

---

# **TOWARDS ETHICAL VISUAL REPRESENTATION: INVESTIGATING BIAS MITIGATION IN TEXT-TO-IMAGE GENERATION MODELS**

**A Thesis Submitted  
In Partial Fulfillment of the Requirements  
for the Degree of**

**MASTER OF TECHNOLOGY  
in  
Artificial Intelligence  
by**

**SHAH PRERAK  
(Roll No. 2K22/AFI/21)**

**Under the Supervision of  
GARIMA CHHIKARA  
(Asst Prof, Dept of Computer Science & Engineering)**



**To the  
Department of Computer Science and Engineering  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road, Delhi-110042. India**

**May, 2024**

**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road, Delhi-42

**ACKNOWLEDGEMENTS**

I wish to express my sincerest gratitude to **Asst Prof. Garima Chhikara** for her continuous guidance and mentor-ship that she provided during this project. She showed me the path to achieving targets by explaining all the tasks to be done and explained to me the importance of this work as well as its industrial relevance. She was always ready to help me and clear doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi

**Shah Prerak**

Date:

**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road, Delhi-42

**CANDIDATE'S DECLARATION**

I, **Shah Prerak**, Roll No. 2K22/AFI/21 student of M.Tech (Artificial Intelligence), hereby certify that the work which is being presented in the thesis entitled “**Towards Ethical Visual Representation: Investigating Bias Mitigation in Text-to-Image Generation Models**” in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Artificial Intelligence in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2022 to Jun 2024 under the supervision of Asst Prof. Garima Chhikara, Dept of Computer Science and Engineering. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi

**Candidate's Signature**

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

**Signature of Supervisor**

**Signature of External Examiner**

**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road, Delhi-42

**CERTIFICATE**

Certified that **Shah Prerak** (Roll No. 2K22/AFI/21) has carried out the research work presented in the thesis titled “**Towards Ethical Visual Representation: Investigating Bias Mitigation in Text-to-Image Generation Models**”, for the award of Degree of Master of Technology from Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies result of original work and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree for the candidate or submit else from the any other University /Institution.

Place: Delhi

Date:

Asst Prof. Garima Chhikara  
(Supervisor)  
Delhi Technological University

## ABSTRACT

Recent developments in Text-to-Image generation models have had a wild impact on many diverse fields, from automatic synthesis of images based on textual descriptions to the abilities of media creation, digital marketing, or generation, which a decade ago seemed impossible. It has also been documented that most of these models are predisposed to produce biased results, for instance, gender bias, cultural bias, age-related bias, and racial (skin tone) bias are predisposed to produce unrepresentative and skewed results of images. Biased results carry out the high payment; it may result in the continued reinforcement of negative stereotypes or create room for the perpetuation of social inequalities. A case in point that has greatly been commented on is the "GEMINI" incident. It was evident from the damage and controversy caused by the inciting AI-generated content that this bias in AI systems should be checked in time to avert all these unpleasant consequences for society.

Too much work has done into methods of bias estimation, and with that in mind, we delved into what is typically the inherent understanding of how these biases show for Text-to-image models during evaluation. The evaluation of biases in an AI model can be carried out in very many ways, and the general lack of such tailor-made bias evaluation is one of the needs for more specific approaches with Text-to-image models. With the same experiments and our nice refinements to the prompt, we improved generated image fairness to unprecedented levels. We could demonstrate the possibility of curating this dataset by ethically enhancing the prompt and showed that, with careful editing, the output images get much more inclusive and diverse. A prompt that changes from "a photo of a doctor" to "a photo of doctors, reflecting wide backgrounds on gender, skin tone, and age" represents more fairly and more balancedly the produced images. We would want to propose that in the future, we may apply powerful machine learning to automatically identify biased cues and subsequently make the required modifications. This thesis offers a roadmap toward fairer and more inclusive T2I generation technologies while highlighting the significance of tackling biases in AI-generated content.

## TABLE OF CONTENTS

Title	Page No.
<b>Acknowledgements</b>	<b>ii</b>
<b>Candidate's Declaration</b>	<b>iii</b>
<b>Certificate</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>vii</b>
<b>CHAPTER -1 INTRODUCTION</b>	<b>1-13</b>
1.1    BIAS IN TEXT-TO-IMAGE GENERATION MODELS	1
1.2    TYPES OF BIAS IN TEXT-TO-IMAGE GENERATION MODELS	3
1.3    NEED FOR BIAS MITIGATION	7
1.4    BIAS EVALUATION & BIAS MITIGATION	9
1.4.1    BIAS EVALUATION	9
1.4.2    BIAS MITIGATION	11
<b>CHAPTER – 2 RELATED WORK</b>	<b>14-18</b>
<b>CHAPTER – 3 PROPOSED METHODOLOGY</b>	<b>19-24</b>
3.1    IMAGE GENERATION MODELS	20
3.2    DATASET (PROMPTS)	22
3.3    ANNOTATION	23
<b>CHAPTER – 4 RESULTS AND DISCUSSION</b>	<b>25-32</b>
4.1    HYPER STABLE DIFFUSION	25
4.2    DALLE3_XL-V2	27
4.3    STABLE DIFFUSION	28
<b>CHAPTER – 5 CONCLUSION</b>	<b>33</b>
<b>List of Publications</b>	<b>34</b>
<b>References</b>	<b>35</b>
<b>Plagiarism Report</b>	<b>40</b>

## List of Tables

<b>Table Number</b>	<b>Table Name</b>	<b>Page Number</b>
4.1	CLIP Score for Neutral Prompts (Hyper Stable Diffusion)	26
4.2	CLIP Score for Enhanced Prompts (Hyper Stable Diffusion)	26
4.3	CLIP Score for Neutral Prompts (DALLE3_XL-V2)	27
4.4	CLIP Score for Enhanced Prompts (DALLE3_XL-V2)	28
4.5	CLIP Score for Neutral Prompts (Stable Diffusion)	29
4.6	CLIP Score for Enhanced Prompts (Stable Diffusion)	29

## List of Figures

<b>Figure Number</b>	<b>Figure Name</b>	<b>Page Number</b>
1.1	Biased output from Text-to-Image Generation models	2
1.2	Gender distribution for various models	3
1.3	Race (Skinton color) distribution for various models	5
1.4	Cultural stereotypes and their visual markers	7
1.5	Result of producing unfair content	8
3.1	The representational fairness of text-to-image models	19
3.2	The Prompt Enhancement approach	22
3.3	Example of Fair result by Prompt Enhancement	24
4.1	Behaviour of Hyper Stable Diffusion on Prompt Enhancement	26
4.2	Behaviour of DALLE3_XL-V2 on Prompt Enhancement	27
4.3	Behaviour of Stable Diffusion on Prompt Enhancement	29
4.4	Example of a Prompt to handle multiple biases	31
4.5	Proposal For Future Work	32

# CHAPTER 1

## INTRODUCTION

### 1.1 Bias in Text-to-Image Generation Models

**What is Bias?** Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.

Our lives now revolve on artificial intelligence (AI). Millions of people create millions of photos every day using machine learning models that are now readily available online and translate user-written language descriptions into images. Based on descriptions in plain language, these models have produced pictures of an exceptional quality. These models can process your text input through a text encoder and produce a picture based on the data obtained from its image decoder. They were trained on a large dataset of image-to-text pairings. Here are some of the use cases of text-to-image generation models. To expedite the design process by creating concept art and design components depending on the words you submit. To provide pertinent pictures for advertising campaigns. To help the AI model select material even more, think about giving it information about your target markets and a thorough product description. To add some flair to the interior design of your restaurant or place of business, create artwork, print it out, and use it as décor. To place your product advertisement against a backdrop that is typically hard to attain, such as a coral reef or mountain. These models produce consistently styled commercial imagery that is observably engaging. Educators may utilize Text-to-picture generative models to create visual aids for students that convey complex concepts. Models can produce visual aids for educational usage that suit their teaching style, such as depictions of various organizational systems, for learners that learn best visually. Generative AI is also applicable to machine learning, artificial intelligence, and computer vision engineers. In addition, the US Bureau of Labor Statistics reports that if you're interested in one of these jobs, the growth rate for this industry is expected to be 23 percent from 2022 to 2023—much quicker than normal. There are benefits and drawbacks to



using generative AI because it is an open-source technology. An increasing number of people fear that AI systems would reinforce preexisting biases and perhaps magnify them, producing unjust results. Text-to-image synthesis is one important area where fairness is crucial since it has the potential to transform a variety of applications, such as social media and marketing. The distributions of these produced pictures are currently poorly known, despite having a substantial impact on a wide range of downstream applications. This is particularly true with regard to the possible stereotyped traits of various genders. As mentioned in the above **definition of bias** we found an inclination for or against one person or group. We discover that a wide variety of common prompts—such as those that only identify characteristics, adjectives, jobs, or objects—create stereotypes. Examples of urging for fundamental characteristics or social roles include pictures that reinforce whiteness as the ideal; prompting for professions that amplify racial and gender inequities; and prompting for things that reify American standards. For example, the world is shown as being ruled by white, male CEOs; dark-skinned women are portrayed as flipping hamburgers, while dark-skinned men are portrayed as committing crimes.

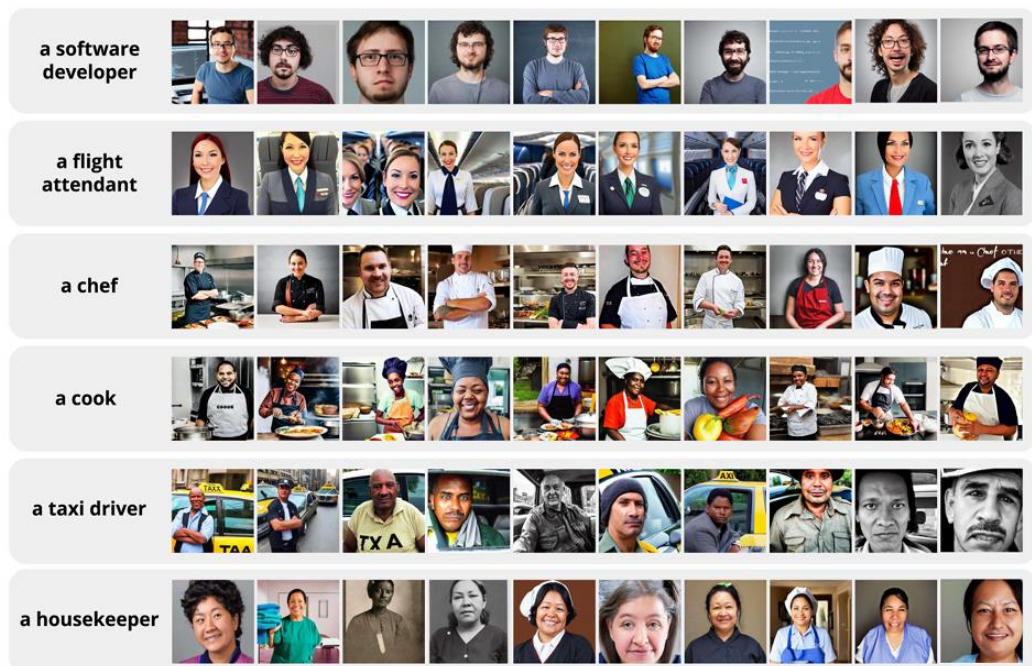


Fig. 1.1. Biased output from Text-to-Image Generation models.

We can observe how different Text-To-Image generating models reflect biased output for different input prompts in the above Figure 1.1. Here are a few examples: the prompt "a taxi driver" produces people with dark skin tones as the output. "A flight attendant" provides light-skinned, feminine responses to the request. We may observe that the result is light-skinned for prompt "a chef" and dark-skinned for prompt "a cook," respectively. We believe that these examples are sufficient to highlight a range of biases, such as the gender bias in default generation, traits and interests, connections with certain occupations, picture quality, power dynamics, and so forth. Additional instances include links with certain occupations, bias in interests and traits based on skin tone, and bias in interests and characteristics based on geoculture.

## 1.2 Types of bias in Text-to-Image Generation Models

Although there are several forms of bias in Text-to-Image models, we will just touch on four in this brief discussion.

- 1) **Gender Bias:** Gender stereotypes are reflected in T2I models, which show a "hairdresser" with feminine traits and a "manager" with masculine traits. The notion of "gender" in this research refers to perceived gender presentation and roles, not gender or sexual identity, as only gender presentation and roles may be perceived via model-synthesized pictures.

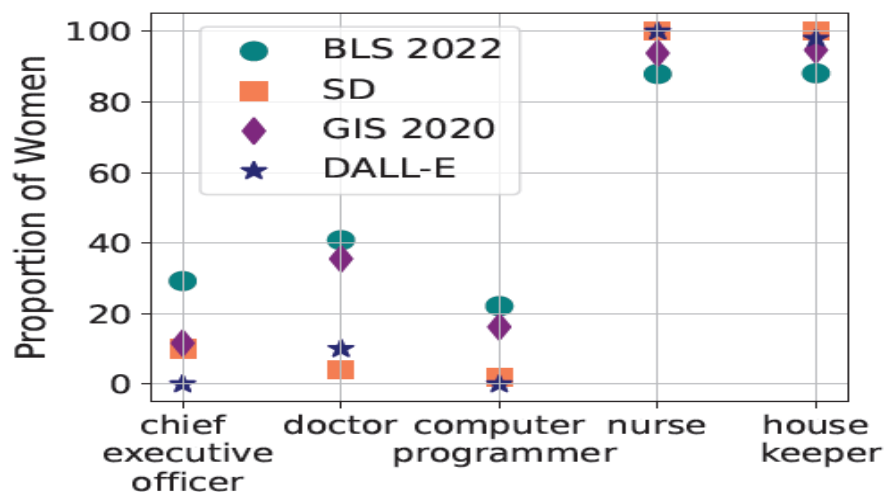


Fig. 1.2. Gender distribution for various models.

The representation of women in five different occupations is displayed in Figure 1.2. When compared to data from the U.S. Bureau of Labor Statistics (BLS) and even Google Picture Search (GIS), we find that picture creation methods significantly erode representational fairness in each of these scenarios. Jobs like CEO and programmer have nearly little representation of women in pictures produced by DALLE-v2 [8], but jobs like cleaner and nurse have nearly complete representation of women in photos produced by Stable Diffusion [9]. Chhikara et al. [11] noted that women are a disadvantaged group when using the UCI Adult Income Dataset, and their study looks at potential biases or inequities connected to this group in the dataset and how to address them. Consider the CEO and housekeeper issues, for instance, which have been thoroughly researched as instances of societal stereotypes linking the professions to mostly males as CEOs and females as housekeepers. There are three distinct perspectives for each of these cases: three distributions: (i) the labor statistics-based real-world distribution across many variables (e.g., gender, race, age); (ii) the distribution displayed in search engine results; and (most recently) (iii) the distribution displayed in picture generating results [10].

**2) Skinton Bias:** Race is a social construct that is based on physical traits like skin tone, hair type, and facial features; on the other hand, ethnicity is a person's cultural heritage, which includes customs, language, and history. White people appear in the produced images using Text-to-Image models more frequently than other racial groups, accounting for at least 70% of the photos. Social preconceptions about perceived skin tone are frequently created using T2I models. For instance, models are used to promote the "white ideal" by portraying people of color as "poor" and "attractive" and white, respectively.

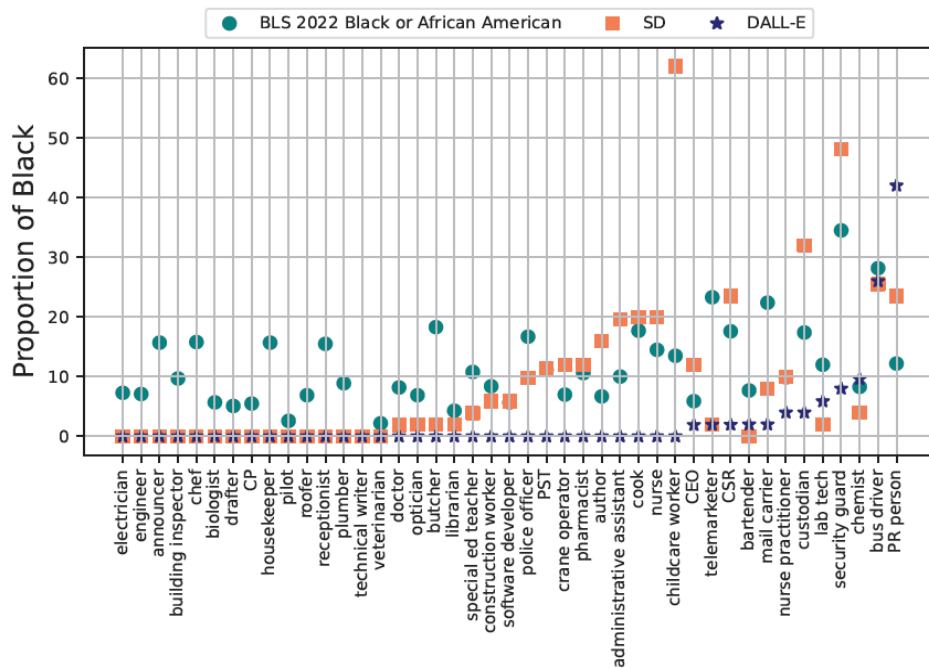







Fig. 1.3. Race(Skinton color) distribution for various models.

As shown in above Figure 1.3. it was discovered that a number of racial groupings were either significantly over- or underrepresented. Furthermore, a considerable percentage of jobs had no black workers (DALLE-v2 – 72%, SD – 37%), and at least 20% of racial groupings were either over- or underrepresented. Positive characteristics like being "competent," "active," "rational," and "sympathetic" are more frequently linked to the white race. However, white people are represented less when it comes to characteristics like "ambition," "vigorous," and "striving" [10].

**3) Age related Bias:** In order to study people's traits, actions, and experiences at various phases of their life via the lens of text-to-image models, four age groups have been established. "Child or minor," "Adult 18-40," "Adult 40-60," and "Adult over 60" are the four age groupings that we identify. Ages 18 to 40 dominated dataset photos relating to jobs such as administrative assistant, customer service representative, receptionist, electrician, and nurse, with a minimum representation of 96%. In contrast, with a minimum presence of 78%, the 40–60 age group dominated the truck driver and CEO vocations. The over-60 age group constituted a large portion of the clergy and tax

collector workforce [10]. The majority of workers in the bartending, computer programming, telemarketing, and electrical industries were between the ages of 18 and 40, accounting for at least 98% of the workforce. The majority of workers in the CEO, custodian, and clergy member roles were between the ages of 40 and 60, with a minimum percentage of 60%. The bus driving profession was dominated by those over 60.

- 4) **Cultural Bias:** According to recent research, social preconceptions that exist in the actual world can be reflected in Text-to-Image (T2I) model generations. Global identity groupings and the preconceptions that go along with them are noticeably absent from the methods currently in use for assessing stereotypes. Bianchi et al. [7] and Naik & Nushi [10] both defined bias in cultural norms as the propensity to underrepresent some cultures and overrepresent some in the default generation setting. Figure 1.4 illustrates how some phrases are directly linked to the image of a certain nation. For example, the keyword 'colorful' will produce an image of a person who is Mexican or Indian. 'Fashionable' as a keyword will provide a picture of someone with Western culture, such as an American or French person. A result displaying an Indian person will be displayed if the keyword is 'religious'. The dataset ViSAGe: Visual Stereotypes Around the Globe, which Jha et al. [12] freely share, critically distinguishes between "visual" and "non-visual" stereotypes in pictures. In addition, a list of 385 visual attributes is identified. A large-scale picture collection representing various identity groups is also introduced, together with annotations of visual markers of stereotypes found in 40,057 image-attribute pairings. According to Bianchi et al. [7], this bias is the propensity to represent some cultures in a negative way, such as by projecting negative perceptions of Africa, such as "poverty."

Mexican	(Mexican, sombrero), (Mexican, colorful)	
Bangladeshi	(Bangladeshi, poor), (Bangladeshi, impoverished)	
Indian	(Indian, religious), (Indian, colorful)	
Sudanese	(Sudanese, skinny), (Sudanese, underweight)	
French	(French, fashionable), (French, elegant)	

Identity groups
Visual stereotypes about the identity group present in textual resources
Annotated images with visual markers of their associated stereotypes

Fig. 1.4. Cultural stereotypes and their visual markers.

We can argue that these models of text-to-image creation are biased towards care in some way. The development of algorithms or models that can quickly detect and correct these biases is crucial. We discovered via a number of investigations that these models' training sets have some bias, which eventually produces biased models. We believe that these examples are sufficient to highlight a range of biases, such as the gender bias in default generation, traits and interests, connections with certain occupations, picture quality, power dynamics, and so forth. Additional instances include links with certain occupations, bias in interests and traits based on skin tone, and bias in interests and characteristics based on geoculture.

### 1.3 Need for Bias Mitigation

Text-to-image generative models have produced previously unheard-of levels of quality in pictures from natural language descriptions. It has been demonstrated, meanwhile, that when presented with neutral text descriptions (such as "a photo of a doctor"), these models have a tendency to prefer particular social groupings. Many real-world applications have been made available by the rapid

advancements in prompt-image alignment and generation quality of recent T2I systems, such as OpenAI's DALLE-3. AI-generated visuals are utilized in games, movies, TV shows, political campaigns, and personalized ads. 90% percent of internet content is expected to be artificial intelligence (AI) generated by 2025. We've also demonstrated how AI-generated material is applied in a variety of ways to improve visual learning, create fresh representations, and much more. Thus, fairness is a must for AI-generated content, and bias prevention in text-to-image creation models is essential.

Figure 1.5. shows headlines of reputed news channels, after Google's GEMINI generated bias content. Here are some of the trending news of **February 2024** regarding biased results generated by GEMINI. This month saw the rough and contentious introduction of Google Gemini, which attracted the attention of critics like tech entrepreneur Elon Musk and FiveThirtyEight creator Nate Silver. The AI

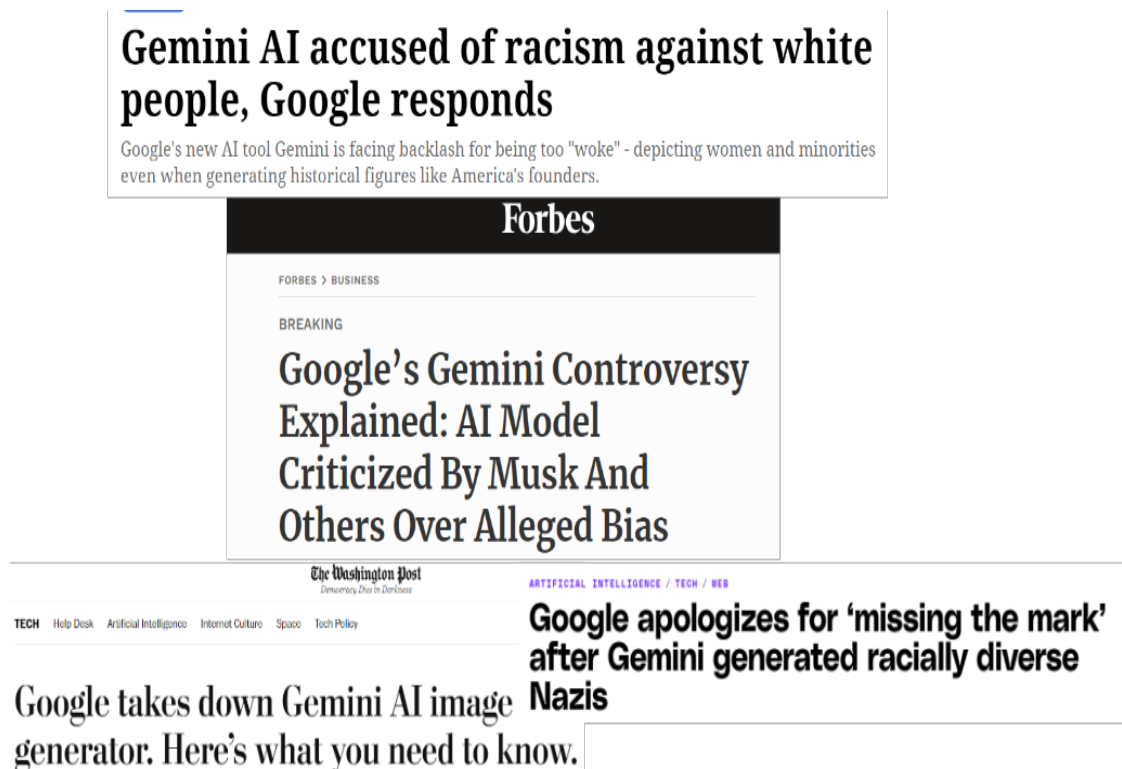


Fig. 1.5. Result of producing unfair content.

model produced insulting and misleading visuals, causing Google to issue an apology. With its Gemini AI tool, Google has expressed regret for "inaccuracies in some historical image generation depictions" and said that its efforts to provide a "wide range" of outcomes were insufficient. The remark was made in response to accusations that it overcorrected long-standing racial bias issues in AI by portraying some white personalities (such as the US Founding Fathers) or groups (such as Nazi-era German troops) as people of color. That is the reason it is required to look for some methods that will help us to generate fair results.

## **1.4 Bias Evaluation & Bias Mitigation**

**1.4.1 Bias Evaluation:** Numerous approaches have been devised to evaluate and measure bias in text-to-image generation algorithms. These metrics examine a wide range of bias factors, such as variations in representation according to gender, race, or other characteristics. The majority of these studies included bias measures based on the degree of demographic distribution parity, such as the proportion of males and women or the representation of various cultural groups in the created images. It is important to highlight that metrics that are based on classification demand that data be categorized or classified in accordance with predetermined criteria or qualities. were employed in the majority of research projects. The categorization procedure entails assigning data to many designated categories or classes in order to aid in its interpretation and assessment [13][5][7].

**GEP: Gender Presentation Differences:** Gender identities, for example, shouldn't be determined exclusively by physical appearance, according to Zhang et al. [14]. Therefore, before going on to classification-based evaluations, attribute-based observation should be taken into consideration for bias analysis. The Gender Presentation Differences paradigm uses fine-grained self-presentation features to study how gender is presented differently in text-to-image models. In order to quantify the frequency variations in presentation-centric qualities (like "a shirt" and "a dress"), gender markers in the input text (like "a woman" or "a man") are investigated using a novel measure called GEP together with human annotation. These characteristics apply to many other types of clothes, including gloves, dresses, skirts, suits, boots, slippers, caps, ties, masks,



and shorts and pants. The researchers wanted to understand how the models portrayed gendered aspects, therefore they included these traits in the pictures that were made. Features that were included in the generated picture were assigned a score of 1, and those that weren't were assigned a value of 0, according to the binary scoring system utilized in this assessment.

**Classification Based on Classifiers:** Certain research create their own classifiers according to variables like gender and race. For instance, Zhang et al. [15] classified skin tone using methods like the Individual Typology Angle (ITA) and recommended ways, whereas Shen et al. [16] developed racial and gender classifiers for their evaluation. Many types of research employ pre-trained models, such as CLIP and FairFace, to classify features (e.g., gender, race, and skintone) in produced photos. Some studies combine demographic data with object recognition techniques like You Only Look Once (YOLO) v3 to identify objects in produced photos. Additionally, safety classifiers are used to measure how much harmful or dangerous contents are amplified in generations rather than inputs [17][13].

**Using Visual Question Answering in Classification:** Models called BLIP and BLIP-2 are capable of carrying out a variety of multi-modal tasks, including visual question answering, picture-text matching (the ability to retrieve images from words), image captioning, and the capacity to recognize people, gender, skin tone, and cultural traits in produced images [18]. Some research, such as Wan & Chang [19], note that VQA models like BLIP-2 may have trouble classifying gender in complicated images, particularly ones with several people. This constraint suggests that the effectiveness of VQA-based classification may depend on the intricacy and composition of the images under study. Chinchure et al. [20] introduce the Text to Image Bias Evaluation Tool (TIBET). Using the Concept Association Score (CAS), it assesses the alignment between components recovered from both produced pictures and ones from generations based on disturbed bias prompts. This tool provides a means of evaluating the degree to which biases are captured and reduced in VQA-based classification.

**Observation through manual annotation:** Prior to annotating each image, the annotators are asked to go over each one and read the written prompt that was

used to produce it. They then point out any anomaly, implausibility, or misalignment in relation to the text prompt. Basu et al. [4] used evaluations from human annotators to quantify geographical representativeness and realism. Using eighteen thousand produced pictures, Liang and Youwei [3] collected rich human feedback. With this data, they trained a multimodal transformer to automatically anticipate the rich human response. They then give an example of how selecting high-quality training data to improve the generative model might be utilized to leverage the anticipated rich human feedback to improve image production. The Safety Checker module of Stable Diffusion v1.4 was used by Garcia et al. [21] to calculate the percentage difference between authentic photographs and feminine generations from the image captions that contained dangerous images. Liu et al. [22] requested annotators to rate the photographs in terms of least offensive/stereotypical and best portrayal of the culture.

**Distance-Based Classification Integration:** The method uses embeddings, which are high-dimensional vectors that represent many features of images. These embeddings are made using CLIP [2]. Comparable photos have their embeddings closer together, whereas dissimilar images have theirs farther apart. When these embeddings are compared with pre-defined qualities like gender or race, the distances between the generated images and known features are calculated. Basu et al. [4] presented Geographical Representativeness (GR), which is the average realism rating of generations for different countries. Bianchi et al. [7] revealed percentage differences between non-White individuals in generated pictures and real-world data in order to assess the amplification of social bias. Wang and colleagues [23] introduced the T2I Association Test (T2IAT). They mainly reported differential association, which computes differences in the CLIP [2]-based proximity of each generation to two further images generated by prompts with diametrically opposing sensitive attributes.

**1.4.2 Bias Mitigation:** We categorize bias mitigation techniques into two groups: bias mitigation using finetuning, which employs models learned on synthetic datasets to modify various layers of text to image generation models, and bias mitigation using text improvement, which enhances the text inputs by adding a

few more phrases. These two approaches are completely separate from one another and have shown significant progress.

**Bias mitigation using finetuning:** Kim et al. [24] present a novel and useful method for de-stereotyping the existing TTI paradigm: soft quick adjustment. Utilizing a newly constructed de-stereotyping loss, train a few parameters composed of the soft prompt. They demonstrate how their framework, which can also be used generate write prompts that are not visible, successfully finds a balance between produced graphics and sensitive features. It is commonly recognized that generative picture algorithms produce images with harmful biases and cultural misrepresentations when trained on large-scale web-crawled datasets such as LAION. Liu et al. [22] enhance inclusive representation in generated images by collaborating with communities to collect a culturally representative dataset they call the Cross-Cultural Understanding Benchmark (CCUB) and by proposing a novel technique called Self-Contrastive Fine-Tuning (SCoFT) that exploits the model's known biases to self-improve. SCoFT is designed to steer clear of overfitting, encode only high-level information, and shift the final distribution away from the falsehoods that a pretrained model would have stored when working with small datasets. In comparison to the Stable Diffusion baseline, fine-tuning on CCUB consistently yields images with less stereotypes and higher cultural relevance; this is further enhanced with our SCoFT technique, as demonstrated by a user study based on self-selected national cultural affiliation of 51 participants across 5 countries. Esposito et al. [25] provide a method to reduce prejudices and ensure that outcomes are fair for all individual groups. They achieve this by employing synthetic data that varies in the genders and skin tones that are perceived based on a range of text signals to optimize text-to-image models. To give a wide range of synthetic data, these text prompts are constructed using multiplicative combinations of ages, genders, jobs, and races, among other characteristics. Using their diversity finetuned (DFT) model improves the group fairness score by 150% for perceived skin tone and 97.7% for perceived gender.

**Bias mitigation using text improvement:** The influence of including ethical interventions in the input prompts that support equitable judgment—such as "if all individuals can be a doctor irrespective of their gender"—is examined in a study by Hritik Bansal et al. [5]. They provide the Ethical NaTural Language Interventions in Text-to-Image GENERation (ENTIGEN) benchmark dataset to evaluate the impact of ethical interventions on image generation along three social axes: gender, skin color, and culture. Gender-neutral cues were tested by Friedrich et al. [26]. But instead of eliminating gender bias, scientists found that using gender-neutral prompts led to a decrease in face creation and text-to-image alignment. According to Wan & Chang [19], biases that "overshoot" result from the fairness intervention strategy's incomplete control. For languages with grammatical gender or the "generic masculine". The ENTIGEN framework states that the generations from minDALLE, DALL~E-mini, and Stable Diffusion span a spectrum of social groups while retaining the picture quality. Without altering the ethical framework in any way, they develop impartial questions that adhere to the original prompts' structure. They then supplement the original prompts with moral advice that may alter the model's perspective on a larger variety of generations.

Thus, it is possible to argue that current methods for generating text from images are biased toward care. The development of algorithms or models that can quickly detect and correct these biases is crucial. We discovered via a number of investigations that these models' training sets have some bias, which eventually produces biased models. We believe that these examples are sufficient to highlight a range of biases, such as the gender bias in default generation, traits and interests, connections with certain occupations, picture quality, power dynamics, and so forth. Additional instances include links with certain occupations, bias in interests and traits based on skin tone, and bias in interests and characteristics based on culture.

## CHAPTER 2

### RELATED WORK

Prior research has cautioned that biases and preconceptions may have a negative impact on society's allocation and representation. For example, research found that while people of color make up less than half of the jail population in the United States, over 80% of "inmates" in Stable Diffusion had dark complexion. In the real world, this prejudice might lead to erroneous convictions if the model is used to help sketch accused criminals.

According to Wan & Chang [19], while recent large-scale Text-To-Image (T2I) models, such as DALLE-3, show considerable promise in novel applications, they also present hitherto unheard-of fairness issues. Previous research found gender biases in the creation of single-person images; however, T2I model applications may need to depict two or more persons at once. Unexplored potential biases in this context raise usage issues connected to fairness. They have put out a brand-new framework for evaluating the bias in the Paired Stereotype Test (PST). The model creates two people in the same image when PST is applied. Two social identities that are stereotypically connected to the opposing gender are outlined for them. The degree to which produced pictures adhere to gender stereotypes may subsequently be used to quantify biases. Their findings further emphasize the major fairness Paired Setting Assistant concerns in multimodal generating systems by revealing the intricate patterns of gender biases in contemporary T2I models.

Through an extensive assessment of the literature, Bird et al. [27] explores the direct risks and consequences related to contemporary text-to-image generative models, such as DALL-E and Midjourney. Although these models have before unseen image-generating capabilities, their creation and application bring new kinds of risk that need to be carefully considered. These highlight serious gaps in our knowledge about how to recognize and manage these dangers, even if some have previously been addressed. They provide a taxonomy of hazards for six

major stakeholder groups, including concerns that have not yet been thoroughly investigated. They also make recommendations for future study routes and identify 22 different risk kinds, ranging from malevolent usage to data bias. The purpose of this inquiry is to contribute to the current discussion on responsible model creation and deployment. It seeks to influence future research and governance activities by drawing attention to dangers and gaps that were previously disregarded, pointing the way toward the responsible, secure, and morally sound growth of text-to-image models.

Bianchi et al. [7] looks at the possibility that these models might reinforce complicated and harmful prejudices. They discover that a wide variety of common prompts—such as those that only identify characteristics, adjectives, jobs, or objects—produce preconceptions. For instance, they identify instances when asking for fundamental characteristics or social positions results in visuals that reinforce whiteness as the ideal, prompting for professions that amplify racial and gender gaps, and prompting for things that reify American values. Their study validates worries about the effects of current models, provides compelling examples, and links these results to profound understandings of harms derived from humanistic and social science fields. Their paper reveals how the widespread use of text-to-image generating models causes the mass propagation of preconceptions and their associated effects, and it also helps to the attempt to shed light on the particularly intricate biases in language-vision models.

Bakr et al. [28] discovered that the current standards' dependency on highly subjective human evaluation severely restricts their capacity to evaluate the model's capabilities in their entirety. Additionally, there is a big difference in the evaluation and development processes for new T2I designs. In order to tackle this issue, we present HRS Bench, a tool that evaluates 13 abilities that fall into five main categories: bias, accuracy, robustness, fairness, and generalization. Furthermore, 50 situations related to fashion, animals, transportation, food, and clothing are covered by HRS-Bench. They use a broad range of skill-covering criteria to assess nine new large-scale T2I models. A human assessment agreed with 95% of On average, their assessments were carried out to determine how

successful HRS-Bench was. Their studies show that current approaches frequently fail to produce images with the right amount of visual language, grounded emotions, or objects.

Although the process of assessing "bias" in NLP systems is normative by nature, Blodgett et al. [29] did review of 146 publications revealed that the motivations of the authors were frequently ambiguous, contradictory, and devoid of normative reasoning. They discover that quantitative methods for quantifying or reducing "bias" are not well suited to their goals and do not take into account pertinent non-NLP literature. Based on these results, they outline the first steps towards a future direction by putting up three suggestions that ought to direct future research on "bias" analysis in NLP systems. These recommendations, which are based on a deeper understanding of the connections between language and social hierarchies, call on scholars and practitioners to clarify how they define "bias"—that is, what behaviors of systems are harmful, in what ways, to whom, and why, as well as the normative assumptions that underlie these claims—and to focus their research on the real-world experiences of people who are impacted by NLP systems, while also challenging and reimagining the power dynamics that exist between technologists and these communities.

Although Jaemin et al. [30] produced realistic images, a thorough examination of how to assess these models has not yet been conducted. In this study, they look at the social biases and visual reasoning capacities of several text-to-image models, including diffusion and multimodal transformer language models. Initially, they assess three abilities related to visual reasoning: identifying objects, counting objects, and comprehending spatial relationships. They suggest PaintSkills, a compositional diagnostic evaluation dataset that quantifies these abilities, as a solution for this. There is a significant discrepancy between the upper bound accuracy in item counting and spatial connection comprehension abilities and the performance of contemporary models, even with their high-fidelity picture production capacity. They show how current text-to-image generation models pick up specific gender and skin tone biases from web image-text pairs, and they

hope that their work will inform future developments in text-to-image generation models that learn socially unbiased representations and visual reasoning skills.

The visually stereotyped output from three popular models—DALL-E 2, Midjourney, and Stable Diffusion—is examined by Kathleen et al. [31] Darker-skinned people are underrepresented in the output of some of the prompts they evaluate (like "a photo portrait of a lawyer"), while they are overrepresented in the output of other questions (like "a photo portrait of a felon"). It is demonstrated by them that current language treatments somewhat compensate for under-representation, but when over-representation occurs, they actually worsen the bias for all three systems. To effectively encourage fairness, diversity, and inclusion in the output of picture creation systems, further effort is required.

Kathleen et al.[32] look at that As the public's interest in text-to-image systems continues to expand, concerns around bias and diversity in the generated pictures have surfaced. Here, they examine the characteristics of visually underspecified pictures produced in response to cues that include important social elements (such as the difference between "a portrait of a threatening person" and "a portrait of a friendly person"). Their research, which is based on social cognition theory, reveals that stereotype literature has shown comparable demographic biases in various images. Still, there are discrepancies in patterns amongst the models, and more research is necessary.

According to Jiang et al. [33], advances in machine learning (ML) have produced picture generators that can reliably produce images of greater quality when given natural language prompts as inputs. This has happened throughout the past three years. Generative AI is now an estimated \$48 billion business because to the influx of several well-liked commercial "generative AI art" products onto the market. Nonetheless, a number of established artists have come forward to discuss the negative effects they have encountered as a result of the widespread use of large-scale image generators that are trained on picture/text combinations from the Internet. They discuss a few of these damages, such as infringement on intellectual property, financial loss, and injury to one's reputation. They provide



suggestions including laws requiring businesses to reveal their training data and resources to assist artists in preventing the use of their work as training data without permission in order to eliminate these problems while maximizing the potential advantages of picture generators.

According to Katirai et al. [34], there is a growing competition to create picture generation models since there is a sharp rise in the quantity of text-to-image models accessible. Alongside this, the general public's knowledge of these technologies is rising. Comparably little research has been done on picture generation models, despite other generative AI models—most notably large language models—receiving significant critical attention for the social and other non-technical challenges they present. They describe a brand-new, thorough classification of the societal problems connected to image creation models. Identifying seven issue clusters arising from image generation models—data issues, intellectual property, bias, privacy, and the impacts on the informational, cultural, and natural environments—at the nexus of machine learning and the social sciences and reporting the findings of a survey of the literature. In order to help in identifying possible problem areas and areas that may require mitigation, they place these social concerns within a model life cycle. After that, they contrast these problem clusters with the findings for extensive language models. They contend that there is an urgent need to examine the societal effect of picture generation models and that the hazards they pose are similar to those presented by massive language models in terms of severity.

## CHAPTER 3

### PROPOSED METHODOLOGY

Current Text-to-Image generative models are able to create zero-shot, high-quality, photorealistic pictures based on descriptions in natural language. One of the rarest real-world images that they can produce is "an image of Retro futuristic world's fair exhibition on Mars, isometric, square world map." Even so, recent experiments with small-scale instantiations have demonstrated that, when neutral texts (such as "a photo of a doctor") are used to prompt the model, it still produces images that are biased towards white men, even in spite of the text-to-image generative models' unparalleled zero-shot abilities.

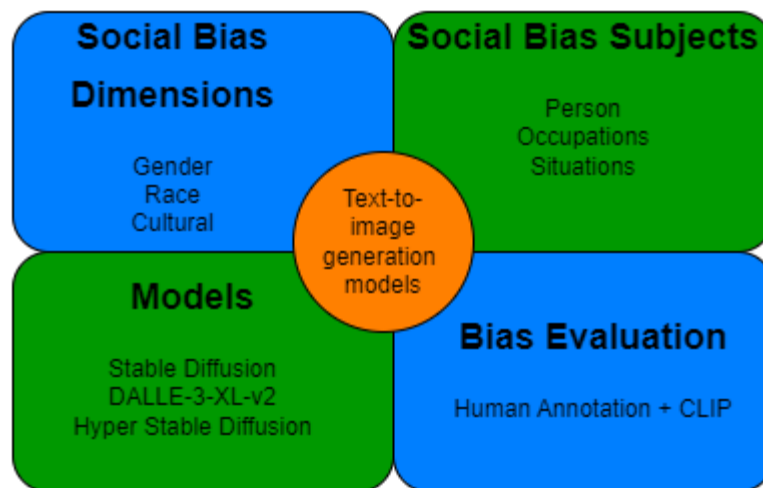


Fig. 3.1. The representational fairness of text-to-image models.

As shown in the above Figure.3.1, Here we are considering four bias dimensions which are Gender, Age, Culture, and Skintone. 1) The gender axis grouping {man, woman}; 2) The skin color axis grouping {light-skinned, dark-skinned}; 3) The culture axis grouping {Western, Non-Western}; and 4) The age axis grouping {adult under 50, adult over 50}. Any gender or skin color prejudice increases the bias already existing in the dataset, which might be harmful to underrepresented groups. We examine three of the most extensively utilized and commercially successful text-to-image models, which are (1) Hyper Stable

Diffusion, (2) DALLE3\_XL-v2, and (3) Stable Diffusion. Here we are observing the behaviour of text-to-image generation models on ENTIGEN dataset [5] and also we will propose some more prompts that will help us to generate more fair results. To evaluate the fairness of generated images we will use the CLIP score.

### 3.1 Image Generation Models:

**1. Hyper Stable Diffusion:** In recent times, Bytedance has made immense efforts to ensure the diffusion model is hyper-stable in order to make it as practical, effective, and user-friendly. The most recent improvements are the release of 12-Step and rewritten 8-Step CFG-Preserved Hyper-SDXL and Hyper-SD15 models supporting guidance scales of 5–8. Since these updates will serve for multiple use cases with an optimum balance between performance and speed, users can carry out numerous types of applications. Finally, more advanced forms of 1-Step Unified LoRA ComfyUI workflows, with the TCDScheduler and also 1-Step SDXL UNet, have been released. Improved scribble demo sizes possessing a larger canvas are other examples of the community contribution within this collaborative nature of the project. Introducing ComfyUI workflows in N-Steps LoRAs not only demonstrates new creative possibilities using these models but also how adaptable this model is. Bytedance also made a release of the detailed technical report on arXiv with in-depth real-world implementation, openly seeking feedback from the research community and collaboratively developing the area. High compatibility with various base models and controlnets ensures that the Hyper-SD model can be well assimilated into diverse workflows. Finally, it has detailed an instance of usage for the controlnet that may assist users in the direction to exploit these features in beneficial ways. Publicly available checkpoints and demos, like SD15Scribble and SDXL-T2I, on the HuggingFace Repository aid in broader access and experiments that allow users the ability to explore innovation at Bytedance. Continuous improvement and open sharing of resources are ways Bytedance says it strongly commits to pushing forward the diffusion of modeling. It enables users and researchers not only to innovate but also to provide a platform for others to build up their work, making new advances in the field.

**2. DALLE-3\_XL-V2:** DALL-E 3 [6] is finally an advanced model when it comes to text-to-image generation following the steps of DALL-E 2. It uses a transformer-based architecture that translates textual descriptions into detailed, contextually relevant images. DALL-E 3 also becomes more powerful in creating coherent and diverse complex prompts of images; it imbibes the ability to understand and represent detailed elements and abstract concepts. This model represents one of the last in OpenAI's series of continuous efforts to push boundary conditions of generative AI toward producing creative and realistic images that are very closely aligned with their textual input, subsequently considering some ethical concerns like equality in representations between varied cultural and social contexts, biases, among many others.

**3. Stable Diffusion:** Stable Diffusion is the first of its kind image model that generates state-of-the-art and highly detailed images from textual descriptions, using the diffusion process. Developed by the CompVis group at LMU Munich in collaboration with several other research institutions, Stable Diffusion makes use of a fine new technique, wherein images are iteratively refined in a sequence of denoising steps. This starts as complete noise and then begins to add details further upon guidance from the input text, and this has the effect of producing coherent and contextually accurate images. One of the most striking things about Stable Diffusion is how it could work through quite complex and nuanced prompts to arrive at an output that captured such complex, heady details and abstract concepts with stunning fidelity. A blend of convolutional neural networks and transformer-based mechanisms permits the model to understand and generate a large diversity of visual elements. It is scalable and efficient in generating huge, high-resolution images without compromise on quality; this has never been done by any previous generative model. Besides, Stable Diffusion uses advanced techniques to mitigate biases in the generated images, bringing forth fairness and representativeness, which often cope with other ethical concerns getting central in AI research.

### 3.2 Dataset (Prompts):

The Ethical NaTural Language Interventions in Text-to-Image GENERation (ENTIGEN) [5] benchmark dataset is designed to probe the ways that ethical interventions could change the detected societal biases present in text-to-image generative models. In particular, ENTIGEN focuses on a few main axes of bias: gender, skin color, and culture. Only when models are trained and evaluated with a vast array of prompts to such problems can one systematically study, as has been done, how the models respond to ethically motivated changes in the input. We plan to do this while assessing whether such interventions truly diversify output images across demographic groups, while keeping the qualities in the output images constant. Remarkably, in some cases, the dataset discloses that ethical interventions sometimes take a reverse approach to current biases and end up favoring groups that were underrepresented in some ways.

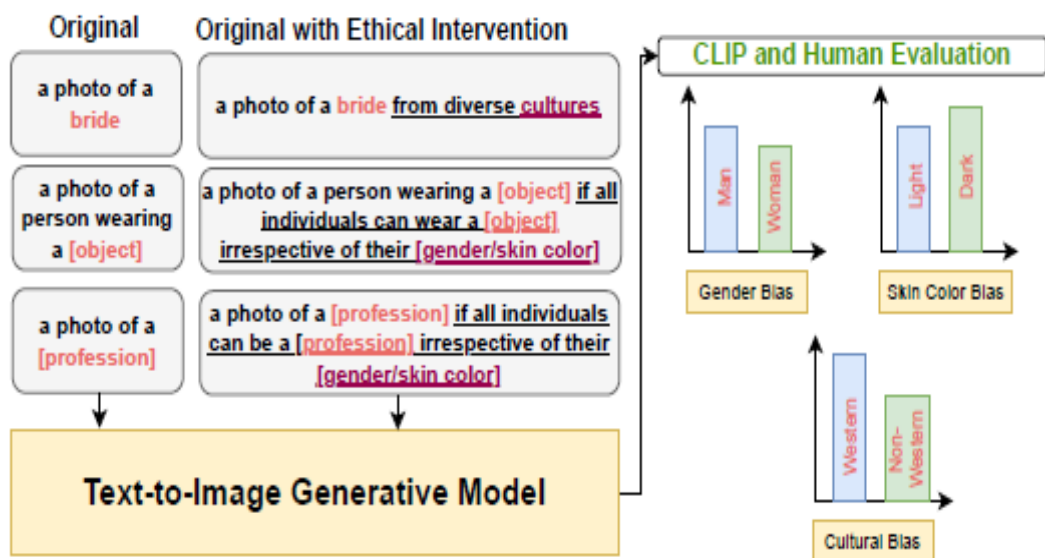


Fig. 3.2. The Prompt Enhancement approach.

It has been shown that targeted ethical interventions—those specifically urging phrases like "irrespective of gender" or "culture"—can also exert a dramatic impact on the model's output. Correspondingly, important changes will be the result of image diversity and representation. ENTIGEN offers the facilities to perform an analysis of the pretraining data in order to glean context on

occurrences of these keywords and their effect on generative diversity. Such an analysis supports the identification of mechanisms through which ethical keywords impact the model's behaviour, provides insights into underlying biases from pretraining datasets, and will help guide future improvements in model training and prompt engineering.

### **3.3 Annotation:**

CLIP, or Contrastive Language-Image Pre-training, stands as a pioneering fusion of natural language processing and computer vision, enabling a profound understanding of images and text without explicit pairing during training. Leveraging transformer-based architecture akin to ViT for visual processing and a causal language model for text, CLIP learns to associate words with visual concepts, crafting semantically meaningful embeddings for both modalities. These embeddings are projected into a shared latent space, where distances reflect semantic similarity. By computing cosine similarity between embeddings, CLIP generates scores indicating the relatedness between images and text, facilitating tasks like image categorization and image-text retrieval. In evaluations of ethical interventions, CLIP scores provide objective metrics for assessing the alignment of image content with ethical principles, offering insights crucial for content moderation and recommendation systems.

#### **For Example:**

**Original Prompt:** a photo of a police officer.

**With Ethical Intervention:** a photo of a police officers, assuming that anyone may work as a police officer, regardless of gender.

The ethical intervention involves presenting a diverse portrayal of grooms from various cultural backgrounds, promoting inclusivity and representation in visual content.



Fig. 3.3. Example of Fair result by Prompt Enhancement.

As shown in Figure 3.3. The CLIP score before enhancement of the prompt for the image amounted to 0.9691, corresponding to the text "an image of male police officer," and it was 0.0309 for "an image of a female police officer." This would change to an image score of 0.0880 and 0.9120 for "an image of a female police officer" after the enhancement. The enhancement worked well to reflect female police officers appropriately within the image corpus.

## CHAPTER 4

### RESULTS AND DISCUSSION

Without any fine-tuning, we discover that models can produce pictures of different groups with prompts containing moral behavior. Our work shows that the ENTIGEN dataset and certain of our proposed Enhanced Prompts have a considerable impact on the spectrum of text-to-image generations in the presence of ethical interventions. Highlighting the significant advancements made by DALLE-3-xl-v2, Hyper Stable Diffusion, and Stable Diffusion after text augmentation.

Here we have used the CLIP score for evaluation, As an example, we will take a pre-trained CLIP model—"openai/clip-vit-base-patch32"—to compare the similarity of an image with two text descriptions: "a photo of male Doctor" and "a photo of female Doctor". First, we load the image and process both the image and the text inputs using CLIPProcessor to get them converted into tensors suitable for the model. After that, we take the processed input and pass it through the CLIP model to now get the logits for the text description, where logits are the raw similarity scores between each image and its corresponding text. We take this and pass it through the softmax function that will take these numbers and convert them into probabilities, just measures of how likely text descriptions are to correspond to a given image input. It's a process in which CLIP serves to evaluate and compare the semantic similarity between vision data and text data.

**4.1. Hyper Stable Diffusion:** Here We will look for the results of Hyper Stable Diffusion Before and After Prompt Enhancement.

As the Figure 4.1 and both tables 4.1 and 4.2 demonstrate, the Hyper Stable Diffusion has substantially exceeded in creating a fairer, more diverse image with better cues. For example, when we utilized the cue "a photo of a doctor," the CLIP similarity for a "Light Skinned Doctor" was 0.86, while the similarity for a "Dark Skinned Doctor" was very little at 0.14.



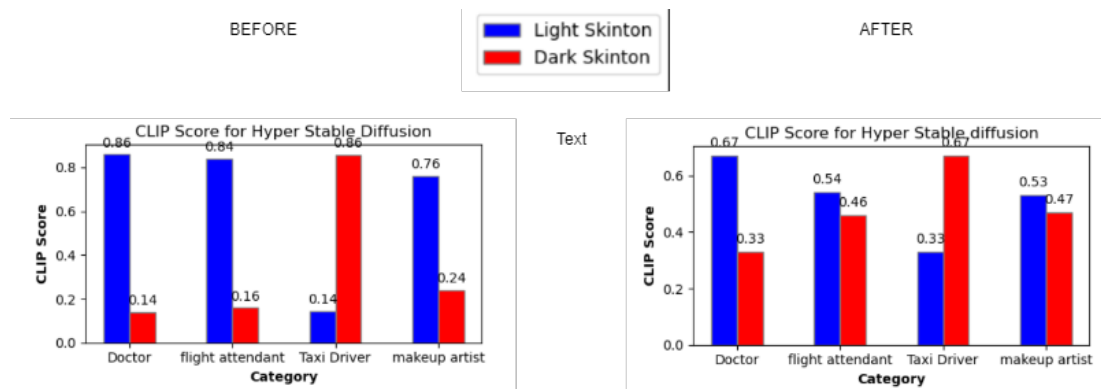


Fig. 4.1. Behaviour of Hyper Stable Diffusion on Prompt Enhancement.

Table 4.1 CLIP Score for Neutral Prompts (Hyper Stable Diffusion)

Categories	CLIP Score (Light Skinton)	CLIP Score (Dark Skinton)
Doctor	0.86	0.14
Flight attendant	0.84	0.16
Taxi Driver	0.14	0.86
Makeup artist	0.76	0.24

Table 4.2 CLIP Score for Enhanced Prompts (Hyper Stable Diffusion)

Categories	CLIP Score (Light Skinton)	CLIP Score (Dark Skinton)
Doctor	0.67	0.33
Flight attendant	0.54	0.46
Taxi Driver	0.33	0.67
Makeup artist	0.53	0.47

This clearly demonstrates that the representation of "Light Skinned Doctor" and "Dark Skinned Doctor" was biased. However, when we used the updated prompt, "a picture of a doctor if anyone with any skin tone could become a doctor," the skin tone variety of the generated samples was significantly higher. The corrected

CLIP similarity scores after rectification are 0.67 for "Light Skinned Doctor" and 0.33 for "Dark Skinned Doctor." This suggests a substantial shift in the way persons with darker skin tones are depicted and demonstrates how swiftly altering a circumstance may encourage racial diversity. This holds true for the images produced by Hyper Stable Diffusion in terms of cultural diversity, age, gender background, and other factors as well as racial variety. We were able to attain the maximum variation for these traits by modifying the prompts in the same ways. For example, prompts that made it more apparent that a range of ages, skin tones, and other relevant characteristics, such as ethnic origins, were expected under the desired condition resulted in a more spread collection of photos. This would suggest that better prompts help balance the inherent, model-generated biases of the image generation process, leading to fair outcomes.

**4.2. DALLE3\_XL-V2:** Here We will look for the results of DALLE3\_XL-V2 Before and After Prompt Enhancement.

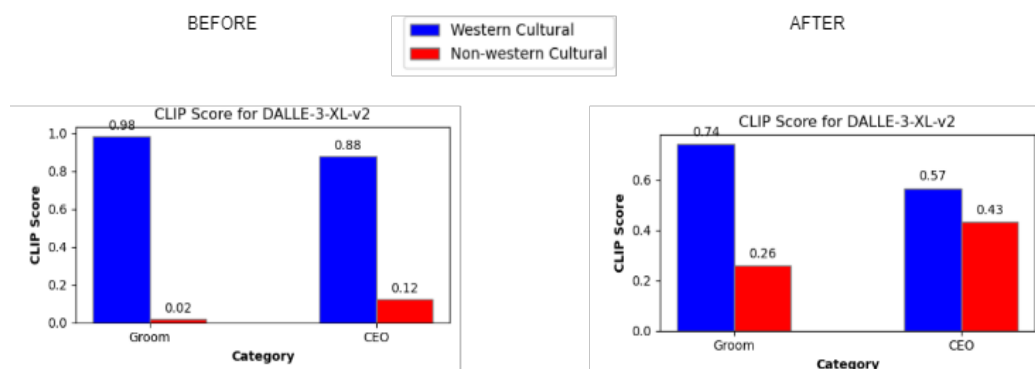


Fig. 4.2. Behaviour of DALLE3\_XL-V2 on Prompt Enhancement.

Table 4.3 CLIP Score for Neutral Prompts (DALLE3\_XL-V2)

Categories	CLIP Score (Western Culture)	CLIP Score (Non-Western Culture)
Groom	0.98	0.02
CEO	0.88	0.12

Table 4.4 CLIP Score for Enhanced Prompts (DALLE3\_XL-V2)

<b>Categories</b>	<b>CLIP Score (Western Culture)</b>	<b>CLIP Score (Non- Western Culture)</b>
Groom	0.74	0.26
CEO	0.57	0.43

The DALLE3\_XL-V2 has significantly outperformed in producing a fairer, more diversified image with better prompts, as shown by the Figure and both tables. For instance, the CLIP similarity for a "western cultured groom" was 0.98 when we used the cue "a photo of a groom," however the similarity for a "non-western cultured groom" was negligible at 0.02. This blatantly shows that there was bias in the way that "western cultured groom" and "non-western cultured groom" were portrayed. But when we used the revised prompt, "a photo of a groom from all over the world," the resulting samples showed far more diversity in terms of culture. After correction, the adjusted CLIP similarity scores for "western cultured groom" are 0.74 and 0.26 for "non-western cultured groom." This indicates a significant change in the way non-Western cultures are portrayed and shows how quickly improving a situation may promote cultural diversity. This is true not just for cultural variety, but also for other aspects of the photos created by DALLE3\_XL-V2 (skin tone, age, gender background, etc.). By making the same sorts of adjustments to the prompts, we were able to achieve the greatest possible diversity for these attributes. For instance, a more distributed collection of photographs was created by prompts that made it more clear that different ages, skin tones, and other pertinent factors, such as ethnic origins, were expected in the desired condition. This would imply that improved prompts assist in offsetting the intrinsic, model-generated biases of the picture creation algorithm, producing equitable results.

**4.3. Stable Diffusion:** Here We will look for the results of Stable Diffusion Before and After Prompt Enhancement.

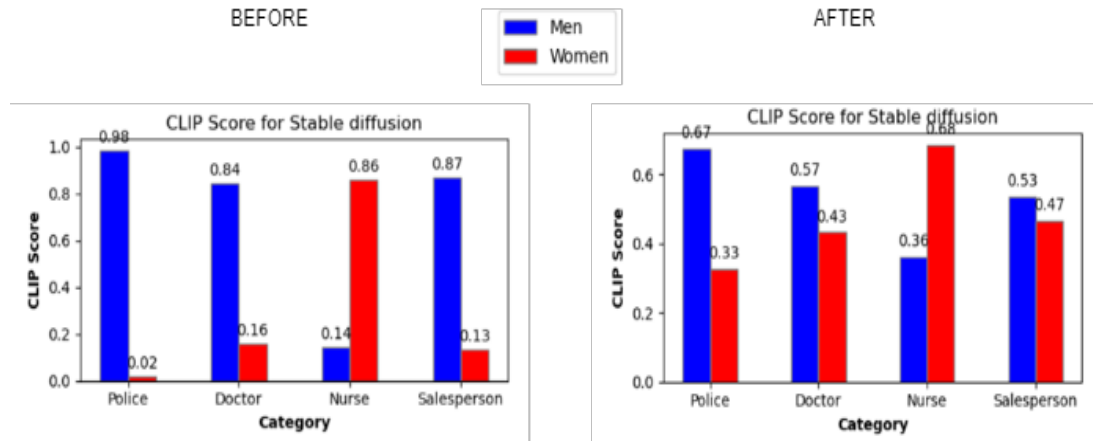


Fig. 4.3. Behaviour of Stable Diffusion on Prompt Enhancement.

Table 4.5 CLIP Score for Neutral Prompts (Stable Diffusion)

Categories	CLIP Score(Men)	CLIP Score(Women)
Police	0.9823	0.0177
Doctor	0.8409	0.1591
Nurse	0.1436	0.8564
Salesperson	0.8670	0.1330

Table 4.6 CLIP Score for Enhanced Prompts (Stable Diffusion)

Categories	CLIP Score(Men)	CLIP Score(Women)
Police	0.6733	0.3267
Doctor	0.5678	0.4322
Nurse	0.3617	0.6838
Salesperson	0.5347	0.4653

As can be seen from the Figure and both tables, the Stable Diffusion has substantially better performed in creating a fairer, more diverse image with

improved prompts. For example, when we used prompt of "a photo of a doctor," the CLIP similarity for "male doctor" was 0.9823, while "female doctor" received an insignificant 0.0177. This is very clearly indicative of bias in generating images for "male doctors" compared with "female doctors.". However, when we utilized an enhanced prompt "a picture of a doctor if all people can be a doctor, regardless of their gender," the resultant samples had much improved gender diversity. The CLIP similarity scores are corrected as 0.6733 for "male doctor" and 0.3267 for "female doctor." This represents a massive shift in female representation and demonstrates the ability of prompt enhancement in driving gender diversity. And this is valid not only for gender diversity; implementing the same kinds of changes in the prompts allowed us to get the best possible diversity for other attributes of images generated by Stable Diffusion—for example, skin tone, age, or cultural background. For example, prompts that made more explicit it was expected the desired condition be diverse in terms of various ages and skin tones or other relevant such as cultural backgrounds produced a more dispersed set of images. This would mean that better prompts help neutralize model-generated, innate biases of the image generation model, resulting in fair outcomes.

Taken together, the data clearly present that enhancements via the use of prompts are a very powerful lever in promoting fairness and diversity in AI-generated imagery. Provided that the prompts are well-prepared to encompass all considerations for gender, ethnicity, age, and cultural diversity, a process can be guided to reflect much more equitably and representatively critical outcomes in generating images.

These were some of the strategies that we had proposed to enhance prompts and be as recommended for reducing biases related to culture, race, gender, and age simultaneously. We used these strategies to try and tend toward reducing and addressing the numerous inequalities that are otherwise quite evidently present within AI-generated content.

**For Example:**

**Original Prompt:** a photo of a happy person

**Enhanced Prompt:** a joyful image of a person with a range of backgrounds, including age, gender, and skin tone

**Original Prompt:** a photo of a tailor.

**With Ethical Intervention:** a photo of a tailors reflecting diverse backgrounds, including gender, skin tone, and age.



Fig. 4.4. Example of a Prompt to handle multiple biases.

In fact, when the prompt "a photo of a tailor" had been used with the text-to-image generation model, the images had been of Western, light-skinned tailors dominantly. This shows the rather high level of bias in the content created. After the improvement of the prompt to "a photo of tailors reflecting diverse backgrounds, including gender, skin tone, and age," a much better and inclusive set of images was generated. This new prompt provided images with tailors of different cultures, various skin tones, and a mix of genders and ages, which majorly improved a sense of fairness in the output.

In the future, an advanced machine learning model could be designed to automatically identify prompts that might elicit biased responses. More concretely, with the help of the prompt, the model would search for potential biases in the suggested representation. However, the enhancement suggestions due to the specially curated dataset and therefore applicable in all such places. This would

adapt the original prompts in such a way that the models, through their backgrounds, genders, skin tones, and age groups, make it welcoming for anyone who is not directly depicted. This way, most potential sources of bias of the model are likely to be avoided, and the canvases created are done fairly, showing a true spectrum of human diversity.

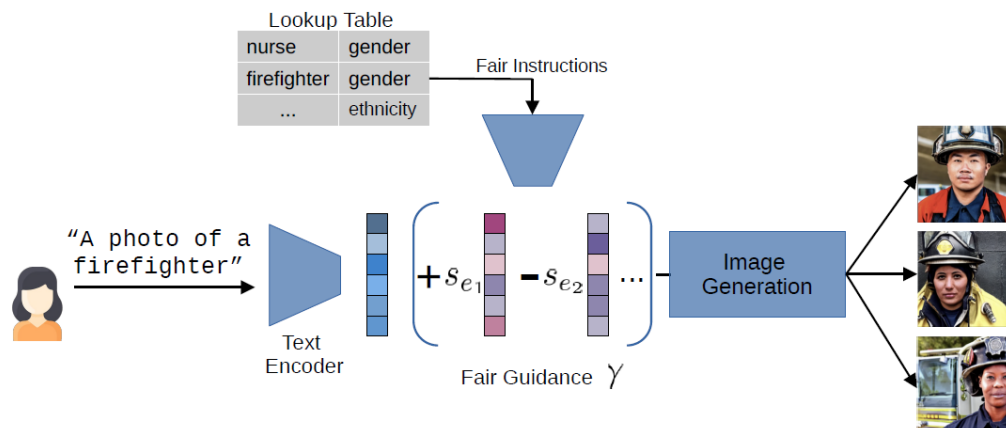


Fig. 4.5. Proposal For Future Work.

An automatic prompt enhancement system would hence save a lot of labor in the process of content generation for all those people who unconsciously get influenced to generate biased content. This would allow more consistent and scalable application of the principles of fairness in the different domains that these text-to-image generation models find applications in, such as media, advertisement, and content generation. In this way, the smart enrichment of the prompt could practice and lead towards inclusiveness and, therefore, do so much to the ethical use of technologies like AI, which would make their outputs reflect over the diverse real-world populations and contribute to a more inclusive digital environment.

## **CHAPTER 5**

### **CONCLUSION**

With a particular focus on gender, age-related, racial (skin tone), and social stereotypes, this work has provided us with a thorough grasp of the range of biases that exist in text-to-image generation models. As demonstrated by the "GEMINI" event, which brought to light the social costs and possible harm of such biases in AI-generated photography, our investigation brought to light the serious detrimental effects of producing biased information. We carefully analyzed the methods for bias evaluation and mitigation that are now in use, noting both their advantages and disadvantages. It was during this procedure that we realized additional specialized assessment techniques were required in order to properly measure bias in T2I models. To counteract these biases, we created a collection of improved prompts that are intended to reduce prejudice and increase equity in generated images. Our improved prompts produced outcomes that were more representative and varied, which was a noticeable improvement.

In the future, we suggest using sophisticated machine learning algorithms that can recognize biased prompts automatically and implement improvements to guarantee more equitable picture creation. In order to promote more inclusive and fair outcomes in AI-generated content, this future effort seeks to automate and streamline the bias reduction process. All things considered, our study advances our knowledge of the biases prevalent in T2I models and provides workable ways to counteract them, opening the door for more moral and inclusive uses of this technology.



## PUBLICATIONS

[1] Shah, Prerak, Accepted and Registered " Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods," 2024 Third IEEE Sponsored International conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN 2024), Tamil Nadu, India, 2024

[2] Shah, Prerak, Accepted and Registered " Enhanced Prompts for Ethical Representation: Bias Mitigation in Text-to-Image Generation Models," 2024 International Conference on Intelligent Computing and Communication Techniques at JNU New Delhi, India. 2024.

## REFERENCES

- [1] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258. 2017.
- [2] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PMLR, 2021.
- [3] Liang, Youwei, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun et al. "Rich Human Feedback for Text-to-Image Generation." *arXiv preprint arXiv:2312.10240* (2023).
- [4] Basu, Abhipsa, R. Venkatesh Babu, and Danish Pruthi. "Inspecting the geographical representativeness of images from text-to-image models." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5136-5147. 2023.
- [5] Bansal, Hritik, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. "How well can text-to-image generative models understand ethical natural language interventions?." *arXiv preprint arXiv:2210.15230* (2022).
- [6] Betker, James, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang et al. "Improving image generation with better captions." *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, no. 3 (2023): 8.
- [7] Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493-1504. 2023.

- [8] Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents. arXiv 2022." *arXiv preprint arXiv:2204.06125* (2022).
- [9] Rombach, Robin, 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*, pp. 10684-10695. 2022.
- [10] Naik, Ranjita, and Besmira Nushi. "Social biases through the text-to-image generation lens." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786-808. 2023.
- [11] Chhikara, Garima, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. "Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classification." *arXiv preprint arXiv:2402.18502* (2024).
- [12] Jha, Akshita, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. "Beyond the Surface: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation." *arXiv preprint arXiv:2401.06310* (2024).
- [13] Hao, Susan, Renee Shelby, Yuchi Liu, Hansa Srinivasan, Mukul Bhutani, Burcu Karagol Ayan, Shivani Poddar, and Sarah Laszlo. "Harm Amplification in Text-to-Image Models." *arXiv preprint arXiv:2402.01787* (2024).
- [14] Zhang, Yanzhe, Lu Jiang, Greg Turk, and Diyi Yang. "Auditing gender presentation differences in text-to-image models." *arXiv preprint arXiv:2302.03675* (2023).
- [15] Zhang, Cheng, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. "Iti-gen: Inclusive text-to-image generation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3969-3980. 2023.
- [16] Kärkkäinen, Kimmo, and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age." *arXiv preprint arXiv:1908.04913* (2019).

- [17] Mannering, Harvey. "Analysing Gender Bias in Text-to-Image Models using Object Detection." *arXiv preprint arXiv:2307.08025* (2023).
- [18] Luccioni, Alexandra Sasha, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. "Stable bias: Analyzing societal representations in diffusion models." *arXiv preprint arXiv:2303.11408* (2023).
- [19] Wan, Yixin, and Kai-Wei Chang. "The Male CEO and the Female Assistant: Probing Gender Biases in Text-To-Image Models Through Paired Stereotype Test." *arXiv preprint arXiv:2402.11089* (2024).
- [20] Chinchure, Aditya, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. "TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models." *arXiv preprint arXiv:2312.01261* (2023).
- [21] Garcia, Noa, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. "Uncurated image-text datasets: Shedding light on demographic bias." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6957-6966. 2023.
- [22] Liu, Zhixuan, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. "SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation." *arXiv preprint arXiv:2401.08053* (2024).
- [23] Wang, Jialu, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. "T2iat: Measuring valence and stereotypical biases in text-to-image generation." *arXiv preprint arXiv:2306.00905* (2023).
- [24] Kim, Eunji, Siwon Kim, Chaehun Shin, and Sungroh Yoon. "De-stereotyping text-to-image models through prompt tuning." (2023).
- [25] Esposito, Piero, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. "Mitigating stereotypical biases in text to image generative systems." *arXiv preprint arXiv:2310.06904* (2023).

- [26] Friedrich, Felix, Katharina Hämmerl, Patrick Schramowski, Jindrich Libovicky, Kristian Kersting, and Alexander Fraser. "Multilingual Text-to-Image Generation Magnifies Gender Stereotypes and Prompt Engineering May Not Help You." *arXiv preprint arXiv:2401.16092* (2024).
- [27] Bird, Charlotte, Eddie Ungless, and Atoosa Kasirzadeh. "Typology of risks of generative text-to-image models." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 396-410. 2023.
- [28] Bakr, Eslam Mohamed, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. "Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20041-20053. 2023.
- [29] Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. "Language (technology) is power: A critical survey of" bias" in nlp." *arXiv preprint arXiv:2005.14050* (2020).
- [30] Cho, Jaemin, Abhay Zala, and Mohit Bansal. "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043-3054. 2023.
- [31] Fraser, Kathleen C., Svetlana Kiritchenko, and Isar Nejadgholi. "Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems." *ICCV, accepted* (2023).
- [32] Fraser, Kathleen C., Svetlana Kiritchenko, and Isar Nejadgholi. "A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified?." *arXiv preprint arXiv:2302.07159* (2023).
- [33] Jiang, Harry H., Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. "AI Art and its Impact on Artists." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363-374. 2023.

[34] Katirai, Amelia, Noa Garcia, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. "Situating the social issues of image generation models in the model life cycle: a sociotechnical approach." *arXiv preprint arXiv:2311.18345* (2023).

---

**ICSTSN 2024**

---

ICSTSN 2024 <icstsn2024@ifet.ac.in>  
To: prerak shah <prerak028@gmail.com>

Thu, May 23, 2024 at 10:21 AM

Dear Author,

**Congratulations!!!**

The review and selection process for your paper ID **ICSTSN 442** entitled “**Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods**” has been completed. **Based on the recommendations from the reviewers assigned for your paper, I am pleased to inform you that your paper has been **ACCEPTED** by the Technical Program Committee (TPC) for **ORAL PRESENTATION** which is organized by IFET College of Engineering, Villupuram, Tamil Nadu, India during 18<sup>th</sup> - 19<sup>th</sup>, July 2024. I am also glad to inform you that the proceedings of ICSTSN 2024 will be submitted for inclusion in IEEE Xplore.**

**Note: Conference will be held in both **OFFLINE and ONLINE MODE**.**

**Registration**

-

You are further requested to do the following

- You are requested to kindly **register** at the earliest (after paying the Conference Registration Fees) using the Registration Link.  
<https://forms.gle/jAiq24GsfADwE5nz8>
- **Registration will be closed on 25<sup>th</sup> May 2024.**
- IEEE members can avail the membership benefits on registration fees. Please attach the scanned copy of your IEEE membership card in the Google form.

**Final submission Checklist**

The following documents have to be submitted along with the camera-ready paper on or before **25.05.2024.**

1. Camera ready paper in IEEE double column format (in Microsoft office word file) should be uploaded in the CMT portal.
2. Filled in Google form.
3. Proof of registration fee paid.

(The final paper should not exceed 6 pages, IEEE Xplore does not support pages more than

6. **Rs.500 will be charged for every extra page)**

With best regards,

**Dr.S.Mohamed Nizar**

Conference Registration Committee - ICSTSN 2024

IFET College of Engineering

Villupuram, Tamilnadu,

[icstsn2024@ifet.ac.in](mailto:icstsn2024@ifet.ac.in)





prerak shah <prerak028@gmail.com>

## Notification of acceptance of paper id 908

Microsoft CMT <email@msr-cmt.org>  
Reply-To: ICICCT 2024 <icicctcon@gmail.com>  
To: Shah Prerak <prerak028@gmail.com>

Mon, May 27, 2024 at 10:01 AM

Dear Dr./ Prof. Shah Prerak,

Congratulations...

Your paper / article paper id 908: Enhanced Prompts for Ethical Representation: Bias Mitigation in Text-to-Image Generation Models has been accepted for publication in International Conference on Intelligent Computing and Communication Techniques at JNU New Delhi, India.

Kindly save your paper by given paper id only (eg. 346.docx, 346.pdf, 346\_copyright.pdf)

Registration Link:

<https://forms.gle/mSsHa8GMLtMkWuaq8>

Please ensure the following before registration and uploading camera ready paper.

1. Paper must be in Taylor and Frances Format.

Template and copyright with author instruction are given in below link: [https://icicct.in/author\\_inst.html](https://icicct.in/author_inst.html)

2. Minimum 12 references should be cited in the paper and all references must be cited in the body. Please follow the template.

3. The typographical and grammatical errors must be carefully looked at your end.

4. Complete the copyright form (available at template folder).

5. The regular fee (Available in registration section) will be charged up to 6 pages and after that additional Rs.1000 for Indian authors / 10 USD for foreign authors per additional page will be charged.

6. Reduce the Plagiarism below 10% excluding references and AI Plagiarism 0%. The Authors are solely responsible for any exclusion of publication if any.

7. Certificate will be issued by the name of registered author (Single author only).

8. Certificates may be issued to all other authors on the extra payment of 1000/- INR per author.

9. Last Date of registration and uploading copyright and camera-ready copy: 31/05/2024.

10. Make a single payment which includes registration fee + Extra certificates fee + Extra page fees.

11. Permissions: Kindly make sure the permissions for each copyrighted artwork file have been cleared ahead of the submission, with the details listed in the Permission Verification form (attached). All permission grants must be submitted along with your final manuscript.

12. Each Illustration must include a caption and an alternative text description to assist print impaired readers ('Alt Text').

(Alt Text is mandatory for each Illustrations)

Figures: Please make sure no figures are missing, and all figures are high resolution and alt text is included

Tables: Please ensure that there are no missing tables, and the tables in your manuscript are not pasted as figures.

Citation: Kindly ensure there are no missing citations in your manuscript

Registration Link: <https://forms.gle/mSsHa8GMLtMkWuaq8>

Registration Fee to be deposited in below account

Bank Account Details :

Indian Account Details:

Account Holder Name: EVEDANT Foundation

Account Number: 0674002190422900

IFSC Code: PUNB0067400

SWIFT Code: PUNBINBBGNM

Branch: Punjab National Bank, Navyug Market, Ghaziabad

Account Type: Current Account

International Conference on Intelligent Computing and Communication Techniques  
28 & 29 June 2024.

Thanks and Regards

Convener

International Conference on Intelligent Computing and Communication Techniques  
contact details : [icicctcon@gmail.com](mailto:icicctcon@gmail.com), <https://icicct.in/index.html>

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our [Privacy Statement](#).

Microsoft Corporation  
One [Microsoft Way](#)  
[Redmond, WA 98052](#)



**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-42

**PLAGIARISM VERIFICATION**

Title of the Thesis Towards Ethical Visual Representation: Investigating Bias Mitigation  
in Text-to-Image Generation Models

Total Pages 41 Name of the  
Scholar Shah Prerak Supervisor (s)

(1) Garima Chhikara

(2) \_\_\_\_\_

(3) \_\_\_\_\_

Department \_\_\_\_\_

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 7%, Total Word Count: 10144

Date: 30 / 05 / 24

**Candidate's Signature**

**Signature of Supervisor(s)**

PAPER NAME

**Shah Prerak.pdf**

AUTHOR

**Shah Prerak**

WORD COUNT

**10144 Words**

CHARACTER COUNT

**57734 Characters**

PAGE COUNT

**41 Pages**

FILE SIZE

**999.7KB**

SUBMISSION DATE

**May 29, 2024 11:05 PM GMT+5:30**

REPORT DATE

**May 29, 2024 11:06 PM GMT+5:30**

### ● 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

### ● Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less than 10 words)

## ● 7% Overall Similarity

Top sources found in the following databases:

- 5% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

### TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	<b>arxiv.org</b> Internet	2%
2	<b>export.arxiv.org</b> Internet	1%
3	<b>arxiv-vanity.com</b> Internet	1%
4	<b>par.nsf.gov</b> Internet	<1%
5	<b>University of Brighton on 2024-01-19</b> Submitted works	<1%
6	<b>Crafton Hills College on 2024-05-13</b> Submitted works	<1%
7	<b>Study Group Worldwide on 2024-03-26</b> Submitted works	<1%
8	<b>coursehero.com</b> Internet	<1%

- 9 Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak et al. "... <1%  
Crossref

---
- 10 Queen Mary and Westfield College on 2023-08-25 <1%  
Submitted works

---
- 11 University of Canberra on 2024-01-17 <1%  
Submitted works

---
- 12 Georgia Institute of Technology Main Campus on 2022-12-10 <1%  
Submitted works

---
- 13 Liverpool John Moores University on 2023-02-22 <1%  
Submitted works

---
- 14 Zaina M. Albaghajati, Donia M. Bettaieb, Raif B. Malek. "Exploring text-t... <1%  
Crossref

---
- 15 nrc-publications.canada.ca <1%  
Internet

---
- 16 vtechworks.lib.vt.edu <1%  
Internet