The Bias Amplification Paradox in Text-to-Image Generation

Preethi Seshadri
UC Irvine
preethis@uci.edu

Sameer Singh
UC Irvine
sameer@uci.edu

Yanai Elazar
Allen Institute for AI
University of Washington
yanaiela@gmail.com

Abstract

Bias amplification is a phenomenon in which models increase imbalances present in the training data. In this paper, we study bias amplification in the text-to-image domain using Stable Diffusion by comparing gender ratios in training vs. generated images. We find that the model appears to amplify gender-occupation biases found in the training data (LAION). However, we discover that amplification can largely be attributed to discrepancies between training captions and model prompts. For example, an inherent difference is that captions from the training data often contain explicit gender information while the prompts we use do not, which leads to a distribution shift and consequently impacts bias measures. Once we account for various distributional differences between texts used for training and generation, we observe that amplification decreases considerably. Our findings illustrate the challenges of comparing biases in models and the data they are trained on, and highlight confounding factors that contribute to bias amplification.[1]

1 Introduction

Breakthroughs in machine learning have been fueled in large part by training models on massive unlabeled datasets (Gao et al., 2020; Raffel et al., 2020; Schuhmann et al., 2022). However, several studies have shown that these datasets exhibit biases and undesirable stereotypes (Birhane et al., 2021; Dodge et al., 2021; Garcia et al., 2023), which in turn impact model behavior. Given that models are trained to represent the data distribution, it is not surprising that models perpetuate biases found in the training data (De-Arteaga et al., 2019; Sap et al., 2019; Adam et al., 2022, among others).

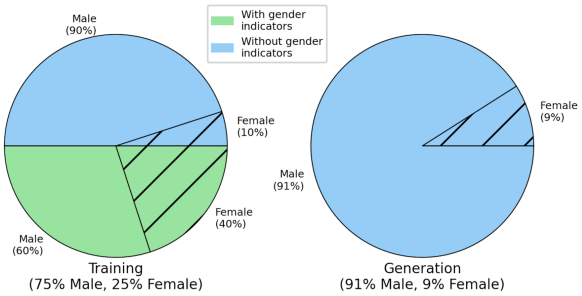


Figure 1: Comparing model generation and training data for different professions (e.g. **engineer**), the model clearly seems to amplify bias by going from 25% female in training images to 9% female in generated images. However, when looking at the subset of training examples **without gender indicators** in text captions (similar to the prompts we use), the model hardly amplifies bias (10% vs. 9% female).

Imagine a dataset where 75% of the depicted engineers are male, as shown in Figure 1. Since models learn to fit the training data, we may expect a model trained on such data to reflect this association when generating images.[2] However, it would be problematic for a model to instead exacerbate existing imbalances by generating images of engineers that are male 90% of the time. This phenomenon, known as *bias amplification* (Zhao et al., 2017), i.e. models intensify biases found in the training data, is concerning because it further reinforces stereotypes and widens disparities. While previous works suggest that models amplify biases (Zhao et al., 2017; Hall et al., 2022; Hirota et al., 2022), these works do not consider whether discrepancies between training data and model usage impact amplification.

In this paper, we investigate how model biases compare with biases found in the training data.

[1]We release the code and data used in this study at: https://github.com/preethiseshadri518/bias-amplification-paradox/

[2]Note that even such bias preservation is undesirable, but challenging to solve.

We focus on the text-to-image domain and analyze gender-occupation biases in Stable Diffusion, (Rombach et al., 2021) as well as its publicly available training dataset LAION (Schuhmann et al., 2022), which consists of image-caption pairs in English (§2). To select training examples, we identify captions that mention an occupation (e.g. "engineer") and obtain corresponding images. We follow previous work (Bianchi et al., 2023; Luccioni et al., 2023) and use prompts that contain a given occupation (e.g. "A photo of the face of an engineer") to generate images. For each occupation, we then classify binary gender to measure bias in corresponding training and generated images, and compare the respective quantities to determine whether the model amplifies biases[3] from its training data (§3).

At first glance, it appears that the model amplifies bias considerably (§4). However, we discover clear distributional differences when comparing how training texts and prompts are written, which consequently impacts amplification measurements. For example, an inherent distinction is that training captions often contain explicit gender information while prompts used to study gender-occupation biases do not.[4] As shown in Figure 1, the gender distribution for captions with gender indicators (green/bottom half) clearly differs from the distribution for captions without such indicators (blue/top half) for the occupation engineer.

To address such differences, we make prompts as close as possible to training captions by simply using the captions themselves to generate images (§5). This approach eliminates differences between captions and prompts, and the results indeed show that amplification is minimal. Then, we move to a more realistic scenario by using standard prompts to generate images, and adjusting the subset of the training data to reduce distribution shifts (§6). We propose two approaches to address distributional differences observed in qualitative evaluation: (1) We employ a nearest neighbor (NN) approach on text embeddings to select training captions that resemble prompts, and (2) We automatically detect captions that contain gender indicators (e.g. pronouns and names) and remove them from our analysis. We find that each of these approaches reduce amplification when applied individually, as well as when combined.

To summarize, we show that amplification reduces substantially when accounting for discrepancies between texts used for training and generation. Furthermore, we demonstrate that naively quantifying bias can inflate amplification measures and provide a misleading depiction of model behavior. Our work highlights that comparing biases in datasets and models is nuanced, and requires careful consideration. We hope that our work encourages future studies that analyze model behavior through the lens of the data.

2 Experimental Setup

2.1 Dataset and Models

To study bias amplification, we use Stable Diffusion (Rombach et al., 2021), a text-to-image model that generates images based on a textual description (prompt). Stable Diffusion is trained on pairs of captions and images taken from LAION-5B (Schuhmann et al., 2022), a public dataset created by scraping data from the web. We focus on two versions, Stable Diffusion 1.4 and 1.5, which are both trained on text-image pairs from the 2.3 billion English portion of LAION-5B. While both models are trained in a similar manner, Stable Diffusion 1.5 is finetuned for a longer duration on LAION-Aesthetics (a subset of higher quality images).

2.2 Gender Classification

In this work, we analyze bias in images with respect to binary gender.[5] To classify gender at scale, we utilize an automated approach. Since we do not rely on manual filtering, it is important to verify that both training and generated images include faces, and that gender is discernible from these images.

We first check whether an image contains a single face using a face detector,[6] and filter out cases where more than one face or no faces are detected. Then, we use CLIP (Radford et al., 2021), a multimodal model with zero-shot image classification capabilities, to predict gender (note that Stable Diffusion also uses CLIP's text encoder to encode

[3]We define bias as a deviation from the 50% balanced (binary) gender ratio. Note that this definition is different than other common bias measures used to measure differences in performance between groups (e.g. TPR difference), which is common in classification setups.

[4]Since we study gender bias, prompts exclude explicit gender information to avoid skewing generations.

[5]We acknowledge that our analysis excludes non-binary individuals. However, inferring non-binary gender from appearance alone leverages problematic assumptions and risks further perpetuating stereotypes against a marginalized group.

[6]https://developers.google.com/mediapipe/solutions/vision/face_detector/python

#	Prompt
1	A photo of the face of a/an [OCCUPATION]
2	A portrait photo of a/an [OCCUPATION]
3	A photo of a/an [OCCUPATION] smiling
4	A photo of a/an [OCCUPATION] at work

Table 1: The four prompts we use to generate images. "[OCCUPATION]" is a placeholder we replace with one of the 62 occupations we use.

prompts). To exclude cases where gender is difficult to infer (e.g. faces might be blurred or partially obscured), we only consider images for which the predicted probability of male or female is greater than or equal to 0.9 in our analysis.

While CLIP is also susceptible to biases (Hall et al., 2023), previous works have shown that CLIP gender predictions align with human-annotated gender labels (Bansal et al., 2022; Cho et al., 2022). In addition, we perform human evaluation with 7 participants on 200 randomly selected training and generated images. We ask participants to provide binary gender annotations (or indicate that they are unsure), and find that Krippendorff's coefficient, which measures inter-annotator agreement, is high ($\alpha = 0.948$). Additionally, 98% of CLIP predictions match the majority vote annotations.[7]

2.3 Occupations

We analyze gender-occupation biases for occupations that exhibit varying levels of bias. These include occupations that skew male (e.g. CEO, engineer, musician), fairly balanced (e.g. attorney, journalist, reporter), and female (e.g. dietitian, receptionist, therapist) according to the training data. We select a subset of common job occupations from previous works that study gender-occupation biases (Rudinger et al., 2018; Zhao et al., 2018; De-Arteaga et al., 2019). In total, we consider 62 occupations, as shown in Table 5.

3 Methodology

3.1 Measuring Model Bias

To measure biases exhibited by the model, we generate images using short prompts that contain the occupation, as shown in Table 1. These prompts deliberately do not contain any gender information since we aim to measure the biases learned by the model. Both prompts 1 and 2 also direct the model to generate faces by including "face" and "portrait". We generate 500 images per occupation and prompt. We define G_{P_o} as the percentage of females in generated images for a prompt P describing an occupation o.

3.2 Measuring Data Bias

Measuring biases in the training data is less clear-cut, because we need to identify examples that pertain to a given occupation without explicit occupation labels. Ideally, we would have access to ground truth labels to select training examples that correspond to a given occupation. Since the training data instead consists of image-caption pairs, we use captions to infer the relevant training examples. In doing so, we assume that training examples relating to a given occupation mention the occupation within the caption. We define T_{S_o} as the percentage of females in training images for a training subset S corresponding to occupation o (we provide details of how examples are chosen in Section 4)

3.3 Evaluating Bias Amplification

We first measure whether occupations skew female or male in both the model (§3.1) and training data (§3.2). Then, we compute bias amplification by comparing the % female in training vs. generated images using the general approach outlined in Zhao et al. (2017). We define amplification for a specific occupation o as the following:

$$A_{P_o,S_o} = |G_{P_o} - 50| - |T_{S_o} - 50| \quad (1)$$

This formulation takes into account that amplification for a given occupation is specific to the prompt P_o used to generate images, as well as the chosen subset of training examples S_o. For a set of occupations O, the expected amplification is computed as follows:

$$\mathop{\mathbb{E}}_{o \in O}[A_{P_o,S_o}] = \frac{1}{|O|} \sum_{o \in O} A_{P_o,S_o}. \quad (2)$$

As shown above, A_{P_o,S_o} is calculated for each occupation and aggregated across occupations in O to obtain $\mathbb{E}[A_{P_o,S_o}]$ for each prompt. We then average $\mathbb{E}[A_{P_o,S_o}]$ across all four prompts. For occupations that skew male in the training data, bias amplifies if bias skews more male in generated images, and vice versa for occupations that skew female. If bias decreases from training to generation, this behavior is considered de-amplification.

[7] We acknowledge that leveraging appearance (e.g. hair, clothing, etc.) to determine gender has fundamental limitations and perpetuates gender stereotypes.

Caption	Details
Portrait of smiling young female **mechanic** inspecting a CV joint on a car in an auto repair shop	Contains person description (smiling young female), activity, and location
Muscular bearded **athlete** drinks water after good workout session in city park	Contains person description (muscular bearded), clues about attire (workout clothes), and activity
Portrait of a **salesperson** standing in front of electrical wire spool with arms crossed in hardware store	Contains activity, information about surroundings, and location

Table 2: Training captions often include additional context and details (e.g. descriptions, location and activity information) that reduce ambiguity, as shown in these examples. In contrast, the prompts we use to generate images, as shown in Table 1, lack specificity and can refer to a larger set of scenarios.

We exclude occupations that exhibit different directions of bias at training and generation from our analysis altogether (i.e. switch from skewing male to female, and vice versa), since this behavior does not adhere to our notion of bias amplification. In total, there are eight occupations (assistant, athlete, author, dentist, graphic designer, painter, supervisor, tutor) that exhibit switching behavior between training and generation on all prompts.

4 Keyword Querying

We start by examining the extent to Stable Diffusion amplifies gender-occupation biases from the data by selecting training examples that contain a given occupation in the caption (e.g. all captions that contain the word 'president'). We refer to this procedure as *keyword querying*. In practice, we sample a subset of 500 training examples as opposed to using all training examples. We present the bias amplification results for each occupation using *keyword querying* in Figure 2. Stable Diffusion seems to amplify bias relative to the training data by 12.57%[8] on average across all occupations and prompts (10.24% for prompt 1, as shown in Figure 2). This behavior is concerning because instead of reflecting the training data and its statistics, the model compounds bias by further underrepresenting groups. However, when qualitatively inspecting examples, we observe discrepancies in how occupations are presented in captions vs. prompts due to varying levels of ambiguity.

Prompts commonly used to study gender-occupation bias are intentionally underspecified, or lack detail. Underspecification results in the model having to generate images from textual inputs that are vague and open to interpretation (Hutchinson et al., 2022; Mehrabi et al., 2023). For example, the prompt "A photo of the face of a/an [OCCUPATION]", does not contain any attributes about the individual or information about what they are doing, what their surroundings are, etc. In contrast, captions may contain context or details that result in less ambiguous descriptions, as shown in Table 2. We specifically showcase examples that include descriptions of the individual and provide information about the activity they are engaged in (e.g. inspecting a CV joint).

Discrepancies in how captions and prompts are written also impact how occupations are depicted in training and generated images. These differences are especially notable for occupations that

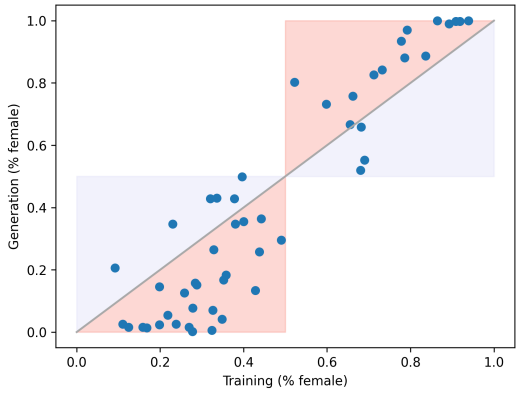
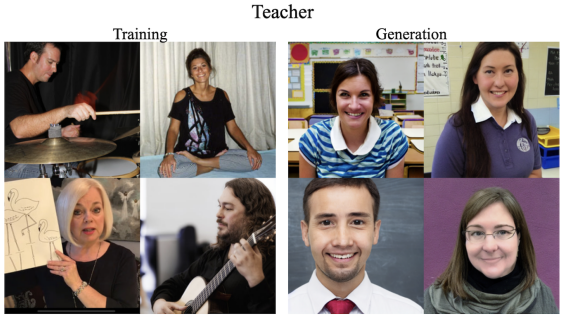


Figure 2: **Bias amplification for keyword querying** There appears to be consistent bias amplification between training and generated images when sampling training examples that mention a given occupation. The x-axis corresponds to the % female in training images, and the y-axis corresponds to the % female in generated images (generated with prompt 1). Each point in the plot represents a different occupation, and regions are shaded based on Amplification and De-Amplification.

[8]We report and discuss values for Stable Diffusion 1.4 in the paper, but results for both model versions are in Table 4.

(a) Training captions for **President**: 1) "Leana Wen, Planned Parenthood president..." 2) "New Schaumburg Business Association President Kaili Harding..." 3) "BCCI president N Srinivasan..." 4) "Larry Bird, Indiana Pacers president of basketball operations..."



(b) Training captions for **Teacher**: 1) "Brad Draper, percussion teacher..." 2) "teacher/author in the 80s sits in yoga lotus pose..." 3) "Jo Anne Young Art Teacher..." 4) "patrick oconnell Classical Guitar Teacher..."

Figure 3: Examples of discrepancies in how occupations are depicted in training (*keyword querying*) vs. generated examples for **President** (left) and **Teacher** (right).

have multiple interpretations. For example, when we query for training examples containing president, the resulting captions refer to various types of presidents, including the president of a company or organization (as shown in Figure 3a). However, when generating images using the prompt "A photo of the face of a president", the model appears to interpret president as a leader of a country, often the United States. Another example that illustrates similar differences is teacher, which broadly refers to anyone whose job is to teach, instruct, or train. In the training data, teacher occurs in both school and non-school contexts, such as a yoga teacher or music teacher[9] (as shown in Figure 3b). In contrast, generated images primarily depict teachers in a classroom or headshots of teachers. Without additional information or context in the prompt, the model seems to leverage common associations when generating images for these occupations.

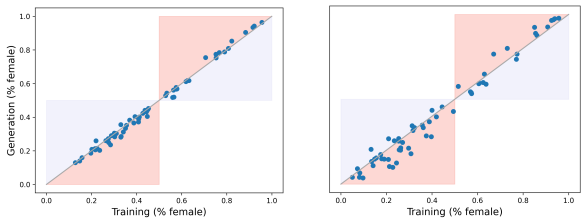
Finally, we also observe that explicit gender indicators are a crucial component of captions. For example, we notice the use of gender indicators to emphasize uncommon co-occurrences, such as male hairdressers or female engineers in our initial example (Figure 1). While gender information is used both to describe overrepresented and underrepresented gender groups for a given occupation, we hypothesize that usage is more common for underrepresented groups. If this hypothesis holds, the gender distribution in resulting training images would shift closer towards balanced in resulting training images. As a result, the decision to focus on all captions vs. captions without any gender indicators can exaggerate bias measures.

In order to make reasonable comparisons between bias at training vs. generation, we should compare gender ratios over similar captions and prompts. Therefore, we cannot conclude whether differences in gender ratios at training and generation are due to the model amplifying bias, or other confounding factors that contribute to amplification. In the next section, we address these distribution shifts by providing training captions to the model as prompts.

5 Removing Distributional Differences: A Lower Bound

Although the model seems to amplify bias with the keyword querying approach (§4), we observe prominent differences when comparing training and generated examples. These discrepancies between captions and prompts can impact bias amplification measures, and therefore conclusions about model behavior. To reduce these discrepancies we can either (1) modify the prompts we use to more closely resemble examples from the training data, or (2) modify the procedure for selecting training examples. We start by following the first approach, and prompt the model with training captions instead of using prompts. The training captions (S_o) remain the same as before, but the prompts (P_o) match the captions, verbatim. For every prompt in P_o, we generate 10 images (using Stable Diffusion 1.4), and then compute amplification using

[9]For both occupations, we hand-pick training examples that are illustrative of mismatches.



(a) All Captions (b) Captions without Gender Indicators

Figure 4: **Bias amplification when prompting with training captions.** If we feed training captions verbatim to the model as prompts, we observe minimal amplification (left). This behavior mostly holds when focusing on captions without explicit gender indicators (right). Regions are shaded based on Amplification and De-Amplification.

$P_o := S_o$ for each occupation.

By design, this approach removes mismatches between captions and prompts, since prompts and captions are now identical. We then ask, does imposing consistency between the texts used for measuring training and generation bias lower amplification? We hypothesize that enforcing prompts and captions to match yields similar bias measurements, which in turn reduces amplification. As shown in Figure 4a, amplification is minimal when $P_o = S_o$ and most occupations reside along the diagonal (no amplification). The average amplification drops to 0.68%, indicating that the model mostly reflects training data.[10] Furthermore, amplification remains consistently low, even for occupations that are highly imbalanced.

It is worth noting that the model achieves near-zero amplification on captions that contain explicit gender information (a substantial fraction of examples). For examples that contain either male or female gender indicators, the model is able to easily generate images that match the gender of corresponding training images. Therefore, we analyze results separately on the subset of captions without gender indicators. As shown in Figure 4b, bias amplification is larger for the no gender indicator subset as compared to all captions. That being said, the average amplification remains low at 2.05% (↓ 84% relative to keyword querying).[10]

Another discrepancy between training and generation is the difference in the directional usage of texts and images (Jin et al., 2021). In the training data, captions accompany images to describe what is depicted (training image → caption). On the other hand, prompts guide the image generation process and direct model output (prompt → generated image). By providing captions as prompts, we bridge this gap (training image → caption → generated image). Although practitioners are unlikely to utilize prompts that exactly match training captions, this experiment highlights the importance of distributional similarity between captions and prompts when comparing biases. In addition, it provides a lower bound to the bias amplification problem. In summary, we conclude that the model primarily mimics biases from the training data when prompted with captions used for training. In the next section, we explore other approaches to reduce distribution shifts between training and generation.

6 Reducing Discrepancies

In our initial approach, *keyword querying* (§4), we do not impose any restrictions or filtering criteria to select training examples beyond the mention of a given occupation, which has clear limitations as we discuss in Section 4. In this section, we introduce and evaluate approaches to restrict the search space of training examples, with the goal of reducing distribution shifts between training and generation.

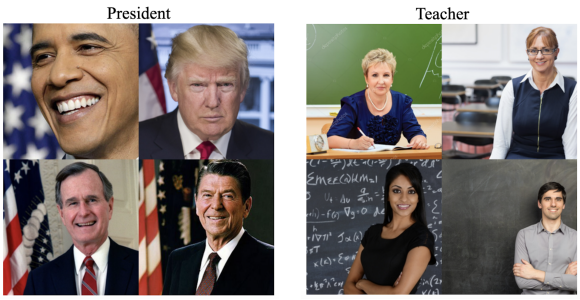
6.1 Nearest Neighbors (NN)

The results from directly prompting the model with captions (§5) indicate that bias amplification is minimal when controlling for distributional differences between captions and prompts. However, our preliminary approach (§4) does not take these differences into consideration. The prompts we use are concise and structured, but lack detail and specification. On the other hand, randomly sampled training captions for a given occupation are more diverse and vary in their usage of the occupation and the amount of contextual information, as highlighted in Table 2 and Figure 3.[11] These qualitative differences are also apparent when comparing caption and prompt text embeddings. We use Sentence-BERT[12] (Reimers and Gurevych, 2019) to compute text embeddings, and calculate the average pairwise cosine similarity between caption and prompt embeddings for each occupation. We find

[10] However, we reject the null hypothesis that the expected amplification is 0 using a one-sample t-test.

[11] For instance, consider the caption presented in Figure 3: "New Schaumburg Business Association President Kaili Harding speaks Tuesday during the association's monthly Good Morning Schaumburg breakfast."

[12] We use the all-MiniLM-L6-v2 model for text embeddings



(a) Training captions for **President**: 1) "The president is pictured smiling." 2) "President Donald J. Trump - Official Photo" 3) "Portrait of President George H. W. Bush" 4) "Official Portrait of President Ronald Reagan"

(b) Training captions for **Teacher**: 1) "Picture of a teacher in the classroom" 2) "Portrait of a smiling teacher in a classroom." 3) "Portrait of teacher woman working" 4) "Teacher smiling in classroom, portrait"

Figure 5: **Training examples chosen with Nearest Neighbors**. Selected training captions and images are more similar to prompts and generated images for both occupations.

that the average cosine similarity across occupations is 0.385, indicating that captions and prompts are highly dissimilar (relative to similarity using NN, which we will see next).

Addressing Similarity Discrepancies To account for these gaps, we propose using nearest neighbors (NN) on text embeddings to select captions that closely resemble prompts. NN considers all captions that contain a given occupation, as done in keyword querying, but selects training examples based on the similarity between captions and prompts instead of randomly sampling a subset. As a result, the chosen captions are closer in structure and wording to prompts. We use Sentence-BERT to obtain text embeddings and compute the cosine similarity between embeddings to measure the similarity between captions and prompts.[13] For a given occupation, we consider the top-k similar captions, where $k = 500$.

Intuitively, the previous section addresses distributional differences by enforcing prompts to be identical to captions, while NN instead selects captions that are similar to prompts. Once we apply NN, the average cosine similarity between caption and prompt embeddings increases to 0.704 (↑ 83% from keyword querying), which happens by design since we are directly targeting examples that resemble prompts. Note however, that the text embedding similarity increase is also reflected in image embeddings. As shown in Table 3, the pairwise similarity of CLIP image embeddings increases with NN (↑ 13% from keyword querying), which indicates that training images corresponding to NN captions are more similar to generated images.

[13]We acknowledge that the text embedding used for computing NN can reinforce certain biases. While perhaps CLIP and Sentence-BERT exhibit similar biases, our rationale for choosing the latter is to avoid leaking biases from Stable Diffusion's text encoder when selecting training examples.

Occupation	**Keyword Querying**	**NN**
Teacher	0.469	0.570
President	0.379	0.444
All Occupations	0.519	0.586

Table 3: Average pairwise cosine similarity between training and generated image embeddings for prompt 1. Embeddings are computed using CLIP's image encoder.

There are noticeable qualitative improvements as well. Going back to our prior example with president and teacher, NN appears to choose captions that are closer in structure and meaning to the prompts. By reducing discrepancies between captions and prompts, we find that training images are also more consistent with generated images. As shown in Figure 5, the training images corresponding NN captions to the word 'president' represent world leaders (often US presidents). This is in contrast to the *keyword querying* approach that often returned presidents of an organization or company (Figure 3). Similarly, training images for teacher depict educators sitting at a classroom desk or standing in front of a blackboard as opposed to art and music teachers, as we saw previously.

Reduced Bias Amplification When selecting training examples S_o using NN, we see that bias amplification reduces considerably across occupations and prompts. The results are described in Table 4. The average amplification drops to 6.76% after applying NN (↓ 46% relative to keyword querying). While NN increases the similarity between training and generated examples, there are still unresolved sources of distribution shift that impact amplification measures.

Approach	Model	Bias Amplification
Keyword Querying	SD 1.4	12.57
	SD 1.5	12.07
Nearest Neighbors	SD 1.4	6.76
	SD 1.5	6.01
No Gender Indicators	SD 1.4	8.66
	SD 1.5	7.97
Nearest Neighbors + No Gender Indicators	SD 1.4	4.35
	SD 1.5	3.59

Table 4: Average Bias Amplification (BA) across occupations and prompts. Results are shown both for Stable Diffuson (SD) 1.4 and 1.5. Bias amplification lowers considerably when using nearest neighbors to select training captions and excluding captions with gender indicators. We see further reductions when combining approaches.

6.2 Captions Without Explicit Gender Indicators

Another notable distinction between training captions and prompts is the use of explicit gender indicators. On average, more than half of the captions (59.5%) contain some form of explicit gender information. Furthermore, gender usage in captions varies depending on which gender is underrepresented for a given occupation. For example, images of female mechanics in the training data frequently accompany captions that explicitly indicate the mechanic is female. However, this specification is less common for male mechanics, since mechanics are often associated with males (30% of male mechanic examples contain explicit gender indicators, as opposed to 68% for female mechanic examples).

To validate these observations, we compute the correlation between the percentage of females in training images and the percentage of captions with female indicators. We expect a negative correlation, since we hypothesize that occupations that skew female are less likely to contain explicit female gender indicators in captions. The Pearson's correlation coefficient is indeed negative, with a coefficient value of -0.458 and statistically significant (significance level < 0.05). These results suggest that including training examples with gender information, used by the naive *keyword querying* approach, may inflate bias amplification measures.

Addressing Gender Indicators To assess whether amplification differs for the subset of captions without indicators, we split our training examples by detecting gender indicators in the captions. We consider explicit gender words,[14] binary gender pronouns, and names[15] to infer gender. We focus on the subset of training captions, S_o, without any male or female indicators in our analysis below.

Reduced Bias Amplification We observe that bias amplification is noticeably lower when focusing on the no-gender indicator subset of training examples. Compared to the initial amplification of 12.57% for keyword querying, the average amplification for captions without gender indicators is 8.66% (↓ 31%), as shown in Table 4. This behavior aligns with the reasoning described above — gender indicators are more likely to delineate the presence of the underrepresented gender, which drives the % female in resulting images closer to 50% and in turn increases amplification measures when naively evaluating amplification.

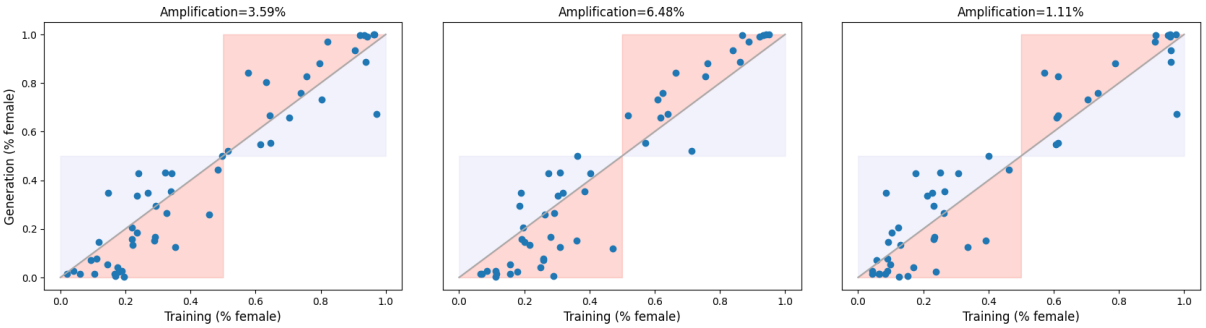
6.3 Combining Approaches

While NN and filtering explicit gender indicators reduce distributional differences when applied individually, perhaps both approaches behave in complementary ways. When combining the no-gender indicator subset with NN, reductions in amplification further compound, as shown in the last rows in Table 4. The average amplification decreases to 4.35%, which is noticeably lower compared to the values for each individual method. Both methods work in tandem to reduce distributional differences in non-overlapping ways (at least partially). We also observe greater reductions for specific prompts; as shown in Figure 6c, the average amplification is just 1.11% for prompt 1 (↓ 89% relative to the keyword querying value of 10.24% for prompt 1).

Note that even after combining methods, we see that points in Figure 6c are dispersed rather than residing along the diagonal (zero amplification). The overall reduction in amplification is largely due to an increase in occupations exhibiting deamplification, instead of most occupations exhibit-

[14] male/female, man/woman, gent/gentleman/lady, boy/girl

[15] We perform named entity recognition using the en_core_web_lg model from spaCy to identify name mentions, and then use the gender-guesser library https://pypi.org/project/gender-guesser/



(a) Nearest Neighbors

(b) Captions without Gender Indicators

(c) Combined Approach

Figure 6: **Bias amplification for various approaches to address discrepancies between training and generation.** The proposed approaches yield lower bias amplification, especially the combined method (c). Results are shown for prompt 1. Regions are shaded based on Amplification and De-Amplification.

ing near-zero amplification. However, this behavior may still align with A_{P_o,S_o} as normally distributed with its expectation equal to zero. We perform a one-sample t-test to test the null hypothesis that the expected amplification is equal to 0 for each of the prompts; we fail to reject the null hypothesis for prompts 1 and 3 and reject the null hypothesis for prompts 2 and 4 (significance level < 0.05) . Our results indicate a portion of amplification is still unexplained for all prompts, particularly prompts 2 and 4, and may involve more complex and subtle confounding factors. Although the proposed methods do not account for all possible discrepancies between training and generation, we are able to bring the bias measures closer together as various differences are addressed.

7 Related Work

Relating pretraining data to model behavior There is a growing body of work focused on studying pretraining data properties and statistics, as well as understanding their impact on model behavior. This type of large-scale data and model analysis provides useful insights into model learning and generalization capabilities (Carlini et al., 2023). Recent work shows that few-shot capabilities of large language models are highly correlated with pretraining term frequencies, and that models struggle to learn long-tail knowledge (Kandpal et al., 2023; Razeghi et al., 2022).

Several works have also explored the relationship between pretraining data and model performance from a causal perspective. Biderman et al. (2023) demonstrate that re-training models with a gender swapping intervention during the latter stages of pretraining reduces gender bias on targeted benchmarks. Elazar et al. (2023) introduce a framework to estimate causal effects between co-occurrences in the pretraining data and model predictions, without requiring any model re-training. Longpre et al. (2023) comprehensively investigate how various data curation choices and pretraining data slices affect downstream task performance.

Bias Amplification Our work is strongly inspired by the findings of Zhao et al. (2017), who show that structured prediction models amplify biases present in the data. However, there are important differences between our works. First, their task involves jointly predicting multiple target labels, including gender, as opposed to generating images. Additionally, they use a pretrained convolutional neural network to extract features, which may contain different biases from the training data for the task. As a result, it is unclear how much of the observed amplification is due to the training data alone, as opposed to biases encoded in the pretrained model. Although Stable Diffusion suffers from a similar problem (since it uses CLIP), we refrain from making definitive conclusions about amplification and instead focus on reducing confounding factors that impact amplification measures. Wang and Russakovsky (2021) highlight that Zhao et al. (2017) conflate different types of bias amplification and propose a new metric, but also do not decouple bias from the pretrained model vs. the data. Wang et al. (2018) follow the same setup as Zhao et al. (2017) and discover that even models trained on balanced datasets in which male and female examples co-occur equally with target variables amplify bias, which is consistent with earlier findings from

Elazar and Goldberg (2018) on a different experimental setup. The authors posit that amplification occurs due to gender-correlated features in the data that behave as spurious correlations.

Bias in text-to-image models While it is well-established that language and vision models are susceptible to biases individually, recent work has shown that text-to-image models are prone to similar biases and often exhibit associations that perpetuate stereotypes. Zhang et al. (2023) propose a method to automatically evaluate the extent to which text-to-image models portray men vs. women differently in various contexts. Building on implicit association tests from the language domain, Wang et al. (2023) introduce a framework to quantify implicit stereotypes in generated images. Furthermore, both Fraser et al. (2023) and Luccioni et al. (2023) highlight intersectional biases in text-to-models along multiple stereotype dimensions. Bianchi et al. (2023) demonstrate that stereotypes persist even after explicitly prompting the model with counter-stereotypes. However, these works focus on evaluating various model biases, and do not focus on the training data.

Friedrich et al. (2023) analyze biases exhibited in LAION and by Stable Diffusion, and demonstrate that the model exhibits bias amplification. Instead of identifying relevant training examples using captions as done in our work, they select training images using text-image similarity between prompts and training images. However, their paper primarily focuses on mitigating biases using their proposed approach, fair diffusion, as opposed to analyzing confounding factors and sources of distribution shift. We encourage future work to evaluate training example selection using captions vs. images, and provide a comparative analysis.

8 Discussion

Generalizability Our work demonstrates that evaluating bias amplification is nuanced, and using naive procedures can lead to exaggerated amplification measures. However, we acknowledge that our analysis does not account for all possible sources of distribution shift that contribute to amplification, since our work is illustrative and not exhaustive. Moreover, it is important to investigate bias amplification for various experimental setups to determine if similar confounding factors are present. Although our findings indicate that confounding factors play a huge role in the amplification of gender-occupation biases in Stable Diffusion, it remains unclear to what extent these findings apply more broadly. We encourage future studies to expand upon our findings by examining different datasets, models, and types of bias, and highlighting similarities and differences. By doing so, we can gain a more comprehensive understanding of bias amplification.

Variation Across Prompts As we highlight in Figure 7, even small changes to prompts can have a resounding impact on conclusions about model bias. For example, "A portrait photo of an attorney" skews heavily male while "A photo of an attorney at work" skews female in generated images. While we observe a consistent decrease in amplification across prompts as we address distributional differences, the relative reduction differs based on the prompt. For example, prompt 1 exhibits an 89% reduction while prompt 2 only exhibits a 49% reduction, which indicates that the confounding factors we have identified have a varying impact. Overall, these results suggest that there may also be prompt-specific sources of distribution shift, which is an important consideration when choosing prompts.

External Biases In addition to the training data, another source of bias is the text embeddings obtained from the text encoder (CLIP) used by Stable Diffusion. By solely comparing biases found in the data vs. those exhibited by the model, our analysis overlooks biases that arise from encoding the prompt. As a result, we cannot disentangle how much this component impacts bias measures and isolate its contribution toward overall amplification. Note that the effect of such an external embedding cannot be easily accounted for, since CLIP's training data is not public. More work is needed to understand the effect of using external, frozen models as a part of self-supervised models such as Stable Diffusion.

Amplification Definition In this work, we define bias amplification as models exacerbating biases found in the training data. However, some works adopt a different interpretation of bias amplification altogether (Kirk et al., 2021; Bianchi et al., 2023) and compare model bias to real-world statistics (e.g. labor force statistics). Both definitions are useful to study but answer fundamentally different questions. Our definition offers insights into what models learn and whether model behavior reflects training data, while the real-world bias amplification approach captures how well the data and model

(a) A photo of the face of an attorney

(b) A portrait photo of an attorney

(c) A photo of an attorney smiling

(d) A photo of an attorney at work

Figure 7: Generated examples for the occupation **attorney** using different prompts. Specific wording and phrasing choices in prompts lead to noticeable differences in the % of female images generated by the specific prompt. Yellow boxes indicate images predicted as female. Although we only include 9 images per prompt here, these proportions are similar to what is exhibited in the 500 generated images.

as a collective system reflect reality.

Connection to Simpson's Paradox The title of our paper alludes to Simpson's Paradox (Simpson, 1951), a phenomenon in which a trend or relationship observed in subgroups within the data reverses or disappears when subgroups are combined. A well-known example of Simpson's Paradox is a study on gender bias in graduate admissions at UC Berkeley (Bickel et al., 1975). The aggregate results from this study showed that men were more likely to be admitted to graduate programs than women. However, when analyzing results at a department level, it turned out women were slightly favored, but applied to departments with lower admission rates more often. Therefore, by disaggregating their results, the researchers accounted for the confounding factor of varying admission rates amongst different departments. Similarly, we draw direct parallels to our analysis and insights. Although we observe substantial amplification in our initial setup, amplification reduces drastically after selecting specific subsets of the training data and decreasing the impact of confounding factors.

9 Conclusion

In summary, we study how data and model biases are related by investigating whether models amplify bias. We discover that distributional differences between training and generated examples impact bias measures, and therefore our understanding of model behavior. We propose approaches to reduce discrepancies between captions and prompts, and find these to be effective at bringing data and model bias closer together. Although amplification is not eliminated altogether, we observe substantially lower amplification measurements. It is important to emphasize our specific approaches may not directly apply to setups studying different datasets, tasks, models, or types of bias. Nevertheless, our findings highlight how confounding factors can inflate bias amplification, which has broader implications. We recommend that evaluations comparing training data and model bias, or any dataset and model properties more generally, account for distribution shifts that can skew the analysis.

Ethics Statement

Bias Definition Our work focuses on a narrow slice of social bias analysis by studying gender-occupation stereotypes. However, since models exhibit various types of discriminatory bias (e.g., racial, age, geographical, socioeconomic, disability, etc.), as well as intersectional biases, it is equally important to perform evaluations for these definitions of bias. Furthermore, we only consider binary gender, which has clear drawbacks. Our analysis ignores how text-to-image models perpetuate biases for non-binary identities and relies on information such as appearance and facial features to infer gender in training and generated images, which can further propagate gender stereotypes.

Gender Classification We automate gender classification using CLIP because previous works have shown that CLIP gender predictions align with human annotations and CLIP gender classification performance on the FairFace dataset[16] is strong

[16]https://github.com/joojs/fairface

($> 95\%$) across various racial categories. We also show that CLIP predictions are consistent with our small-scale human evaluation study. Nevertheless, we recognize the limitations of using a model to classify gender in images, since CLIP inherits biases from its training data.

Geographical Diversity The captions and prompts used to study bias are solely written in English. We hope future work will shed light on multilingual bias amplification in text-to-image models. It is also worth noting that the gender-guesser library (infers gender from names) likely performs worse on non-Western names. The documentation mentions that the library supports over 40,000 names and covers a "vast majority of first names in all European countries and in some overseas countries (e.g. China, India, Japan, USA)". Therefore, the name coverage (or lack thereof) impacts our ability to identify captions with gender information.

References

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 7–21, New York, NY, USA. Association for Computing Machinery.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA. Association for Computing Machinery.

Peter J. Bickel, Eugene a. Hammel, and J W O'connell. 1975. Sex bias in graduate admissions: Data from berkeley. *Science*, 187:398 – 404.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. *arXiv preprint arXiv:2202.04053*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2023. Measuring causal effects of data statistics on language model's 'factual' predictions.

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2023. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified?

Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *ArXiv*, abs/2302.10893.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6957–6966.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-language models performing zero-shot tasks exhibit gender-based disparities.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. 2022. A systematic study of bias amplification.

Y. Hirota, Y. Nakashima, and N. Garcia. 2022. Quantifying societal bias amplification in image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13440–13449, Los Alamitos, CA, USA. IEEE Computer Society.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1172–1184, Online only. Association for Computational Linguistics.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Neural Information Processing Systems*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14367–14388, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294.

English Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the royal statistical society series b-methodological*, 13:238–241.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10882–10893. PMLR.

Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. 2023. T2IAT: Measuring valence and stereotypical biases in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2560–2574, Toronto, Canada. Association for Computational Linguistics.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2018. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318.

Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. 2023. Auditing gender presentation differences in text-to-image models.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. 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, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 List of Occupations

A full list of occupations is shown in Table 5.

A.2 Average Amplification Plots

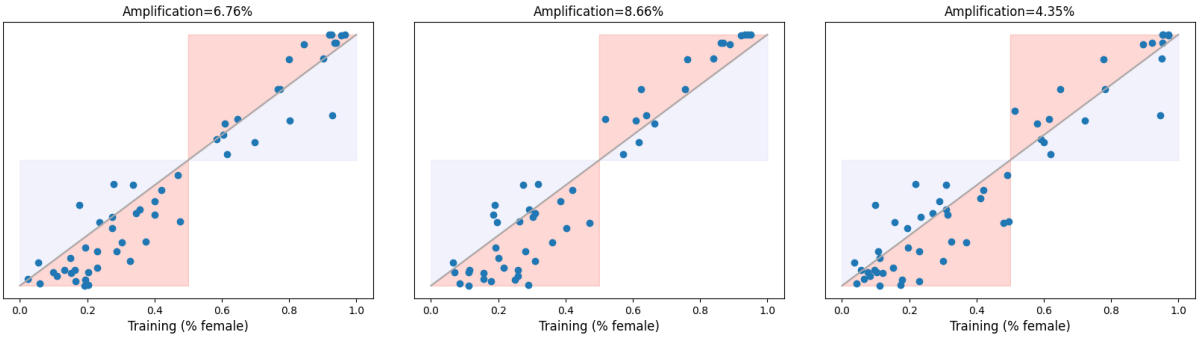
Instead of visualizing amplification for only prompt 1 (as seen in Figure 6), we show amplification for our proposed approaches averaged across all prompts using Stable Diffusion 1.4 in Figure 8. Although points are dispersed for both the former and latter cases, we see that (1) average amplification is higher and (2) fewer points exhibit de-amplification when averaging.

A.3 Prompt Specific Results

We show results on all our approaches for each prompt, instead of averaging across prompts in Table 6. We observe consistent reductions in amplification for each prompt when using NN and excluding captions with gender indicators, with prompt 1 resulting in the lowest amplification. In addition, we see differences in which model exhibits lower amplification based on the prompt. For example, Stable Diffusion 1.4 exhibits lower amplification consistently on prompt 1, while Stable Diffusion 1.5 exhibits lower amplification consistently on prompts 2 and 4.

Occupations				
accountant	dentist	journalist	poet	singer
architect	dietitian	lawyer	politician	student
assistant	doctor	librarian	president	supervisor
athlete	engineer	manager	prime minister	surgeon
attorney	entrepreneur	mechanic	professor	teacher
author	fashion designer	musician	programmer	technician
baker	filmmaker	nurse	psychologist	therapist
bartender	firefighter	nutritionist	receptionist	tutor
ceo	graphic designer	painter	reporter	veterinarian
chef	hairdresser	pharmacist	researcher	writer
comedian	housekeeper	photographer	salesperson	
cook	intern	physician	scientist	
dancer	janitor	pilot	senator	

Table 5: List of 62 occupations used to study gender-occupation biases.



(a) Nearest Neighbors

(b) Captions without Gender Indicators

(c) Combined Approach

Figure 8: **Bias amplification for various approaches to address discrepancies between training and generation.** The proposed approaches yield lower bias amplification, especially the combined method (c). Results are averaged across all prompts. Regions are shaded based on Amplification and De-Amplification.

Approach	Model	Prompt 1	Prompt 2	Prompt 3	Prompt 4
Keyword Querying	SD 1.4	10.24	17.57	10.77	11.68
	SD 1.5	10.87	16.36	11.15	9.91
Nearest Neighbors	SD 1.4	3.59	12.62	5.58	5.27
	SD 1.5	4.01	11.14	5.21	3.65
No Gender Indicators	SD 1.4	6.49	13.58	7.09	7.49
	SD 1.5	6.76	12.41	6.82	5.87
Nearest Neighbors + No Gender Indicators	SD 1.4	1.11	8.72	3.06	4.05
	SD 1.5	1.55	7.29	2.78	2.72

Table 6: Average Bias Amplification (BA) across occupations, for each prompt. Results are shown both for Stable Diffuson (SD) 1.4 and 1.5. For each prompt, bias amplification lowers considerably when using nearest neighbors to select training captions and excluding captions with gender indicators. We see further reductions when combining approaches.