Can Prompt Modifiers Control Bias? A Comparative Analysis of Text-to-Image Generative Models

Philip Wootaek Shin* , **Jihyun Janice Ahn*** , **Wengpeng Yin** , **Jack Sampson** and **Vijaykrishnan Narayanan**

The Pennsylvania State University

{pws5345, jfa5672, wenpeng, jms1257, vxn9}@psu.edu

Abstract

It has been shown that many generative models inherit and amplify societal biases. To date, there is no uniform/systematic agreed standard to control/adjust for these biases. This study examines the presence and manipulation of societal biases in leading text-to-image models: Stable Diffusion, DALL·E 3, and Adobe Firefly. Through a comprehensive analysis combining base prompts with modifiers and their sequencing, we uncover the nuanced ways these AI technologies encode biases across gender, race, geography, and region/culture. Our findings reveal the challenges and potential of prompt engineering in controlling biases, highlighting the critical need for ethical AI development promoting diversity and inclusivity.

This work advances AI ethics by not only revealing the nuanced dynamics of bias in text-to-image generation models but also by offering a novel framework for future research in controlling bias. Our contributions—spanning comparative analyses, the strategic use of prompt modifiers, the exploration of prompt sequencing effects, and the introduction of a bias sensitivity taxonomy—lay the groundwork for the development of common metrics and standard analyses for evaluating whether and how future AI models exhibit and respond to requests to adjust for inherent biases.

1 Introduction

Within the dynamic realm of artificial intelligence, the advent of text-to-image generation models [Li *et al.*, 2023; Yang *et al.*, 2023; Avrahami *et al.*, 2023] marks a significant leap forward. Leveraging deep learning, these models convert text descriptions into detailed images, captivating users and pioneering new avenues in artistic creation, design, and communication [Brooks *et al.*, 2023; Couairon *et al.*, 2023]. These models, powered by vast datasets [Schuhmann *et al.*, 2022] and advanced algorithms [Ho *et al.*, 2020; Sohl-Dickstein *et al.*, 2015; Song *et al.*, 2021], promise a new era of creativity and efficiency. However, with great power comes great responsibility, particularly in ensuring that these innovations do not perpetuate or amplify societal biases [Naik and Nushi, 2023].

Unfortunately, initial observations highlight a significant variance in the depiction of culturally and geographically nuanced concepts within existing text-to-image models. Consider, for instance the archetype of the "monk," traditionally associated with Asian cultures and male roles: A preliminary analysis of image outputs for a generic "monk" prompt across various models unveils a marked inclination towards representing monks as Asian males, as detailed in Tab. 1. This tendency, while possibly reflective of historical accuracies, prompts scrutiny over the data and algorithms that inform these models, particularly in how they navigate cultural and gender biases. Interestingly, the Firefly (FF) model showcases a notably more balanced gender and racial representation, indicating a distinct internal approach to bias attenuation.

Model	Male / Female	Asian / Others	Total Samples
SD	50 / 0	50 / 0	50
DallE	36 / 0	35 / 1	36
FF	28 / 24	5 / 47	52

Table 1: Distribution of Gender and Race for "Monk" Prompt

Model	Asian	Black	Others	Total Samples
SD	50	0	0	50
DallE	35	3	15	53
FF	14	26	12	52

Table 2: Distribution of Race for "Monk Who is Black" Prompt

The complexity of this issue deepens when examining the models' responses to compound prompts aimed at eliciting non-traditional representations, such as a "Monk who is black," shown in Tab. 2. Notably, despite explicit instructions, Stable Diffusion (SD) and Dall·E 3 (DallE/DE) continued to predominantly produce imagery tied to Asian cultural markers, highlighting **a proclivity to default to historical and cultural stereotypes over direct prompt cues**. The divergent responses to these prompts, particularly Firefly's shift towards equitable representation, spotlight the nuanced challenge of bias within AI systems. Such variance raises pivotal questions about the objective of these models in reflecting the diversity of human experience. Should they aim to accurately mirror historical and sociodemographic realities, or aspire towards an idealized inclusivity that may diverge from factual representation? While Firefly's inclusive approach is laudable, it ignites debate on the validity of achieving balance at the potential expense of demographic authenticity.

Motivated by these observations, this study aims to dissect and understand the bias embedded within these AI technologies. It undertakes a thorough analysis of bias across three forefront text-to-image models: Stable Diffusion [Rombach *et al.*, 2022], OpenAI's DALL·E 3 [Betker *et al.*, 2023], and Adobe Firefly [Adobe Systems Incorporated, 2023]. Our structured examination employs singular prompts to compare and contrast biases and statistical variations within these models. We navigate this research through three critical phases. Initially, we perform an analysis of each model using standardized prompts to identify biases related to gender, race, geography, and religion/culture, providing a baseline for bias assessment. Subsequently, we investigate the use of "modifiers" in prompts, integrating various bias aspects into a singular prompt to see if biases can be mitigated. This exploration into "Base Prompt + Modifier" configurations reveals the potential of prompt engineering to create more equitable AI applications. Lastly, we assess the impact of prompt sequencing—whether placing the modifier before or after the base prompt affects image generation—suggesting that even minor adjustments in prompt structure can significantly alter outcomes, thereby illustrating the complex dynamics of bias within text-to-image models.

By examining gender, race, geography, and religion/culture biases with the aid of base prompts and modifiers, this study aims to deepen the understanding of bias in AI. Through comparative analysis, we illuminate each model's specific biases and underscore the role of prompt engineering in bias reduction. Specifically, the paper highlights:

- **Prompt Modifiers as a Tool for Bias Adjustment**: We introduce the use of prompt modifiers as a means of adjusting bias within image generation models. Importantly, our experiments with this form of prompt engineering do not yield uniform results, highlighting the fundamental nature of this challenge and the need for more complex strategies.
- **Demonstration of Control-resistant Biases**: While prompt engineering may seem to be a direct and nearly trivial fix for overcoming model biases, we demonstrate both several examples of inherent biases that are not overcome by adding prompt modifiers and several more where the behavior with respect to modifer addition is fragile (i.e. sensitive to ordering).
- **Impact of Prompt Sequencing on Bias Control**: By analyzing how the sequence of base prompts and modifiers influences image generation, we highlight the importance of prompt structure in bias control within AI-driven processes.
- **Introduction of a Taxonomy and Validation Method**: We introduce a taxonomy to gauge models' sensitivity to prompt engineering and validate this approach through a quantitative metric of distributional shift, based on modifier application. Providing this structure enhances our understanding of bias control mechanisms in AI models and yields a framework for future characterizations and cross-comparisons in measuring both bias and attempts at its adjustment in AI models.
- **Broad Comparative Analysis Across Multiple Models and Bias Categories**: Our investigation expands on the scope of prior work by providing a comparative analysis of four bias categories over three leading text-to-image generation models: Stable Diffusion, DALL·E 3, and Firefly, and their entanglement with LLMs via prompt processing.

2 Related Work

A growing body of scholarly work has begun to explore the various dimensions of bias present in these models, providing a foundation for the comparative analysis we undertake in this study. The summary of the bias categories and the corresponding models examined in the related literature is presented Tab. 3

2.1 Biases in Text-to-Image Model

Significant strides in understanding these biases were made by the DALL·Eval project [Cho *et al.*, 2023], which introduced a diagnostic dataset to assess visual reasoning in AI and pinpoint gender and skin tone biases. The research conducted by Seshadri *et al.* [Seshadri *et al.*, 2023] shifts the lens towards the amplification of gender-occupation biases within Stable Diffusion, advocating for a thoughtful consideration of how biases are evaluated, particularly in relation to the discrepancies between training datasets and generated outputs. Struppek *et al.* [Struppek *et al.*, 2023] delve into the inadvertent reflection of cultural biases by models trained on diverse internet-sourced image-text pairs. In the realm of ethical AI development, Fair Diffusion[Friedrich *et al.*, 2023] charts a course towards fairness, spotlighting the gender and racial biases prevalent in the training data of Stable Diffusion. Lastly, Naik *et al.* [Naik and Nushi, 2023] provide a thorough evaluation of biases across DALL·E 2 and Stable Diffusion v1, utilizing both human judgment and algorithmic assessments. Oppenlaneder *et al.* [Oppenlaender, 2023] explored modifiers to enhance the style and quality of generated images, yet did not examine how modifiers affect the distribution shift of bias.

Building on these insights, our investigation seeks to further elucidate the biases embedded within the leading text-to-image generation models. As shown in Tab. 3, our analysis spanning gender, race, geography, and religion/culture biases across multiple models covers a superset of the interactions covered by prior works. By investigating the use of uniform and modified prompts in effecting specific desired output distribiutions we aim to enrich the discourse on AI ethics and creativity with respect to controlling biases as well as quantifying their presence.

2.2 Biases in Large Language Model

In the rapidly evolving domain of artificial intelligence, significant strides have been made not only in text-to-image

Prior Work	Bias Category				Model Used			
	Gender	Race	Geography	Cultural/Religion	SD	DallE	FireFly	LLM
Cho *et al.*[Cho *et al.*, 2023]	✓	✓			✓	✓		
Seshadri *et al.*[Seshadri *et al.*, 2023]	✓				✓			
Struppek *et al.*[Struppek *et al.*, 2023]		✓	✓	✓	✓	✓		
Friedrich *et al.*[Friedrich *et al.*, 2023]	✓	✓			✓			
Naik *et al.*[Naik and Nushi, 2023]	✓	✓	✓		✓	✓		
Dong *et al.*[Dong *et al.*, 2024]	✓							✓
Yeh *et al.*[Yeh *et al.*, 2023]	✓	✓		✓				✓
Our Paper	✓	✓	✓	✓	✓	✓	✓	✓

Table 3: Summary of biases and models used in related works for LLMs and Text-to-Image Generation Models(SD,DallE, FireFly)

Base	**Bias Type**	**SD**	**DallE**	**Firefly**
Nurse	Gender(M/F)	0/50	9/71	20/32
	Race	-		
	Geography	-		
	Culture/Religion	-		
Seasons in January	Gender	-		
	Race	-		
	Geography(S/W)	0/50	19/21	0/52
	Culture/Religion	-		

Table 4: Comparative Bias Analysis Across Text-to-Image Generation Models. M/F represent Male/Female and S/W represent Summer/Winter. "-" indicate the field which is not applicable

generation technologies but also in the realm of large language models (LLMs). Dong *et al.* [Dong *et al.*, 2024] shed light on the gender biases present in LLMs, even in the absence of explicitly biased inputs, questioning the realism of template-based probes for bias assessment. Yeh *et al.* [Yeh *et al.*, 2023] examine the impact of contextually rich inputs on LLM behavior, demonstrating that the lack of detailed auxiliary information in ambiguous contexts can hinder the generation of unbiased and precise responses. The Rainbow Teaming [Samvelyan *et al.*, 2024] approach employs numerous mutators to create adversarial prompts, focusing primarily on simulating criminal planning and role-playing rather than investigating the generation of visual content.

Collectively, this body of work highlights the pervasive nature of bias in AI and emphasizes the necessity for holistic strategies to confront and rectify these biases. The shared challenges and solutions identified in LLM research are invaluable to our comparative study on text-to-image models, reinforcing the vital part of advanced prompt engineering.

3 Bias Evaluation

Tab. 4 provides an illuminating snapshot of the complexities involved in mitigating biases across various categories within text-to-image generation models. Turning to the 'Season in January' category, a notable distinction arises in the geographical representation of seasons. Stable Diffusion and Firefly revealed a Northern Hemisphere winter bias, which inadvertently reflects the demographic and climatic realities of more than 85% of the global population residing in the Northern Hemisphere. Conversely, DallE showcased a more balanced depiction of both summer and winter scenes, thus acknowledging the seasonal contrasts between hemispheres.

This balance raises an intriguing question regarding the role of AI in mirroring versus moderating real-world disparities. While DallE's balanced output may seem fair and inclusive at face value, it may also inadvertently gloss over the demographic predominance of the Northern Hemisphere, suggesting that a truly balanced AI model must navigate the fine line between representational fairness and demographic fidelity. These contrasting approaches underscore the complexity of bias in AI, where the pursuit of balance must be carefully weighed against the representation of statistical realities, such as the population distribution across hemispheres, which directly impacts the prevalence of seasonal experiences worldwide. These findings compel a deeper consideration of how text-to-image models encapsulate and convey societal norms and raise fundamental questions about the benchmarks for unbiased AI representations.

In examining the presence of biases across the specified categories, it becomes evident that not all bias types manifest uniformly or are even applicable to each category. This is reflective of the nuanced reality that certain societal constructs and roles carry specific historical and cultural biases [Buolamwini and Gebru, 2018], while others may be more universally recognized and less prone to subjective bias [Noble, 2018]. To anchor our investigation in empirical rigor, we have leveraged prior scholarly work and widely acknowledged consensuses to establish our base prompts and categories that have historically exhibited strong biases [Barocas *et al.*, 2019]. These informed baselines serve as a critical reference point for assessing whether the models merely replicate known biases [Mehrabi *et al.*, 2021] or whether they have the capacity to transcend these limitations [Mitchell *et al.*, 2019], potentially yielding a more diverse range of outputs as required by the user.

For instance, the nurse category across Stable Diffusion, DallE, and Firefly did not display any overt racial biases, as the models generated diverse racial representations in the absence of a clear skew towards any particular group, but did exhibit gender skew. The lack of overt racial biases could be seen as a positive step toward unbiased AI, reflecting an equitable cross-section of racial identities in the nursing profession. Cultural and geographical factors were similarly nondescript, indicating that these models may not strongly encode or perpetuate biases along these dimensions within the scope of the tested prompts. However, the gender bias observed,

BasePrompt (TOPIC)	Bias	BasePrompt Category	Modifier	Modifier Category
Male Housekeeper	Asian	Gender	Black	Race
Korean Soldier	Male	Geography	Female	Gender
American football player	Male	Geography	Female	Gender
Gay couple	White	Gender	Asian	Race
Lesbian couple	White	Gender	Black	Race
US Politician	White	Geography	Asian	Race
Person in Africa	Black	Geography	Asian	Race
Person in East Asia	East Asian	Geography	Hispanic	Race
Pastor	Male/White	Religion/Culture	Female/Asian	Gender
Monk	Asian/Male	Religion/Culture	White/Female	Race
Tanning Man	Male	Religion/Culture	Black	Race
Lunar New Year	Chinese	Religion/Culture	US/Hispanic	Geography
Vegan	NonEastAsian	Religion/Culture	Korean	Geography

Table 5: Base prompt that we generated to conduct study for different text to image model

with a skew towards female representations, resonates with societal associations of the nursing profession. Firefly's more balanced gender output, intimates the potential for mitigating such biases, although it also prompts further scrutiny into the methods and training data employed for such counter-bias modeling efforts: As demonstrated in Sec. 5, the opacity of counter-bias modeling can impact the ability to understand and alter distributional outcomes via prompt engineering.

4 Methodology

In our experimental setup, we engaged three distinct models—Stable Diffusion, DallE, and Firefly—to create images from a set of base prompts, aiming to uncover any inherent biases. With Stable Diffusion, we generated a suite of 50 unique images for each prompt to ensure a robust sample size. In the case of Firefly, we leveraged its functionality to differentiate between real and stylized characters, opting for the generation of real-person images. For each prompt, Firefly produced images of four distinct individuals, culminating in a total of 52 images per prompt. Meanwhile, our use of DallE was facilitated through the ChatGPT4 interface, which serves as a gateway to the DallE image generation backend. Due to operational constraints for ChatGPT, we were limited to crafting 40 prompts every three hours. To circumvent this and maximize output, we utilized compound prompts requesting the creation of images in a grid format, specifically instructing the model to "generate `A` with 3 rows and 3 columns" where `A` is a prompt of interest. While there was no strict limit on the number of images generated, we aimed for upwards of 30 images per prompt to ensure a statistically significant sample that could provide a meaningful analysis of distribution trends across the models.

In our study, we employed 16 distinct base prompts, intentionally chosen to span the breadth of biases commonly associated with gender, geography, religion/culture, and race. These categories, as detailed in Tab. 5 and discussed in Sec. 3 , do not encompass the entire scope of possible biases, yet they offer a representative cross-section of biases that are visually identifiable within the images produced by the models. A comprehensive list of the base prompts utilized for this study is available in the supplemental materials.

When these prompts were deployed across three distinct models—Stable Diffusion, DallE, and Firefly—we were able to detect certain biases that these base prompts seemed to induce in the model outputs. Delving deeper, our analysis involved the introduction of modifiers to these base prompts, which effectively altered the bias distribution observed initially. This modification approach not only provides a straightforward means of disrupting the detected biases but also opens up new avenues for understanding the dynamics of bias within AI-generated imagery. Moreover, we explored how the sequencing of these prompts and modifiers (either 'Base + Modifier' or 'Modifier + Base') might impact the models' image generation, probing the influence of prompt structure on the visual representation of societal categories.

5 Results

In Fig. 1, we illustrate the outputs generated by the three models using the base prompt "US Politician" in conjunction with the modifier "Asian." The figure presents a side-by-side comparison of images produced from the base prompt alone, followed by the combined prompt with the modifier preceding the base ("Modifier+Base"), and finally, the base prompt followed by the modifier ("Base+Modifier"). This structured comparison across the three different models offers insights into the influence of prompt structure on the distribution of image generation.

Through a comparative analysis of images generated by each model, we identified distinct characteristics inherent to each image generation algorithm. Fig. 2 shows one example of the generated image by each model:

- **Stable Diffusion**: This model frequently produced images of lower resolution. Particularly for underrepresented subjects, such as a "Korean Soldier," the model predominantly generated images in black and white. When prompted without specific instructions, the emergence of bias was notably apparent. Moreover, in instances involving sensitive themes (e.g., "Tanning Man" or "Gay Couple"), the model defaulted to generating a black image should it deem the content sensitive.
- **DallE**: Of the models evaluated, DallE was most inclined to produce images that leaned towards the unrealistic. Similar to Stable Diffusion, bias was significantly apparent in basic prompts. For sensitive subjects (such as "Tanning Man," "Gay Couple," and "Lesbian Cou-

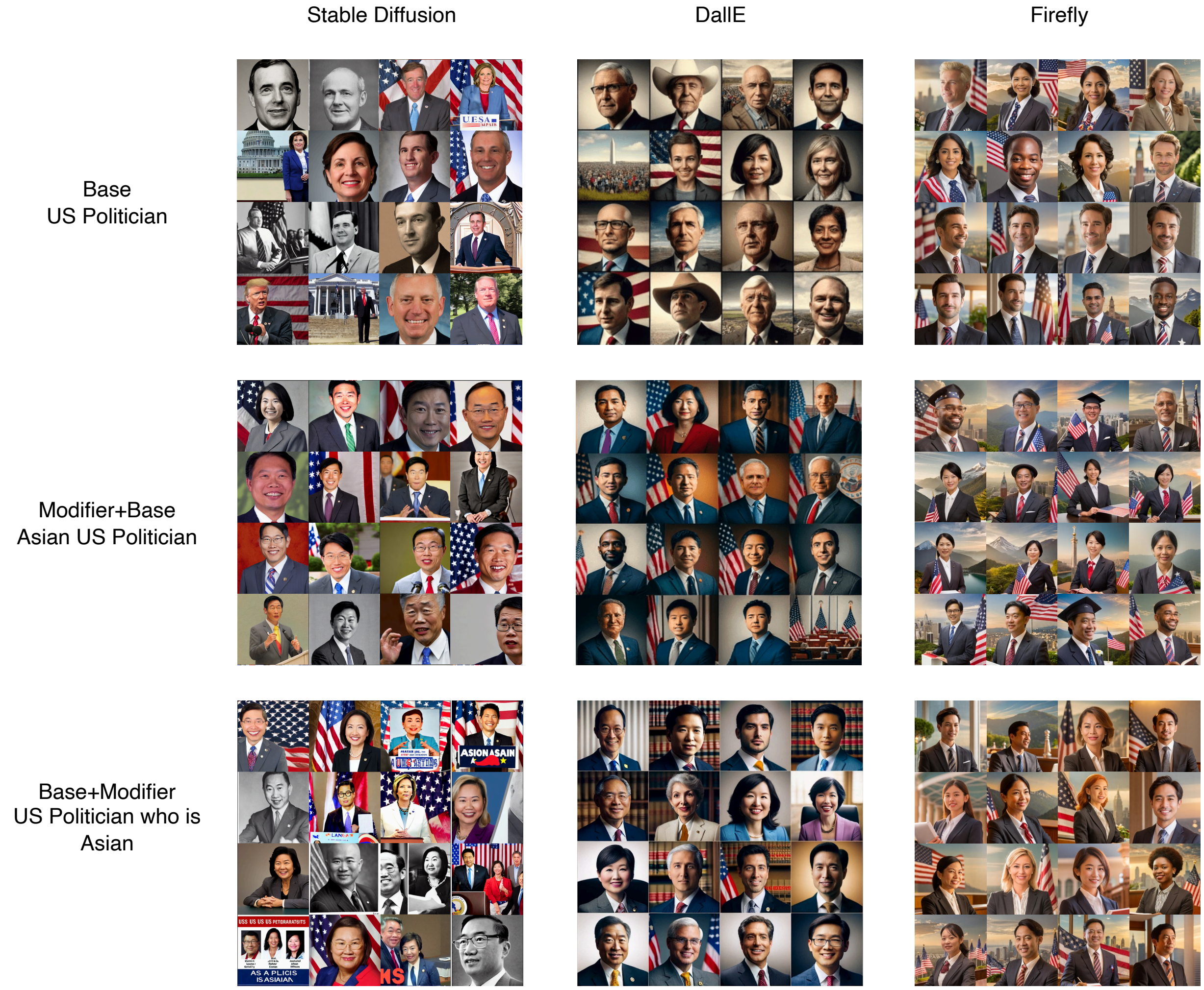

Figure 1: Example of images in different model. Note that we tried to maintain the percentage of Asian presented by our prompt

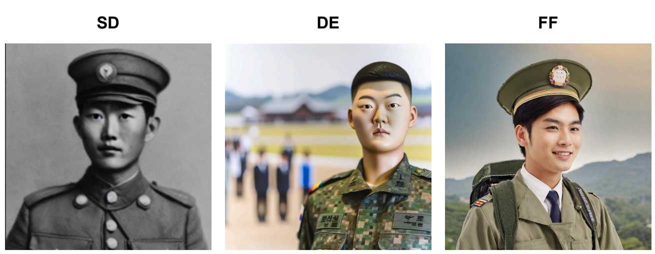

Figure 2: Example of images Generated by Stable Diffusion(SD), DallE(DE), Firefly(FF) with prompt "Korean Soldier"

ple"), it either abstained from generating images or produced representations more reminiscent of artistic drawings than realistic depictions.

- **Firefly**: This model was observed to generate the highest quality images, showcasing the least amount of bias when prompted without modifications. For instance, when analyzing the output of each model in generating images of U.S. Politicians (referenced in Fig. 1), Firefly displayed a commendable diversity in ethnicity and a balanced gender representation. However, it exhibited a strict refusal to generate content for topics even mildly sensitive, such as "Tanning Man."

In the investigation of our combined prompt experiment, results were consolidated in Tab. 6, focusing on the alteration in distribution from the base prompt when modified (denoted as "Change of Distribution (Yes/No)") and the impact of prompt sequencing on outcomes ("Order Matters (Yes/No)"). This analysis substantiated our hypothesis that incorporating a modifier within the prompt could mitigate the biases observed in base prompt scenarios. For ease of comprehensive visualization, the applicability of each model to the test scenarios is denoted using abbreviations and color codes.

In examining images generated from prompts specifying 'Asian,' we observed a predominance of East Asian imagery, sidelining the vast diversity within Asia, such as South Asian representations. This trend is evident in experiments like 'Asian US Politician,' highlighted in Fig. 1 Notably, Firefly exhibited a broader interpretation of 'Asian,' attempting to diversify beyond East Asian characteristics. This disparity un-

Triplet (Base, Modifier, Model)	Order Matters (Yes)	Order Matters (No)
Change of Distribution (Yes)	(Male Housekeeper, Black, FF) (Korean Soldier, Female, SD) (American football player, Woman, SD) (Gay couple, Asian, FF) (Lesbian couple, Black, FF) (US Politician, Asian, DE) (Person in Africa, Asian, SD) (Person in East Asia, Hispanic, SD FF) (Monk, Woman, FF) (Monk, Black, SD DE FF) (Lunar New Year, Hispanic, SD DE) (Vegan, Korean, FF)	(Male Housekeeper, Black,SD DE) (Korean Soldier, Female, DE FF) (American football player, Woman, DE FF) (Gay couple, Asian, SD DE) (Lesbian couple, Black, SD DE) (US Politician, Asian, SD FF) (Person in Africa, Asian, FF) (Pastor, Woman, SD DE FF) (Pastor, Asian, SD DE FF) (Monk, Woman, SD DE) (Tanning Man, Asian, SD DE) (Lunar New Year, Hispanic, FF) (Lunar New Year, US, SD DE FF) (Vegan, Korean, SD DE)
Change of Distribution (No)		(Person in Africa, Asian, DE) (Person in East Asia, Hispanic, DE)

Table 6: Analysis for change of distribution respect to order of prompt

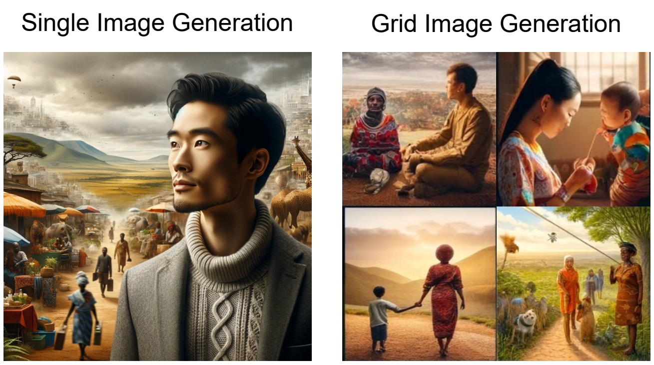

Figure 3: Example of images Generated by DallE with prompt "An Asian person living in Africa"

derscores the necessity for AI models to encompass a more comprehensive understanding of Asian diversity, reflecting the true range of cultures and identities within the continent.

For instance, the experiment employing the base prompt "US Politician" with the modifier "Asian" indicated a shift in the distribution of generated images across all three models. Interestingly, the sequence of the prompt notably influenced the results with DallE, whereas such an effect was not pronounced in the other models. Specifically, as depicted in Fig. 1, both Stable Diffusion and Firefly maintained a consistent proportion of images depicting Asians, irrespective of the prompt sequence. Conversely, DallE demonstrated a higher propensity to generate images of individuals from diverse ethnic backgrounds when the modifier "Asian" preceded the base prompt. This phenomenon, however, was relatively rare, with DallE's results being affected by prompt ordering in merely three out of twelve tested scenarios, including that involving US Politicians, contrasting with the more frequent influence observed in the other models.

A notable observation about DallE pertains to scenarios classified under "Change of Distribution (No)," such as (Person in Africa, Asian, DE) and (Person in East Asia, Hispanic, DE). These cases aimed to modify the distribution to favor images matching the modifier, thereby addressing the bias inherent in the base prompt. Despite this intent, the desired shift towards images corresponding to the modifiers was not achieved significantly in these instances, with DallE producing a substantial number of ambiguous images. Despite efforts to categorize these images, many were found too complex for clear ethnic identification. Yet, when generating images independently rather than in a grid, the model's outputs, though detailed, were more discernible in terms of racial representation. Fig. 3 shows an example of a generated image by DallE. In contrast, the other models favored simplicity, focusing on a singular, easily identifiable subject against a symbolic background, thereby aligning more closely with the expectations set by the base and modifier prompts. Given these observations, incorporating sample images for this analysis might be beneficial for clarity.

5.1 Quantitative Analysis

In this quantitative observation, we scrutinized the standard deviation across two prompt configurations ('Base+Modifier') and ('Modifier+Base') across three distinct models: Stable Diffusion (SD), DALL·E (DE), and Firefly (FF). With modified prompts, designed to specify and limit the distribution, the expected outcomes were predetermined.

Consider the example prompt "A Female American Football Player," where we anticipate that generated imagery conforming to the requested prompt will prominently feature a female figure, equating the expected outcome to a 100%/0%(F/M) gender distribution. Similar logic can apply to our other prompt+modifier pairs and their expected outcomes. Utilizing our dataset, we calculated variances for each category and then computed an average variance across 16 base prompts, as shown in Tab. 5. This process led to determining the average standard deviation for these prompts (range: 0 to 1), which are summarized in Tab. 7. In this table, lower values indicate closer conformity with the expected distribution.

Determining expected values for base prompts presents a significant challenge, as illustrated by the example prompt "Pastor." Specifically, the ambiguity in expected gender dis-

	SD	DE	FF
B+M	0.6498	**0.5067**	0.5602
M+B	**0.2597**	0.4129	0.3577

Table 7: Standard Deviation of 3 different models (SD,DE,FF) on 16 prompts of ordering B+M (Base+Modifier) and M+B (Modifier+Base)

tribution for this prompt highlights the complexity of establishing a clear expectation. Three potential scenarios emerge: a gender parity assumption (50:50), alignment with the actual demographic distribution of males and females (50.4:49.6) [United Nations and Social Affairs, 2022], or adherence to the real-world ratio of males to females within the pastoral occupation(80:20) [CNN, 2023]. This variance underscores the difficulty in defining a singular expectation for gender representation. Extending this dilemma to all 16 prompts, it becomes evident that establishing universally applicable expected values is fraught with challenges, reflecting the broader difficulty in applying a consistent expectation framework across diverse contexts.

Our analysis revealed that the 'Modifier+Base' configuration generally yielded more consistent results than the 'Base+Modifier' approach. We posit this could be due to the modifier's enhanced emphasis when positioned at the start of the prompt. Notably, the variance among standard deviations was minimal for DALL·E, suggesting this model's resilience to prompt order. However, DALL·E's performance dipped notably with the Modifier+Base setup, attributed to ChatGPT4's expansion of the prompts, which sometimes resulted in a focus on background elements over the main subject, leading to ambiguous outcomes. This phenomenon, as discussed in Section 4, could also be linked to generating image grids rather than individual images per prompt when using ChatGPT4.

6 Discussion

Bias is an inherent characteristic of models trained on real-world data, which inevitably contain biases. Our approach—utilizing modifiers as a form of prompt engineering to influence bias distribution—represents an unexplored method of bias adjustment within the field. This preliminary strategy did not yield consistently effective results, indicating that simplistic applications of modifiers are insufficient. This finding points to the necessity for a more nuanced approach, potentially involving a larger-scale, subjective analysis to tailor bias distribution when the intent is to generate data points from the extremes of a distribution.

Reflecting on the challenges faced by the Gemini case [CNBC, 2024; CNN, 2024], we recognize that any attempts to correct biases in models are fraught with complexity. Gemini's failures—oversights in presenting a diverse range of individuals and an overly cautious response to benign prompts—exemplify the difficulties in achieving balance. [Google, 2024] The question of whether to align model outputs with geographical or demographic realities remains open. More concerning, however, is the presence of unacknowledged biases within models, as unrecognized biases that are not addressed pose a significant issue.

In our investigation, a limited number of images were produced and analyzed. The images were generated through the ChatGPT interface rather than directly using DallE's API, The assessment of model-generated images was carried out solely by the authors, constrained by resources and foregoing external human studies. To maintain analytical rigor, the authors collectively verified each evaluation to reach a unanimous agreement. Our investigation rigorously evaluated quantitative metrics such as Image Text Alignment [Xu *et al.*, 2018a; Xu *et al.*, 2018b] and Image Quality [Salimans *et al.*, 2016a; Salimans *et al.*, 2016b] and determined that they do not adequately measure the specific tasks we are examining. Additionally, we attempted to apply the DallEval [Cho *et al.*, 2023] framework to our generated data, but the visual reasoning metrics utilized by DallEval were not appropriate for our analysis.

The study demonstrates that the LLM frontend, as utilized in this context, exhibits a robustness against manipulation attempts through prompt engineering, irrespective of prompt ordering. This stability suggests that the LLM frontend effectively mitigates the risk of generation failures that might arise from the sequence of the prompt components.

Furthermore, we establish a framework for subsequent research focused on refining models to address and control rare yet impactful biases that risk distorting data representation. This work highlights a crucial discourse on the reconciliation of biases—whether models should be aligned with an idealized vision of inclusivity or adhere to factual representations drawn from demographic and historical contexts.

7 Conclusion

This study explores biases in text-to-image models, revealing how societal biases are embedded and can be mitigated within these AI systems. Our characterization experiments showed that while Stable Diffusion and DallE often reproduce biases from their training data, Firefly shows the potential for less biased outputs, pointing to differences in data handling and model design. Meanwhile, our study of prompt modification highlights the uneven success of using modifiers for bias adjustment and the importance of prompt structure in shaping outputs, demonstrating that direct approaches to prompt engineering are not sufficient to reliably overcome intrinsic model biases in all cases.

The observed complexity in model responses to even these relatively straightforward adjustments in stimuli underscores the ethical imperative for AI developers to balance innovation with sensitivity, advocating for transparency and inclusivity in AI development to prevent the reinforcement of societal inequalities. This work introduces a taxonomy for categorizing model robustness to prompt modification and a quantitative, expectation-based metric for conformity with supplied prompt modifies that can be utilized by future work for similar cross-comparative studies. Both the limitations and opportunities highlighted by this research point to the necessity for ongoing efforts to understand and correct biases in AI, suggesting future exploration into more effective bias-controlling strategies and diverse AI development approaches.

References

[Adobe Systems Incorporated, 2023] Adobe Systems Incorporated. Adobe firefly: Generative ai for creative processes. https://firefly.adobe.com, 2023. Accessed: 2024-03-24.

[Avrahami *et al.*, 2023] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18370–18380, June 2023.

[Barocas *et al.*, 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org, 2019.

[Betker *et al.*, 2023] Jason Betker, Greg Goh, Li Jing, Tim Brooks, Jianyu Wang, Liang Li, Lucy Ouyang, Jie Zhuang, Jason Lee, Yuxuan Guo, Waseem Manassra, Prafulla Dhariwal, Chenxi Chu, and Yong Jiao. Improving image generation with better captions. *OpenAI Blog*, 2023.

[Brooks *et al.*, 2023] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Cho *et al.*, 2023] Jaemin Cho, 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*, pages 3043–3054, 2023.

[CNBC, 2024] CNBC. Google pauses gemini ai image generator after it created inaccurate historical pictures. https://www.cnbc.com/2024/02/22/google-pauses-gemini-ai-image-generator-after-inaccuracies.html, February 2024. Accessed: 2024-03-29.

[CNN, 2023] CNN. Women and church leadership in the united states. CNN, July 2023. Available online.

[CNN, 2024] CNN. Google halts ai tool's ability to produce images of people after backlash. https://www.cnn.com/2024/02/22/tech/google-gemini-ai-image-generator/index.html, February 2024. Accessed: 2024-03-29.

[Couairon *et al.*, 2023] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023.

[Dong *et al.*, 2024] Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*, 2024.

[Friedrich *et al.*, 2023] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[Google, 2024] Google. Gemini image generation got it wrong. we'll do better. https://blog.google/products/gemini/gemini-image-generation-issue/, 2024. Accessed: 2024-03-29.

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

[Li *et al.*, 2023] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, June 2023.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[Naik and Nushi, 2023] Ranjita Naik 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*, pages 786–808, 2023.

[Noble, 2018] Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press, 2018.

[Oppenlaender, 2023] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pages 1–14, 2023.

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

[Salimans *et al.*, 2016a] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[Salimans *et al.*, 2016b] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[Samvelyan *et al.*, 2024] Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022.

[Seshadri *et al.*, 2023] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023.

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

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

[Struppek *et al.*, 2023] Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, et al. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068, 2023.

[United Nations and Social Affairs, 2022] Department of Economic United Nations and Population Division Social Affairs. World population prospects 2022. https://population.un.org/wpp/, 2022. Accessed: 2024-04-03.

[Xu *et al.*, 2018a] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[Xu *et al.*, 2018b] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[Yang *et al.*, 2023] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14246–14255, June 2023.

[Yeh *et al.*, 2023] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299, 2023.