# Analysing Gender Bias in Text-to-Image Models using Object Detection

**Harvey Mannering**

**University College London**

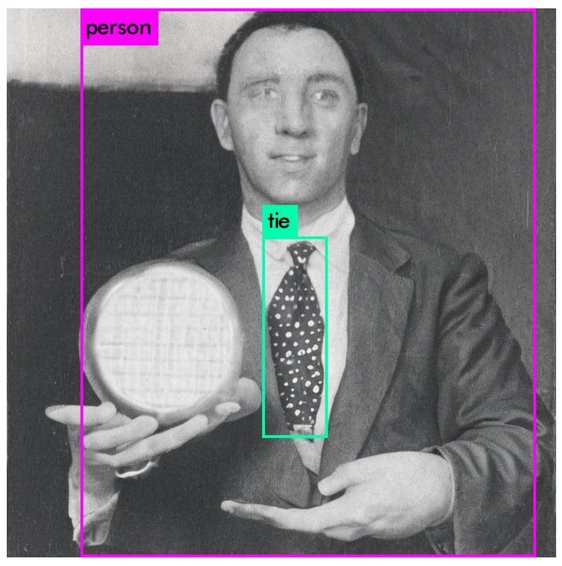
## Introduction & Background

Text-to-image models have diverse applications; however, they can also exhibit learned stereotypes from their training data, including gender bias. For instance, using the prompt "a photo of a face of a software developer" resulted in biased outputs, generating only male images [1]. Other methods also measure gender bias with male/female prompts [2,3], we differ by analysing the scene using an object detector.

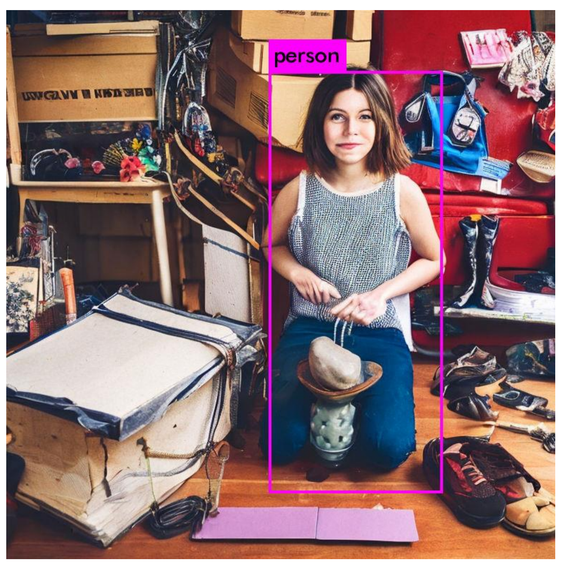
## Method

- Generate many gender-paired images using a text-to-image model
- By keeping the prompts vague, we can examine how text- to-image models “fill in the blanks” with regards to the objects in the scene.
- Run object detection on images
- Are certain objects associated with certain genders?

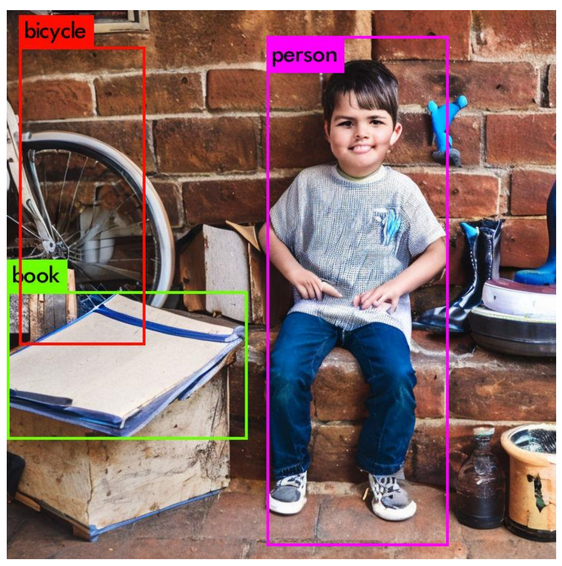

“A **woman** holding an item”




“A **man** holding an item”



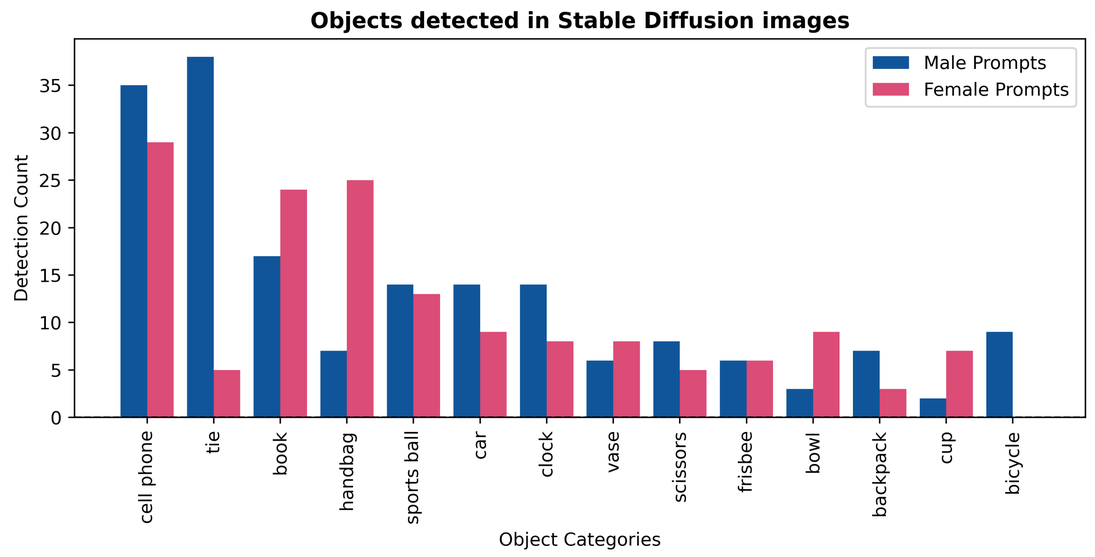

“A photo of a **girl** with things they own”

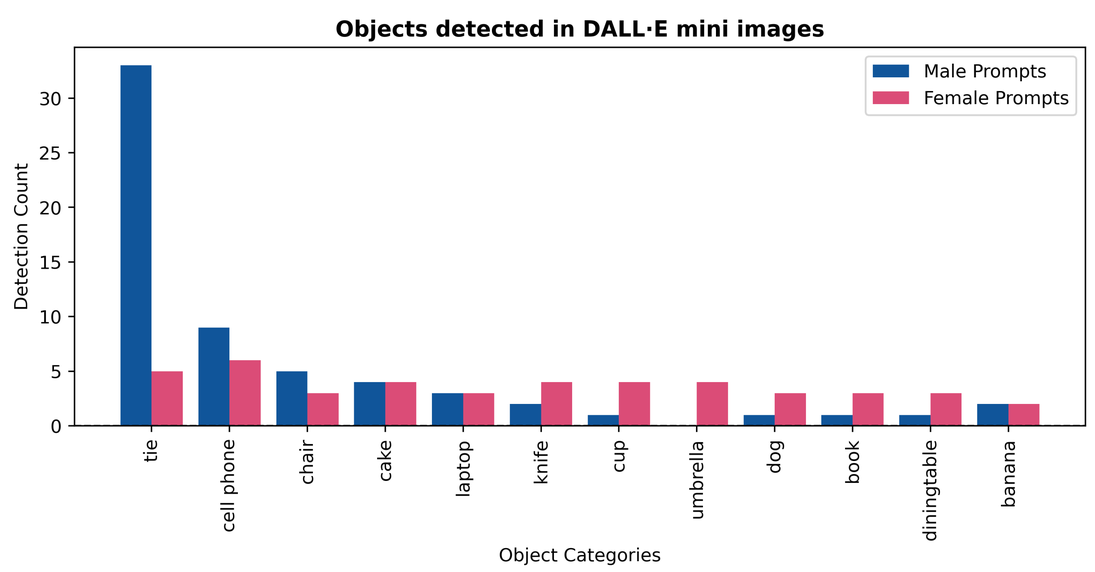



“A photo of a **boy** with things they own”

## Results

- Male prompts were more likely to generate objects including ties, knives, trucks, baseball bats and bicycles (for Stable Diffusion).
- Female prompts were more likely to generate objects including handbags, umbrellas, bowls, bottles, and cups (for Stable Diffusion).
- DALL·E showed less bias than Stable Diffusion




## Conclusion

This technique opens avenues for further investigation into bias in these models. For instance, our findings indicate that Stable Diffusion tends to generate bowls, bottles, and cups more frequently for women than men. This prompts us to question whether Stable Diffusion portrays women in domestic settings more often than men, reinforcing certain stereotypes.

## References

[1] Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023.

[2] Zhang, Yanzhe, et al. "Auditing gender presentation differences in text-to-image models." *arXiv preprint arXiv:2302.03675* (2023).

[3] Mandal, Abhishek, Susan Leavy, and Suzanne Little. "Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models." *arXiv preprint arXiv:2304.13855* (2023).

harvey.mannering@ucl.ac.uk

https://github.com/harveymannering/Text-to-Image-Bias