文字編碼器適應進行條件擴散模型中的語義編輯和去偏見

Semantic Editing and Debiasing in Conditional Diffusion Models by CLIP Text-Encoder Adaptation

鄭廷瑋

Ting-Wei Cheng

指導教授: 吳家麟 博士

Advisor: Ja-Ling Wu Ph.D.

中華民國 113 年 6 月

June, 2024

doi:10.6342/NTU202403591

Acknowledgements

在我完成這篇碩士論文的過程中，我受到了許多人的幫助和支持。首先，我要感謝我的指導教授吳家麟教授。感謝他在我研究期間給予的悉心指導和不懈支持。無論是在研究方向的選擇上，還是在論文的撰寫和修改過程中，他都給予了我寶貴的建議和鼓勵。

其次，我要感謝實驗室中的各位學長姐，在每次論文進度報告中都給我很多新的方向，以及告訴我需要補足的知識。同時，我還要感謝我的同學和朋友們。在我遇到困難和瓶頸時，他們總是願意提供幫助和分享他們的見解。

最後，我要感謝國立台灣大學的所有老師和同學，感謝你們在我學習期間給予的幫助和支持。感謝 CM Lab 提供的優越的學習和研究環境，使我能夠專注於我的研究工作。

再次感謝所有在這個過程中給予我幫助和支持的人，沒有你們的幫助，我不可能完成這篇論文。

謝謝你們！

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

摘要

本論文研究了 CLIP 文本編碼器在條件擴散模型中的適應，以解決語義編輯和去偏見相關的挑戰。我們探討了透過適應 (Adaptation) 在增強生成圖像語義屬性控制方面的有效性，同時減少內在偏見。研究利用了各種解耦策略，並對文字編碼器進行修改，以評估緩解與性別和種族相關偏見的潛力。我們的研究結果表明，通過針對性適應微調文本編碼器可以顯著提高語義控制的精確性和去偏見的有效性。本研究為圖像合成領域的更公平和可控的生成模型的發展做出了貢獻。

關鍵字：語意編輯，去偏見，圖片生成、LoRA

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Abstract

This thesis investigates the adaptation of the CLIP text encoder for use in conditional diffusion models to address challenges related to semantic editing and debiasing. We explore the effectiveness of low-rank adaptations in enhancing the control over semantic attributes of generated images while simultaneously reducing inherent biases. The study utilizes various disentanglement strategies and introduces modifications to the text encoder to evaluate the potential for mitigating biases related to gender and ethnicity. Our findings indicate that fine-tuning the text encoder with targeted adaptations can significantly improve semantic control's precision and debiasing effectiveness. This work contributes to the development of more fair and controllable generative models in the field of image synthesis.

Keywords: Semantic Editing, Debiasing, Image Generative, LoRA

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Contents

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

List of Figures

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

List of Tables

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Denotation

α	Adaptation Strength Rate
C	Concept, a prompt after being tokenized
η	Learning Rate
W_0	Pretrained Model Weights
ΔW	Adaptation Weights
$(W)(C)$	Output of Model W Given Input C

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

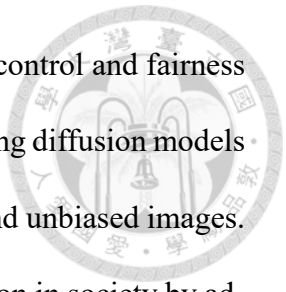
Chapter 1 Introduction

With the recent emergence of text-to-image generative models, users can access technologies like Midjourney and DALL-E 3 [7], powered by OpenAI, through websites, or use Stable Diffusion [8], which can be run locally on a user's GPU . More and more people are now using them as an alternative tool to help them create graphic design materials for advertisements and presentations. The advent of these new tools is expected to bring higher productivity and profoundly affect our daily experiences. However, current text-to-image generative models are not as perfect as imagined. Users might find that when they try to make subtle modifications to an image by updating small parts of the prompt, it also affects the parts they want to preserve or generate a completely different image. [10] This leads to unpredictable results with each editing, thereby reducing user productivity. Additionally, biases are commonly present in various pre-trained models, [3] which could further reinforce societal stereotypes in all images produced by generative AI.

1.1 Research Objective and Questions

The rapid evolution of text-to-image generative models promises significant advancements in digital content creation, poised to boost productivity and foster innovation across various sectors. However, the current limitations of these models, including their unpre-

doi:10.6342/NTU202403591



dictability and inherent biases, highlight a critical need for improved control and fairness in generated outputs. This work is motivated by the challenge of refining diffusion models to enhance user control over text encoders to produce more reliable and unbiased images. Our proposed enhancements seek to ensure their responsible application in society by addressing these technical and ethical issues.

1.2 Methodology Overview

In this work, we utilize LoRA, a parameter-efficient adaptation method for fine-tuning large models, as our adaptation approach. We attempt to adjust only the Text Encoder while maintaining the UNet configuration to validate our hypothesis that text embeddings contain both subtly modifiable semantics and inherent stereotypes related to gender and ethnic biases. This is achieved by updating in ΔW rather than directly adjusting the original pre-trained parameters W. This approach offers the following three benefits: (1) It is only necessary to share a smaller portion of parameters, not the entire model; (2) It allows for simultaneous adjustments of different aspects by pairing with multiple ΔW; (3) It provides a scalar α that enables the tuning of the strength or weakness of a single concept during the inference phase."

doi:10.6342/NTU202403591

Chapter 2 Related Works

2.1 Disentanglement in Image Generative Models

An image generative model is considered to have good disentanglement capability if it meets two criteria [10]: (1) modifying the target concept does not affect other unrelated features, and (2) the direction of modification for the same concept can be applied across different images. Early research found that GAN-based methods *** possess strong disentanglement capabilities, and can achieve concept modification and style conversion by altering the direction of vectors during the generation process. These capabilities can also be demonstrated mathematically, such as through Principal Component Analysis (PCA) [5], which identifies directions orthogonal to other concepts, thereby effectively proving the existence of disentanglement. However, such properties have not yet been discovered in stable diffusion models. Previous works tried discovering the directions or latent vectors for individual attributes in conditional diffusion models.

Previous work [10] have aimed to achieve partial modification of images by adjusting the semantic concepts contained within the text descriptions. This is accomplished by modifying the user's prompt, which, after passing through the CLIP text encoder, produces an embedding, also referred to as a neutral concept. The goal of semantic editing is achieved by either adding a target concept or reducing a negative concept within this

doi:10.6342/NTU202403591

vector. This modified vector then guides the synthetic processing in diffusion models.

2.2 Bias in Diffusion Models

Bias in generative models refers to systematic and unfair discrepancies in model outputs. These biases often reflect social and cultural prejudices embedded in the dataset.**??** The biases in diffusion models can manifest in various forms, such as stereotyping or underrepresenting specific demographics in generated images.

2.3 Guidance Based Methods

Guidance-based methods employ additional information to enable diffusion models to produce results users desire more accurately. Classifier Guidance [2], for example, utilizes a pre-trained classifier to direct the sampling process of the diffusion model. This classifier delivers gradients that adjust the generation process to ensure the results conform to the desired category. Conditional diffusion models generate results users anticipate by incorporating supplementary information, such as ControlNet [11] using text embeddings and different types of images as additional information, to produce images under specific conditions. Additional text conditioning variables are provided when training conditional models like Stable Diffusion [9], These variables then guide the denoising process during sampling, enhancing the model' s ability to generate targeted outputs.

doi:10.6342/NTU202403591

2.4 Semantic Control

Brack et al. [1] introduce Semantic Guidance (SEGA), a novel approach that allows interaction with concepts in diffusion models during image generation. In their work, they manipulate concepts directly in the latent space by adding and subtracting semantic vectors to regenerate images with adjusted semantic attributes.

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Chapter 3 Background

3.1 Diffusion Models

Diffusion models are generative models that generate images by iterative denoising an initially Gaussian random picture. These models operate through a forward and backward process, often referred to as the diffusion and denoising processes, respectively. The diffusion process gradually adds noise to the original image over a series of steps, transforming it into a simple Gaussian noise distribution. Mathematically, this can be explained as shown in Equation 3.1: let x_0 represent the initial image. The forward process then creates a sequence of noisy images $x_1, x_2, \ldots, x_T$, where T is the final step and x_T is almost a pure Gaussian noised image. Here, a_t controls the variance of the noise added at each step is controlled.

$$x_t = \sqrt{a_t} x_{t-1} + \sqrt{1 - a_t}\epsilon \tag{3.1}$$

The Denoising process, also known as the backward process, involves learning to gradually remove the noise added during the forward process, conditioned on a text prompt, to generate an image from a purely noisy image. Mathematically, as shown in Equation 3.2, the backward process is modeled as another Markov chain but in reverse, starting

doi:10.6342/NTU202403591

from a noisy image x_T and gradually removing predicted noise at each step, aiming to reconstruct x_0. Here, C_t represents the text conditioning information or the text embedding encoded by the text encoder from a conditional prompt, and t represents the current time step. The objective is to learn the reverse transitions that denoise the image step by step. Our work maintains image generation capability by partially altering the text encoder to guide the denoising process in a frozen U-Net.

$$x_{t-1} = x_t - \epsilon_\theta(x_t, t, C) \tag{3.2}$$

3.2 Low-Rank Adaptation

The Low-Rank Adaptation (LoRA) [6] method provides the ability to train a model with relatively fewer parameters that the original model by decomposing the weights for each targeted layer. Given a layer of a pre-trained model with weights $W_0 \in \mathbb{R}^{m \times n}$,where *m* and *n* are input and output dimensions, respectively. The LoRA method first selects a small number $r \ll min(m, n)$, and decomposes weights to be trained ΔW as

$$\begin{gathered} \Delta W = BA \\ where\ \ B \in \mathbb{R}^{m \times r} and\ \ A \in \mathbb{R}^{r \times n} \end{gathered} \tag{3.3}$$

we can form a new weight during inference time by computing $W = W_0 + \Delta W$, and in this work we can further manipulate a scale to enhance or suppress the effect of trained weights by giving a variable α such that

$$W = W_0 + \alpha \Delta W \tag{3.4}$$

doi:10.6342/NTU202403591

Chapter 4 Method

Inspired by Gandikota et al. [4], they introduced Concept Sliders, a method for fine-tuning LoRA on layers of UNet to guide concepts in generative process.As shown in Figure 4.1, we add our work adds LoRA onto the text encoder instead of UNet to significantly reduce the training time and number of required parameters and achieve similar results. Furthermore, we propose a method that trains LoRA to solve the bias issue in diffusion models.

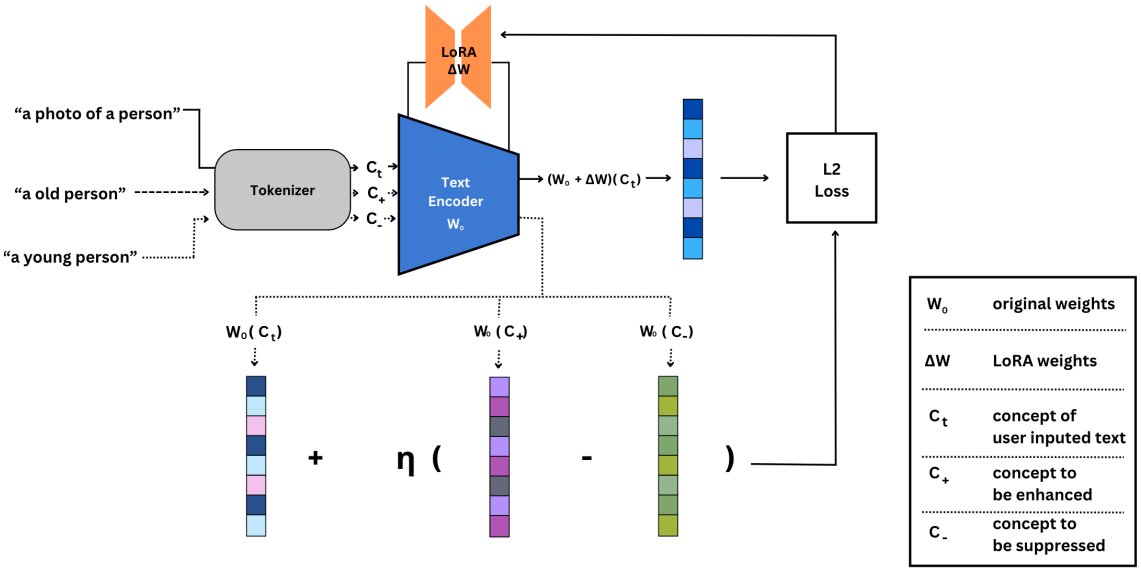


Figure 4.1: Our method demonstrates that fine-tuning LoRA makes it possible to maintain the concept of C_t while strengthening C_+ and weakening the concepts contained in C_-. This method can be executed just once during the inference phase instead of three times, and during training, it can also provide rich C_+ and C_- pairs to achieve the goal of disentanglement.

doi:10.6342/NTU202403591

4.1 Concepts Editing

Previous studies have demonstrated that adjusting the vector direction in latent space can effectively influence the outcomes of the final images. Consequently, incorporating this characteristic into the LoRA fine-tuning method enables users to modify images by adjusting weights. Our method enhances disentanglement by learning from multiple concept pairs

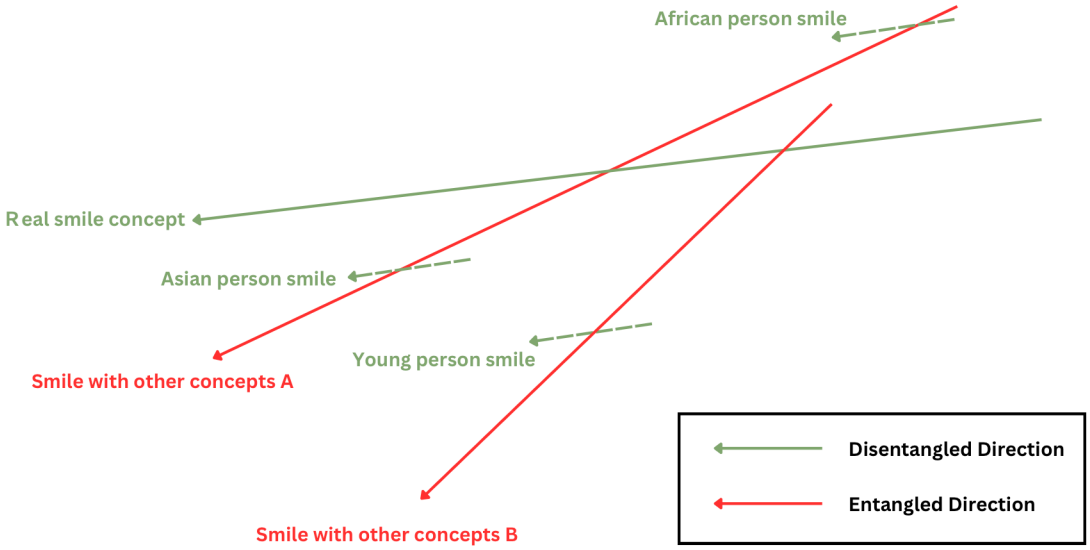


Figure 4.2: This figure demonstrates the concept of entanglement in the latent space. Green arrows represent disentangled directions, and red represents a direction containing other concepts. In this case, adjusting this direction may also alter other unexpected concepts

As shown in Algorithm 1, to obtain the final trained concept LoRA weights ΔW. Firstly, we encode the ground truth embedding which is designed to strengthen the C_+ concept that the user wants to enhance and weaken the C_- concept that the user wants to suppress, which is achieved by adding eta times the enhanced embedding to the current neutral embedding and subtracting eta times the suppressed embedding. Subsequently the

doi:10.6342/NTU202403591

target embedding is calculated by adding ΔW to W_0 and using these combined weights to encode the neural concept exclusively. We aim for the target embedding to approximate the ground truth embedding closely. To achieve this, we utilize the L2 loss function to quantify the discrepancy and update the weights of ΔW.

Additionally, as shown in Figure 4.2, directly adjusting the direction of concepts may inadvertently affect other unintended concepts. For example, selecting the direction of 'smile with concept A' in the figure may alter the ethnic features of a person while adjusting the smile, and choosing 'smile with concepts B' might adjust the person's age. Therefore, by selecting ground truth, we can find a disentangled direction that exclusively alters the concept the user intended in image editing by using more concept pairings.

Algorithm 1 Update_Text_Encoder

INPUT :
Pre-trained SD weights W_0,
(C_t, C_+,C_-) : (target concepts, enhancing concepts, suppressing concepts) pairs
OUTPUT :
Concept LoRA weights ΔW
for epoch in 1..N **do**
  $(c_t, c_+, c_-) \leftarrow$ randomly selected a pair from (C_t, C_+, C_-)
  $Embed_{ground}$ = $(W_0)(c_t)$ - η($(W_0)(c_+)$ - $(W_0)(c_-)$)
  $Embed_{target}$ = $(W_0 + \Delta W)(C_t)$
  loss = L2_loss($Embed_{ground}$, $Embed_{target}$)
  $\Delta W \leftarrow$ update_weights(loss)
end for

4.2 Debiasing

This work aims to mitigate the biases inherent in the pre-trained text encoder through this approach. Experiments in Section 5 reveal that in the original Stable Diffusion XL, the input prompt "a photo of a doctor" tends to generate images of males and lighter skin tones, whereas the prompt "a photo of a librarian" often results in females and lighter skin tone. Thus, we suspect that these professions' representations in the latent space

doi:10.6342/NTU202403591

carry strong gender and racial information or biases, causing the U-Net to produce specific gender and ethnic outcomes during the denoising step even without additional gender or ethnic information. To avoid potential reverse unfairness due to excessive intervention, we generate sample images during training to verify current gender ratios and ethnic distributions. Based on this information, we adjust η to ensure the appropriate strength and direction of learning among various concepts.

4.2.1 Gender Fairness

Algorithm 2 Binary Concept Debias

INPUT :
pre-trained SD weights W_0,
binary classes C
(C_t, C_+,C_-) : (target concept, enhancing concept, suppressing concept)
OUTPUT :
debiased LoRA weights ΔW
while True **do**
  images= $generator(W_0, \Delta W, p, nums = 20)$
  results = $Binary\_Classifier(images, C)$
  if results[0] == results[1] **then**
    break # Debiasing completely
  else
    if results[0] > results[1] **then**
      $\eta = -(1 - \frac{results[1]}{results[0]})$
    else
      $\eta = (1 - \frac{results[0]}{results[1]})$
    end if
  end if
  $\Delta W = Update\_Text\_Encoder(W_0, \Delta W, \eta, C_t, C_+, C_-)$
end while

Algorithm 2 proposes a method using adaptation on the text encoder to eliminate binary gender bias from the latent space, thus producing more equitable distributions in images. This approach defines male-related concepts as C_+ and female-related concepts as C_- We aim to let stable diffusion models generate gender-balanced results in line with

doi:10.6342/NTU202403591

the user-inputted prompt related to a specific profession. Each epoch begins by sampling 20 images and classifying each as male or female to assess the current distribution ratio. If the proportion of male representations in the images is lower than average, we enhance C_+ and suppress C_-, and vice versa. We constrain η within the range of -1 to 1 by adjusting to 1 − smaller proportion ÷ larger proportion.

4.2.2 Ethnic Fairness

Algorithm 3 General Concept Debias

INPUT :
pre-trained SD weights W_0,
multiple classes C
(C_t, C_+,C_-) : (target concept, enhancing concept, suppressing concept)
OUTPUT :
debiased LoRA weights ΔW
while True **do**
  Generate 20 images $images = generator(W_0, \Delta W, p)$
  Classified results $result = Binary\_Classifier(images, C)$
  if results[0] == results[1] **then**
    # Debias complete.
    break
  else
    if results[0] > results[1] **then**
      $\eta = -(1 - \frac{results[1]}{results[0]})$
    else
      $\eta = (1 - \frac{results[0]}{results[1]})$
    end if
  end if
  $\Delta W = Update\_Text\_Encoder(W_0, \Delta W, \eta, C_t, C_+, C_-)$
end while

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Chapter 5 Experiments

5.1 Concept Adjustment

In this section, we demonstrate how adjusting the weights of LoRA can steer the outcome of image generation. Our experiments have shown that we can influence the generation intensity of specific concepts by adjusting the scalar α to construct new weights of text encoder by the equation $W = W_0 + \alpha\Delta W$. Furthermore, we have highlighted the characteristics of disentanglement, enabling the simultaneous modification of the same image using multiple LoRAs as represented by the equation $W = W_0 + \sum_{i=0}^{n} \alpha_i \Delta W_i$.

5.1.1 Linearly Adjustable

As shown in Figure 5.1, our method successfully modifies the image without altering unwanted details in the original image. After training on various concepts, we achieve the effect of sliders by adjusting the weights of that LoRA. Providing larger values (more to the right) produces more intense results while providing negative values (more to the left) can suppress the generation of that concept in the image.

doi:10.6342/NTU202403591

Figure 5.1: This figure illustrates the results of implementing our LoRA-based text sliders.From top to bottom row, they represent the concepts of smile, age and surprise respectively, and from left to right, they represent the strength of these concepts. The picture in the middle represents the original image, It can be seen that under different concepts, adjusting the strength of the values can influence the outcome of the image

5.1.2 Multiple Concepts Adjustable

As shown in Figure 5.2 our method can also be implemented simultaneously, and here we provide some examples to demonstrate. Horizontal and vertical axes represent different concepts. Since we have found disentangled directions, unrelated concepts do not affect each other in image editing.

5.1.3 Parameter Efficiency

Our work demonstrates improved parameter efficiency compared to the Concept Slider. As shown in Table 5.2, the Concept Sliders method involves adding LoRA to

doi:10.6342/NTU202403591

Figure 5.2: This figure demonstrates that we can adjust different concepts simultaneously. The horizontal direction shows the degree of age adjustment, while the vertical direction displays the adjustment of the smile.

Benchmark	**Concept Sliders [4]**	**Ours**
Number of training Layers	346	88
Training time	1 hr 50 mins - 2 hrs	$\sim$ 2 mins
Weights size	9.1MB	1.7MB

Table 5.1: A comparison of Concept Sliders and Our Work across benchmarks shows that our approach reduces the training time by a factor of 60 and requires only about 18% of the space to save parameters.

UNet and requires learning the noise individually for each denoising step. In my method, adjustments are made only to the text encoder, which uses fewer parameters and layers. Moreover, it eliminates the need for different adjustments at each denoising step, thus reducing the computational complexity by 50 times per epoch

5.2 Debiasing

This section demonstrates the result of implementing trained LoRA weights to promote gender equality and ethnic diversity. We generate 250 images for each profession to

doi:10.6342/NTU202403591

compare the original Stable Diffusion XL model with the modified model incorporating LoRA weights.

5.2.1 Gender Equality

Figure 5.3 and Figure 5.4 4 show the results between the original and modified models. Our method can successfully generate more equally distributed results.

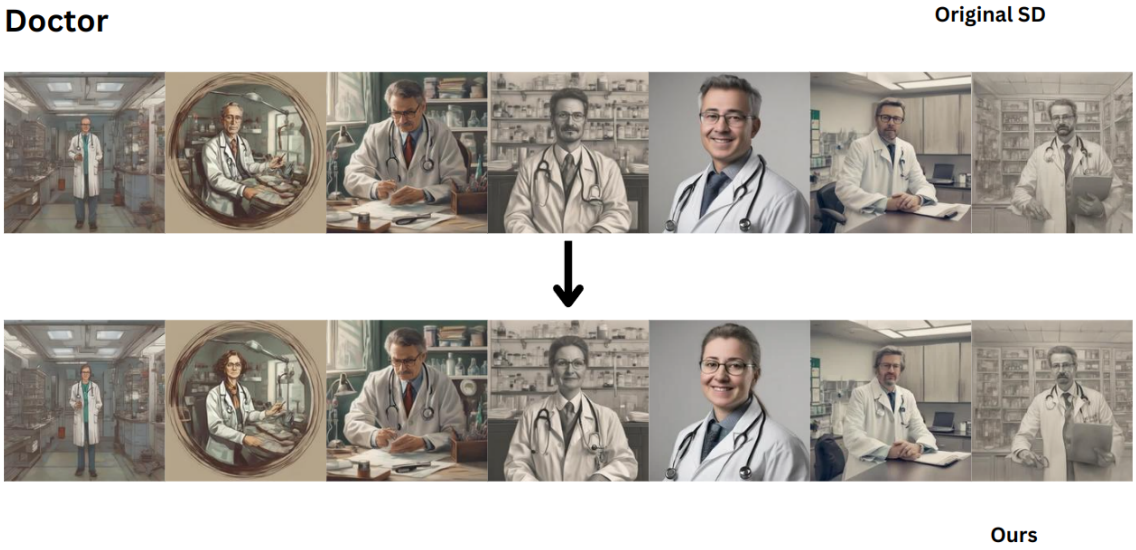


Figure 5.3: This figure compares the original stable diffusion model and the results after our debiasing adaptation. Initially, the proportion of males decreased from 97.2% to 54.4%.

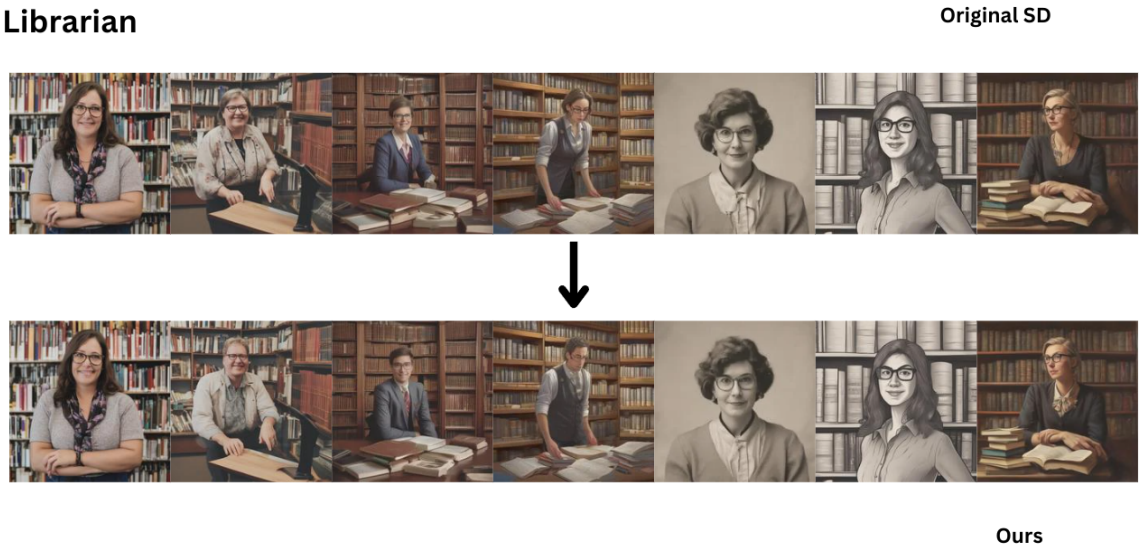


Figure 5.4: This figure compares the original stable diffusion model and the results after our debiasing adaptation. Initially, the proportion of females decreased from 82% to 52%.

doi:10.6342/NTU202403591

5.2.2 Ethnic Diversity

Figure 5.5 shows the result between the original model and the modified one. Because there is no universally agreed-upon method for categorizing race, to avoid generating subjective classification biases, we only present results that demonstrate diversity.

5.2.3 Training Cost

Figure 5.5: This figure compares the original stable diffusion model and the results after our debiasing adaptation, where the original single ethnicity has become more diverse.

Benchmark	**Concept Sliders [4]**	**Ours**
Training time	1 hr 50 mins - 2 hrs	30 - 90 mins

Table 5.2: Comparison of Concept Sliders and Our Work for debiasing method across benchmarks. The training time varies due to different sampling frequencies, however our approach remains faster than Concept Sliders.

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

Chapter 6 Conclusion and Discussion

6.1 Conclusion

In our work, we successfully addressed the two practical application scenarios initially set by adding adaptation to the text encoder. In the first use case, image editing, we demonstrated the ability to increase or reduce designated concepts through methods of enhancement and weakening. Thanks to the disentanglement capability, we can adjust various concepts to different degrees simultaneously, giving users more control over image generation. Concerning bias elimination, we discovered that occupational terms in the text encoder exhibit inherent gender and ethnic distribution imbalances. Our experiments demonstrate that the proposed learning methods can effectively reduce these biases, producing more equitable and diverse outcomes.

6.2 Discussion

Although we attempted to resolve the entanglement issues between different concepts using many sets of prompt pairs, our method can effectively modify concepts but still cannot achieve completely non-interfering results. Additionally, in the part of debiasing, modifications to address ethnic issues tend to be more drastic compared to gender

doi:10.6342/NTU202403591

modifications. We suspect this may be due to the difficulty of expressing ethnicity as clearly as binary gender using text alone, as ethnic concepts might inherently entangle multiple concepts (for example, pupil eye color, hair types, and different aesthetic preferences). Therefore, as shown in the Figure 6.1 , while maintaining the same gender, it is possible to retain more original features. However, under the same ethnicity, there have been noticeable changes towards darker glasses, pupils, and hair under the same ethnicity. Thus, attempting to eliminate bias solely through text remains a challenging task.

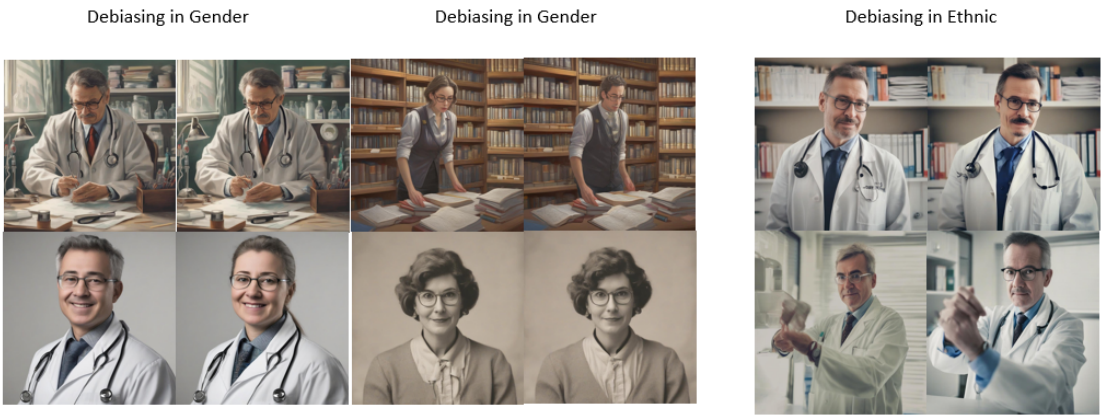


Figure 6.1: This figure shows that debiasing for gender preserves more original image features, while debiasing for ethnicity alters more, even though it maintains the same ethnic group

6.3 Future Work

We have initially demonstrated that partially modifying the text encoder has the potential to address the issues presented in diffusion models. As discussed above, the related problems still need to be resolved. Therefore, in the future, we plan to continue to delve deeper to solve the entanglement issue in the latent space. Thanks to the low cost of training in our proposed method, multiple different concepts can be trained simultaneously in feasible training time. During the training process, their cosine similarity is calculated

doi:10.6342/NTU202403591

to ensure that the vectors between them are orthogonal, potentially finding the proper direction of modification. In terms of optimizing ethnic classification, future efforts will attempt to incorporate a visual question-answering model to capture more detailed appearance descriptions and verify whether more precise debiasing results can be improved.

doi:10.6342/NTU202403591

doi:10.6342/NTU202403591

References

[1] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting. Sega: Instructing text-to-image models using semantic guidance. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 25365–25389. Curran Associates, Inc., 2023.

[2] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021.

[3] M. D'Incà, E. Peruzzo, M. Mancini, D. Xu, V. Goel, X. Xu, Z. Wang, H. Shi, and N. Sebe. Openbias: Open-set bias detection in text-to-image generative models. ArXiv, abs/2404.07990, 2024.

[4] R. Gandikota, J. Materzynska, T. Zhou, A. Torralba, and D. Bau. Concept sliders: Lora adaptors for precise control in diffusion models, 2023.

[5] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9841–9850. Curran Associates, Inc., 2020.

doi:10.6342/NTU202403591

[6] 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.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Dall·e 2: A modern text-to-image generation model. arXiv preprint arXiv:2204.06125, 2022.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[10] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. M. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang. Uncovering the disentanglement capability in text-to-image diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1900–1910, 2022.

[11] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

doi:10.6342/NTU202403591