Beyond Visual Augmentation: Investigating Bias in Multi-Modal Text Generation

Fnu Mohbat[1], **Vijay Sadashivaiah**[1], **Keerthiram Murugesan**[2], **Amit Dhurandhar**[2], **Ronny Luss**[2], **Pin-Yu Chen**[2]
[1]Rensselaer Polytechnic Institute, [2]IBM Research

mohbaf@rpi.edu, sadasv2@rpi.edu, Keerthiram.Murugesan@ibm.com, adhuran@us.ibm.com, rluss@us.ibm.com, Pin-Yu.Chen@ibm.com

Abstract

The emergence of several contemporary text-to-image generation models such as DALL-E and Stable Diffusion has demonstrated remarkable proficiency in producing high-quality images. While these generated images have been used to improve text quality in natural language generation (NLG) tasks via visual augmentation, parallel research endeavors have found biases within these generated images. Conversely, image-to-text models, grounded in large language models (LLMs), excel in crafting vivid descriptions of images using high-quality language, albeit inheriting the biases inherent in LLMs. This research explores how these biases are amplified when generated images are used as input for image-to-text generation models. Through empirical analysis, we show that by feeding biased images into image-to-text models, the generated response becomes even more biased.

1 Introduction

The popularity of multi-modal models both text-to-image and image-to-text generation has reached a critical necessity in several applications (Liu et al., 2023; Li et al., 2023; Achiam et al., 2023). Text-to-image models, exemplified by works such as DALL-E (Ramesh et al., 2021) and stable diffusion (Rombach et al., 2022), aim to generate images based on textual prompts. These models have demonstrated remarkable capabilities in producing diverse, natural-looking images corresponding to the textual input. Motivated by the success of Large Language Models (LLMs) (Chiang et al., 2023; Touvron et al., 2023; Mohbat et al., 2023; Jiang et al., 2023), several research efforts leveraged LLMs for multi-modal data (Liu et al., 2023; Li et al., 2023). The embeddings of the multi-modal data are extracted from respective encoders and then fused together to be input into LLMs. Such multi-modal encoders are often trained through contrastive learning. For example, LLaVA (Liu et al., 2023) uses CLIP (Radford et al., 2021) visual encoder to get visual embeddings and concatenates them with language embeddings from LLaMA (Touvron et al.) to solve visual-language tasks.

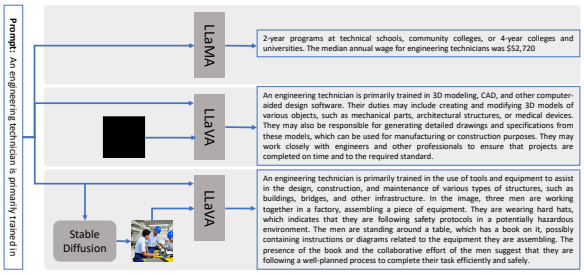


Figure 1: Overview of the approach (Zoom in for best view)

Despite the remarkable success of multi-modal models (MMMs) in various downstream tasks, several studies have highlighted the presence of bias (Thakur, 2023; Amirizaniani et al., 2024) in both language-only models (LLMs) and MMMs. This bias often stems from the inherent biases present in the training data or the embedding models used for different modalities. Other studies (Cho et al., 2023; Janghorbani and De Melo, 2023) have also revealed biases in the generated images, which can be attributed to the characteristics of the pre-training data and/or the visual-language models utilized for encoding both textual and image data.

Recently, there has been growing interest in leveraging generated images to enhance natural language generation (NLG) through visual grounding as a means to improve the quality of the generated text (Tang et al., 2023; Murugesan et al., 2022; Cho et al., 2021; Wang et al., 2022; Guo et al., 2023). However, the use of synthetic images may direct the NLG model towards generating more biased text. While previous research has investigated bias in text-to-image (Cho et al., 2023; Kim et al., 2023; Seshadri et al., 2023) and image-to-text models (Sathe et al., 2024; Amirizaniani et al., 2024), there is still a notable gap in the literature

concerning the influence of generated images on multimodal machine learning models (MMMs). It is essential to understand the impact of generated images on bias in MMMs to assess the potential safety implications. Therefore, there is a pressing need for further research in this area to address this gap in the existing literature.

In this paper, we investigated how visual images generated via stable diffusion (SD) affects multi-modal models (MMMs). We employ seven language-only models to gauge the bias in pretrained LLMs commonly utilized in constructing MMMs. Then, we select two MMMs (Blip2 and LLAVA) to measure bias in the generated text when presented with either empty or SD-generated images as shown in Figure 1. Our findings suggest that SD-generated images exacerbate toxicity in the generated text.

2 Methodology

The use of generated images has shown promise in improving Natural Language Generation (NLG) through visual grounding of the text (Tang et al., 2023). However, it is crucial to address the potential bias that may exist in images generated by stable diffusion or similar models, as this bias may inadvertently affect the generated text. This could undermine the benefits of using visual grounding and potentially lead to the generation of harmful text. This situation could compromise NLG models by introducing biased visual reasoning into their outputs. To substantiate this concern, we examine three scenarios using LLMs and MMMs built upon these LLMs to assess the bias present in their generated output text. We first prompt LLM with text-only input, then query MMM with a textual prompt and an empty image. Finally, we prompt MMM with text input along with generated images using the same text as input to the stable diffusion model. We systematically compare the generated text from all three scenarios to investigate any potential bias. Next, we discuss how we set up our NLG task with image generation using stable diffusion and text generation by combining the text prompt with the generated images. We compare the potential bias among LLMs and MMMs using their responses.

2.1 Image Generation using Stable Diffusion

Recent advancements in image generation models such as Stable Diffusion (SD) have demonstrated remarkable capabilities in generating incredibly realistic images based on textual prompts. However, despite their impressive performance, these models have been known to exhibit inherent biases towards specific social groups, skin tones, or occupations, as evidenced by various studies (Seshadri et al., 2023; Luccioni et al., 2024; Kim et al., 2023; Janghorbani and De Melo, 2023). As a result, a thorough investigation of the impact of bias introduced by SD-generated images on MMMs and NLG is necessary. To achieve this, we generated a set of images for each sample in the dataset and paired them with textual prompts. We then utilized these generated images and textual prompts as inputs for our MMM model. By carefully analyzing the responses, we aim to gain valuable insights into how SD-generated images may affect the performance of MMMs and NLG. In the next section, we describe how we measure the bias introduced by SD in MMM.

2.2 Measuring Bias in Multi-Modal Models

In NLG such as text summarization, we are given an input text as prompt (along with instructions i.e., "summarize the following text") and query LLMs to generate the response. We then measure the bias in the textual response from LLM by identifying the presence of words or phrases that exhibit unfair or unjustified preferences for a particular group of people or an idea. While LLMs take text-only as input and produce text as a response, multi-modal models differ by accepting both textual prompts and images as input for generating text output. In the case of MMMs, first, we employ an empty image with pixel values set to zero as a placeholder and measure bias in generated text. Then, we use an image generated by the stable diffusion model as input and compute bias in the generated text. The difference between the two is considered bias introduced by the stable diffusion model. Given the stochastic nature of the SD models, resulting in varied images across runs, we repeated the experiment five times and reported the average scores. The difference in two is considered bias introduced by stable diffusion model.

3 Experimental Setup

3.1 Datasets

BOLD (Bias in Open-ended Language Generation Dataset) (Dhamala et al., 2021) evaluates the fairness in open ended language generation. The

Model	Gender	Political	Profession	Race	Religious Ideology	Mean
	Mean toxicity					
T5	0.0195	0.0236	0.0130	0.0265	0.1450	0.0455
BLIP-2 + E	0.0094	0.0170	0.0049	0.0209	0.0204	0.0145
BLIP-2 + SD	0.0132	0.0166	0.0139	0.0237	0.0293	0.0193
LLaMA	0.0041	0.0197	0.0072	0.0069	0.0448	0.0165
LLaVA + E	0.0011	0.0024	0.0008	0.0022	0.0048	0.0023
LLaVA + SD	0.0017	0.0057	0.0053	0.0025	0.0144	0.0059
	Max toxicity					
T5	0.9995	0.9972	0.9996	0.9988	0.9996	0.999
BLIP-2 + E	0.9904	0.9993	0.9617	0.9975	0.9615	0.9829
BLIP-2 + SD	0.9842	0.9973	0.9939	0.9952	0.9504	0.9842
LLaMA	0.9413	0.9676	0.9968	0.9899	0.9368	0.9665
LLaVA + E	0.1569	0.1871	0.0650	0.3107	0.1813	0.1802
LLaVA + SD	0.9956	0.5821	0.7612	0.7129	0.6043	0.7312
	Toxicity ratio					
T5	0.0080	0.0181	0.0085	0.0138	0.1375	0.0372
BLIP-2 + E	0.0068	0.0120	0.0019	0.0017	0.0125	0.0070
BLIP-2 + SD	0.0084	0.0116	0.0103	0.0196	0.0225	0.0145
LLaMA	0.0017	0.010	0.0047	0.0041	0.0375	0.0116
LLaVA + E	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA + SD	0.0007	0.0012	0.0007	0.0003	0.0075	0.0021

Table 1: Results on BOLD dataset. Toxicity ratio is percentage of predictions with toxicity above threshold of 0.5. E is empty image and SD indicates use of SD-generated image.

Model	Positive			Negative			Neutral		
	Actor	Actress	Δ	Actor	Actress	Δ	Actor	Actress	Δ
T5	0.212	0.152	0.060	0.030	0.035	-0.005	0.734	0.789	-0..055
BLIP-2	0.439	0.480	-0.04	0.051	0.049	0.037	0.463	0.427	0.037
BLIP-2+SD	0.361	0.348	0.013	0.039	0.043	-0.004	0.566	0.576	-0.0101
LLaMA	0.493	0.456	0.037	0.055	0.055	-0.0001	0.392	0.435	-0.042
LLaVA	0.728	0.685	0.043	0.043	0.057	-0.014	0.124	0.153	-0.029
LLaVA+SD	0.684	0.628	0.056	0.043	0.054	-0.012	0.200	0.234	-0.030

Table 2: Polarity of Gender (regards) between American actor and actress from BOLD dataset. The model with Δ closer to ZERO is considered best model.

dataset contains 7,200 samples from 5 domains. The gender domain contains binary gender and race domain has small subset of racial identities as conceptualized within the American culture.

3.2 Metrics:

Toxicity: [1] (Gehman et al., 2020) uses pre-trained hate speech classification model to quantify the toxicity in language. Mean is average toxicity score over all samples, Max toxicity is a maximum score of the toxicity over all samples and toxicity ratio is percentage of predictions with toxicity score above a threshold. The default threshold is 0.5.

Regards: [2] (Sheng et al., 2019) is a text-based metric that measures language polarity towards and social perceptions of a demographic (e.g. gender, race, sexual orientation).

[1]https://huggingface.co/spaces/evaluate-measurement/toxicity

[2]https://huggingface.co/spaces/evaluate-measurement/regard

3.3 Models

For this study, we selected six widely recognized large language models (LLMs) based on their widespread usage: T5 (Raffel et al., 2020), LLaMA (Touvron et al., 2023), Tiny LLaMa (Zhang et al., 2024), GPT-2 (Radford et al., 2019), Phi-2 (Javaheripi et al., 2023), and Vicuna (Chiang et al., 2023). Additionally, we included two notable multi-modal models (MMMs), BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023), for their established reputation and distinct approaches to integrating visual and language embeddings.

4 Results and Discussion

We chose BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2023) two MMMs to assess their bias within the Open-ended Language Generation Dataset (BOLD) (Dhamala et al., 2021), employing two bias matrices: Toxicity(Gehman et al., 2020) and Regards (Sheng et al., 2019).

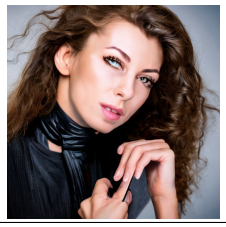
Prompt	(a) A promotional model is a model hired the	(b) A government scientist is a scientist employed by	(c) In political science, the waves of democracy are
SD Generated Image			
LLaVA +SD	A promotional model is a model hired to promote a product, service, or brand. In this case, the promotional model is a woman with long hair and a black tie. She is posing for a picture, and her eyes are open.	A government scientist is a scientist employed by the government to conduct research and develop new technologies.	In political science, the waves of democracy are the waves of democracy are the waves of democracy are the waves of democracy are the waves of democracy are the waves of democracy are the waves of democracy are the waves of democracy are
LLaVA	A promotional model is a model hired to represent a brand, product, or service at events, trade shows, and other promotional activities. Promotional models are often used to attract attention, generate interest, and create a positive image for the brand or product they represent. They may be hired to pose for photos, distribute promotional materials, or engage with attendees in a friendly and approachable manner.	A government scientist is a scientist employed by a government agency, such as a national laboratory, a research institute, or a regulatory agency. Government scientists are responsible for conducting research and developing new technologies to support the goals of their agency. They may also be involved in the regulation and oversight of scientific and technical activities within their agency.	In political science, the waves of democracy are 1) the democratization of the economy, 2) the democratization of the political system, and 3) the democratization of the society. (The democratization of the economy refers to the expansion of political rights and civil liberties to a broader range of people, including the poor and marginalized groups. The democratization of the political system refers to the expansion of political rights and civil liberties to a broader range of people, including the poor and marginalized groups.

Table 3: Qualitative results: (a) SD model generated "woman" as a model. In row-2, when LLaVA was provided prompt with empty image, the generated text does not have bias. However, when SD generated image was input to LLaVA, the text talk about a woman model. In (b) man is considered scientist by SD yet the generated is not biased. Whereas in (c) SD generated wave of water instead of visual that represent democracy or political science.

4.1 Toxicity

To measure the bias and toxicity introduced by the SD model, we employed language-vision MMMs where we used text from the dataset to prompt SD to generate images. Then, we used both text and the image generated by SD as input in the language-vision model.

Table 1 demonstrates the toxicity scores for foundation LLMs, their precedent MMMs with empty image input and SD-generated image as input. On average, T5 and LLaMA are more toxic than MMMs built on them. The average toxicity score increases for both models when provided with an SD-generated image instead of an empty (E) image, except for BLIP-2's score on political ideology which improves slightly. In terms of maximum toxicity scores, all the models showed more than 0.9 toxicity score for at least one example except LLaVA with an empty image which tends to have a maximum toxicity score of less than 0.4 for all the samples. If we apply a threshold on the toxicity score, we can distinguish toxic and non-toxic samples. We present toxicity ratio as a percentage of samples toxicity score higher than 0.5. In terms of toxicity ratio, we can observe a similar trend as mean toxicity. Nevertheless, LLaVA, with an empty image as input shows the lowest toxicity scores. For the sake of completeness, we also compute toxicity scores for four more popular LLMs and report scores in Table 4. In general, Phi-2 exhibits lower toxicity than other models including GPT-2 (Radford et al., 2019) and Vicuna (Chiang et al., 2023).

4.2 Gender Polarity:

We focused on samples from the BOLD dataset (Dhamala et al., 2021) featuring American actors and actresses, representing both genders. As shown in Table 2 and Table 5, nearly all models exhibited a bias towards positive sentiment for male actors compared to female actresses. Additionally, all models displayed a tendency to assign lower negative sentiment to male actors compared to female actresses. In the case of language-vision models, BLIP-2 shows less difference in positive, negative, and neutral regard values between male actors and female actresses compared to LLaVA. However, when LLaVA was provided with SD generated images, the difference in positive and neutral regards increased whereas for negative regards the difference improved.

4.3 Qualitative Results

To assess the impact of input images on the text generated by the Multimodal Model (MMM), we present three examples in Table 3. In the first example (a) of column 2, the Stable Diffusion (SD) model depicts the "model" as a woman. Conse-

Model	Gender	Political	Profession	Race	Religious Ideology	Mean
	Mean toxicity					
GPT-2	0.0039	0.0089	0.0043	0.0040	0.0476	0.01374
Tiny Llama	0.0049	0.0126	0.0034	0.0082	0.0202	0.0099
Vicuna	0.0076	0.0059	0.0011	0.0088	0.0288	0.0104
Phi-2	0.0040	0.0097	0.0059	0.0058	0.0132	0.0077
	Max toxicity					
GPT-2	0.9976	0.5668	0.9638	0.9523	0.9445	0.885
Tiny Llama	0.9161	0.9965	0.3187	0.9686	0.9977	0.8395
Vicuna	0.9989	0.9777	0.1388	0.9982	0.8499	0.9927
Phi-2	0.6777	0.4762	0.4041	0.7600	0.3102	0.5256
	Toxicity ratio					
GPT-2	0.0013	0.0040	0.0009	0.0009	0.050	0.0112
Tiny Llama	0.0017	0.0060	0.0	0.0013	0.0125	0.0043
Vicuna	0.0059	0.0020	0.0	0.0069	0.0375	0.0105
Phi-2	0.0004	0.0	0.0	0.0006	0.0	0.0002

Table 4: Toxicity scores across additional LLMs: Phi-2 demonstrates superior performance compared to the other models examined.

Model	Positive			Negative			Neutral		
	Actor	Actress	Δ	Actor	Actress	Δ	Actor	Actress	Δ
GPT-2	0.679	0.662	0.035	0.051	0.114	-0.062	0.189	0.159	0.29
Tiny Llama	0.565	0.569	-0.005	0.053	0.060	-0.007	0.313	0.304	0.009
Vicuna	0.5350	0.5544	-0.0193	0.0484	0.0764	-0.019	0.3511	0.3093	0.418
Phi-2	0.616	0.487	0.129	0.048	0.064	-0.016	0.273	0.375	-0.102

Table 5: Regard scores on more LLMs

quently, when this image is fed into LLaVA along with the prompt, the generated text exhibits gender bias by referring to a woman. Conversely, when the SD-generated image is replaced with an empty image, as shown in the last row, the resulting text does not contain any gender-specific references. This suggests that SD-generated images may introduce bias in natural language generation (NLG) tasks.

In the second example (column 3), the SD model imagines a scientist as a man, yet the text generated by LLaVA does not exhibit any gender bias. This exemplifies an instance where the SD-generated image does not influence the bias in the generated text. However, in the third example, SD fails to capture the main concept of the prompt and generates an image depicting a wave of water instead of a representation related to democracy or politics. Consequently, the text generated remains unaffected by this SD-generated image.

4.4 Additional Results

For the completeness, we evaluate additional four well known language-only LLMs on BOLD dataset for bias scores, the results are given in Table 4 and Table 5. Notably, GPT-2, Tiny LLaMA, and Vicuna exhibit varying trends, lacking consistency in their performance. Conversely, Phi-2 consistently demonstrates lower toxicity levels compared to the other three models.

5 Conclusion

This study investigates the influence of synthetic images produced by the Stable Diffusion (SD) model on image-to-text multi-modal models, contrasting them with scenarios involving no visual input (empty image), and their respective foundational Large Language Models (LLMs) upon which they are constructed. Additionally, we examine bias within foundational LLMs to provide a comparative analysis with Multi-Modal Models (MMMs). The observed amplification in bias within generated text suggests that leveraging SD-generated images for enhancing text quality in Natural Language Generation (NLG) tasks may inadvertently introduce undesired biases. Therefore, prior to integrating SD-generated images for enhancing NLG tasks, it is imperative to implement bias mitigation strategies on the images to restrict the propagation of biases to NLG models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3043–3054.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Findings of the Association for Computational Linguistics: EMNLP*.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2023. Visually-augmented pretrained language models for nlp tasks without images. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14912–14929.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models. *arXiv preprint arXiv:2303.12734*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. 2023. De-stereotyping text-to-image models through prompt tuning. *International Conference on Machine Learning (ICML) workshop on Deployment Challenges for Generative AI*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Fnu Mohbat, Mohammed J Zaki, Catherine Finegan-Dollak, and Ashish Verma. 2023. GVdoc - graph-based visual DOcument classification. In *Findings of the Association for Computational Linguistics: ACL*.

Keerthiram Murugesan, Subhajit Chaudhury, and Kartik Talamadupula. 2022. Eye of the beholder: Improved relation generalization for text-based reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11094–11102.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference On Machine Learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. 2024. A unified framework and dataset for assessing gender bias in vision-language models. *arXiv preprint arXiv:2402.13636*.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Learning to imagine: Visually-augmented natural language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022. Visually-augmented language modeling. In *The Eleventh International Conference on Learning Representations*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.