Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Stable Diffusion Exposed: Gender Bias from Prompt to Image

Yankun Wu, Yuta Nakashima, Noa Garcia

Osaka University
{yankun@is., n-yuta@, noagarcia@}ids.osaka-u.ac.jp

Abstract

Several studies have raised awareness about social biases in image generative models, demonstrating their predisposition towards stereotypes and imbalances. This paper contributes to this growing body of research by introducing an evaluation protocol that analyzes the impact of gender indicators at every step of the generation process on Stable Diffusion images. Leveraging insights from prior work, we explore how gender indicators not only affect gender presentation but also the representation of objects and layouts within the generated images. Our findings include the existence of differences in the depiction of objects, such as instruments tailored for specific genders, and shifts in overall layouts. We also reveal that neutral prompts tend to produce images more aligned with masculine prompts than their feminine counterparts. We further explore where bias originates through representational disparities and how it manifests in the images via prompt-image dependencies, and provide recommendations for developers and users to mitigate potential bias in image generation.

1 Introduction

Text-to-image generation models have gained significant attention due to their remarkable generative capabilities. Cutting-edge models, such as Stable Diffusion (Rombach et al. 2022) and DALL-E 2 (Ramesh et al. 2022), have demonstrated outstanding success in generating high-fidelity images based on natural language inputs. However, due to their widespread applications across different domains and their easy accessibility, concerns about the social impact of data (Birhane, Prabhu, and Kahembwe 2021; Garcia et al. 2023; Birhane et al. 2023), bias (Bianchi et al. 2023; Luccioni et al. 2023; Ungless, Ross, and Lauscher 2023), privacy (Carlini et al. 2023; Katirai et al. 2023), or intellectual property (Somepalli et al. 2023; Wang et al. 2023b) have surfaced. This work focuses on the automatic evaluation of gender bias in Stable Diffusion models.

Previous studies have shown that certain adjectives (Luccioni et al. 2023) or professions (Luccioni et al. 2023) can lead to the generation of stereotypes regarding the demographic attributes of faces. However, beyond the regions depicting faces, the remaining areas of the generated image may also exhibit disparities between different genders (Bianchi et al. 2023). Figure 1 shows triplets of generated images with prompts differing by only one word in the gender indicators.[1] The representation of the person in the images adapts accordingly, but the context surrounding the individual (e.g., different musical instruments on the left) and the layout of the image (e.g., on the right) also undergo alterations, even when these changes are not explicitly mentioned in the prompt. This reveals that gender bias does not only manifest within areas depicting people but is sustained in the broader context of the entire image. Thus, while demographic bias in text-to-image generation models has been consistently reported (Ungless, Ross, and Lauscher 2023; Bianchi et al. 2023; Garcia et al. 2023; Cho, Zala, and Bansal 2023; Luccioni et al. 2023; Wang et al. 2023a; Seshadri, Singh, and Elazar 2023; Naik and Nushi 2023), there is a need for automatic evaluation protocols regarding 1) the **entire image** and 2) the **generation process**. Although previous work has explored bias on the final generated images (Luccioni et al. 2023; Bianchi et al. 2023; Cho, Zala, and Bansal 2023; Lin et al. 2023), there is still a lack of analysis about what text-to-image model generates to fill the unguided regions of the image and why it responds distinctly to different gender indicators.

To the best of our knowledge, this is the first work that analyzes the internal components of Stable Diffusion to study where gender bias originated and how it is propagated. We suggest that such disparities arise from the interplay of representational disparities and prompt-image dependencies during image generation: the process involves transitioning from prompt space to image space, potentially treating genders differently and resulting in representational disparities. Using template-free natural language prompts, we further study the dependencies between the prompt and the generated image with the inherent cross-attention mechanism, and categorize objects in prompt and image into five dependency groups. These dependencies/independencies can be modulated by representational disparities. To systematically explore the two intertwined factors, we generate images from a set of (*neutral*, *feminine*, *masculine*) triplet prompts as in Figure 1, aiming to quantify representational disparities (Sec. 4) and prompt-image dependencies (Sec. 6).

[1] Gender indicators refer to words that indicate the gender of a person.

Figure 1: We use free-form triplet prompts to analyze the influence of gender indicators on the overall image generation process. We show that 1) gender indicators influence the generation of objects (left) and their layouts (right), and 2) the use of gender *neutral* words tends to produce images more similar to those prompted by *masculine* indicators rather than *feminine* ones.

Our evaluation protocol allows us to formulate and answer the following research questions (RQ):

RQ1 Do images generated from neutral prompts exhibit greater similarity to those generated from masculine prompts than to images generated from feminine prompts, and if so, why?

RQ2 Do object occurrences in images significantly vary based on the gender specified in the prompt? If there are differences, do these object occurrences from neutral prompts exhibit greater similarity to those from masculine or feminine prompts?

RQ3 Does the gender in the input prompt influence the prompt-image dependencies in Stable Diffusion, and if so, which prompt-image dependencies are more predisposed to be affected?

We conduct experiments on three Stable Diffusion models spanning four caption datasets as well as a text set generated by ChatGPT (Brown et al. 2020). Through the generation of triplet prompts with only gender indicators differing, we observe a consistent trend across Stable Diffusion models. Our key findings indicate:

- **For Stable Diffusion, person = man**
  - Quantitatively, neutral prompts consistently produce images that look more similar to those from masculine prompts than feminine prompts.
  - The neutral representations are closer to the masculine representations *for all the internal stages* of the generation process.
- **Explicit objects are consistent across genders**
  - Objects generated explicitly from prompts exhibit similar co-occurrence for different genders.
- **Unguided objects are gendered**
  - Objects not explicitly mentioned in the prompt are generated at different rates for each gender.
  - Co-occurrences of objects in images from neutral prompts consistently exhibit greater similarity to those from masculine prompts.

Our findings show that gender bias extends beyond people's representations, permeating through the entire image and affecting the generated objects. We conclude the paper with recommendations for both model developers and users, aimed at mitigating this effect.

2 Related Work

Text-to-image models There are three main types of text-to-image generation models: GANs (Goodfellow et al. 2020; Tao et al. 2022; Reed et al. 2016), autoregressive (Ramesh et al. 2022, 2021; Ding et al. 2021, 2022; Yu et al. 2022), and diffusion (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Saharia et al. 2022). Within diffusion models, Stable Diffusion (Rombach et al. 2022) has emerged as the preferred testbed due to its high-quality generations and open-source nature. As diffusion models rely on cross-attention to connect text and image modalities, it enables the examination of the image generation process at the word level (Hertz et al. 2023). The cross-attention module assists in tasks such as editing (Hertz et al. 2023; Lu, Liu, and Kong 2023; Epstein et al. 2023; Gandikota et al. 2023a,b) and segmentation (Tang et al. 2023; Wu et al. 2023; Pnvr et al. 2023). By leveraging this property, we can investigate the relationship between gender and prompt-guided generations.

Social bias Text-to-image generation models often reproduce demographic stereotypes tied to gender and race across various factors, including but not limited to occupations (Bianchi et al. 2023; Cho, Zala, and Bansal 2023; Luccioni et al. 2023; Wang et al. 2023a; Mandal, Leavy, and Little 2023; Lin et al. 2023; Seshadri, Singh, and Elazar 2023), adjectives (Luccioni et al. 2023; Naik and Nushi 2023; Berg et al. 2022), objects (Mannering 2023), outfits (Zhang et al. 2023c), and nationalities (Bianchi et al. 2023; Wolfe and Caliskan 2022a). Analysis of prompt templates like "`a photo of the face of [OCCUPATION]`" reveals that certain occupations, such as *software developers*, are predominantly represented as white men, while *housekeepers* tend to be associated with women of color. Additionally, Wolfe et al. (Wolfe et al. 2023) showed that models are more inclined to generate sexualized images in response to prompts containing "`a [AGE] year old girl`". Moreover, Zhang et al. (Zhang et al. 2023a) argued that unfairness extends to images depicting underrepresented attributes like *wearing glasses*, highlighting the per-

vasive nature of biases in the generation process. In addition to biases concerning humans, previous studies have explored geographical-level differences in objects (Hall et al. 2023) and the correctness of cultural context (Basu, Babu, and Pruthi 2023; Liu et al. 2024).

Bias evaluation A fundamental aspect in the study of bias is the evaluation protocol. Most previous approaches rely on prompts that fill attributes (e.g., profession) with a template (Luccioni et al. 2023; Bakr et al. 2023; Teo, Abdollahzadeh, and Cheung 2023; Lee et al. 2023; Cho, Zala, and Bansal 2023; Bianchi et al. 2023; Wang et al. 2023a; Zhang et al. 2023c; Naik and Nushi 2023), leading to constrained scenarios and limited additional details in the prompts. Moreover, these methods evaluate bias on the proxy presentation of the generated images but do not examine presentations in the generation process. Besides, these methods mainly focus on people's attributes (Luccioni et al. 2023; Teo, Abdollahzadeh, and Cheung 2023; Lee et al. 2023; Cho, Zala, and Bansal 2023; Bianchi et al. 2023; Wang et al. 2023a; Chinchure et al. 2023; Naik and Nushi 2023), such as the gender of faces, thereby overlooking biases in the generated visual elements as well as the entire image context. Except for the method that exclusively on gender bias evaluation, there are traditional evaluation criteria for text-to-image models measuring image fidelity and text-image alignment with automated metrics (Salimans et al. 2016; Heusel et al. 2017; Vedantam, Lawrence Zitnick, and Parikh 2015; Papineni et al. 2002) or human evaluation (Otani et al. 2023).

Overall, there is an absence of automated methods for nuanced bias evaluation that conveys bias at the different stages of the generation process. Using free-form prompts, our work proposes a method to uncover prompt-image dependencies, disclosing how objects are generated differently according to gender indicators in the prompt.

3 Preliminaries

Triplet prompt generation Let $\mathcal{P}_{\text{n}}$ be a set of *neutral* prompts, which do not specify the gender of the person. As shown in Figure 1, from these neutral prompts, we generate two counterpart prompt sets, $\mathcal{P}_{\text{f}}$ and $\mathcal{P}_{\text{m}}$, as *feminine* and *masculine* prompt sets, respectively. The only difference among these three prompt sets is the gender indicator, while all other words remain unchanged. Our bias evaluation is based on analyzing distinctions between pairs of generated images from the triplet $\{\mathcal{P}_{\text{n}}, \mathcal{P}_{\text{f}}, \mathcal{P}_{\text{m}}\}$.

We generate neutral prompts from natural language sentences, consisting of captions from four vision-language datasets (GCC validation set (Sharma et al. 2018), COCO (Lin et al. 2014), TextCaps (Sidorov et al. 2020), and Flickr30k (Young et al. 2014)), as well as a profession prompt set generated by ChatGPT 3.5 (Brown et al. 2020).[2] From the vision-language datasets, we generate neutral prompts by choosing *neutral captions* that meet two criteria: (1) they contain the word *person* or *people*, and (2) they do not include other words that indicate humans. To generate feminine and masculine prompts, we swap *person/people* in the neutral captions with the gender indicators *woman/women* and *man/men*, respectively. For the profession prompt set, we generate neutral prompts with ChatGPT based on professions, such as `ecologist` or `doctor`, across 16 topics. For example, `an ecologist studies the ecosystem in a lush green forest.` To create feminine and masculine prompts, we prepend *female/male* before the profession. Further details can be found in the supplementary materials.

Data	Triplets	Prompts	Seeds	Images
GCC (val)	418	1,254	5	6,270
COCO	51,219	153,657	1	153,657
TextCaps	4,041	12,123	1	12,123
Flickr30k	16,507	49,521	1	49,521
Profession	811	2,433	5	12,165

Table 1: Number of generated triplets, prompts, and images for each dataset.

Image generation Given prompt p as input, Stable Diffusion transforms it into a text embedding $\mathbf{t}$ in the *prompt* space using the text encoder. This text embedding is fed into the cross-attention module in UNet (Ronneberger, Fischer, and Brox 2015), which performs the denoising operations from an initial noise $\mathbf{z}_T$ in the latent space. After T denoising steps, the embedding $\mathbf{z}_0$ in the *denoising* space is obtained. Finally, image x in the *image* space is generated from $\mathbf{z}_0$ by the image decoder. In this work we evaluate Stable Diffusion models: v1.4,[3] v2.0-base,[4] and v2.1-base[5] (denoted as SD v1.4, SD v2.0, and SD v2.1, respectively). The same generation pipeline is used in all the models.

Table 1 reports the details of image generation for each dataset. The seed is the same within each triplet, ensuring the same initial noise $\mathbf{z}_T$. To address data scarcity in GCC and Profession sentences, we produce five images per prompt with five different seeds. In the following, when mentioning a dataset, we are referring to the generated images whose prompts originate from the corresponding dataset.

Gender bias definition The interpretation of gender bias varies across literature, resulting in different work attributing different meanings to the term. In this paper, we define gender bias as:

- Within the triplet, images generated from *neutral* prompts consistently display greater similarity to those from either *feminine* or *masculine* prompts.
- Specific objects tend to appear more frequently in the generated images associated with a specific gender.

Whereas objects are not equally distributed in the real world or across cultures, and recognizing that not all dis-

[2] Accessed November 2023.

[3] https://github.com/CompVis/stable-diffusion

[4] https://huggingface.co/stabilityai/stable-diffusion-2-base

[5] https://huggingface.co/stabilityai/stable-diffusion-2-1-base

Pairs	Prompt $\mathbf{t}$	Denoising $\mathbf{z}_0$	Image *SSIM* ↑	*Diff. Pix.*↓	*ResNet* ↑	*CLIP* ↑	*DINO* ↑	*split-product* ↑
SD v1.4								
(neutral, feminine)	0.909	0.770	0.516	42.61	0.848	0.794	0.543	0.956
(neutral, masculine)	**0.931**	**0.798**	**0.543**	**39.34**	**0.859**	**0.808**	**0.576**	**0.961**
SD v2.0								
(neutral, feminine)	0.980	0.767	0.543	39.00	0.847	0.797	0.545	0.957
(neutral, masculine)	**0.982**	**0.790**	**0.571**	**35.82**	**0.864**	**0.817**	**0.581**	**0.963**
SD v2.1								
(neutral, feminine)	0.980	0.755	0.522	41.48	0.842	0.805	0.527	0.952
(neutral, masculine)	**0.982**	**0.782**	**0.552**	**37.96**	**0.856**	**0.820**	**0.566**	**0.959**

Table 2: Representational disparities between neutral, feminine, and masculine prompts in GCC in the three spaces on Stable Diffusion models.

parities regarding genders are inherently problematic (i.e., the association of *dress* with *women* may not be an issue, whereas *kitchen* might), we argue that it is essential to have a methodology for recognizing and quantifying these differences. Our proposed evaluation protocol is not envisaged to identify objects that perpetuate discrimination and gender stereotypes but to *highlight significant gender disparities*, regardless of whether they are deemed problematic.

We apply our evaluation protocol in three Stable Diffusion models, and analyze gender bias by addressing our research questions.

4 Gender Disparities in Neutral Prompts

RQ1 *Do images generated from neutral prompts exhibit greater similarity to those generated from masculine prompts than to images generated from feminine prompts, and if so, why?*

In this section, we address the above research question through the use of representational disparities.

Representational Disparities

We use representational disparities to analyze how images generated by different gender indicators compare with respect to neutral prompts. For a given triplet, the analysis consists on comparing the similarity between *neutral* embeddings and *feminine* and *masculine* embeddings. To measure the extent of gender disparities in the generative process, we examine the representational disparities throughout the entire generation, tracking embeddings from the prompt space to the denoising space and the image space, offering insights into when bias is introduced.

Prompt space The prompt space is defined as the space in which all text embeddings lie. Different points in this space provide different semantics to the following image generation process. To measure the disparity between a pair prompt set $\mathcal{P}$ and $\mathcal{P}'$ in the triplet, we compute cosine similarity as

$$s_{\mathrm{P}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p_i'} \cos(\mathbf{t}, \mathbf{t}'), \tag{1}$$

where $|\cdot|$ is the number of elements in the given set, $\cos(\cdot, \cdot)$ gives cosine similarity, the summation is computed over all prompts p_i from $\mathcal{P}$ and p_i' from $\mathcal{P}'$,[6] text embeddings $\mathbf{t}$ and $\mathbf{t}'$ correspond to prompts p_i and p_i', respectively.

Denoising space The embedding $\mathbf{z}_0$ after the last denoising process lies in the denoising space. Similarly to the prompt space, we compute cosine similarity as

$$s_{\mathrm{D}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p_i'} \cos(\mathbf{z}_0, \mathbf{z}_0') \tag{2}$$

where $\mathbf{z}_0$ and $\mathbf{z}_0'$ are derived from p_i and p_i', respectively.

Image space As bias often involves more in the semantics rather than pixel values, we adopt a spectrum of metrics computed from the generated images. To measure image structural differences, we use the average of SSIM scores over all pixels as one of our disparity metrics *SSIM*. Additionally, the ratio of the number of pixels in the contours with higher SSIM scores is used as another disparity metric *Diff. Pix.* To quantify differences in higher-level semantics, we apply latent vectors of pre-trained neural networks, adopting the last fully-connected layer of ResNet-50 (He et al. 2016), the CLIP image encoder (Radford et al. 2021), and the last layer of DINO (Caron et al. 2021), referred to as *ResNet*, *CLIP*, and *DINO*, respectively. For all metrics, we compute the cosine similarity between the latent vectors from image pairs as in Eq. 1 and Eq. 2. Additionally, we adopt *split-product* (Somepalli et al. 2023), computing the maximum cosine similarity among corresponding patches between image pairs.

Pairs	GCC	COCO	TextCaps	Flickr30k	Profession
SD v1.4					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.379	0.486	0.413	0.424	0.350
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.414**	**0.516**	**0.444**	**0.457**	**0.374**
SD v2.0					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.382	0.512	0.420	0.445	0.362
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.425**	**0.531**	**0.448**	**0.476**	**0.376**
SD v2.1					
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$	0.380	0.499	0.388	0.426	0.349
$s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$	**0.419**	**0.522**	**0.419**	**0.451**	**0.382**

Table 3: Co-occurrence similarity on Stable Diffusion models.

Results Analysis

By analyzing the representational disparities on (*neutral*, *feminine*) and (*neutral*, *masculine*) pairs, we can provide some answers for **RQ1**.

In the image space, regardless of whether considering the entire image holistically (*SSIM*, *Diff. Pix*, *ResNet*, *CLIP*, and *DINO*) or the highest similarity on corresponding patches (*split-product*), images generated from *neutral* prompts consistently demonstrate greater similarity to those from *masculine* prompts. Results on GCC-derived prompts are shown in Table 2, whereas results on other datasets and models can be found in the supplementary material. This trend is consistently observed in all datasets and all models.

Tracing back to the prompt space and denoising space to explore where and when gender bias emerges in the generated images, results in Table 2 show that embeddings from *neutral* prompts are closer to the embeddings from *masculine* prompts both in the prompt space and the denoising space. Although Stable Diffusion models apply different text encoders (OpenCLIP-ViT/H for SD v2.0 and SD v2.1, while CLIP ViT-L/14 for SD v1.4), the same trend is observed across all three models and all datasets. This indicates that gender bias originates from the text embedding and perpetuates through the generation process, leading to the disparities observed in the generated images.

5 Influence of Gender in Objects

RQ2 *Do object occurrences in images significantly vary based on the gender specified in the prompt? If there are differences, do these object occurrences from neutral prompts exhibit greater similarity to those from masculine or feminine prompts?*

The representational disparities reflect the holistic similarity between gender groups, but they do not convey fine-grained differences, i.e., why a certain object appears in the generated image given a gender-specific prompt. In this section, we address **RQ2** by investigating the relationship between gender and the objects in the generated images. To do so, we extract objects with a visual grounding model and study their co-occurrence with each gender.

[6]Subscript i is the index of the prompt to clarify p_i and p_i' are corresponding prompts, derived from the same one.

Detecting Generated Objects

To detect objects in the generated images we use the assembled model RAM-Grounded-SAM. Given a generated image, RAM (Zhang et al. 2023b) predicts plausibly objects, which are used by Grounded DINO (Liu et al. 2023) to propose bounding boxes around the candidate objects. Then, Segment Anything Model (SAM) (Kirillov et al. 2023) extracts object regions m_o within the bounding box of the object o. For each image, a set of object names and a set of regions are obtained.

Evaluation Metrics

Our evaluation protocol involves measuring the differences in object co-occurrences for different genders. Let $\mathrm{cnt}(o, p)$ denote the number of occurrences of the object o in the image generated from the prompt p in the prompt set $\mathcal{P}$. The total number of co-occurrence $C(o, \mathcal{P})$ is given by:

$$C(o, \mathcal{P}) = \sum_{p \in \mathcal{P}} \mathrm{cnt}(o, p) \tag{3}$$

With the above definition and a set of triplet prompts, we use the following three methods to evaluate the influence of gender in the generated objects:

1) Statistical tests We use the chi-square test to check whether there are statistical differences in the object co-occurrence among two or three image sets. This test is applicable to the triplet and any pairs in the triplet. If the resulting p-value is below 0.05, we interpret significant differences in the object distribution in the pair or triplet.

2) Co-occurrence similarity We compute the similarity of the co-occurrences of detected objects between two image sets. Formally, let the vector $\mathbf{v}_p$ denote the object occurrences in the image generated from prompt p, and each element in $\mathbf{v}_p$ is the occurrence $\mathrm{cnt}(o, p)$ for the object o in

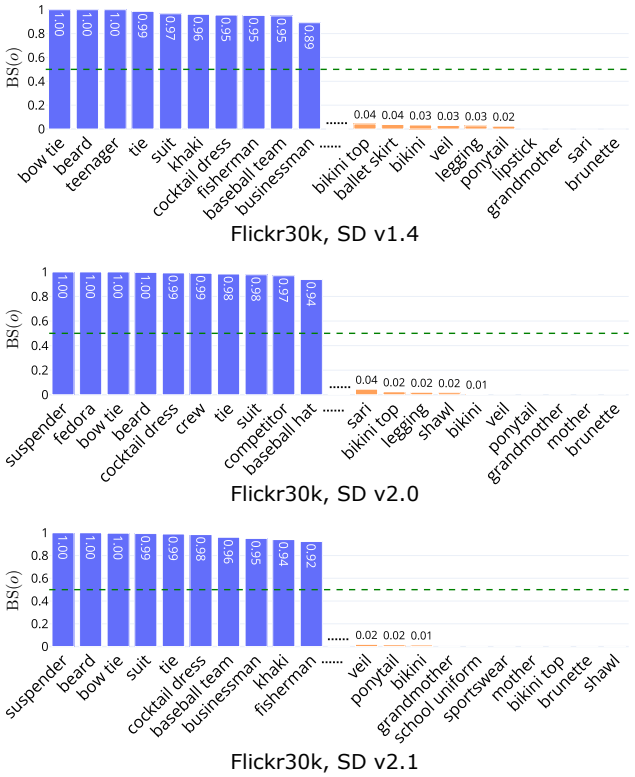

Figure 2: Bias score in Flickr30k. The higher values (in blue) suggest an object is biased toward masculine prompts, while lower values (in orange) indicate a preference toward feminine prompts. $\mathrm{BS}(o) = 0.5$ (green line) shows the object does not skew toward a certain gender. We filter objects if the maximum co-occurrence is less than 20.

the image. Similarly to Eq. 1 and Eq. 2, we compute cosine similarity on object co-occurrences as

$$s_{\mathrm{O}}(\mathcal{P}, \mathcal{P}') = \frac{1}{|\mathcal{P}|} \sum_{p_i, p_i'} \cos(\mathbf{v}_i, \mathbf{v}_i'), \tag{4}$$

where prompt sets $\mathcal{P}$ and $\mathcal{P}'$ are in the triplet. $\mathbf{v}_i$ and $\mathbf{v}_i'$ are derived from prompt p_i in $\mathcal{P}$ and p_i' in $\mathcal{P}'$, respectively. A higher co-occurrence similarity means that objects are detected with the same-level frequency in two image sets, whereas a low similarity means that objects are detected at different rates.

3) Bias score Following (Zhao et al. 2017), we compute the bias score $\mathrm{BS}(o)$ for a certain object o as:

$$\mathrm{BS}(o) = \frac{C(o, \mathcal{P}_{\mathrm{m}})}{C(o, \mathcal{P}_{\mathrm{m}}) + \frac{|\mathcal{P}_{\mathrm{m}}|}{|\mathcal{P}_{\mathrm{f}}|} C(o, \mathcal{P}_{\mathrm{f}})}. \tag{5}$$

$\mathrm{BS}(o)$ ranges from 0 to 1, with 1 meaning the object is skewed towards *masculine* prompts and 0 towards *feminine* prompts. If $\mathrm{BS}(o) = 0.5$, object o does not favor any gender.

Results Analysis

All the p-values from chi-square tests among the triplets and pairs are below 10^{-5}, implying significant differences in the object distributions of each gender across all datasets and models. This shows that according to gender, not only the person in the image may change, but also the objects generated in the image are statistically different.

To investigate whether the object co-occurrences of neutral images exhibit larger similarity to a certain gender image set, we compute co-occurrence similarity on pairs (*neutral, feminine*) and (*neutral, masculine*). Results in Table 3 indicate that object co-occurrences in *neutral* consistently exhibit greater similarity to those in *masculine* prompts than in *feminine* prompts across all datasets and models, corroborating the observations in Section 4. This, again, indicates that prompts that use gender neutral words tend to generated objects that are more commonly generated for masculine prompts than for feminine prompts.

Subsequently, we examine specific examples by computing the bias score for each object in the generated images. Figure 2 shows results on Flickr30k. Results for other datasets and models can be found in the supplementary material. We can observe that objects with higher or lower bias scores in different versions of Stable Diffusion show a similar pattern. Thus, we analyze results on SD v2.0 as an example. Notably, clothing exhibits a high bias: for example `suspender`(1), `fedora`(1), and `bow tie`(1) lean towards *masculine*, while `veil`(0), `bikini`(0.01), and `shawl`(0.02) lean towards *feminine*. This is not surprising, considering that clothing elements are traditionally gendered. Other than clothing, we find a strong association between `family`(0.11) and `child`(0.31) with *feminine* prompts, potentially associating *feminine* with caregiver, while *masculine* prompts exhibit greater alignment with words related to sports such as `baseball team`(0.91), `skateboarder`(0.89), and `golfer`(0.86), a phenomenon that has been previously observed in VQA datasets (Hirota, Nakashima, and Garcia 2022). Another observation is that *feminine* prompts also have a high association with food, such as `salad`(0.22), `meal`(0.25), and `cotton candy`(0.31). Results on other versions of Stable Diffusion and other datasets show similar trends, and additionally reveal that `businessman` tends to be skewed towards *masculine* whereas `kitchenware` tends to be associated with *feminine* prompts.

6 Gender in Prompt-Image Dependencies

RQ3 *Does the gender in the input prompt influence the prompt-image dependencies in Stable Diffusion, and if so, which prompt-image dependencies are more predisposed to be affected?*

To answer to this question, we need to know not only which objects are generated for each gender but also how each object is generated in the diffusion process. To do so, we propose to classify objects into prompt-image dependency groups according to their relationship with the input prompt and the generated image. First, we conduct an *extended object extraction* by detecting not only the objects in the generated image as in Section 5, but other objects also involved in the generative process. Then, we classify each object according to five *prompt-image dependency groups*,

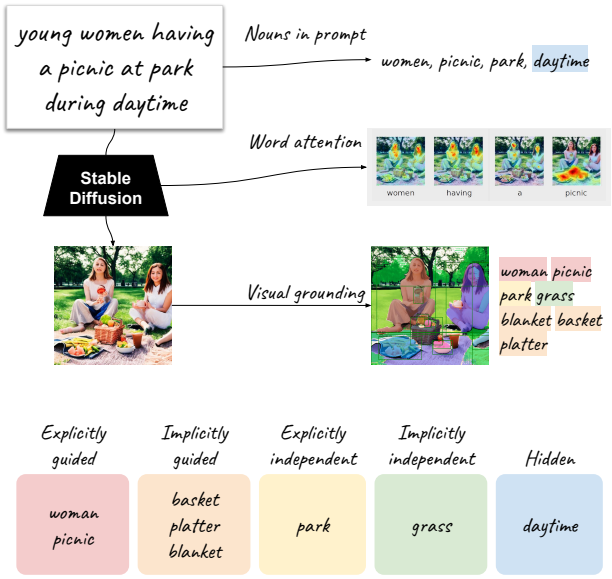

Figure 3: Prompt-image dependency groups.

which allows us to study how gender influences objects according to their generative process.

Extended Object Extraction

To detect extended objects involved in the generative process, we conduct three extraction processes.

1) Nouns in prompt Prompts, designed by users, are a direct cue of what they wish to see in the generated image. The generated image, on the other hand, is required to be faithful to the prompt. The first extraction process targets nouns within the prompt, recognizing their importance in directly shaping the occurrence of objects in the generated image. For each prompt, we obtain a noun set including all lemmatized nouns n in the prompt by using NLTK (Bird, Klein, and Loper 2009).

2) Word attention Verifying whether objects in the noun set are faithfully generated in the image is demanding, as it requires locating the region where the noun guides. Fortunately, cross-attention has proven to be effective in exploring the word guidance during the generation process (Hertz et al. 2023; Tang et al. 2023). Our second extraction process is the word attention masks generated by the cross-attention module via DAAM (Tang et al. 2023). For each word, we first compute the normalized attention map,[7] where a higher value indicates that the pixel is more associated with the word. Then, we binarize the attention map with a threshold θ to obtain a set of masks a_n, responding to the region of an object specified by the word n. In each prompt, we obtain a mask set containing the mask a_n for each word n.

3) Visual grounding Nouns and the corresponding object regions cover only a small subset of objects in the generated image; there should be many other objects that are not explicitly described in the prompt but are still included in the image to complete the scene. We aim to enumerate as many objects as possible for comprehensive object-level analysis. To spot regions of arbitrary objects, the last extraction process is the same visual grounding process as in Section 5.

[7]Details are in the supplementary material, which can be found in https://arxiv.org/abs/2312.03027.

Prompt-Image Dependency Groups

Next, we classify each detected object according to its generative process. On one hand, the generated image should align with its prompt, which can be verified using the noun set and the mask set. On the other hand, the image may have other visual elements beyond the prompt, listed in the object set and the object region set. To define prompt-image dependency groups, we consider the dependency among objects, the noun set, and the mask set based on its membership.

Definition 6.1 (Explicitly)**.** If the object o is in the noun set, it is *explicitly* described in the prompt.

Definition 6.2 (Guided)**.** If object region m_o *sufficiently* overlaps with at least one mask in the mask set, the object o is *guided* by cross-attention between the prompt and the image. Sufficiency is determined by the coverage of object region m_o by the mask a:

$$\text{coverage}(m_o, a) = \frac{|m_o \cap a|}{|m_o|}, \tag{6}$$

where $|\cdot|$ is the number of pixels. Thus, if $\text{coverage}(m_o, a)$ is larger than a certain threshold σ, the object region m_o sufficiently overlaps with the mask a.

With these definitions, we cluster objects in the object set into five groups, as illustrated in Figure 3 with the example prompt `young women having a picnic at the park during daytime`:

Explicitly guided The object is *explicitly* mentioned in the prompt and *guided* by cross-attention. Faithful image generation may require each noun to be associated with the corresponding object.

Implicitly guided The object is *not explicitly* mentioned in the prompt but *guided* by cross-attention. The object may be strongly associated with or pertain to a certain noun in the noun set, e.g., the object `basket` for the noun `picnic`.

Explicitly independent The object is *explicitly* mentioned in the prompt but *not guided* by cross-attention. e.g., `park`.

Implicitly independent The object is *not explicitly* mentioned in the prompt and *not guided* by cross-attention. The object is generated solely based on contextual cues, e.g., `grass`.

Hidden The noun has no association with objects in the object set, i.e., the noun is *not included* in the images, e.g., `daytime`.

Figure 3 illustrates the object extraction processes and the resulting dependency groups. Dependency groups are important as they depict if an object tends to appear, for example, in relation to the prompt (*explicitly guided*) or just for

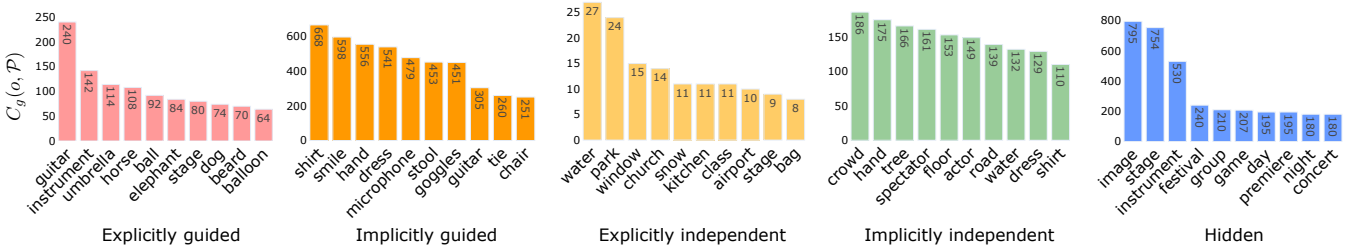

Figure 4: Top-10 most frequent objects in each prompt-image dependency group in GCC on SD v2.0.

filling the scene (*implicitly independent*). Together with the gender-specific sets of prompts, they vividly provide essential insights into how an image generation model behaves for different genders.

Result Analysis

We denote co-occurrence $C_g(o, \mathcal{P})$ as the number of occurrences of object o in each dependency group g. To clarify, given that there are nouns included in the hidden group, the computation of occurrence should be adjusted from $C_g(o, \mathcal{P})$ to $C_g(n, \mathcal{P})$ for n in the hidden group.

Objects in dependency groups To answer RQ3, we first investigate objects in the prompt-image dependency groups, aiming to identify which types of objects are generated under the influence of the prompt, the cross-attention, or the context of the generated image. As shown in Figure 4, we look into the prevalent objects within each dependency group on SD v2.0.[8] Although the specific generated objects align with the prompt's domain, and their frequencies may vary across datasets, we observe consistent trends.

Objects in the *explicitly guided* group include animals and tangible items commonly encountered in daily life, such as `guitar` and `umbrella`. The *implicitly guided* group contains objects surrounding human beings, such as clothing and personal belongings like `shirt` and `microphone`. The *explicitly independent* group comprises words related to the surrounding environment, such as `park` or `church`. Objects in the *implicitly independent* group are typically part of the background that can be detected, like `crowd` and `tree`, along with attire accompanying individuals. Lastly, the *hidden* group comprises words challenging to detect in images, such as `game` and `day`.

Gender and dependency groups Next, we investigate the relationship between gender and the objects in each prompt-image dependency group. To discern whether object differences are statistically significant, we conduct chi-square tests[9] on the object co-occurrence for each dependency group. While we find significant differences (p-value < 0.05) across all datasets in the *implicitly guided* and *implicitly independent* groups, we do not find significant differences in most datasets in the *explicitly guided*, *explicitly independent* and *hidden* groups. This suggests that while Stable Diffusion may consistently generate the nouns explicitly mentioned in the prompt, it may rely on gender cues for generating elements that are not specified in the prompt, such as the background and surroundings of the individuals.

To explore further into the text-image dependencies and their correlation with gender, we calculate the bias score on object co-occurrence in the *implicitly guided* and the *implicitly independent* groups, both of which exhibit statistically significant differences. Figure 5 shows the top-10 objects skewed toward *masculine* and *feminine* in TextCaps and GCC datasets on SD v1.4 and SD v2.0.[10] We filter objects if maximum co-occurrence is less than 20 in TextCaps, and 5 in GCC. We analyze results on SD v2.0 as examples. For the *implicitly guided* group in TextCaps, we observe high bias scores for clothing items, such as `cocktail dress`(1), `suit`(1), `bow tie`(0.98), and `tie`(0.97) for *masculine* and `ponytail`(0.03), `dress`(0.09) and `boot`(0.14) for *feminine*, aligning with observations in previous work (Zhang et al. 2023c). Another prominent observation, consistent with the findings on Flickr30k in RQ2, is the strong association of `child`(0.27) with *feminine*, and *masculine* with sports-related terms such as `player`(0.8) and `football player`(0.72). Similar gendered associations are observed across different datasets and models. For the *implicitly independent* group in GCC, words related to sports such as `bodybuilder`(1) and `football team`(1) are again skewed toward *masculine*, while `instrument`(0.17) and `apron`(0.33) are skewed to *feminine*. There are also disparities in the words indicating backgrounds, such as `backdrop`(0.15) and `dirt field`(0.17) for *feminine* and `stone building`(1) and `tennis court`(0.63) for *masculine*. Other datasets and models report similar results. Furthermore, it is observed that `smile` and `flower` are skewed towards *feminine*.

7 Additional Experiments

To further evaluate our protocol, we conduct intro-prompt evaluation and human evaluation.

Intra-Prompt Evaluation

To eliminate the influence of randomness, we investigate the research questions using images generated from the same

[8]To focus on the differences between generated objects, we remove individuals (person, people, women, woman, men, man, female, male, girl, boy).

[9]Details and results can be found in the supplementary material.

[10]Results on other datasets and models are in the supplementary material.

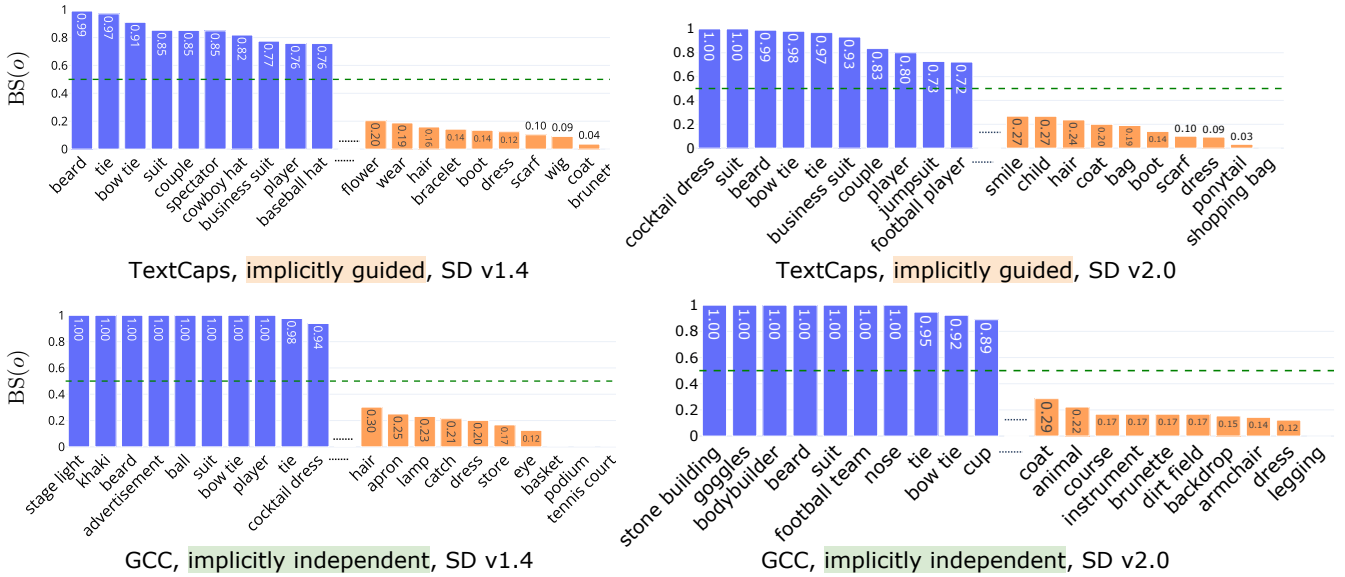

Figure 5: Bias score by groups on SD v1.4 and SD v2.0. Top: *implicitly guided* group in TextCaps on SD v1.4 and SD v2.0. Bottom: *implicitly independent* group in GCC on SD v1.4 and SD v2.0.

triplet prompts. We generate a total of $3,000$ images on $1,000$ seeds with SD v2.0, from triplet prompts derived from a caption in GCC: "*person looks at the falling balloons at the conclusion*". We use the same settings as conducted in the experiments above.

For RQ1, results provided in the supplementary material show that neutral is consistently closer to masculine across all spaces. For RQ2, the chi-square tests on the object occurrences among the triplets and every pair within the triplets, p-value is consistently less than 10^{-5}, indicating statistically significant differences. For RQ3, the chi-square tests also reveal significant differences in the groups *implicitly guided* and *implicitly independent* ($p < 10^{-5}$). However, we do not apply chi-square test on *explicitly guided*, *explicitly independent*, and *hidden*, as the numbers of objects in these groups are less than 5. The co-occurrence similarity $s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{f}})$ between the neutral and feminine is 0.733, while the similarity $s_{\mathrm{O}}(\mathcal{P}_{\mathrm{n}}, \mathcal{P}_{\mathrm{m}})$ between the neutral and masculine is **0.773**. This indicates that the object co-occurrences in images generated from *neutral* prompts are closer to those from *masculine* prompts than *feminine* prompts. These findings correspond to the above results.

Human Evaluation

To evaluate the reliability of the visual grounding model, we randomly select 100 generated images from SD v2.0 along with the nouns from the corresponding prompts and conduct a human evaluation to determine whether the nouns are present in the images. The 100 prompts contain 346 nouns, from which 227 (65.61%) are correctly identified both by humans and the automated vision grounding. Out of the remaining 119 nouns, only 8 nouns are detected by the model but not observed by humans. These nouns are `frisbee`(2), `women`(1), `people`(1), `kite`(1), `scooters`(1), `tennis`(1) and `speaker`(1). For the nouns not detected by the model but identified by humans, the most frequent ones are `woman`(10), `street`(7), `people`(6), and `snowy`(4). The absence of the noun `street` in the model's detection might be attributed to the strict alignment between nouns and objects. Even if the model successfully identifies `street scene`, the specific noun `street` might be placed in one of the *implicitly guided*, *implicitly independent*, or *hidden* groups. These results indicate that the visual grounding model has reasonable accuracy in detecting nouns appearing in the generated images, though there is still room for improvement on abstract nouns and scene-level nouns.

8 Recommendations

Our methodology revealed significant disparities in the objects generated by three Stable Diffusion models according to the gender in the input prompt. While these discrepancies may seem harmless, they can potentially reinforce gender stereotypes. With this in mind, we propose a series of suggested practices aimed at mitigating these concerns, both for model developers and for users:

Model Developers

Debias text embeddings We have identified that gender bias originates in the text embedding, with *neutral* prompts consistently being more similar to *masculine* prompts than to *feminine* prompts, and propagates through the entire generation process. Given the documented presence of gender bias in CLIP (Wolfe et al. 2023; Wolfe and Caliskan 2022b; Agarwal et al. 2021; Wolfe, Banaji, and Caliskan 2022), it comes as no surprise that text-to-image generation models relying on CLIP also exhibit such biases. The first mitigation technique should focus on debiasing the text embedding

space, aiming for more equitable representations.

Identify problematic representations While some associations of certain objects with specific genders may not immediately raise concerns, others could potentially do so. Therefore, researchers must meticulously assess these associations, taking into account the cultural context in each instance. It is crucial to examine the co-occurrence of objects across genders and check whether neutral prompts tend to exhibit a preference toward a particular gender.

Investigate modules that complete the scene Significant differences were observed in the *implicitly* generated objects, underscoring the need to investigate how the model completes the scene. Future research could explore other modules, probing fine-grained control over the regions not guided by the input.

Users

Explicitly specify objects Our results showed that there are no significant differences in the objects explicitly mentioned in the input prompts concerning gender. This suggests that Stable Diffusion models can adhere to the simple instructions in the prompt regardless of gender. Therefore, expanding the number of objects in the input could offer greater control over broader guided regions and potentially lead to the generation of images with less gender disparity.

Explicitly specify gender Considering that *neutral* prompts consistently produced images more similar to those from *masculine* prompts, we advise refraining from using neutral prompts if targeting a balanced distribution across genders. Instead, using prompts with specified gender indicators may be more reliable.

9 Limitations

We acknowledge that our proposed evaluation protocol has limitations, and we emphasize them here for transparency and to inspire the community to propose enhancements in future studies. Firstly, our evaluation protocol focuses on binary genders, neglecting to evaluate gender from a broader spectrum perspective. To enhance inclusivity, future research could extend the analysis to encompass a more diverse range of genders. Secondly, our protocol relies on a stringent alignment between nouns and objects, assuming their identity after lemmatization, which may overlook variations and synonyms. Thirdly, the objects segmented in visual grounding may encounter errors, possibly perpetuating issues in the classified groups. Additionally, if gender bias exists in the visual grounding model, where certain objects may be more challenging to detect in specific genders, this bias could transfer to the final results. Besides, when the object comprises more than one word (e.g., "picnic basket"), each noun in the phrase has its own word attention rather than being considered as a single entity. Last but not least, our study only examines the presence of objects not differentiating with distinct attributes, such as color or shape.

10 Conclusion

We introduced an automated evaluation protocol to study gender bias in image generation by probing the internal components of Stable Diffusion models. We investigated both representational disparities and prompt-image dependencies to uncover the origin of bias and how it manipulated image generation. Through the generation of free-form triplet prompts with only gender indicators differing, our findings indicate that:

1. Prompts that use *neutral* words to refer to people (`a person in a park`) consistently yield images more similar to the ones generated from prompts with *masculine* words (`a man in a park`) than from prompts with *feminine* words (`a woman in a park`).
2. There are statistically significant differences in the type of objects generated in the image based on the gender indicators in the prompt.
3. The frequency of objects generated explicitly from prompts exhibit similar behavior for different genders.
4. Objects not explicitly mentioned in the prompt exhibit significant differences for each gender.
5. We particularly observed significant statistical disparities in generated objects based on gender in items related to clothing and traditional gender roles such as sports, which are highly skewed towards images generated from *masculine* prompts, and food, which are skewed towards images generated from *feminine* prompts.

Based on these observations, we provided recommendations for developers and users to reduce such representational disparities and gender bias in the generated images. We hope these insights contribute to underscoring the nuanced dynamics of gender bias in image generation, offering a new and valuable perspective to the growing body of research on this topic.

Acknowledgments

This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI Nos. JP22K12091 and JP23H00497, JST SPRING Grant No. JPMJSP2138.

References

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Bakr, E. M.; Sun, P.; Shen, X.; Khan, F. F.; Li, L. E.; and Elhoseiny, M. 2023. HRS-Bench: Holistic, reliable and scalable benchmark for text-to-image models. In *ICCV*.

Basu, A.; Babu, R. V.; and Pruthi, D. 2023. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. In *ICCV*.

Berg, H.; Hall, S.; Bhalgat, Y.; Kirk, H.; Shtedritski, A.; and Bain, M. 2022. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning. In *AACL-IJNCLP*.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT*.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*.

Birhane, A.; Han, S.; Boddeti, V.; Luccioni, S.; et al. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. In *NeurIPS Datasets and Benchmarks Track*.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *USENIX Security Symposium*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*.

Chinchure, A.; Shukla, P.; Bhatt, G.; Salij, K.; Hosanagar, K.; Sigal, L.; and Turk, M. 2023. TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models. *arXiv preprint arXiv:2312.01261*.

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; and Tang, J. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *NeurIPS*.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. In *NeurIPS*.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A. A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. In *NeurIPS*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023a. Erasing concepts from diffusion models. In *ICCV*.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2023b. Unified concept editing in diffusion models. *arXiv preprint arXiv:2308.14761*.

Garcia, N.; Hirota, Y.; Wu, Y.; and Nakashima, Y. 2023. Uncurated Image-Text Datasets: Shedding Light on Demographic Bias. In *CVPR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*.

Hall, M.; Ross, C.; Williams, A.; Carion, N.; Drozdzal, M.; and Soriano, A. R. 2023. DIG In: Evaluating Disparities in Image Generations with Indicators for Geographic Diversity. *arXiv preprint arXiv:2308.06198*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *ICLR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Hirota, Y.; Nakashima, Y.; and Garcia, N. 2022. Gender and racial bias in visual question answering datasets. In *FAccT*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Katirai, A.; Garcia, N.; Ide, K.; Nakashima, Y.; and Kishimoto, A. 2023. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *arXiv preprint arXiv:2311.18345*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*.

Lee, T.; Yasunaga, M.; Meng, C.; Mai, Y.; Park, J. S.; Gupta, A.; Zhang, Y.; Narayanan, D.; Teufel, H. B.; Bellagente, M.; et al. 2023. Holistic Evaluation of Text-to-Image Models. In *NeurIPS Datasets and Benchmarks Track*.

Lin, A.; Paes, L. M.; Tanneru, S. H.; Srinivas, S.; and Lakkaraju, H. 2023. Word-Level Explanations for Analyzing Bias in Text-to-Image Models. *arXiv preprint arXiv:2306.05500*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *arXiv preprint arXiv:2303.05499*.

Liu, Z.; Schaldenbrand, P.; Okogwu, B.-C.; Peng, W.; Yun, Y.; Hundt, A.; Kim, J.; and Oh, J. 2024. SCoFT: Self-Contrastive Fine-Tuning for Equitable Image Generation. In *CVPR*.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *ICCV*.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Analyzing societal representations in diffusion models. In *NeurIPS*.

Mandal, A.; Leavy, S.; and Little, S. 2023. Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models. *arXiv preprint arXiv:2304.13855*.

Mannering, H. 2023. Analysing Gender Bias in Text-to-Image Models using Object Detection. *arXiv preprint arXiv:2307.08025*.

Naik, R.; and Nushi, B. 2023. Social Biases through the Text-to-Image Generation Lens. In *AIES*.

Otani, M.; Togashi, R.; Sawai, Y.; Ishigami, R.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; and Satoh, S. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Pnvr, K.; Singh, B.; Ghosh, P.; Siddiquie, B.; and Jacobs, D. 2023. LD-ZNet: A Latent Diffusion Approach for Text-Based Image Segmentation. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurIPS*.

Seshadri, P.; Singh, S.; and Elazar, Y. 2023. The Bias Amplification Paradox in Text-to-Image Generation. *arXiv preprint arXiv:2308.00755*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: a dataset for image captioning with reading comprehension. In *ECCV*.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Ture, F. 2023. What the DAAM: Interpreting stable diffusion using cross attention. In *ACL*.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *CVPR*.

Teo, C.; Abdollahzadeh, M.; and Cheung, N.-M. M. 2023. On measuring fairness in generative models. *NeurIPS*.

Ungless, E.; Ross, B.; and Lauscher, A. 2023. Stereotypes and Smut: The (Mis) representation of Non-cisgender Identities by Text-to-Image Models. In *ACL*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.

Wang, J.; Liu, X. G.; Di, Z.; Liu, Y.; and Wang, X. E. 2023a. T2IAT: Measuring Valence and Stereotypical Biases in Text-to-Image Generation. In *ACL*.

Wang, S.-Y.; Efros, A. A.; Zhu, J.-Y.; and Zhang, R. 2023b. Evaluating Data Attribution for Text-to-Image Models. In *ICCV*.

Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022. Evidence for hypodescent in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1293–1304.

Wolfe, R.; and Caliskan, A. 2022a. American== white in multimodal language-and-image ai. In *AIES*.

Wolfe, R.; and Caliskan, A. 2022b. Markedness in visual semantic ai. In *FAccT*.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *FAccT*.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *ACL*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *TMLR*.

Zhang, C.; Chen, X.; Chai, S.; Wu, C. H.; Lagun, D.; Beeler, T.; and De la Torre, F. 2023a. ITI-GEN: Inclusive Text-to-Image Generation. In *CVPR*.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2023b. Recognize Anything: A Strong Image Tagging Model. *arXiv preprint arXiv:2306.03514*.

Zhang, Y.; Jiang, L.; Turk, G.; and Yang, D. 2023c. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *EMNLP*.