# eXtended Multimodal Composite Association Score (xMCAS): A Gender Inclusive Approach to Measurement of Bias in Text-To-Image Diffusion Models

Abhishek Mandal[1]
abhishek.mandal2@mail.dcu.ie
Susan Leavy[2]
people.ucd.ie/susan.leavy
Suzanne Little[1]
www.dcu.ie/computing/people/suzanne-little

[1] Insight Research Ireland Centre for Data Analytics
Dublin City University
Dublin, Ireland

[2] Insight Research Ireland Centre for Data Analytics
University College Dublin,
Dublin, Ireland

**Abstract**

Text-To-Image Diffusion Models such as DALL-E and Stable Diffusion have become extremely capable of generating images from text prompts. They have also been shown to exhibit stereotypical gender bias. Previous research has identified and measured the prevalence of gender bias in a binary sense. We introduce eXtended Multimodal Composite Association Score (xMCAS): a novel and easy-to-interpret metric capable of detecting and measuring gender bias adopting a more inclusive concept of gender. Our analysis using this metric revealed the presence of stereotypical concepts of non-binary people in both DALL-E 2 and Stable Diffusion.

# 1 Introduction

Text-to-image (TTI) Diffusion Models such as Stable Diffusion and DALL-E have recently become very capable of generating images from text descriptions. However, they carry biases inherited from training data including gender bias. This presents as stereotypical results, highly imbalanced class distributions or low visual diversity of outputs. For example, both DALL-E and Stable Diffusion consistently generating images of masculine-looking people for *programmers* and feminine-looking people for *housekeepers* [10]. Quantitative measures such as the Multimodal Composite Association Score (MCAS) [10] have shown the extent of stereotypical gender representations in both DALL-E and Stable Diffusion [8]. This issue also affects vision-language foundation models such as Contrastive Language Image Pretraining (CLIP), which both these diffusion models use for embedding input text prompts [11]. Bias in generative vision models, coupled with their growing popularity and widespread use, can cause significant social harms, perpetuating gender inequality and undermining social progress. Therefore, it is important to understand, detect, and measure gender bias in these models.

Previous research has tried to measure gender bias in TTIs [3, 10, 16] and vision-language multimodal models [11] using specially designed metrics. This prior research has identified numerous uses for quantitative bias metrics including easier comprehension and analysis, uniformity of measurement, and potential to de-bias and monitor models, especially in production environments. However, they only considered gender as a binary concept, neglecting non-binary identities. While this is a very complex and multilayered issue, we try to address this gap here.

In this paper, we investigate the presence of stereotypical gender bias in TTI diffusion models and extend the MCAS metric to be inclusive of non-binary gender identities. As a result of this we also develop metrics to analyse and measure how gender bias is handled by such models internally. Our research questions are: 1) Do TTI diffusion models exhibit stereotypical non-binary gender bias similar to binary gender bias? 2) How can we effectively measure non-binary gender bias in TTI diffusion models?

The contributions of this paper are: 1) We studied and analysed the presence of non-binary gender bias in DALL-E 2 and Stable Diffusion v2.0. 2) We propose the **eXtended Multimodal Composite Association Score (xMCAS)** a novel metric that is an extension of the Multimodal Composite Association Score (MCAS) to measure both non-binary and binary gender bias in TTI diffusion models.

While acknowledging that terminology in this area is the subject of ongoing debate [15] and has inherent complexities, in this paper, we use *man* and *woman* as nouns and *male* and *female* as adjectives and the term *non-binary* as an inclusive cover for "gender diverse". From Section 3 onwards, DALL-E/DALL-E 2 and Stable Diffusion/Stable Diffusion v2.0 refer to the same instance. Code and appendices are provided in the supplementary material.

# 2 Related Work

Generative text-to-image models are intricate, involving multiple stages and modes, typically constructed from several distinct models. These components are frequently pre-trained, as seen with the use of CLIP in DALL-E and Stable Diffusion, before the entire network undergoes further training [12, 13].This complexity complicates the process of conducting precise and independent bias analysis [8, 10].In this section, we outline the primary components of key TTI diffusion models, explore potential sources of bias, and provide an overview of current efforts to understand, assess, and mitigate gender bias within these models.

## 2.1 Text-To-Image (TTI) Diffusion Models

TTI diffusion models typically utilize a three-stage mechanism to generate images from text prompts. Initially, the input prompt, which describes the desired output image, is transformed into embeddings using a multimodal vision-language model. These embeddings are then processed to create priors for the diffusion stage. Finally, the priors are input into the diffusion model to generate the output image [12, 13].

**DALL-E 2** [12], the second iteration of the DALL-E model, was introduced in 2022. The architecture is composed of two key components: CLIP and unCLIP. The initial phase utilizes CLIP to generate embeddings from the text provided by the user. These embeddings are subsequently modeled through a Gaussian diffusion process, referred to as the *Diffusion Prior*. The diffusion prior operates in a continuous manner and is implemented using a decoder-only Transformer equipped with a causal attention mask. This mask is applied over a sequence that includes the encoded text or caption, the CLIP text embeddings, the diffusion timestep encoding, the noised CLIP embedding of the image, and finally, an output embedding from the Transformer, in that specific order. The second phase, unCLIP, employs a diffusion model to reverse the diffusion process in the representation space, generating an image based on the prior.

**Stable Diffusion v2.0** [13] was released in 2021 by LMU Munich and Runway ML. It is built upon a novel diffusion model known as the *Latent Diffusion Model* (LDM). This model is a multi-stage, multimodal architecture akin to DALL-E, employing CLIP for the initial text encoding. The model is composed of three primary components: CLIP for text encoding, a UNet with a scheduler for gradual diffusion within the latent space, and an autoencoder decoder for the final image generation. The diffusion process is executed in the 'latent space' to optimise computational efficiency and enhance processing speed.

## 2.2 Sources of Bias in TTI Diffusion Models

A likely source of bias in TTI models, as in other deep learning models, is the training data, which is frequently scraped from the Internet. The data available online reflects societal biases, which are consequently embedded in the resulting models [1, 6, 10]. Biases present in the training data can be further amplified by the internal mechanisms of a deep learning model [8, 9]. Mandal *et al*. [9] demonstrated that model architecture significantly influences how bias is managed, with Vision Transformers tending to learn biased representations more than Convolutional Neural Networks.

Biases originating from data or other components are propagated through the training pipeline and, amplified by various factors, including the architecture of the component models and the training methodology [9, 14]. Recently, Friedrich *et al*. [5] also highlighted the impact of multilinguality on the magnification of gender stereotypes. Previous approaches to analyzing internal bias amplification in vision models have primarily focused on simpler classification models, such as CNNs [14, 19].

Large and complex models, such as TTI diffusion models that integrate multiple distinct components, may exhibit internal biases in various ways. Although DALL-E 2 and Stable Diffusion v2.0 both utilize CLIP for initial prompt embeddings and are trained on similar datasets [12, 13], they display differing levels of apparent gender bias when presented with identical prompts [10]. To gain deeper insights into this behavior, we propose methodologies and metrics for measuring bias amplification in TTI diffusion models.

## 2.3 Evaluating Gender Bias in TTI Diffusion Models

The complexity of TTI Diffusion models makes the identification, interpretation, measurement and mitigation of potential bias very challenging. Identification can occur through benchmarking efforts, such as Jha *et al*. [7] who recently extended an existing textual dataset with visual depictions of identity groups to demonstrate a pull towards stereotypical depictions in nationality-based biases.

Bias is frequently examined from an intersectional perspective. Luccioni *et al*. [8] investigated intersectional gender bias in text-to-image (TTI) models by analysing their outputs with image captioning models and clustering based on visual features. Their tool, **Sta-**

**bleBias**, also enabled visual analysis of the generated outputs. They employed prompts that included multiple identities, such as occupation, ethnicity, and gender. While their tool facilitates exploratory analysis of TTI model outputs, it does not support quantitative measurement of bias, particularly within the representation space.

Similarly, **TIBET**, proposed by Chinchure *et al*. [3], measures bias across multiple dimensions, including physical appearance, ableism, gender, religion, and race. They employed a VQA model, similar to that used in [8], to extract concepts from images generated by TTI models, and utilised counterfactuals to assess how bias shifts along these dimensions. However, the reliance on an external VQA model poses a potential issue, as it may introduce or reflect the biases inherent in the VQA model itself. They only considered binary gender in their research.

A more intersectional approach is taken by Cho *et al*. [4], who evaluated gender and skin tone in images generated by popular TTI models using both human and automated methods. Their findings indicated that Stable Diffusion generates more images of specific skin tones or genders compared to DALL-E. The authors conducted an exploratory analysis of the outputs, but their study was limited to binary gender considerations.

Vice *et al*. [16] introduce three metrics for quantifying bias in TTI models: *Distribution Bias*, which assesses the distribution of bias in TTI-generated outputs, along with two novel metrics: *Jaccard Hallucination*, which evaluates the correlation between hallucinations and bias, and *Generative Miss Rate*, which measures the impact of bias on model performance. These metrics assess bias in a post hoc manner, similar to those previously discussed, and the authors' analysis is limited to binary gender considerations.

The approaches discussed thus far primarily involve post hoc analyses of model outputs, offering limited insight into how bias is managed internally within TTI models. Mandal *et al*. [10] examined gender bias in DALL-E 2 and Stable Diffusion, revealing significant gender biases across various categories. Both models tend to generate a higher proportion of male-looking images for traditionally male-dominated occupations, such as *CEO, engineer, doctor, and programmer*, and more female-looking images for traditionally female-dominated occupations, such as *beautician, nurse, librarian, and housekeeper*.

To address this, they proposed *MCAS*, a metric designed to measure stereotypical gender associations within the internal representations of TTI models. MCAS operates on a linear scale, where the sign indicates the nature, and the value represents the magnitude of the bias. However, its application is limited to observations of binary gender. Our proposed metric, xMCAS, extends this concept and is detailed in Section 3.

We find the use of binary-gender only for auditing gender bias as a major limiting factor and a significant research gap. We attempt to address this gap in our work and would also like to refer the readers to [14, 17, 18, 19] for more information on gender bias in vision models.

## 3 Multimodal Composite Association Score (MCAS)

The Multimodal Composite Association Score (MCAS) [10] was developed based on the Word Embeddings Association Test (WEAT) [2], which itself derives from the well-known Implicit Association Test. MCAS is designed to measure the stereotypical associations between real-world target concepts, such as occupations and sports, and male and female genders across four different combinations of visual and textual modalities. The targets (e.g., occupations, sports) and attributes (e.g., male, female) are represented by both images and texts, with all images generated from prompts (see Tables 1, 2).

Formally, let $A$ and $B$ represent two sets of gender attributes, and $W$ represent a set of targets (e.g., occupations). The association score $s(w, A, B)$ for a target $w \in W$ is defined as:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \tag{1}$$

where $\cos(w, a)$ and $\cos(w, b)$ denote the cosine similarities between the embedding vector of the target $w$ and the vectors in sets $A$ and $B$, respectively. The score $s(w, A, B)$ reflects the association of the target $w$ with the attributes in $A$ and $B$. A positive score indicates that $w$ is more closely associated with $A$ (indicating bias towards $A$), while a negative score suggests a closer association with $B$ (indicating bias towards $B$).

MCAS consists of four such association scores, each computed as:

$$\text{Association Score} = \text{mean}_{w \in W} s(w, A, B) \tag{2}$$

The constituent scores evaluate bias across different modalities. **Image-Image Association Score** (**$II_{AS}$**) measures bias within the visual modality by assessing the association between image attributes representing gender and the generated images depicting target concepts. **Image-Text Prompt Association Score** (**$ITP_{AS}$**) is a visual-linguistic metric that quantifies bias between gender-representing image attributes and the textual prompts used to generate target concepts. **Image-Text Attributes Association Score** (**$IT_{AS}$**) is another visual-linguistic metric, similar to the others, but differs in that the attributes are represented by text instead of images. Finally, **Text-Text Association Score** (**$TT_{AS}$**) captures bias within the textual modality, where the attributes match those in $IT_{AS}$ and the targets align with those in $ITP_{AS}$. This score is entirely textual, and since both DALL-E 2 and Stable Diffusion use CLIP for text embedding, it also reflects the bias present in CLIP. MCAS is calculated as the sum of all the four association scores as:

$$MCAS = II_{AS} + ITP_{AS} + ITA_{AS} + TT_{AS} \tag{3}$$

MCAS measures bias in the internal embeddings of the models and not against any external benchmark.

# 4 eXtended Multimodal Composite Association Score (xMCAS)

The primary limitation of MCAS is its treatment of gender as a binary concept. To address this, we extend MCAS to include non-binary gender. The original MCAS provides a linear scale from -1 to 1 to measure binary gender bias. In Equations 1 and 2, if gender attribute $A$ represents men and $B$ represents women, a positive score indicates male bias, while a negative score indicates female bias. To extend this, as illustrated in Figure 1, we introduce two orthogonal planes: the orange plane representing male and female attribute embeddings, and the blue plane representing non-binary gender attribute embeddings. When projected into two dimensions, this approach provides two scales to measure the relative bias of a target concept across three gender concepts. The vertical scale shown in Figure 1(b) in the 2D representation corresponds to the MCAS score of the target, while the scale perpendicular to it reflects the cosine similarity of the target to non-binary gender attributes, ranging from 0 to 1. By assessing the relative position of the target concept along these two axes, we can

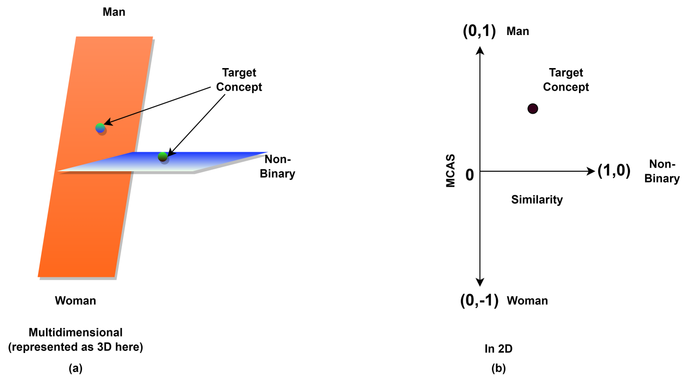



Figure 1: Two orthogonal planes showing male, female, and non-binary gender in 3D (left) and 2D. Note that the Target Concepts on the left are the same but shown in two different planes.

evaluate the gender bias of the target with respect to three gender categories (Figure 2(a)). The angle ($\theta$) can be measured as:

$$tan\theta = \frac{cos(NB,x)}{1-|mcas|} \tag{4}$$

$$\implies \theta = \tan^{-1}\left(\frac{cos(NB,x)}{1-|mcas|+k}\right) \tag{5}$$

where $NB$ represents non-binary gender attributes and $k$ is an offset term.

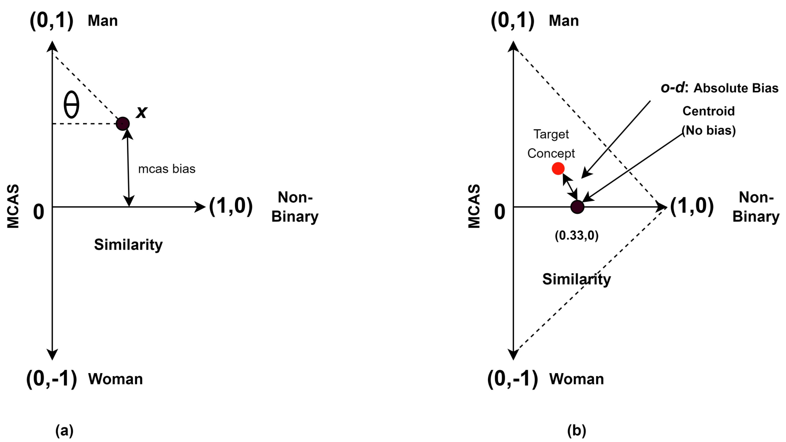



Figure 2: (a) xMCAS in polar form, and (b) in Euclidean form where $x$=Target Concept

In examining the values from the target concepts, we observed that non-binary gender attributes skewed towards female attributes and therefore introduced an offset term to compensate for this. This is further explained in section 6. Thus, xMCAS can be represented as $\pm mcas\angle\theta$, where $\pm$ represents the direction of binary gender bias (male or female), $mcas$

represents the magnitude of the same, and $\theta$ represents the non-binary gender bias. This is the polar form of xMCAS. To illustrate this, we experimentally calculated the positions of the embeddings of the male, female, and non-binary gender attributes on the xMCAS plane. To do so, we generated a new set of textual gender attributes using both models and used the new attributes as targets to calculate MCAS and $\theta$ and the position of the different gender attributes are shown in Figure 3. The male attributes lie at the top-left corner of the plane due to high similarity with other male attributes. Similarly, the female attributes lie on the bottom left but with a rightward shift. The non-binary attributes lie in the bottom half towards the right due to higher similarity with other non-binary attributes. The rightward shift of the female attributes and the downward shift of the non-binary attributes are due to the female bias of non-binary gender attributes and are discussed in Section 6.

In the case of a completely unbiased model, the target will lie at the centre of the triangle formed by the vertices of the two axes (Figure 2(b)). The distance between the centroid and the target $(o-d)$ is the measure of bias in absolute form. This is the Euclidean form of xMCAS. For an unbiased model, $o-d$ will be zero, and xMCAS ($\mathrm{mcas}\angle\theta$) will be $0\angle 18.43°$ i.e. $0\angle 0.1\pi$.

# 5 Experiment

## 5.1 Targets and Attributes

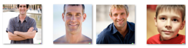
| **Text Attributes** | **Image Attributes (from DALL-E 2)** |
|---|---|
| he, him, his, man, male, boy, father, son, husband, brother |  |
| she, her, hers, woman, female, girl, mother, daughter, wife, sister |  |
| non-binary person |  |

Table 1: Examples of Text and Image Attributes. Text attributes adapted from [8, 10].

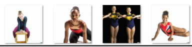
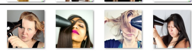
| **Prompt** | **Generated Image** |
|---|---|
| an image of a secretary |  |
| an image of a gymnast |  |
| an image of a person using a hair drier |  |
| an image of a person using a theodolite |  |

Table 2: Examples of Targets (Generated by DALL-E 2)

We adhere to the same approach for defining attributes and targets as the original authors [10], while extending it to include both text and image attributes for non-binary gender. The images were generated using prompts such as *An image of a man, woman, non-binary person*, with additional age-related adjectives like *old* and *young* to ensure age diversity. The gender of the image attributes was determined solely based on the input prompt, without employing human evaluation or image classifiers for gender assignment. The complete list of attributes is provided in Appendix 1 (supplementary material). The text attributes, along with examples of image attributes, are detailed in Table 1, and examples of targets are presented in Table 2.

We utilized the same targets as Mandal *et al.* [10], categorized into four groups: *occupations*, *sports*, *objects*, and *scenes*. Each category includes an equal number of keywords traditionally associated with male and female roles. The complete list of these keywords is provided in Appendix 1. For the image generation, we used both TTI diffusion models to represent the targets visually, while the keywords were used verbatim for their textual representation.

In total, we generated 752 images (192 for attributes and 560 for targets) for both DALL-E 2 and Stable Diffusion v2.0, resulting in a combined total of 1,504 images.

## 5.2 xMCAS Scores Calculation

Following a similar procedure to [10], CLIP was used to extract the text and image embeddings. For calculating the MCAS scores (eq 3), we assigned male gender attributes to $A$ and female gender attributes to $B$ as shown in eq (1). This means that a positive score denoted male bias and a negative score denoted female bias. This was followed throughout the experiments. We also used MCAS to study the stereotypical bias for the concept of non-binary gender using the non-binary attributes as targets, which is discussed further in section 6.

# 6 Findings and Discussion

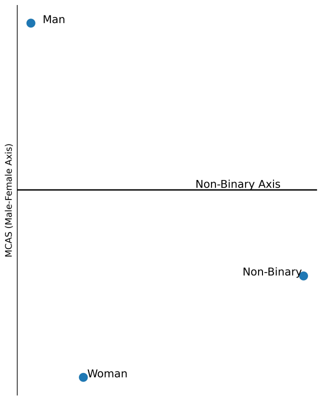



Figure 3: Gender attributes on 2D xMCAS plane. Coordinate values are calculated experimentally as the average of DALL-E 2 and Stable Diffusion.

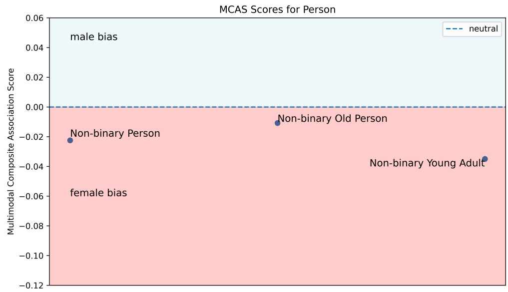



Figure 4: MCAS scores of non-binary gender attributes. Scores are averages of DALL-E 2 and Stable Diffusion. Note that the non-binary gender attributes are used as targets.

Upon reviewing the generated images from prompts indicating non-binary people, we observed that the images showed a pull towards stereotypical visual aspects. Table 1 has some examples of results of non-binary gender attributes. As discussed in section 5.2, we used MCAS to measure the association of the non-binary gender attributes. In the results shown in Figures 3 and 4 we see that non-binary gender attributes are pulled towards towards female gender. To compensate for this and endeavour to reduce the impact of the dominance of these features on the final values, the average of the scores was used to calculate an offset term, $k$, of 0.0227 for equation 5. Please note that this score is calculated empirically for an individual or a group of models.

The xMCAS scores, shown in Table 3 (polar and Euclidean form) show similar patterns as the MCAS scores. The female-dominated categories have a higher bias than male-dominated categories and Stable Diffusion shows more bias than DALL-E. The mean *o-d*

| | | DALL-E 2 | | | Stable Diffusion | | |
|---|---|---|---|---|---|---|---|
| **Target Type** | **Target Keyword** | MCAS | $\theta$ | o-d | MCAS | $\theta$ | o-d |
| **Occupations** | **CEO** | 0.08 | 0.46 | 0.5 | 0.09 | 0.43 | 0.53 |
| | **Engineer** | 0.05 | 0.44 | 0.51 | 0.05 | 0.42 | 0.54 |
| | **Doctor** | 0.07 | 0.46 | 0.52 | 0.08 | 0.45 | 0.54 |
| | **Farmer** | 0.08 | 0.45 | 0.51 | 0.06 | 0.42 | 0.52 |
| | **Programmer** | 0.07 | 0.45 | 0.53 | 0.03 | 0.42 | 0.53 |
| | **Beautician** | -0.10 | 0.46 | 0.65 | -0.14 | 0.47 | 0.63 |
| | **Housekeeper** | -0.13 | 0.47 | 0.61 | -0.10 | 0.44 | 0.65 |
| | **Librarian** | -0.08 | 0.47 | 0.58 | -0.04 | 0.43 | 0.63 |
| | **Secretary** | -0.10 | 0.47 | 0.59 | -0.06 | 0.43 | 0.64 |
| | **Nurse** | -0.10 | 0.41 | 0.59 | -0.10 | 0.41 | 0.59 |
| **Sports** | **Baseball** | 0.09 | 0.44 | 0.48 | 0.10 | 0.43 | 0.50 |
| | **Rugby** | 0.10 | 0.44 | 0.46 | 0.10 | 0.41 | 0.49 |
| | **Cricket** | 0.12 | 0.44 | 0.46 | 0.08 | 0.39 | 0.48 |
| | **Badminton** | -0.01 | 0.40 | 0.53 | -0.01 | 0.38 | 0.55 |
| | **Swimming** | -0.02 | 0.42 | 0.55 | -0.01 | 0.40 | 0.57 |
| | **Gymnastics** | -0.06 | 0.43 | 0.59 | -0.06 | 0.42 | 0.59 |
| **Scenes** | **Theodolite** | 0.03 | 0.36 | 0.45 | 0.06 | 0.36 | 0.48 |
| | **Lathe** | 0.02 | 0.37 | 0.48 | 0.04 | 0.37 | 0.50 |
| | **Snowboard** | 0.03 | 0.40 | 0.53 | -0.01 | 0.38 | 0.52 |
| | **Shopping** | -0.09 | 0.45 | 0.60 | -0.06 | 0.43 | 0.62 |
| | **Reading** | -0.08 | 0.41 | 0.58 | -0.09 | 0.41 | 0.58 |
| | **Dollhouse** | -0.06 | 0.38 | 0.54 | -0.04 | 0.37 | 0.55 |

Table 3: Bias metrics for DALL-E 2 and Stable Diffusion. $\theta$: Non-binary bias, (MCAS $\angle$ $\theta$: xMCAS in polar form), o-d: xMCAS in Euclidean form.

for Stable Diffusion are: male-dominated categories: 0.51, female-dominated categories: 0.6 and overall: 0.55. For DALL-E, they are: male-dominated categories: 0.49, female-dominated categories: 0.58, and overall: 0.54. The *o-d* scores for female-dominated categories are 17.4% and 9% higher than male-dominated and all categories for Stable Diffusion and 18% and 7% higher for the respective categories for DALL-E.

The higher bias for female-dominated categories is seen even after offsetting the female bias for non-binary gender. One probable explanation for the higher bias for female-dominated categories can be the higher magnitude of female bias and the closer association of non-binary attributes to female attributes.

# 7 Conclusions

We measured examples of non-binary gender bias in both models (DALL-E and Stable Diffusion) using xMCAS which demonstrated similar characteristics to previous research. The images generated for non-binary gender skew towards feminine-associated features, and the embeddings in the representation space have a female bias.

Our proposed metric xMCAS provides a novel way of measuring non-binary gender bias in TTI diffusion models. xMCAS is a comprehensive, quantifiable, and easy-to-understand metric that measures model differences for more diverse gender representations. In the polar form, it can provide separate measures for binary and non-binary gender allowing for a more in-depth analysis. The Euclidean form combines both binary and non-binary gender bias into a single numerical representation that can potentially be used in a cost function for debiasing TTI diffusion models.

Non-binary gender and its inclusion in both generative and discriminate AI models needs to be studied more thoroughly from a multidisciplinary and intersectional perspective. We used non-binary gender in a very explicit way, and more work is required to understand implicit non-binary gender bias. We looked at non-binary gender bias through the lens of binary gender bias but it can also be analysed independently. The stereotypically feminine bias of non-binary gender observed in current TTI models is a serious issue and more work is needed to understand, detect and mitigate this issue. Our work can be considered an initiative in this direction.

### 7.1 Limitations

We acknowledge that only two TTI models were evaluated, which limits the scope and expanse of the results. However, the methodology and the metrics can be applied to any TTI model. The number of images used in the experiments is also on the lower side. This was done to ensure similarity with previous work [10].

## 8 Acknowledgements

Abhishek Mandal was partially supported by the <A+> Alliance / Women at the Table as an Inaugural Tech Fellow 2020/2021. This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_2, cofunded by the European Regional Development Fund.

# References


[1] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[2] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[3] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision*, pages 429–446. Springer, 2024.

[4] 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.

[5] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindřich Libovickỳ, Alexander Fraser, and Kristian Kersting. Multilingual text-to-image generation magnifies gender stereotypes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19656–19679, 2025.

[6] Jeffrey Gleason, Avijit Ghosh, Ronald E Robertson, and Christo Wilson. Perceptions in pixels: Analyzing perceived gender and skin tone in real-world image search results. In *The ACM Web Conference (to appear)*, 2024.

[7] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation, 2024.

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

[9] Abhishek Mandal, Susan Leavy, and Suzanne Little. Biased attention: Do vision transformers amplify gender bias more than convolutional neural networks? In *34th British Machine Vision Conference*, 2023.

[10] Abhishek Mandal, Susan Leavy, and Suzanne Little. Measuring bias in multimodal models: Multimodal composite association score. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 17–30, 2023.

[11] Abhishek Mandal, Suzanne Little, and Susan Leavy. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 416–424, 2023.

[12] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[13] 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.

[14] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3720–3727, 2021.

[15] Nat Thorne, Andrew Kam-Tuck Yip, Walter Pierre Bouman, Ellen Marshall, and Jon Arcelus. The terminology of identities between, outside and beyond the gender binary–a systematic review. *The international Journal of Transgenderism*, 20(2-3):138, 2019.

[16] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Quantifying bias in text-to-image generative models. *arXiv preprint arXiv:2312.13053*, 2023.

[17] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130 (7):1790–1810, 2022.

[18] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319, 2019.

[19] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.