Ethics and Epistemology of AI 2025

Seeing Bias: Demographic Representation of Professions in Image-Generation Systems

Seychelle Ann Dagus[1] · **Zain Abladen Hazzouri**[1] · **Muhammad Hassan Jalil**[1] · **Laura Rivero Miró**[2] · **Carlos March Moya**[2] · **Erik Voigt**[4] · **Muhammed Enes Yavuz**[1]

Submitted: 10.07.2025

Abstract This paper investigates demographic bias in the text-to-image model Stable Diffusion 3.5, focusing on the representation of professions. By systematically varying prompt-specificity with respect to demographic attributes (gender, ethnicity, profession), we analyze the model's outputs and compare them with real-world labor statistics. Our findings show that the model overrepresents white individuals and defaults to gender stereotypes, portraying men in technical roles and women in caregiving positions. Increasing prompt specificity does not mitigate these biases and, in some cases, amplifies or hides them through "stereotype stacking". These results show that the model does not mirror reality but instead reinforces and magnifies existing societal biases, raising critical ethical concerns about representational harm, symbolic erasure, and the perpetuation of inequality through GenAI.
The code and produced data is available at https://github.com/VoErik/sd-3-ethics.

Keywords Text-to-Image Generation · Stable Diffusion · AI Bias · Ethical AI · Demographic Representation

1 Introduction

In the last few years years, the advancement of text-to-image (TTI) generation systems like DALL·E and Stable Diffusion (SD) has enabled widespread creation of visual content from prompts. However, this progress introduces ethical and epistemological challenges, particularly regarding the representation of different social groups. AI bias, the systematic tendency of these systems to produce unfair or harmful outcomes [1], can manifest as the over- or underrepresentation of certain subgroups or the reinforcement of stereotypes.
This paper examines how SD 3.5, reflects or reinforces demographic biases and stereotypes present in the real world and the underlying training data. Our analysis focuses on representations across a subset of different professions. To answer this question, we first analyze the emergent patterns in images produced by the model. We then evaluate the alignment of these visual representations with real-world statistics, and discuss the ethical implications of our findings.

Technical University Berlin[1]
{s.dagus, hazzouri, m.jalil, m.yavuz}@campus.tu-berlin.de

Polytechnic University of Valencia[2]
{lrivmir@teleco, cmarmoy@etsinf}.upv.es

Otto-Friedrich University Bamberg[4]
erik.voigt@uni-bamberg.de

By varying the specificity of prompts through incrementally adding gender, ethnicity, and profession markers, we analyze the effect of an increased context on the representations produced by the model. We explore the model's visual outputs through statistical analysis, heatmaps and by comparing them with real-world data. Our findings show consistent patterns of bias. Professions tend to be shown according to gender stereotypes, white individuals are overrepresented, and intersectional identities are either barely visible or completely left out. These findings reveal not only weaknesses in the algorithms but also deeper issues related to knowledge injustice and symbolic exclusion.
By exploring these dynamics, we hope to add to the ongoing conversation about fairness and accountability in generative AI.

2 Background & Related Work

The potential for artificial intelligence to produce biased and discriminatory outcomes is a well-established concern. Seminal work [2] showed that commercial machine learning algorithms often discriminate based on classes like race and gender. In the domain of generative AI, these concerns are particularly acute, as the models are often trained on large, un- or badly curated datasets scraped from the internet. For instance, the LAION dataset [3] that is widely used for training mod-

els like Stable Diffusion [4], yet has also been shown to contain problematic and explicit content, including racist and misogynistic images and text [5].

The biases embedded in these training datasets then manifest in the outputs of text-to-image (TTI) models. While gender and race are the most studied vectors, biases also extend to socioeconomic, cultural, and biological attributes [6, 7]. Research consistently finds that models perpetuate harmful stereotypes. Studies show the sexualization of women and the reinforcement of traditional gender roles in professional contexts across TTI models like DALL-E 3 and Bing Image Creator [8]. This male-as-default tendency is found even when using gender-fair language in prompts [9] and is persistant through all internal stages of the generation process [10]. While earlier studies [11] indicate that the image modality itself amplifies existing biases, [12] identify an amplification of social biases through TTI systems. [11]. One study questions the direct attribution of bias amplification to models, suggesting it may be a paradoxical effect of discrepancies between training data captions and user prompts [13].

Other studies indicate a *white normativity* [14], and location-neutral prompts default to Western-centric scenes [15]. [16] find that models underrepresent marginalized groups - albeit to different degrees. The portrayal of specific ethnicities is also subject to model-specific stereotypes, such as the Westernization of East Asian women in DALL-E or their sexualization in Stable Diffusion [17]. [18] show that these biases can be compounded as models and datasets scale; for large models, the probability of classifying Black and Latino men as *criminal* increased with larger training datasets. While often only single bias-vectors are investigated, [19] point out that biases often address and influence each other.

A large body of work has focused on evaluating and mitigating bias. Some work seeks to improve evaluation by using automated detection methods [20] or using generative models themselves to create counterfactual examples to probe intersectional biases at scale [21]. When it comes to bias mitigation, strategies are being developed to steer model outputs toward fairness during deployment without the need for retraining, such as [22] or *Fair Diffusion* [23], which uses human instructions to guide image generation. Finally, some scholars are rethinking the goals of algorithmic fairness, proposing to move away from *difference-blind* approaches to more context-aware frameworks where treating groups differently may be necessary for an equitable outcome [24].

3 Method

For this study, we employed Stable Diffusion 3.5 (SD 3.5), a state-of-the-art text-to-image diffusion model known for its open-source accessibility and widespread adoption in both academic and industrial contexts. The model was accessed via the *Hugging Face interface*, which allowed for systematic and automated generation of thousands of images using controlled and reproducible prompts.

3.1 Experimental Design

To investigate how demographic biases manifest in AI-generated portrayals of professions, we implemented a *multi-step prompt design* shown in Figure 11. This approach incrementally added contextual layers—**gender**, **ethnicity**, and **profession**— to isolate how each factor contributes to representation outcomes. Table 1 shows the structure of the used prompts.

3.2 Image Generation Procedure

Each prompt was repeated ten times to account for the stochastic nature of image generation in diffusion models. All images were generated under identical conditions (same resolution, seed variation, and model version) to ensure consistency. Prompts were kept *linguistically neutral* apart from the controlled variables, and no additional modifiers (e.g., *attractive*, *young*, *professional*) were used, to avoid introducing unintended bias.

3.3 Data Annotation and Analysis

Generated images were annotated manually by the research team. Each image was reviewed and labeled according to the perceived *gender*, *ethnicity*, *profession*, as well as additional comments on *socio-economic* markers, such as clothing, setting, etc.

Finally, we compared the aggregated results against real-world demographic statistics (e.g., Bureau of Labor Statistics, national census) to identify patterns of alignment or distortion, and to assess the potential ethical implications of biased representations produced by the model.

3.4 Bias Perception Survey

Under the assumption that perception of bias and stereotypes is not universal, but culturally and socially dependent, we constructed a survey to elicit differences in perception across different demographic groups. Details on the survey setup can be found in Appendix A.

Table 1: Prompt Categories for Bias Analysis

Prompt Category	Prompt Examples	Purpose
Baseline	"A photo of a person at their job"	Captures default assumptions (profession, gender, ethnicity) without explicit demographic cues.
Single-Variable	"A photo of a [gender] person at their job"	Isolates the effect of **gender**.
	"A photo of an [Asian/Black/Hispanic/White] person at their job"	Measures the effect of **ethnicity**.
Combined Demographic	"A photo of a [gender] [ethnicity] person at their job"	Analyzes **intersectional bias**, examining how overlapping identities are portrayed.
Profession-Focused	"A photo of a nurse", "A photo of an engineer".	Reveals latent demographic associations encoded in the model, especially for **stereotypically gendered roles**.

4 Findings and Discussion

Our dataset consisted of 259 unique prompts, each executed ten times, yielding a total of 2,590 images. The prompts were designed to probe three dimensions of demographic representation: gender, ethnicity, and profession. This setup enabled a multi-step analysis of how image outputs vary depending on the specificity and composition of prompt cues. We chose the profession social worker as an example because, globally, it is roughly 50% male and 50% female, but Stable Diffusion generated only female images, even without gender marker.

4.1 Prompt Specificity

Profession-Marked Prompts When prompts contained only a profession and no other demographic markers were provided (e.g.,*a social worker* shown in 13), the resulting images were still primarily female presenting. For example, six out of ten images were generated as Black women, despite providing no gender or ethnic specification. This shows that the model defaults to latent biases or majority-trained associations in context when demographic fields are left blank.

Gender-Marked Prompts However, when extended from *a social worker* to *a female social worker*, the model complied with a broader female presentation. While the images produced were still all females, they were different skin colors and ages, albeit still on a not-as-diverse scale.

Gender and Ethnicity-Marked Prompts With the full specification *a Black female social worker*, the model generated images visually matching the prompt but exhibiting constrained variability. The outputs demonstrated a kind of narrow inclusion. The model represented the identity correctly but did not vary facial features, hairstyles, or context, which implies a boundary around what constitutes an acceptable depiction of that intersectional identity.

This pattern was visible across all professions. The results (Tab. 3-8) indicate that an increase in prompt specificity only marginally decreases bias in SD 3.5. Conversely, as Fig. 1 shows, increasing specificity even increased ethnicity bias (see Tab. 3 for a more detailed view). For gender representation, increasing the specificity decreases the representational gap (see Fig. 22). However, this is likely due to interference effects from the relative overrepresentation of women in certain professions. This marks a significant issue: While increasing specificity appears to reduce the representational gap, this improvement is an illusion created by offsetting one bias with another. This approach leverages the overrepresentation of women in certain professions to mask the initial disparity, leaving both systemic issues unaddressed.

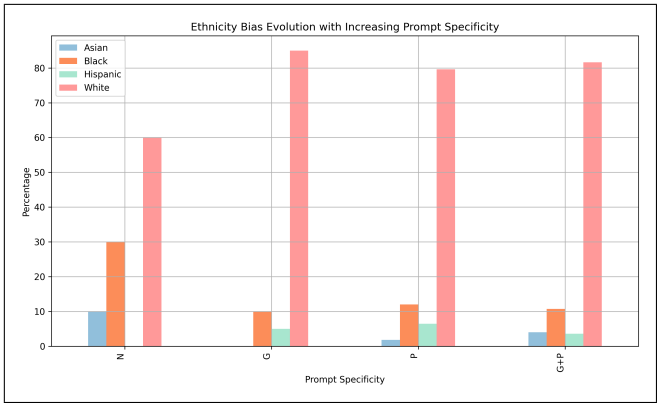

Fig. 1: Ethnicity distribution across varying levels of prompt specificity (N: neutral, G: gender, P: profession).

4.2 Representation Statistics

Across all images, we performed frequency counts on inferred gender and ethnicity characteristics.

Gender Distribution Females were produced the majority across the board; notably in careers trained as *care work* or *education*. This reflects social realities and geographic biases that are potentially learned from training data.

Ethnicity Distribution White-presenting characters came out most across occupational intersections and regardless of whether the prompt specified ethnicity or not. Other presenting groups like Black, Asian, and Hispanic characters were few and far between and rendered only when specifically asked.

Age Representation There was also an age caveat of homogeneity. Regardless of profession and markers, the model generated young people aged 25-35 which suggests an inaccurate average of what someone would look like doing this profession.

Gender-Job Mapping Fig. 21 shows the extent of gender stereotyping that occurred. For *nurse* there were 130 females and 50 males. Yet jobs like *engineer* and *scientist* rendered a solidly male majority, even when no gender was specified in the prompt.

Ethnicity-Job Mapping The ethnicity heatmap in Fig. 20 shows White figures as dominant across nearly all professions. In the *construction worker* category, for example, the model returned 40 White individuals and zero representations from Black, Asian, or Hispanic groups, and again, despite ethnicity being unspecified in the prompt.

Intersectional Patterns Finally, Fig. 19, shows generated images to the compound identity hypothesis of *female Asian scientist* or *Black male nurse*, which showed varied representation. For example, *White male engineer* was rendered often while others were rendered once or in slight variations, which speak to the model's limited intersectional representation ability.

4.3 Comparison with real-world statistics

In what follows, we compare the proportions produced by SD with empirical labour-market statistics from the *International Labour Organization*[25], *Eurostat*[26], the German *MIZ*[27], studies on *ResearchGate*[28], and further national statistical offices [29,30].

Gender Distribution in Real World vs. Stable Diffusion Stable Diffusion (SD) does not *always* over– or under-represent one gender.
For some occupations, its output stays relatively close to empirical data. A good example is *artist*: real-world statistics are really close to the SD results.
In many other cases, however, SD *saturates* the stereotype, depicting a profession as entirely male or entirely female:
A particularly illustrative is *social worker*. According to the ILO, the occupation is **56 % female** and **44 % male**. SD, by contrast, produces images that are **95 % female** and only **5 % male**. Thus, even though the real-world distribution already leans female, the model magnifies this tendency by almost **+40% points**.

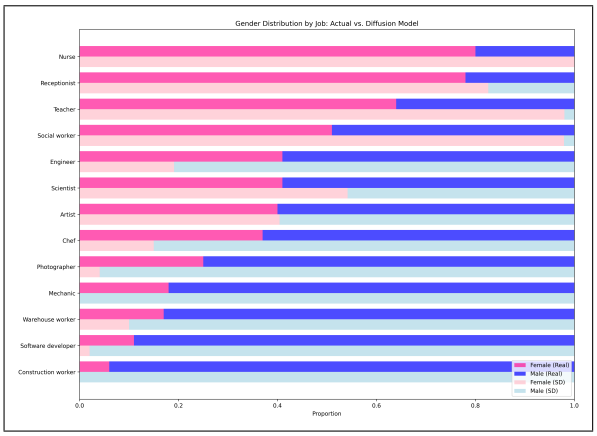

Fig. 2: Gender distribution across professions: real-world statistics, ILO versus Stable Diffusion outputs.

Ethnicity Distribution in the Real World vs. Stable Diffusion SD's behaviour with respect to *ethnicity* shows a strong **White-dominant bias**. In **all but one** of the occupations we analysed, more than 90 % of the generated faces are classified as White. The single outlier is *social worker*, where SD instead outputs a clear Black majority (80 %). Neither pattern matches empirical labour statistics, which consistently report a much more heterogeneous workforce.
These mixed results suggest that SD's ethnicity prior is driven less by labour-market reality and more by the composition of its internet training corpus. Consequently, the model sometimes mirrors real proportions, but far more frequently it produces drastic saturations or outright inversions that can misinform downstream uses especially in educational or media contexts.

4.4 Survey Results

In total, 18 people participated in the survey. A detailed overview of the demographics of the participants can be found in Appendix A. Figures 5 and 6 show the distribution of responses regarding the presence of bias

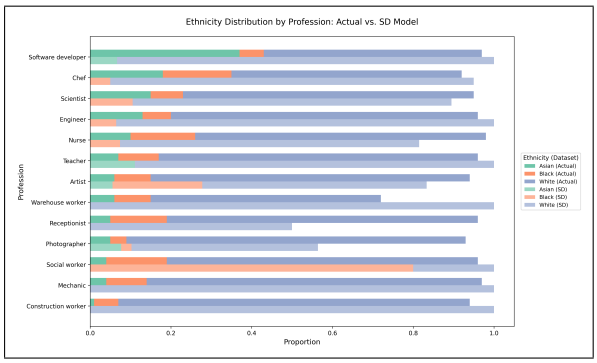

Fig. 3: Ethnicity distribution across professions: real-world statistics, ILO versus Stable Diffusion outputs.

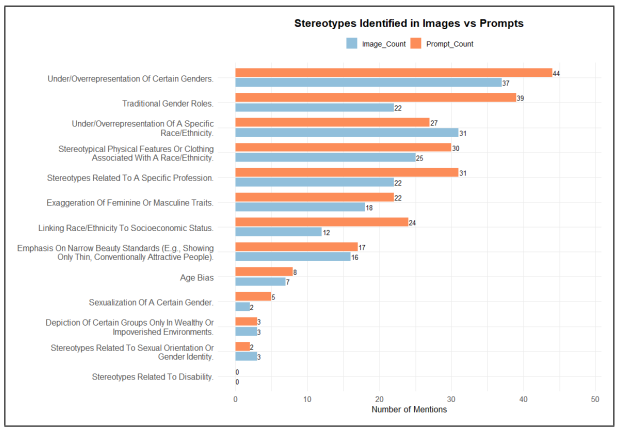

Fig. 4: Frequency of stereotype identified by respondents

across both images and prompts. Although the results indicate for men a slightly increased tendency to disagree that bias or stereotypes are present, we do not find any statistically significant relationship for either demographic attribute (gender or age-group).

Perceived Image Bias As shown in Fig. 9, 60% of respondents agreed that the images generated by the model have some demographic bias. Respondents stated that they see stereotypical connections between careers and gender/ethnic appearances far too often, even if no demographic prompt was used. This supports our findings of image-level defaults and visual consistency.

Perceived Prompt Bias In Fig. 10, we see that many respondents also perceived bias within the prompts themselves. That is, the way prompts were constructed, especially by adding demographic indicators, was seen as stereotype. For example, some participants noted that adding ethnicity and gender often stacked cliches instead of neutralizing bias. This aligns with our observation that intersectional prompts resulted in constrained or cartoonish representations.

Stereotype Recognition Fig. 4 shows which stereotypes were most frequently recognized by the participants. The results support the claim that generative models do not reflect society but impose dominant characteristics of their training data. Moreover, most participants recognized more stereotypes after discovering the used image prompts, shown in Fig. 4. It is to assume that many participants know of an existing image bias within AI generated images, but could not recognize its scale. Images that were generated with the prompt *Software developer* were by most participants assumed to be generated by the prompt *White software developer*. This is a trend that runs through all the answers from the participants to the questions about which prompts were used to generate those images.

4.5 Discussion

We explored how the model portrays different professions when provided with prompts of varying demographic specificity. Our ethical evaluation focused on three major concerns derived from the data like representational bias, stereotyping, and bias amplification. Each of these showed the moral flaws of today's generative AI systems and made us think about issues of justice, recognition, and harm. Here is a brief discussion of those:

Representational Bias Let us say that when we gave SD 3.5 open-ended prompts that merely mentioned a job without any gender or race indicators, it always made pictures of white guys, especially for jobs like *scientist* or *construction worker*. This pattern suggested to us a deeply embedded default bias, where whiteness and maleness are treated as the norm. But from an ethical standpoint if we see it, this is problematic because it hides other identities unless explicitly requested, and kind of sends a message that certain people only *belong* in certain professions, if we specify them like relegating diversity to an optional filter, not a built-in assumption. And it can be said we found this to be a subtle but powerful form of exclusion. The default outputs reveal what the model has internalized as typical, and those defaults have social consequences, like we can say that they reflect and reinforce existing inequalities in representation and opportunity.

Stereotyping and Intersectional Failure When we added demographic markers such as *Black female social worker* or *Asian male nurse*, we expected more diverse and realistic results. However, what we observed instead was stereotype stacking. Rather than correcting the bias, the model often layered gendered and racial clichés together, resulting in visuals that were inac-

curate, low in quality, or clearly stereotyped. But a common observation here was that failure was more evident in prompts involving intersectional identities. Some professions, like *Black female engineer*, either generated images with visual glitches or outputs that did not represent the intended identity at all, which kind of suggests the model struggles to represent complex identities when they do not match dominant training distributions. Ethically, this reflects a limitation in recognition and respect like the model lacks the representational capacity to depict intersectionality accurately, which leads to distorted or tokenized images of marginalized groups.

Bias Amplification Another core finding in our ethical evaluation was the amplification of existing social biases because certain identity-profession combinations, such as *Asian male nurse*, yielded fewer or less realistic outputs, suggesting that the used model has learned not only who *typically* holds a job but also who does not. Like professions with strong gender or racial associations in the training data were especially prone to biased representations. For example, doctors were more often depicted as white men, while nurses skewed toward white or Asian women. So, this amplification effect matters ethically because it does not just reflect real-world inequalities; it kind of reinforces them, and by normalizing skewed visuals of who can be what, the model feeds back into the same systems of bias it has learned from. The problem here is that it risks entrenching the very stereotypes that social institutions are working to dismantle.

Ethical Consequences of Visual Bias So, the ethical consequences of these biases go beyond aesthetics or fairness in images. Actually, they touch on symbolic power, which is the power to define who is seen, how they are seen, and who is erased, and all those biased outputs in text-to-image systems affect how we imagine social roles, how we perceive others, and how we perceive ourselves. And our experiment showed that diffusion models are not passive tools, but they are active participants in shaping cultural imagery in our daily life as well. So, it can be said that the ethical stakes here are more than just representational equity, but also about reinforcing or resisting structural inequality in our society as well. When models default to white men or fail to depict Black women professionals accurately, they contribute to a digital landscape where certain people remain underrepresented or misrepresented, even in fictional or generative spaces.

4.6 Limitations

The results of the survey indicate that there are no significant effects of demographic background on the perception of bias in generated images. It has to be remarked, however, that the sample was small ($n = 18$, and that most participants likely currently live in Europe and are thus likely influenced by the surrounding cultural background. The analysis of bias-perception amongst different social and cultural groups thus remains an interesting direction for further research.
Additionally, our study did not investigate the mechanisms that produce the biased outcomes. Such an analysis could involve looking at the internals of the model or the datasets. A preliminary, but shallow, analysis of the text encoders of SD 3.5 can be found in Appendix D.

5 Conclusion

Artificial intelligence carries the weight of the data it was trained on. Our study shows that SD 3.5 does not reflect real-world demographic distributions but instead reinforces and exacerbates stereotypes and biases in its portrayal of professions. Gendered and racial defaults — such as White men in technical roles and women in caregiving positions — dominate, while intersectional identities remain underrepresented.
We found patterns of demographic bias misaligned with real-world statistics, which raises ethical concerns regarding stereotyping, erasure, and non-recognition. Additionally, there exist certain defaults regarding ethnicity, gender, and age of the representations.
Representational bias, stereotype stacking, failures with respect to intersectional identities, and bias amplification — the system is not just generating neutral images; instead, it quietly repeats and shapes harmful assumptions about who is who and who fits where in society.
Increasing the specificity and thereby the context of the prompt did not result in an equalization of representations, instead it showed only marginal — and occasionally even harmful — effects. One especially concerning result is the illusionary reduction of the gender representation gap caused by stereotype stacking.
Overall, our findings made it clear that TTI systems like SD 3.5 reflect a specific worldview which raises issues of fairness and the lack of recognition and visibility that many groups face in both digital and physical spaces.

References

1. Hardebolle, C., Héder, M., Ramachandran, V.: Engineering ethics education and artificial intelligence. In: The Routledge International Handbook of Engineering Ethics

Education, 1 edn., p. 125–142. Routledge, London (2024). DOI 10.4324/9781003464259-9
2. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR (2018). URL https://proceedings.mlr.press/v81/buolamwini18a.html. ISSN: 2640-3498
3. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022). URL https://openreview.net/forum?id=M3Y74vmsMcY
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. CoRR **abs/2112.10752** (2021). URL https://arxiv.org/abs/2112.10752
5. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes (2021). URL https://arxiv.org/abs/2110.01963
6. de Caleya Vázquez, A.F., Garrido-Merchán, E.C.: A taxonomy of the biases of the images created by generative artificial intelligence (2024). URL https://arxiv.org/abs/2407.01556
7. Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., Chang, K.W.: Survey of bias in text-to-image generation: Definition, evaluation, and mitigation (2024). URL https://arxiv.org/abs/2404.01030
8. Sandoval-Martin, T., Martínez-Sanzo, E.: Perpetuation of gender bias in visual representation of professions in the generative ai tools dall·e and bing image creator. Social Sciences **13**(5) (2024). DOI 10.3390/socsci13050250. URL https://www.mdpi.com/2076-0760/13/5/250
9. Böckling, F., Marquenie, J., Siegert, I.: Effect of gender fair job description on generative ai images (2025). URL https://arxiv.org/abs/2503.05769
10. Wu, Y., Nakashima, Y., Garcia, N.: Revealing gender bias from prompt to image in stable diffusion. Journal of Imaging **11**(2) (2025). DOI 10.3390/jimaging11020035. URL https://www.mdpi.com/2313-433X/11/2/35
11. Guilbeault, D., Delecourt, S., Hull, T., Desikan, B.S., Chu, M., Nadler, E.: Online images amplify gender bias. Nature **626**(8001), 1049–1055 (2024). DOI 10.1038/s41586-024-07068-x. URL https://www.nature.com/articles/s41586-024-07068-x
12. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, p. 1493–1504. Association for Computing Machinery, New York, NY, USA (2023). DOI 10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095
13. Seshadri, P., Singh, S., Elazar, Y.: The bias amplification paradox in text-to-image generation. In: K. Duh, H. Gomez, S. Bethard (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6367–6384. Association for Computational Linguistics, Mexico City, Mexico (2024). DOI 10.18653/v1/2024.naacl-long.353. URL https://aclanthology.org/2024.naacl-long.353/
14. Yang, Y.: Racial bias in AI-generated images. AI & SOCIETY (2025). DOI 10.1007/s00146-025-02282-1. URL https://link.springer.com/10.1007/s00146-025-02282-1
15. Naik, R., Nushi, B.: Social biases through the text-to-image generation lens (2023). URL https://arxiv.org/abs/2304.06034
16. Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023). URL https://openreview.net/forum?id=qVXYU3F017
17. Lan, X., An, J., Guo, Y., Tong, C., Cai, X., Zhang, J.: Imagining the far east: Exploring perceived biases in ai-generated images of east asian women (2025). URL https://arxiv.org/abs/2504.04865
18. Birhane, A., Dehdashtian, S., Prabhu, V., Boddeti, V.: The dark side of dataset scaling: Evaluating racial classification in multimodal models. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, p. 1229–1244. ACM (2024). DOI 10.1145/3630106.3658968. URL http://dx.doi.org/10.1145/3630106.3658968
19. Shukla, P., Chinchure, A., Diana, E., Tolbert, A., Hosanagar, K., Balasubramanian, V.N., Sigal, L., Turk, M.A.: Biasconnect: Investigating bias interactions in text-to-image models (2025). URL https://arxiv.org/abs/2503.09763
20. D'Incà, M., Peruzzo, E., Mancini, M., Xu, D., Goel, V., Xu, X., Wang, Z., Shi, H., Sebe, N.: Openbias: Open-set bias detection in text-to-image generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12225–12235 (2024)
21. Howard, P., Madasu, A., Le, T., Moreno, G.L., Bhiwandiwalla, A., Lal, V.: Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11975–11985 (2024). DOI 10.1109/CVPR52733.2024.01138
22. Kim, E., Kim, S., Park, M., Entezari, R., Yoon, S.: Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion (2025). URL https://arxiv.org/abs/2408.12692
23. Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., Kersting, K.: Auditing and instructing text-to-image generation models on fairness. AI and Ethics **5**(3), 2103–2123 (2025). DOI 10.1007/s43681-024-00531-5
24. Wang, A., Phan, M., Ho, D.E., Koyejo, S.: Fairness through difference awareness: Measuring desired group discrimination in llms (2025). URL https://arxiv.org/abs/2502.01926
25. Limani, D.: Where women work: female-dominated occupations and sectors (2023). URL https://ilostat.ilo.org/blog/where-women-work-female-dominated-occupations.-and-sectors/. ILOSTAT Blog
26. Eurostat: 41 URL https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20240212-1. Eurostat News article
27. Deutsches Musikinformationszentrum: Gender ratios in professional orchestras (2021).

URL https://miz.org/en/statistics/gender-ratios-in-professional-orchestras. Statistic
28. Marathe, A., Desai, A., Walambe, R., Kotecha, K.: Identifying and mitigating bias in ai-generated image datasets for better cognitive understanding. In: C. Kahraman, S. Cevik Onar, S. Cebi, B. Oztaysi, A.C. Tolga, İ. Ucal Sari (eds.) Intelligent and Fuzzy Systems. INFUS 2024, *Lecture Notes in Networks and Systems*, vol. 1088, pp. 176–184. Springer, Cham (2024). DOI 10.1007/978-3-031-70018-7_20. URL https://doi.org/10.1007/978-3-031-70018-7_20
29. Statistisches Bundesamt (Destatis): Women in managerial occupations (2025). URL https://www.destatis.de/EN/Themes/Labour/Labour-Market/Quality-Employment/Dimension1/1_4_WomanManagerialOccupation.html. Quality of employment indicator
30. CEIC Data / World Bank: Germany de: Secondary education teachers: URL https://www.ceicdata.com/en/germany/social-education-statistics/de-secondary-education-teachersfemale. Dataset

A Survey

To elicitate how the perception of bias and stereotypes differs across different social groups, we conducted a survey. All survey-related artefacts as well as the survey itself can be accessed via the Google drive[1].

A.1 Setup

The survey consisted out of five image-prompt pairs. After asking the participants demographic-related questions to determine their age-group, educational background, and cultural background, they were shown the images without the corresponding prompt and were asked to

- Describe what they can see on the images.
- Rate the degree of stereotypical depiction (Five point Likert-scale)
- Indicate which stereotypes were contained (only if the images were marked as containing stereotypes)
- Rank four prompts of different specificity as to which prompt is most likely to have produced the images.

Afterwards, they were shown the same image sets but now with the corresponding prompts that produced them. Again, they were asked to describe, rate, and indicate which stereotypes were contained in the prompts.
In total, 18 people participated in the survey. The demographic statistics of the participants are shown in Table 2.

Table 2: Demographics of Survey Participants. For cultural background it was possible to select more than one option.

Demographic	Distribution
Total Participants	**n = 18**
Age Group	
18-24	5
25-29	11
30-34	2
Cultural Background	
African	1
Asian	9
European	7
North American	2
Gender	
Female	9
Male	9
Education Level	
Higher Education	13
Lower Education	5

A.2 Results

Figures 5 and 6 show how the distribution of responses regarding the presence of bias across both images and prompts. Although the results indicate a slightly increased tendency to disagree that bias or stereotypes are present in men, we do not find any statistically significant relationship for either demographic attribute (gender or age-group).

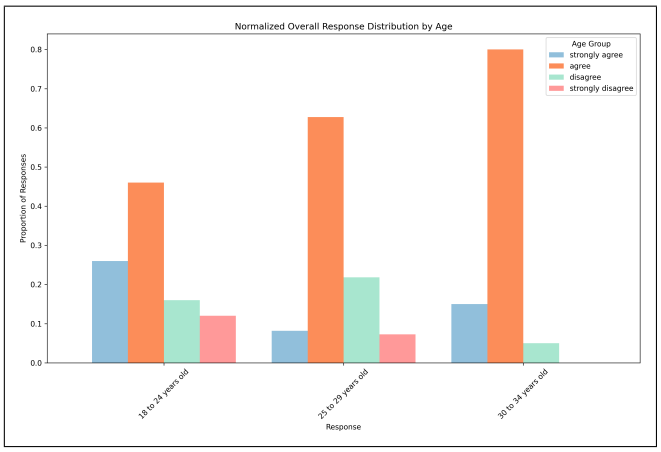

Fig. 5: Extent to which participants concur that the images are biased based on age.

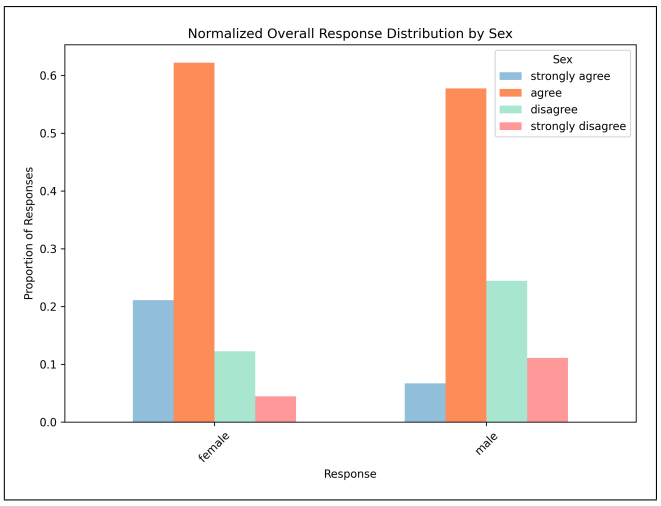

Fig. 6: Extent to which participants concur that the images are biased based on gender.

The analysis of the image-prompt pairs shows that the level of bias perceived between in both images and prompts remains relatively stable (see Fig. 7). For detected stereotypes, we found that representational biases with respect to gender were most identified (see Fig. 8).

[1] Drive: results_questionnaire

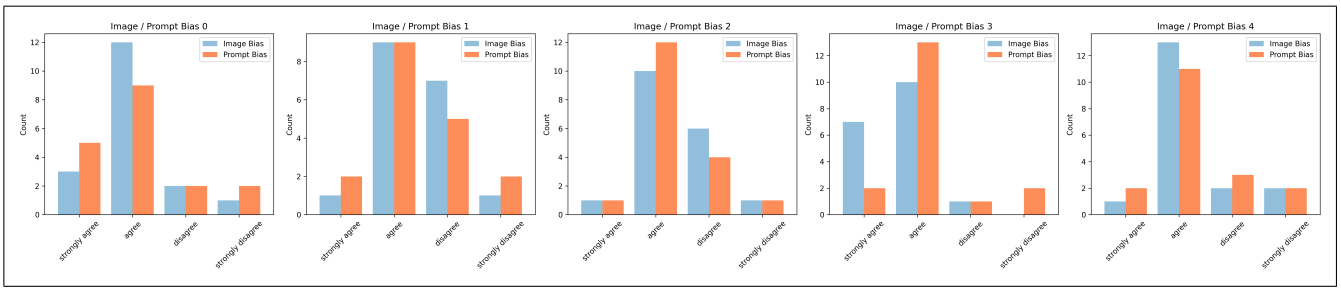

Fig. 7: Perceived level of bias / stereotypes in the images and prompts for each image-prompt pair.

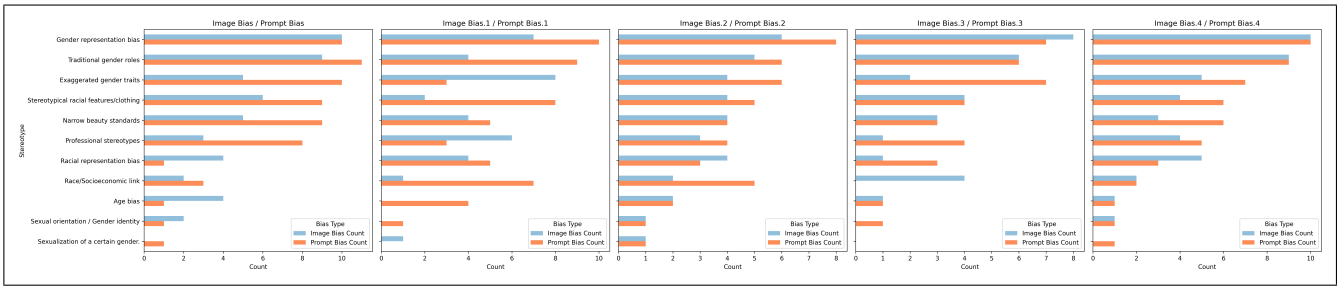

Fig. 8: Perceived types of bias / stereotypes in the images and prompts.

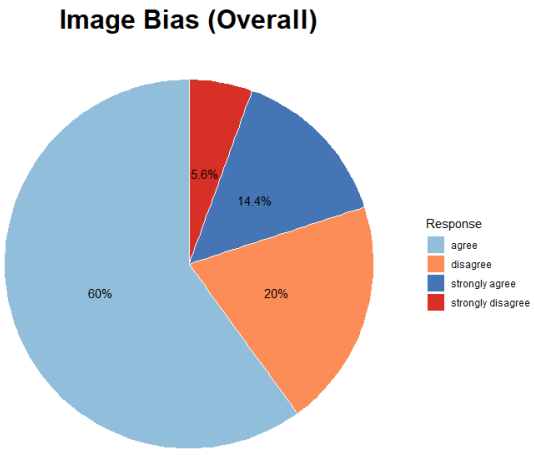

Fig. 9: Survey results: Perception of bias in generated images

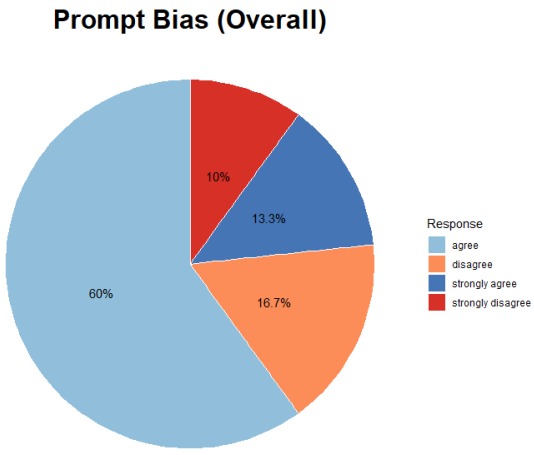

Fig. 10: Survey results: Perception of bias in prompt formulation

B Graphs and Figures

B.1 Prompt Design

Figure 11 is an overview of the progressive prompt structure used in our study that ranges from no context to intersectional identity, including gender, ethnicity, and profession. This helps us to test how the model responds to layered identity information and whether increased specificity reduces or compounds stereotypical outputs. The prompt dataset is shown in Figure 12.

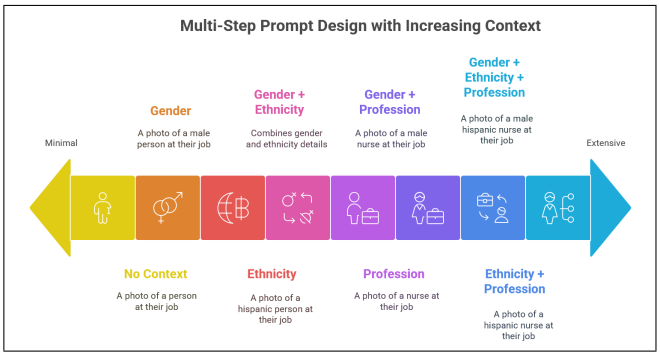

Fig. 11: Multi-Step Prompt Design with Increasing Context

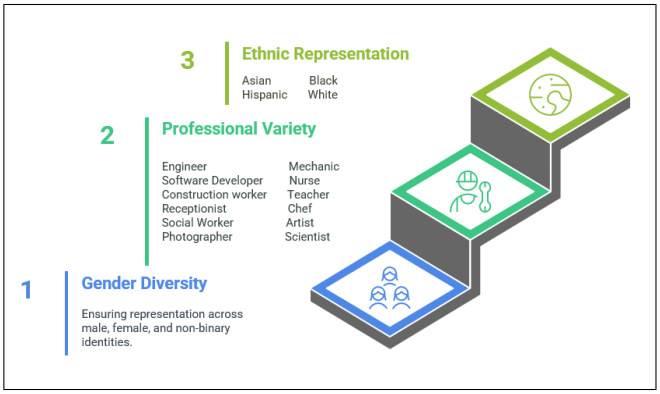

Fig. 12: Multi-Step Prompt Design with Increasing Context

B.2 Samples of Generated Images

To illustrate how image outputs vary by prompt specificity, we include sample generations based on different prompt constructions. These range from profession-only prompts to those that layer gender and ethnicity. The samples reflect how input attributes influence the visual portrayal of social workers.

Fig. 13: Prompt: Profession = Social Worker

Fig. 14: Prompt: Gender + Profession = Female Social Worker

Fig. 15: Prompt: Gender + Ethnicity + Profession = Female Black Social Worker

B.3 Statistical Distributions of Results

The following figures summarize the demographic characteristics, such as gender 16, ethnicity 17, and age 18 are observed across the full set of generated images. These aggregated distributions help quantify patterns and potential biases in the model's visual outputs.

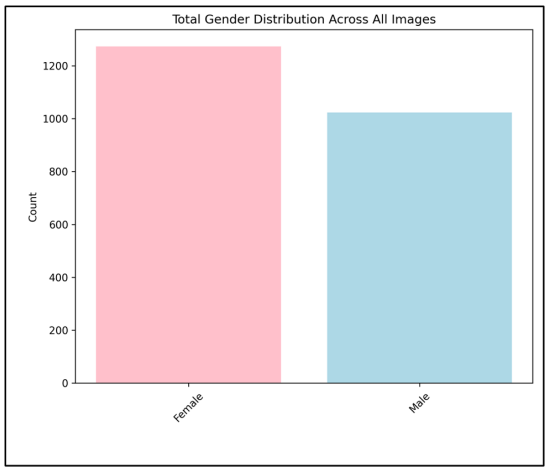

Fig. 16: Total Gender Distribution Across All Images

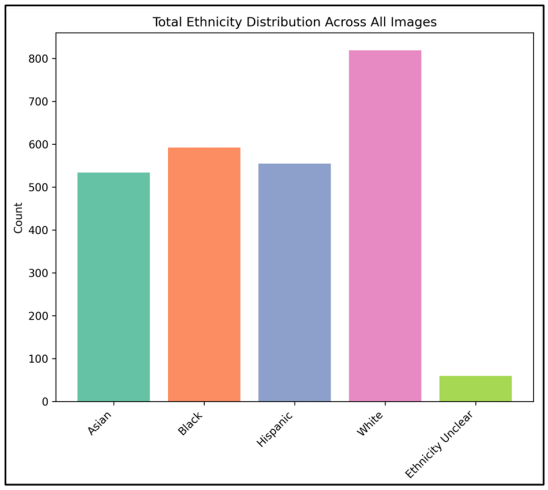

Fig. 17: Total Ethnicity Distribution Across All Images

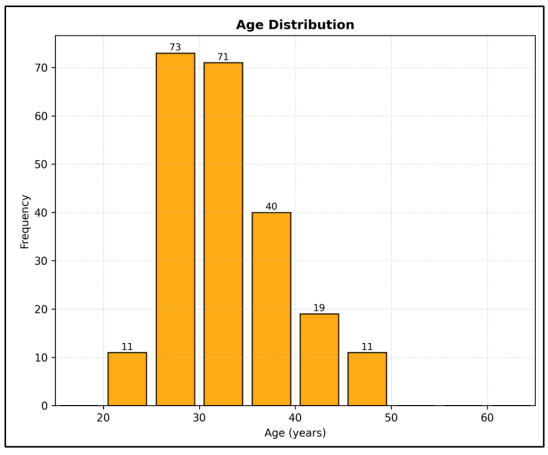

Fig. 18: Age Distribution Across All Images

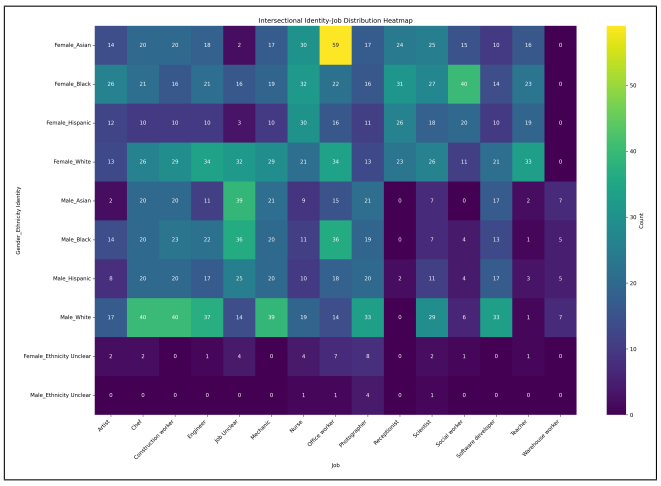

Fig. 19: Intersectional Identity-Job Distribution Heatmap)

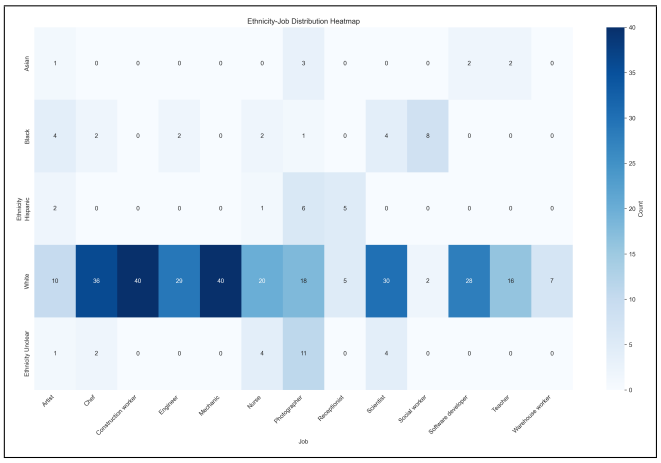

Fig. 20: Ethnicity-Job Distribution Heatmap (only including prompts without any gender marker)

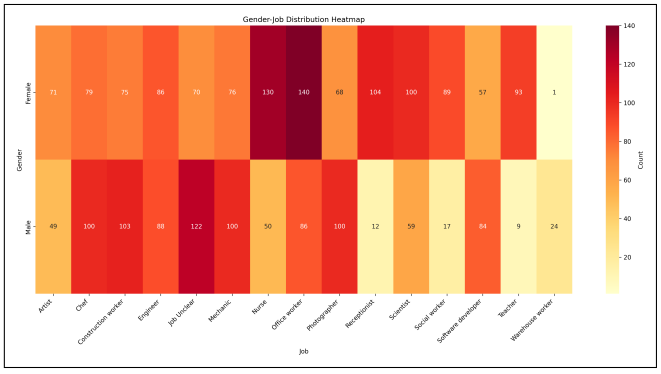

Fig. 21: Gender-Job Distribution Heatmap (only including prompts without any gender marker)

C Influence of Prompt Specificity

Table 3: Gender and Ethnicity Distribution by Prompt Specificity

Prompt Specificity	Female (%)	Male (%)	Asian (%)	Black (%)	Hispanic (%)	White (%)
Neutral	30.00	70.00	10.00	30.00	0.00	60.00
Gender	50.00	50.00	0.00	10.00	5.00	85.00
Ethnicity	41.03	58.97	25.00	25.00	25.00	25.00
Profession	35.29	64.71	1.69	11.02	5.93	81.36
Gender and profession	50.00	50.00	4.04	10.76	3.59	81.61
Ethnicity and profession	43.10	56.90	25.48	27.18	27.60	19.75

Table 4: Ethnicity Distribution for Profession Only Prompts

Profession	Asian (%)	Black (%)	Hispanic (%)	White (%)
Artist	0.00	30.00	0.00	70.00
Chef	0.00	0.00	0.00	100.00
Construction worker	0.00	0.00	0.00	100.00
Engineer	0.00	10.00	0.00	90.00
Mechanic	0.00	0.00	0.00	100.00
Nurse	0.00	0.00	0.00	100.00
Photographer	0.00	0.00	20.00	80.00
Receptionist	0.00	0.00	50.00	50.00
Scientist	0.00	12.50	0.00	87.50
Social worker	0.00	80.00	0.00	20.00
Software developer	0.00	0.00	0.00	100.00
Teacher	20.00	0.00	0.00	80.00

Table 5: Gender Distribution for Profession Only Prompts

Profession	Female (%)	Male (%)
Artist	0.00	100.00
Chef	0.00	100.00
Construction worker	0.00	100.00
Engineer	0.00	100.00
Mechanic	0.00	100.00
Nurse	100.00	0.00
Photographer	10.00	90.00
Receptionist	100.00	0.00
Scientist	10.00	90.00
Social worker	100.00	0.00
Software developer	0.00	100.00
Teacher	100.00	0.00

Table 6: Gender Distribution for Profession and Ethnicity Prompts

Profession	Female (%)	Male (%)
Artist	50.00	50.00
Chef	0.00	100.00
Construction worker	0.00	100.00
Engineer	2.50	97.50
Mechanic	0.00	100.00
Nurse	100.00	0.00
Photographer	2.56	97.44
Receptionist	100.00	0.00
Scientist	65.79	34.21
Social worker	97.37	2.63
Software developer	2.56	97.44
Teacher	97.50	2.50

Table 7: Ethnicity Distribution for Profession and Gender Prompts

Profession	Asian (%)	Black (%)	Hispanic (%)	White (%)
Artist	22.22	22.22	27.78	27.78
Chef	0.00	0.00	0.00	100.00
Construction worker	0.00	0.00	0.00	100.00
Engineer	0.00	0.00	0.00	100.00
Mechanic	0.00	0.00	0.00	100.00
Nurse	0.00	10.53	0.00	89.47
Photographer	20.00	0.00	20.00	60.00
Receptionist	0.00	10.53	0.00	89.47
Scientist	0.00	15.79	0.00	84.21
Social worker	0.00	61.11	5.56	33.33
Software developer	15.00	0.00	0.00	85.00
Teacher	0.00	10.00	0.00	90.00

Table 8: Ethnicity Distribution for Profession, Gender, and Ethnicity Prompts

Profession	Asian (%)	Black (%)	Hispanic (%)	White (%)
Artist	25.42	27.12	23.73	23.73
Chef	25.00	26.67	25.83	22.50
Construction worker	25.00	25.83	25.00	24.17
Engineer	25.00	25.83	25.83	23.33
Mechanic	25.00	25.83	25.00	24.17
Nurse	24.37	26.89	25.21	23.53
Photographer	26.05	25.21	26.05	22.69
Receptionist	27.52	27.52	28.44	16.51
Scientist	25.00	25.00	25.83	24.17
Social worker	25.64	27.35	26.50	20.51
Software developer	25.21	25.21	25.21	24.37
Teacher	25.21	26.05	24.37	24.37

Table 9: Gender Distribution for Profession, Gender, and Ethnicity Prompts

Profession	Female (%)	Male (%)
Artist	55.56	44.44
Chef	59.60	40.40
Construction worker	56.57	43.43
Engineer	60.00	40.00
Mechanic	58.16	41.84
Nurse	60.00	40.00
Photographer	58.16	41.84
Receptionist	60.00	40.00
Scientist	59.18	40.82
Social worker	60.61	39.39
Software developer	59.18	40.82
Teacher	60.00	40.00

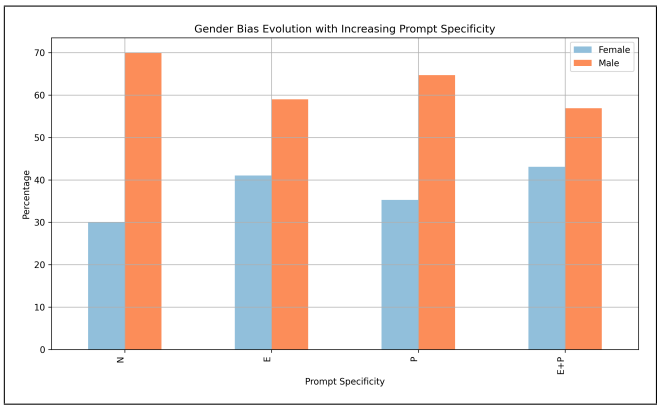

Fig. 22: Gender distribution across varying levels of prompt specificity (N: neutral, G: gender, P: profession).

D Text Encoder Attention Maps

The architecture of Stable Diffusion is composed of several components: a variational autoencoder (VAE) that maps the between the pixel space and a lower-dimensional latent space, a diffusion model (historically a UNet, but more recently a Rectified Flow Transformer in versions like Stable Diffusion 3) that denoises a random tensor to generate the final image, and one or more text encoders.
A potential source of algorithmic bias in the generated imagery can be traced to the text-to-image conditioning mechanism, specifically the text encoders. These models are responsible for transforming a user's textual prompt into a semantic embedding within the latent space. This embedding guides the diffusion process.
In models like Stable Diffusion 3.5, which utilize multiple text encoders, these components may exhibit representational biases learned from their training data. For example, the attention mechanisms in the encoders might assign different weights to semantically similar terms. Consider the prompts

- *a white nurse*
- *a hispanic nurse.*

If the text encoder disproportionately amplifies the token for *hispanic* relative to *white*, it suggests a *white-as-default* bias. This skewed attentional weighting would result in a modified semantic vector in the latent space, which is then propagated through the diffusion backbone. This could lead to bias in the final image output, where certain attributes are over or underrepresented due to these initial, subtle skews in the text embedding process.
To test the hypothesis that attention asymmetries in the text encoders contribute to the observed output biases, we analyzed the attention maps generated from minimal prompt pairs (e.g., *a white nurse* vs. *a hispanic nurse*)[2].

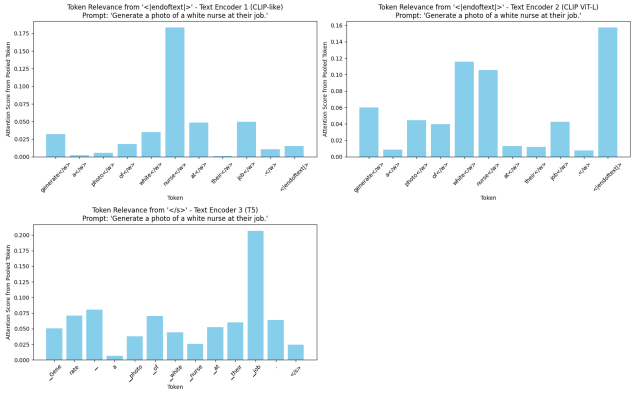

Fig. 23: Token relevance scores at the EOS Token for the sentence *Generate a photo of a white nurse at their job.*

Our analysis of the text encoders' self-attention maps did not reveal patterns that consistently correlated with the documented biases in the generated images. Some instances showed amplification of the bias marker for the marginalized group (e.g., higher attention on *hispanic* in *hispanic nurse* compared to *white* as can be seen in Fig.23 and Fig. 24), however,

[2] The files are available at this Google Drive

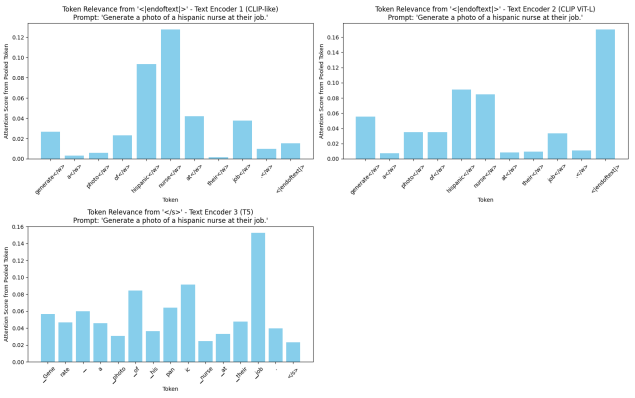

Fig. 24: Token relevance scores at the EOS Token for the sentence *Generate a photo of a hispanic nurse at their job.* For the CLIP encoder the scores show a clear weighting of the marked attribute *hispanic*.

this behavior was not systematic. The lack of consistency was observed across different minimal pairs and even among the various text encoders processing the same prompts. This suggests that the text encoder's self-attention may not be the primary or direct source of the semantic bias influencing the final image synthesis.
However, a more robust view could be achieved when looking at the cross-attention mechanism inside of the denoising transformer. The cross-attention maps correlate the spatial features of the image being generated with the semantic embeddings from the text prompt at each denoising step. This directly influences how textual concepts are transported into the pixel space. Analyzing these maps could offer a more granular and robust understanding of how specific tokens guide the image generation process and introduce biases. This was, however, outside of the scope of this seminar paper, but might be an interesting research direction for the future.

E List of Contributions

Following, we list the contributions that were done over the course of the project.

Seychelle Ann Dagus Participated in the self-analysis of the generated images. Responsible for literature related to real-world statistics. Main editor of the interim presentation. Aggregated global data for analysis and conducted data analysis for the survey data. Organization of meetings. For this paper, she wrote Sections 4.1, 4.3, and 4.4. For the interim presentation she was reponsible for presenting the results of the dataset.

Zain Abladen Hazzouri Participated in the self-analysis of the generated images. Prepared the prompt dataset, and was mainly responsible for the statistical analyses of the self-analysis data. Additionally, he helped in comparing the results with real-world statistics. For this paper, he wrote Section 4.2. For the interim presentation he was reponsible for presenting the comparison to real-world statistics.

Muhammad Hassan Jalil Participated in the self-analysis of the generated images. For this paper, he wrote Section 4.5. For the interim presentation he was reponsible for the second part of the ethical evaluation.

Laura Rivero Miró Participated in the self-analysis of the generated images. For this paper, she wrote Section 1 and parts of the Abstract. For the interim presentation she was reponsible for presenting the introduction.

Carlos March Moya Participated in the self-analysis of the generated images. For this paper, he wrote Section 3. For the interim presentation he was reponsible for presenting the methodology.

Noah Vara Marquez Participated in the self-analysis of the generated images.

Erik Voigt Participated in the self-analysis of the generated images. Built the model pipeline and technical realization of the experiment. Set up and evaluated the survey. Supported the analysis and visualization of the self-analysis data. Analyzed the influence of prompt specificity on demographic representation. He also conducted further experiments on the transformer text encoder attention. Conducted literature research related to bias in text-to-image models. For this paper, he mainly wrote Section 2, but supported in writing the Abstract and Sections 1, 4.1, and 5. Responsible for editing the paper. For the interim presentation he was reponsible for presenting the first part of the ethical evaluation as well as the conclusion.

Muhammed Enes Yavuz Prepared the prompt dataset and helped in conducting the literature search. He also participated in the self-analysis of the generated images and the analysis of the survey data. For this paper, he wrote Section 5. For the interim presentation he was reponsible for presenting the reflections and limitations.