A Large Scale Analysis of Gender Biases in Text-to-Image Generative Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the increasing use of image generation technology, understanding its social biases, including gender bias, is essential. This paper presents a large-scale study on gender bias in text-to-image (T2I) models, focusing on everyday situations. While previous research has examined biases in occupations, we extend this analysis to gender associations in daily activities, objects, and contexts. We create a dataset of 3,217 gender-neutral prompts and generate 200 images per prompt from five leading T2I models. We automatically detect the perceived gender of people in the generated images and filter out images with no person or multiple people of different genders, leaving 2,293,295 images. To enable a broad analysis of gender bias in T2I models, we group prompts into semantically similar concepts and calculate the proportion of male- and female-gendered images for each prompt. Our analysis shows that T2I models reinforce traditional gender roles, and reflect common gender stereotypes in household roles. Women are predominantly portrayed in care- and human-centered scenarios, and men in technical or physical labor scenarios. Code and prompts to evaluate models will be released upon acceptance.

1 Introduction

Rapid advances in image generation technology make it easier than ever to automatically generate large amounts of synthetic images. State-of-the-art text-to-image (T2I) models [10, 92] can generate high-quality images from arbitrary text instructions. Their capabilities are further enhanced through editing [12, 57, 110] and personalization [9, 38, 58, 83] techniques. Synthetic images are not only used in everyday applications such as advertisements [60] and presentation slides [75] but are also increasingly used as training data for other foundation models [34, 58, 95, 96].

As the availability and proliferation of synthetic images increase, so does their power to influence society and amplify any harms originating from the underlying models [17]. In their seminal work, [14] discovered intersectional gender and racial biases in image recognition systems. Within the research community, the list of known biases has only grown: [5, 7, 43, 90, 93] identified social biases in CLIP [77], such as associating men with words related to criminal activities. [45, 50, 51] found social biases in automatic image captioning. [36, 40, 82, 105, 109] uncovered various social biases in multimodal large language models (MLLMs). These examples demonstrate that social bias pervades all aspects of modern generative AI systems.

Research on social bias in T2I models has led to a large body of work covering all computational aspects of social bias, including bias analysis [8, 23, 63], open bias detection [22, 27, 28], and model debiasing [6, 32, 37, 108]. However, most research in this area has focused on gender-occupation bias [98]. While this is an important issue, other aspects of daily life, such as everyday activities and stereotypical contexts, also contribute to perpetuating or amplifying harmful social biases and require

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

careful analysis. Furthermore, many studies on social bias in T2I models use a limited set of prompts and only a few images per prompt, reducing their representativeness.

We present a large-scale, in-depth study on gender bias in T2I models concerning everyday scenrios and address gaps in the literature by analyzing everyday activities complemented by gender-object and gender-context associations. Our main contributions are: (1) We compile 3,217 gender-neutral prompts over four categories to probe T2I models for gender bias and generate 200 images from five state-of-the-art T2I models for each prompt and filter unsuitable images, leaving 2,293,295 images for analysis; (2) We design a carefully structured experimental setup to analyze gender bias in large image datasets and systematically examine the observed gender biases and relate them to known human gender biases; (3) We analyze bias amplification in activities compared to LAION-400m and confirm bias amplification in occupations wrt. U.S. labor statistics. Through this work, we take an important step toward addressing gender stereotypes in T2I models, helping to understand and document perpetuation of gender inequality in AI technology [2, 94].

2 Related Work

Prior work on evaluating social bias in text-to-image models can be categorized into bias analysis and open bias detection. Notable works on open bias detection include TIBET [22], OpenBias [28], and OASIS [27]. Rather than proposing a method for open bias detection, we conduct a careful, large-scale study of gender bias in state-of-the-art diffusion models, analyzing gender bias in T2I models by exploring gendered outputs in gender-neutral prompts.

Work analyzing social bias in text-to-image (T2I) generation has traditionally focused on a few categories [98], most notably perceived gender and race, which are protected under U.S. federal anti-discrimination laws. Analyses have primarily considered occupation-related prompts. In their seminal work, [8] find that T2I models amplify biases in occupations and personal attributes. However, their study includes only 20 manually examined prompts. Our work presents a large-scale, automatic analysis that offers broader insights into specific biases. [21] examine gender and racial bias in occupations and report that generated images are more gender-imbalanced than actual U.S. labor statistics. We confirm and extend their findings by also analyzing gender bias and bias amplification with respect to activities, not just occupations. However, [89] provide evidence that bias amplification may result from specific prompt design choices. To ensure the validity of our results, we incorporate prompt variations and use bounding boxes for automatic gender labeling. [63] propose a method to analyze gender and racial bias without explicitly labeling images by gender or race. They identify distributional differences in 4,380 images generated from 146 occupation prompts. Since their evaluation uses a limited set of prompts, they explicitly call for more in-depth studies of gender bias in T2I models. This is a gap we address in this work. Similarly, [23] observe distributional differences in 737 images from 83 occupation prompts using gendered and gender-neutral phrasing. In contrast, we adopt a more rigorous experimental setup by sampling more images and filtering and cropping unsuitable ones. Moreover, we consider a broader range of everyday scenarios beyond occupations.

Previously, [67, 102, 111] investigate gender bias in person-object co-occurrence, though these insights are limited to a narrow set of objects, mainly clothing. Our study demonstrates gender bias in contexts such as traditional household roles, which hold significant societal relevance. [97] find that non-binary identities are poorly represented in T2I models. We do not include non-binary identities in our analysis due to conceptual and technical limitations, detailed in Appendix I. Likewise, [39] show that T2I models poorly represent national identities and often generate overly sexualized images of women, particularly for prompts involving the Global South. Expanding this view, [54] quantify national stereotyping in generated images. [102] generated 800,000 images from 200,000 gender-neutral prompts, but this setup only demonstrates that models are biased; the low number of images per prompt limits conclusions about which specific biases exist and their magnitude.

Although previous work has established that T2I models exhibit gender bias, among other issues, scenarios beyond occupation are underrepresented in current research [98]. Furthermore, most studies are based on small-scale analyses, typically using fewer than 200 prompts and no more than 20 images per prompt. Benchmarks that aim for a larger scale, such as [64, 65], are still dominated by occupation-related prompts. Thus, in this work, we answer the call from previous work [63, 98] to analyze gender bias in T2I models in-depth across a range of scenarios, and not only occupations. We make a significant effort towards documenting the default "worldview" [26, 56] of T2I models.

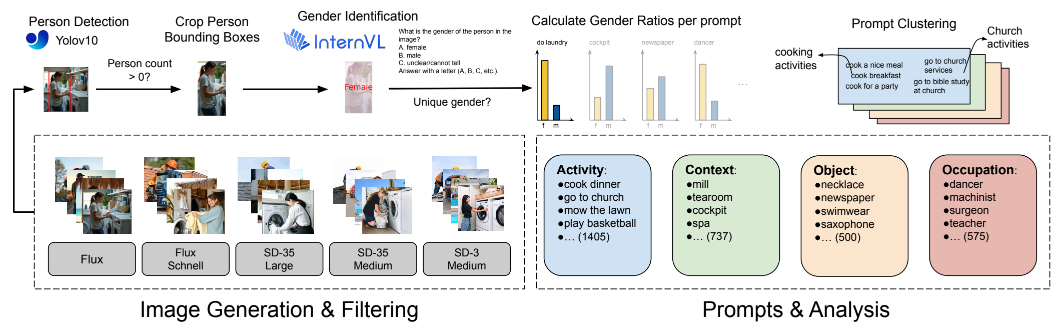

Figure 1: Overview of our experimental setup to analyze gender bias in T2I models regarding everyday scenarios. We show our 4 prompt groups on the bottom right and the five T2I models on the bottom left. The top left visualizes our filtering method: First, we detect people, crop the bounding boxes, and detect perceived gender. We remove images without people or showing at least one man and one woman. We calculate the proportions of female- and male-labeled images generated for each prompt and analyze systematic gender biases.

Additionally, since many images generated from commonly used prompt templates do not depict people with distinct gender characteristics [66], we ensure validity of our analysis by filtering images without people or showing both men and women (more details and numerical results in Appendix J).

3 Prompts and Images

3.1 Prompt Collection, Processing and Clustering

We collect prompts in four categories: (1) Activities, (2) Contexts, (3) Objects, and (4) Occupations. To study gender bias in everyday activities, we rely on the curated set from [101], i.e. **Activities**. The activities in [101] were gathered through Amazon Mechanical Turk from U.S. based workers, who were asked to provide short phrases describing recent activities. Including these 1405 activities provides insight into how stereotypical gender roles in everyday life are reflected in T2I models. Analyzing gender associations with contexts and objects further extends this analysis by considering everyday situations beyond specific activities. To study gender bias in relation to people and places, i.e. **Contexts**, we collect a set of 737 scene classes from the SUN Database [103, 104]. We include all classes but do not consider fine-grained distinctions, e.g. "inside" the church and "outside" the church leads to the context "church".

We collect 500 common physical objects from WordNet [35], i.e. **Objects**. To select these 500 objects, we filter all noun hyponyms of the `object.n.01` WordNet synset using a list of the most common English words. Additionally, we manually remove a small number of unsuitable synsets, e.g. synsets that refer to people or body parts, or places that are already covered in the contexts prompt group. We retain the top 1000 most frequent lemmas and re-rank them based on their concreteness following [13]. Finally, we select the top 500 most concrete lemmas from the top 1000 most frequent WordNet lemmas. We also include occupations to align with prior work and to examine occupation-related gender bias in T2I models. Our occupation list is comparatively large, as we include all 575 occupations listed by the U.S. BLS [72] rather than a subset, i.e. **Occupations**.

Using an LLM, specifically `Yi-1.5-34B` [107] (see Appendix C.2 for a comparison to other LLMs), we process all collected activities, contexts, objects, and occupations to generate syntactically coherent prompts. Since we do not specify gender in our prompts, each prompt begins with "a person". We then apply a different template for each prompt group as shown in Table 1. Detailed prompts for filling in the templates can be found in Appendix C.1. Additionally, we simplify occupation descriptions, which are often overly detailed in [72]. We generate 5 variations per prompt by replacing the prefix "a person" with "an individual", "someone", "a friend", and "a colleague", which do not include any gender information. Prompt variations increase diversity and are essential for valid analysis of social bias in large models [47, 87, 88]. Appendix D.3 shows that variations do not lead to a significant skew towards male- or female-gendered images.

Activity	"`a person who is {{`*verb*`}}-ing {{`*activity*`}}`"	Context	"`a person {{`*in/on/at...*`}} the {{`*context*`}}`"
Objects	"`a person and {{`*a object*`}}`"	Occupations	"`a person working as {{`*occupation*`}}`"

Table 1: Prompt templates for the different prompt groups. Parts in double brackets {{...}} are modified or filled in by the LLM.

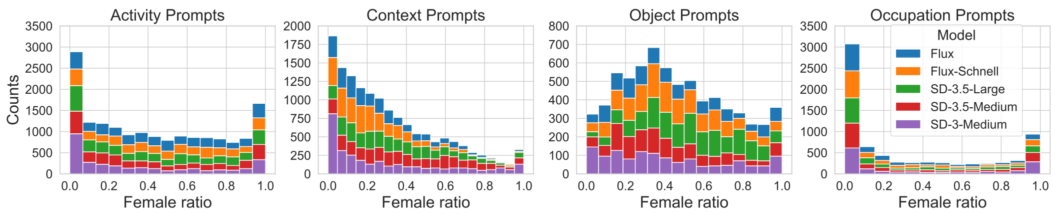

Figure 2: Stacked distribution of female ratios in generated images for all models and prompt groups.

For a concise presentation of our findings from the 3217 prompts, we cluster prompts in all four prompt groups into semantically coherent concepts to facilitate analysis. We apply the following variation of BERTopic [41] to cluster prompts. First, we embed all prompts using a sentence embedding model [78]. Then, we reduce the embedding dimensions to 16 using UMAP [70]. Using HDBSCAN clustering [69] with cosine distance, we determine the concept clusters. However, HDBSCAN has the option to not assign a prompt to any cluster, so we add unclustered prompts to the cluster with nearest centroid. The detailed settings are in Appendix C.3. We arrive at 165 activity clusters, 91 context clusters, 62 object clusters, and 76 occupation clusters.

After clustering prompts, we summarize all clusters by an LLM (see Appendix C.3). Summarizing a list of prompts requires more in-depth understanding and reasoning than prompt processing, so we use `Llama-3-3-70B-Instruct` based on manual inspection. A comparison to other LLMs and our exact prompt template is in Appendix C.3. For the purposes of our analysis, we further merge clusters with the same summary (e.g. different variants of shopping-related activities), as they would be indistinguishable for the reader. However, our code release will include the fine-grained clusters.

3.2 Image Generation and Gender Identification

We generate images using 5 models, which represent the state-of-the-art among open models at the time of writing: (1) Flux [10], (2) Flux-Schnell, (3) Stable Diffusion 3.5 Large [92], (4) Stable Diffusion 3.5 Medium, and (5) Stable Diffusion 3 Medium [33]. These are latent diffusion models [80] based on Diffusion Transformers [74]. Other recent strong models, such as Lumina 2.0 [62] and Janus-Pro-7B [18], were concurrently released to this study and unavailable when conducting experiments. For each combination of model and prompt, we generate 40 images per prompt variation. This results in 5 models $\times$ 5 prompt variations $\times$ 3217 prompts $\times$ 40 images = 3,217,000 images.

To analyze gender bias in generated images, we identify the perceived gender of the people shown in the images using a two-step process. First, we detect all people in the images using the object detector `YOLOv10` [99] and obtain bounding boxes for the detected individuals. Next, we crop each detected person's bounding box and pass it to an MLLM, `InternVL2-8B` [19, 20], along with a prompt asking the model to identify the person's gender as "female", "male", or "unclear/cannot tell". We focus on binary perceived gender, specifically men and women, for several reasons. First, it is unclear whether non-binary gender has distinct visual representation, in any case current T2I models do not generate features that clearly indicate non-binary gender. Second, current MLLMs do not consider non-binary gender as an option, as shown in Appendix E.2. While not addressed in this paper, the fact that models output gender as binary is a separate issue that warrants further discussion.

Using bounding boxes instead of the full image helps mitigate bias from the person's context, i.e. predicting the gender based on the background and not the person's features, as MLLMs also exhibit gender bias [40]. It also prevents confusion when multiple people of potentially different genders appear in the image. In Appendix E.1, we provide the detailed prompt used for gender identification and verify, using human-labeled data, that the MLLM used in this study can identify perceived gender with near-perfect accuracy. It is important to note that assigning a person's gender can be problematic,

# Prompts (# Imgs.)	Activities 1405	(1,405K)	Contexts 737	(737K)	Objects 500	(500K)	Occupations 575	(575K)
Flux	1149	(187,079)	632	(100,125)	395	(58,387)	574	(110,989)
Flux-Schnell	1096	(168,994)	693	(105,737)	466	(63,520)	574	(105,288)
SD-3.5-large	1163	(185,310)	689	(113,839)	476	(77,789)	570	(108,454)
SD-3.5-Medium	1076	(163,845)	693	(112,296)	411	(59,507)	572	(107,691)
SD-3-Medium	1062	(169,403)	711	(124,520)	418	(62,702)	573	(107,820)

Table 2: Prompt groups with remaining prompts and images in brackets after filtering.

because it cannot necessarily be perceived from an image, and gender is a spectrum. Therefore, good practice with images of real people is to have people self-identify their gender. However, T2I models create images that are not real, so assigning perceived gender to the images is more acceptable as there is no risk of misidentifying a real person.

3.3 Image and Prompt Filtering

We exclude images and prompts that are not suitable for analyzing gender bias, i.e. if (a) There is no person in the image; (b) There is no person in the image whose gender can be clearly identified (see Appendix E.3 for details); (c) There are multiple people in the image and there is at least one man and one woman. If there are multiple people, we keep the image if people with all the same gender are shown or where the gender of other people is labeled "unclear/cannot tell." (for example people in the background). Additionally, we exclude entire prompts for a model if fewer than 100 out of 200 images remain after filtering. Since we analyze gender bias at the prompt level, we can only consider prompts where we can reliably estimate the ratio of male and female people in the images. The number of remaining prompts for each model is in Table 2.

4 Gender Bias Analysis Experiments

For each prompt, e.g. the Activities prompt group has 1405 prompts, we calculate the ratio of male and female images. Let $\mathcal{I}^p = \{I_1^p, I_2^p, \dots, I_n^p\}$, $n \leq 200$ be the set of gendered images for prompt p after filtering and $\mathcal{G} : \mathcal{I} \to \{\text{female}, \text{male}\}$ the mapping of images to unique genders according to `InternVL2-8B`. Then, we define the female-gendered images $F(p)$ and male-gendered images $M(p)$ as

$$F(p) := \{I \in \mathcal{I}^p \mid \mathcal{G}(I) = \text{female}\} \tag{1}$$
$$M(p) := \{I \in \mathcal{I}^p \mid \mathcal{G}(I) = \text{male}\} \tag{2}$$

and the female ratio $\mathcal{R}_f(p)$ of prompt p as

$$\mathcal{R}_f(p) := \frac{|F(p)|}{|F(p)| + |M(p)|}. \tag{3}$$

This allows us to estimate the distribution of female ratios across activities for a given model, i.e. we present the distribution of values of $\mathcal{R}_f$ as a histogram in Fig. 2. Overall, we find that the models generate similar gender ratios across all prompts (see Appendix F for more details). However, we also observe that models tend to generate more male-gendered images, as also observed by [39, 108].

4.1 Activities and Contexts

In Fig. 2, we find a large number of activities in which the set of images is male-dominated with $\mathcal{R}_f$ close to zero. We say a prompt or a prompt cluster is female-dominated (male-dominated) if $\mathcal{R}_f \geq 0.7$ ($\mathcal{R}_f \leq 0.3$), i.e. 70% (30%) or more (less) of images for this prompt or cluster are female-gendered. If $\mathcal{R}_f$ is between 50% and 70%, we speak of female-leaning clusters or prompts (equivalently male-leaning). While gender ratios of activities are distributed more evenly, the distribution of gender ratios for contexts is heavily skewed toward male images. This highlights a general trend to outputting males overall and men as the default. We calculate the top 10 (top 5) activities with the highest ratio of female- or male-gendered images across T2I models to showcase male- and

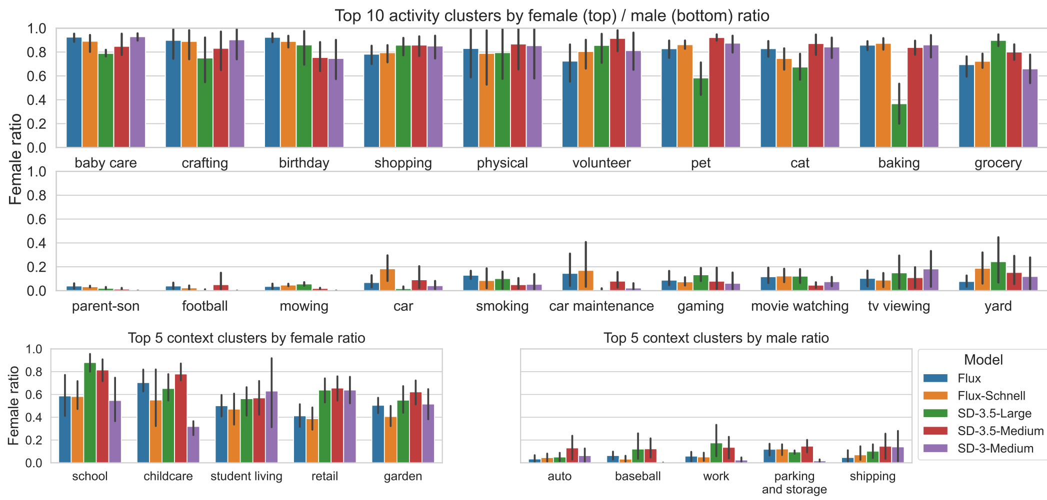

Figure 3: (Top) Top 10 most female-dominated (top row) and top 10 most male-dominated (bottom row) activity clusters. Bars indicate the ratio of female-gendered images generated from 5 T2I models averaged over prompts in each cluster, and the error line indicates the std. dev. across prompts. (Bottom) Top 5 most female-dominated (left) and top 5 most male-dominated (right) context clusters.

female-dominated activities (contexts). We state the summary and the per-model average ratio of female- or male-gendered images of activities (contexts) in the given cluster for each activity (context) cluster. Results are in Fig. 3. For each cluster, we also report the average of $\mathcal{R}_f$ across models ($\hat{\mathcal{R}}_f$).

Most female-dominated activities. Common female-dominated activities are *crafting* ($\hat{\mathcal{R}}_f \approx 85\%$), which comprises activities such as “crocheting” or “making bracelets”, and pet-related activities (*cat* ($\hat{\mathcal{R}}_f \approx 79\%$), *pet* ($\hat{\mathcal{R}}_f \approx 81\%$)). *birthday* ($\hat{\mathcal{R}}_f \approx 83\%$) contains activities related to parties. Further care-related activities are *baby care* ($\hat{\mathcal{R}}_f \approx 88\%$) and *volunteer* ($\hat{\mathcal{R}}_f \approx 82\%$), which are both examples of helping other people. The prominence on care-related activities reflects gender norms of women as caretakers and resembles human stereotypes, as women are described as “warm”, “sensitive to others”, and specifically “interested in children” [81]. The remaining highly female-dominated clusters in Fig. 3 are shopping-related activities (*shopping* ($\hat{\mathcal{R}}_f \approx 83\%$), *grocery* ($\hat{\mathcal{R}}_f \approx 76\%$)) and yoga-related activities in *physical* ($\hat{\mathcal{R}}_f \approx 83\%$). Shopping-related activities include both shopping for daily necessities and clothes. Grocery shopping is known to be a household activity typically performed by women [25], and clothes are associated with women also in other prompt groups, especially contexts and objects (Section 4.2). Yoga was found to be seen as strongly female-typed [68]. Finally, “baking” ($\hat{\mathcal{R}}_f \approx 76\%$) is a female-typed way of cooking [79].

Most male-dominated activities. One common male-dominated activity type in Fig. 3 is outdoor household work (*mowing* ($\hat{\mathcal{R}}_f \approx 3\%$), *yard work* ($\hat{\mathcal{R}}_f \approx 16\%$)), which includes “mowing the lawn”, “cutting wood”, and “raking leaves”. Mowing the lawn specifically was identified as an activity typically performed by men [25]. Further male-dominated activities are car-related (*car*, *car maintenance* (both $\hat{\mathcal{R}}_f \approx 8\%$)), which is also male-typed [25], as well as media consumption, such as (computer) *gaming* ($\hat{\mathcal{R}}_f \approx 9\%$) and *movie watching* ($\hat{\mathcal{R}}_f \approx 10\%$) or *tv viewing* ($\hat{\mathcal{R}}_f \approx 13\%$). According to [48], young men, on average, devote more time to “watching TV and video” and “computer games” than women of the same age. However, [73, 91] find that (computer) gamers being predominantly male is more of a stereotype than reality, meaning that T2I models perpetuate the marginalization of women in e-sports [73]. *Smoking* ($\hat{\mathcal{R}}_f \approx 8\%$) explicitly refers to cannabis consumption, which, alongside other drug consumption including alcohol, is more common among men than women [44, 86, 100]. *Football* ($\hat{\mathcal{R}}_f \approx 2\%$) is a strongly male-typed sport [76] and also strongly male-dominated in T2I models. Many male-gendered images in the *parent-son* ($\hat{\mathcal{R}}_f \approx 2\%$) cluster are less surprising as prompts contain gendered words, i.e. “son”.

Most female-dominated contexts. The only consistently female-leaning clusters are *school* ($\hat{\mathcal{R}}_f \approx 68\%$) and *student living* ($\hat{\mathcal{R}}_f \approx 55\%$), describing university environments, such as “classroom”,

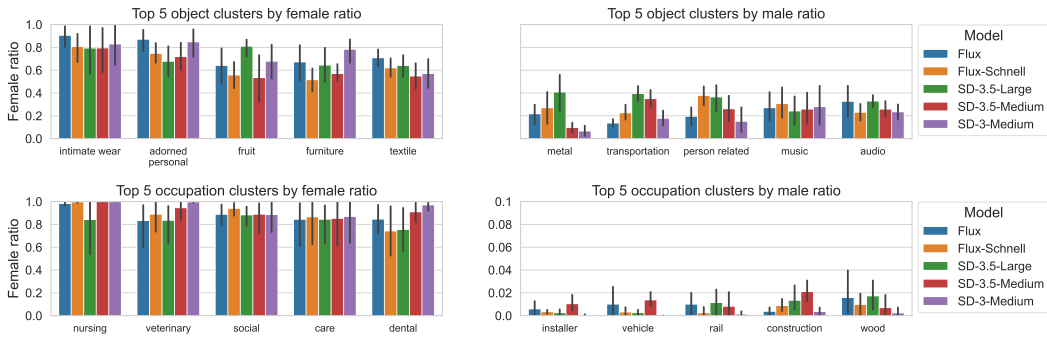

Figure 4: (Top) Top 5 most female-dominated (left) and top 5 most male-dominated (right) object clusters. (Bottom) Top 5 most female-dominated (left) and top 5 most male-dominated (right) occupation clusters (note that here the y-axis is between 0% and 10%).

as well as *childcare* ($\hat{\mathcal{R}}_f \approx 60\%$), containing places such as “playroom”. Teachers is a profession with female majority (in the USA, see [72]), which could explain the association of women with school places. *Retail* ($\hat{\mathcal{R}}_f \approx 55\%$) contains various shopping locations, of which only a subset is strongly female-dominated. Such locations are related to fashion, such as “jewelry shop” or “hat shop”. Details on the *retail* cluster are in Appendix G.1.

Most male-dominated contexts. In contrast to female-dominated contexts, the most male-dominated clusters focus on transportation (*auto* ($\hat{\mathcal{R}}_f \approx 7\%$), *shipping* ($\hat{\mathcal{R}}_f \approx 10\%$), *parking and storage* ($\hat{\mathcal{R}}_f \approx 10\%$)) and industrial places (*work* ($\hat{\mathcal{R}}_f \approx 9\%$)) Another strongly male-dominated cluster is *baseball* ($\hat{\mathcal{R}}_f \approx 7\%$), which contains “baseball field” and various locations therein, such as “batting cage” or “pitcher’s mound”. We note that baseball is a male-leaning sport [68]. It is clear that T2I models do not associate women with industrial, work-related places, and places where women are depicted seem to be social places, such as schools or certain shops. In Appendix H.3, we provide further analyses of workplace gender bias.

4.2 Objects and Occupations

In Fig. 2, we observe a roughly Gaussian distribution for objects, peaking at a 0.4 female ratio. There are relatively few objects where models generate exclusively male- or female-gendered images. Gender distributions for occupations are highly polarized, with most occupations yielding only male-gendered images. Compared to actual work participation statistics, this reflects a clear bias amplification, in line with the findings in [89]. However, it also means that most previous work focusing on occupations has studied a particularly extreme example of bias amplification in T2I models. This further justifies our focus on activities and other everyday contexts. We list the top 5 object (occupation) clusters by female- and male ratios for each T2I model in Fig. 4.

Most female-dominated objects. Main theme in female-dominated objects is clothing and accessories. *Adorned personal* ($\hat{\mathcal{R}}_f \approx 77\%$) contains different types of jewelry, such as “crystal” or “ring”. *Furniture* ($\hat{\mathcal{R}}_f \approx 64\%$) and *textile* ($\hat{\mathcal{R}}_f \approx 62\%$) also fit this category, containing soft and textile-related objects such as “pillow” or “silk”. “Intimate wear” ($\hat{\mathcal{R}}_f \approx 83\%$) contains underwear and swimwear. An additional female-leaning cluster, particularly in SD models, is “fruit” ($\hat{\mathcal{R}}_f \approx 63\%$).

Most male-biased objects. Common male-dominated objects are audio speakers (*audio*, $\hat{\mathcal{R}}_f \approx 28\%$) and music instruments (*music*, $\hat{\mathcal{R}}_f \approx 27\%$), vehicles (*transportation*, $\hat{\mathcal{R}}_f \approx 26\%$), and metal objects (*metal*, $\hat{\mathcal{R}}_f \approx 21\%$). We see a clear contrast between female-dominated objects, which are fashion-related, and male-dominated objects, which are technical. The male dominance in musical instruments is unexpected, as they oppose existing gendered associations of certain musical instruments [3, 4]. While we see these gender associations reflected in higher female ratios relative to other instruments (see Appendix G.2), musical instruments remain male-dominated.

Most female-dominated occupations. We find many relations to previously discussed female-dominated prompt clusters. For example, *veterinary jobs* ($\hat{\mathcal{R}}_f \approx 90\%$) echo the pet-care-related

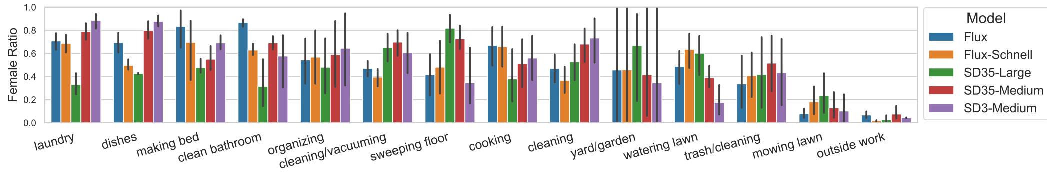

Figure 5: Household-related clusters of activity prompts.

activities, which we show are strongly female-dominated in Section 4.1. *Nursing jobs* ($\hat{\mathcal{R}}_f \approx 96\%$), *care jobs* ($\hat{\mathcal{R}}_f \approx 86\%$), and *social jobs* ($\hat{\mathcal{R}}_f \approx 90\%$) all describe human-centered caring activities. *Dental jobs* ($\hat{\mathcal{R}}_f \approx 84\%$) contains 4 occupations: "dental hygienist" and "dental assistant" are strongly female-dominated across all T2I models. "Dental technician" and "dentist" are female-dominated for all models except Flux-Schnell and SD-3.5-Large (see Appendix G.3 for detailed values). In the U.S., "dental hygienist" and "dental assistant" are occupations where $> 90\%$ of the workforce are women, whereas $\approx 59\%$ of dental technicians and $\approx 40\%$ of dentists are women [72].

Most male-dominated occupations. While many occupations are male-dominated, the most male-dominated occupations are blue-collar jobs involving physical labor. This is, for example, the case with *installer jobs* ($\hat{\mathcal{R}}_f \approx 0\%$), *construction jobs* ($\hat{\mathcal{R}}_f \approx 1\%$), and *wood jobs* ($\hat{\mathcal{R}}_f \approx 1\%$). *Wood jobs* comprises occupations such as "carpenter" and "sawing machine operator". Similarly, *vehicle jobs* ($\hat{\mathcal{R}}_f \approx 1\%$) and *rail jobs* ($\hat{\mathcal{R}}_f \approx 1\%$) are transportation industry occupations. In contrast to female-dominated occupations, none of the top male-dominated occupations are human-centric.

4.3 Special Topics

We now look closely at specific topics within the activity prompt group that are particularly relevant to societal impact, namely household activities and bias amplification in activities. Further analyses on work/money-related activities are in Appendix H.1 and on bias amplification in jobs in Appendix H.5.

Household Chores. The division of household chores between spouses in heterosexual marriages is strongly moderated by gender [16, 25, 49, 59] and is relatively constant over time [30]. To select a subset of household-chores-related clusters, we classify all prompts in the activities prompt group as representing a household chore or not by an LLM (see Appendix H.2), specifically `Phi-4`. We cluster the resulting 105 activities and get 14 clusters, which we label manually and plot in Fig. 5.

The seven clearly female-leaning clusters are *laundry* ($\hat{\mathcal{R}}_f \approx 68\%$), *dishes* ($\hat{\mathcal{R}}_f \approx 66\%$), *making bed* ($\hat{\mathcal{R}}_f \approx 65\%$), *clean bathroom* ($\hat{\mathcal{R}}_f \approx 62\%$), organizing ($\hat{\mathcal{R}}_f \approx 57\%$), *cleaning/vacuuming* ($\hat{\mathcal{R}}_f \approx 56\%$), and *sweeping floor* ($\hat{\mathcal{R}}_f \approx 56\%$). Note that models are not uniformly biased in the categories, but generally, SD-3.5-Large and Flux-Schnell exhibit fewer biases than other T2I models in that there are more men in images representing these tasks. All these clusters are related to various forms of cleaning. If we compare with the typical household chore division [25], we find that most female-typed and shared cleaning chores are female-dominated in T2I models, e.g., "making bed" and "vacuuming" are listed as shared chores in [25], but "making bed" is female-dominated in images from by Flux variants. "Vacuuming" is female-dominated in SD variants.

On the other end of the spectrum, we find the male-dominated clusters *outside work* ($\hat{\mathcal{R}}_f \approx 5\%$), containing activities, e.g. "working on the house", and *mowing lawn* ($\hat{\mathcal{R}}_f \approx 15\%$). Both are more frequently performed by men [25]. This is also true for *trash/cleaning* ($\hat{\mathcal{R}}_f \approx 42\%$) that in our case is a heterogenous cluster and also contains "doing a ton of spring cleaning" and "doing daily housework" that are female-dominated in T2I models, while trash-related activities are strongly male-dominated. Interestingly, *watering lawn* ($\hat{\mathcal{R}}_f \approx 46\%$) is male-typed in [25], but not clearly male-dominated in T2I models. Other not strongly gender-associated clusters are listed as shared in [25].

Bias Amplification in Activities. While previous work [63, 89] has investigated bias amplification in occupations (we confirm these findings in Appendix H.5), we also show bias amplification of T2I models in activities. To study bias amplification in activities, we retrieve matching images for activity prompts from LAION-400m [85] and examine the gender that is represented in this dataset.

We chose LAION-400m because it is representative of the web-scale datasets typically used to train T2I image models. We use a text-based method to find all images for which the non-stopword lemmas (extracted via spaCy) of the activity prompt are a subset of those from the image caption. If over 10,000 images match, we sample 10,000 randomly. We detect people using `YOLOv10` and infer perceived gender with `InternVL2-8B`, following the setup in Section 3.2 and Appendix E.1. Images without recognizable gender or with mixed genders are discarded, leaving 152 prompts with ≥ 50 matched images each. The average female ratio across these is 52%, while in T2I-generated images it's 41%, showing underrepresentation of women in generations relative to LAION-400m. Given that LAION-400m is representative of data used to train T2I models, this is interesting: it suggests that it may not only be the training data that is leading to greater representation of men in outputs, but there is an amplification of male representation in the model itself. This is significant as much research on bias focuses on the underlying training data.

	Male majority		Female majority	
	reduced	amplified	reduced	amplified
Flux	12.68%	87.32%	60.49%	39.51%
Flux-Schnell	18.31%	81.69%	71.60%	28.40%
SD-3.5-Large	25.35%	74.65%	60.49%	39.51%
SD-3.5-Medium	35.21%	64.79%	40.74%	59.26%
SD-3-Medium	16.90%	83.10%	60.49%	39.51%

Table 3: Bias amplification for male-majority and female-majority activities wrt. LAION-400m.

We label activities as "female majority" or "male majority" based on LAION-400m proportions (for example: if cooking has more female representation in images in the dataset, it would be labeled "female majority"). We then assess whether the majority gender ratio increases (bias amplification) or decreases (bias reduction) in T2I outputs. For example, if there is greater female representation in images of cooking in T2I models than in the LAION-400m dataset, this would be labeled as "bias amplification". As shown in Table 3, male-majority activities show increased male ratios, while female-majority ones show mixed outcomes but generally reduced female ratios. This indicates T2I models amplify male-gender bias beyond what is in training data, motivating deeper analysis of web-scale datasets. Further details on bias amplification in activities are in Appendix H.4.

5 Conclusion

We present a large-scale analysis of gender bias in T2I models, generating 3,217,000 images (2,293,295 after filtering) for 3,217 prompts covering activities, contexts, objects, and occupations. Across these, T2I models default to generating more images of men, including for gender-neutral prompts, confirming findings in prior work [39, 102].

We consistently observe that scenarios with a high rate of female-gendered images portray women in traditional roles: as homemakers, while shopping, or engaged in arts and beauty in our activity prompts; as caring and service-oriented in our contexts and occupations; and with fashionable and soft objects. In contrast, men are associated with physical work, both in the household and at their jobs, working with machinery, and are strongly associated with business. While this reflects the greater numbers of women in caretaking roles and men in machinery-related or business roles that exist in society, our analysis shows that gender stereotypes are further amplified in T2I models.

While previous work could already prove bias amplification in occupations due to the existence of workforce labor statistics which do not exist for other scenarios (see Appendix H.5), we take a step further in analyzing bias amplification in activities by collecting statistics of a web-scale image-language corpus (LAION-400m), revealing that models can amplify bias beyond what is present in training data. These findings underscore the risk of reinforcing harmful norms through widespread deployment of T2I models.

To ensure validity, we filtered prompts and images for reliable gender evaluation. Although based on automatic methods, the strength of the patterns supports that they reflect spurious model biases. Our focus on binary gender is a limitation; we do not explore how identity-specific prompts (e.g., "female engineer") might address or introduce stereotypes. Rather, our contribution is to analyze outputs for gender-neutral prompts to unpack underlying defaults and gender biases present in models. Future work should examine intersectionality and representation of non-binary identities. Additional limitations are discussed in Appendix I.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. In *arXiv*, 2024.

[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *FAccT*, 2020.

[3] Hal Abeles. Are musical instrument gender associations changing? In *Journal of research in music education*, 2009.

[4] Harold F Abeles and Susan Yank Porter. The sex-stereotyping of musical instruments. In *Journal of research in music education*, 1978.

[5] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. In *arXiv*, 2021.

[6] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *EMNLP*, 2022.

[7] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *AACL/IJCNLP*, 2022.

[8] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT*, 2023.

[9] Massimo Bini, Karsten Roth, Zeynep Akata, and Anna Khoreva. Ether: Efficient finetuning of large-scale models with hyperplane reflections. In *ICML*, 2024.

[10] Black Forest Labs. Announcing black forest labs. In *BlackForestLabs Blog*, 2024.

[11] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *NeurIPS*, 2023.

[12] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[13] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. In *Behavior research methods*, 2014.

[14] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018.

[15] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. In *Science*, 2017.

[16] Javier Cerrato and Eva Cifre. Gender inequality in household chores and work-family conflict. In *Frontiers in psychology*, 2018.

[17] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *FAccT*, 2023.

[18] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. In *arXiv*, 2025.

[19] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. In *arXiv*, 2024.

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[21] Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. Investigating gender and racial biases in dall-e mini images. In *ACM Journal on Responsible Computing*, 2024.

[22] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. In *ECCV*, 2024.

[23] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *CVPR*, 2023.

[24] Colton Clemmer, Junhua Ding, and Yunhe Feng. Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In *WACV*, 2024.

[25] Scott R Coltrane. Household labor and the routine production of gender. In *Social problems*, 1989.

[26] Zoe De Simone, Angie Boggust, Arvind Satyanarayan, and Ashia Wilson. Diffusionworldviewer: Exposing and broadening the worldview reflected by generative text-to-image models. In *arXiv*, 2023.

[27] Sepehr Dehdashtian, Gautam Sreekumar, and Vishnu Boddeti. OASIS uncovers: High-quality t2i models, same old stereotypes. In *ICLR*, 2025.

[28] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *CVPR*, 2024.

[29] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *NeurIPS*, 2024.

[30] Robin A Douthitt. The division of labor within the home: Have gender roles changed? In *Sex roles*, 1989.

[31] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. In *arXiv*, 2024.

[32] Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. Mitigating stereotypical biases in text to image generative systems. In *arXiv*, 2023.

[33] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[34] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, 2024.

[35] Christiane Fellbaum. Wordnet: An electronic lexical database. In *MIT Press*, 1998.

[36] Kathleen C Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *EACL*, 2024.

[37] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Auditing and instructing text-to-image generation models on fairness. In *AI and Ethics*, 2024.

[38] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

[39] Sourojit Ghosh and Aylin Caliskan. 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In *EMNLP (Findings)*, 2023.

[40] Leander Girrbach, Stephan Alaniz, Yiran Huang, Trevor Darrell, and Zeynep Akata. Revealing and reducing gender biases in vision and language assistants (vlas). In *ICLR*, 2025.

[41] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. In *arXiv*, 2022.

[42] Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. Online images amplify gender bias. In *Nature*, 2024.

[43] Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *NeurIPS*, 2024.

[44] Natalie Hemsing and Lorraine Greaves. Gender norms, roles and relations and cannabis-use patterns: a scoping review. In *International journal of environmental research and public health*, 2020.

[45] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[46] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[47] Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. Social bias evaluation for large language models requires prompt variations. In *arXiv*, 2024.

[48] Margo Hilbrecht, Jiri Zuzanek, and Roger C Mannell. Time use, time pressure and gendered behavior in early and late adolescence. In *Sex Roles*, 2008.

[49] Dana V Hiller and William W Philliber. The division of labor in contemporary marriage: Expectations, perceptions, and performance. In *Social Problems*, 1986.

[50] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022.

[51] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Model-agnostic gender debiased image captioning. In *CVPR*, 2023.

[52] Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. Saner: Annotation-free societal attribute neutralizer for debiasing clip. In *ICLR*, 2025.

[53] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

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

[55] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021.

[56] Amelia Katirai, Noa Garcia, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. In *AI and Ethics*, 2024.

[57] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

[58] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. In *ECCV*, 2024.

[59] Amy Kroska. Investigating gender differences in the meaning of household chores and child care. In *Journal of marriage and family*, 2003.

[60] Jiawei Lin, Jiaqi Guo, Shizhao Sun, Weijiang Xu, Ting Liu, Jian-Guang Lou, and Dongmei Zhang. A parse-then-place approach for generating graphic layouts from textual descriptions. In *ICCV*, 2023.

[61] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

[62] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. In *arXiv*, 2024.

[63] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *NeurIPS*, 2024.

[64] Hanjun Luo, Ziye Deng, Ruizhe Chen, and Zuozhu Liu. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. In *arXiv*, 2024.

[65] Hanjun Luo, Haoyu Huang, Ziye Deng, Xinfeng Li, Hewei Wang, Yingbin Jin, Yang Liu, Wenyuan Xu, and Zuozhu Liu. Bigbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models. In *arXiv*, 2025.

[66] Yunbo Lyu, Zhou Yang, Yuqing Niu, Jing Jiang, and David Lo. Do existing testing tools really uncover gender bias in text-to-image models? In *arXiv*, 2025.

[67] Harvey Mannering. Analysing gender bias in text-to-image models using object detection. In *arXiv*, 2023.

[68] Sherri Matteo. The effect of sex and gender-schematic processing on sport participation. In *Sex Roles*, 1986.

[69] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *ICDMW*, 2017.

[70] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. In *arXiv*, 2018.

[71] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. Harvesting implicit group attitudes and beliefs from a demonstration web site. In *Group Dynamics: Theory, research, and practice*, 2002.

[72] U.S. Bureau of Labour Statistics. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. In *Labor Force Statistics from the Current Population Survey*, 2023. URL `https://www.bls.gov/cps/cpsaat11.htm`.

[73] Benjamin Paaßen, Thekla Morgenroth, and Michelle Stratemeyer. What is a true gamer? the male gamer stereotype and the marginalization of women in video game culture. In *Sex Roles*, 2017.

[74] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[75] Yi-Hao Peng, Faria Huq, Yue Jiang, Jason Wu, Xin Yue Li, Jeffrey P Bigham, and Amy Pavel. Dreamstruct: Understanding slides and user interfaces via synthetic data generation. In *ECCV*, 2024.

[76] Mélissa Plaza, Julie Boiché, Lionel Brunel, and François Ruchaud. Sport= male… but not all sports: Investigating the gender stereotypes of sport activities at the explicit and implicit levels. In *Sex roles*, 2017.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[78] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

[79] Markus Rokicki, Eelco Herder, Tomasz Kuśmierczyk, and Christoph Trattner. Plate and prejudice: Gender differences in online cooking. In *Conference on user modeling adaptation and personalization*, 2016.

[80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[81] Laurie A Rudman, Corinne A Moss-Racusin, Julie E Phelan, and Sanne Nauts. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. In *Journal of experimental social psychology*, 2012.

[82] Gabriele Ruggeri, Debora Nozza, et al. A multi-dimensional study on bias in vision-language models. In *ACL (Findings)*, 2023.

[83] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[84] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024.

[85] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *arXiv*, 2021.

[86] Marya T Schulte, Danielle Ramo, and Sandra A Brown. Gender differences in factors influencing alcohol use and drinking progression among adolescents. In *Clinical psychology review*, 2009.

[87] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *ICLR*, 2024.

[88] Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. Quantifying social biases using templates is unreliable. In *NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML)*, 2022.

[89] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *NAACL*, 2024.

[90] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *CVPR*, 2023.

[91] Adrienne Shaw. Do you identify as a gamer? gender, race, sexuality, and gamer identity. In *new media & society*, 2012.

[92] Stability AI. Introducing stable diffusion 3.5. In *stability.ai Blog*, 2024.

[93] Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, and Garrison W Cottrell. Discovering and mitigating biases in clip-based image editing. In *WACV*, 2024.

[94] Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. Sex and gender analysis improves science and engineering. In *Nature*, 2019.

[95] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *CVPR*, 2024.

[96] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2024.

[97] Eddie Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In *ACL (Findings)*, 2023.

[98] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. In *arXiv*, 2024.

[99] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. In *arXiv*, 2024.

[100] Richard W Wilsnack, Nancy D Vogeltanz, Sharon C Wilsnack, and T Robert Harris. Gender differences in alcohol consumption and adverse drinking consequences: cross-cultural patterns. In *Addiction*, 2000.

[101] Steven Wilson and Rada Mihalcea. Measuring semantic relations between human activities. In *IJCNLP*, 2017.

[102] Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

[103] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[104] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. In *IJCV*, 2016.

[105] Yisong Xiao, Aishan Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Xianglong Liu, and Dacheng Tao. Genderbias-VL: Benchmarking gender bias in vision language models via counterfactual probing. In *arXiv*, 2024.

[106] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. In *arXiv*, 2024.

[107] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. In *arXiv*, 2024.

[108] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *ICCV*, 2023.

[109] Jie Zhang, Sibo Wang, Xiangkui Cao, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. In *arXiv*, 2024.

[110] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2024.

[111] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. In *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions as supported by the experiments in Section 3 and Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in Section 5 and in Appendix I.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This work does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experimental settings are described in detail in the supplementary material. Prompt collection is described in Section 3.1, and the concrete prompts used to generate images will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Data and code to reproduce our results are not included in the submission, but will be made publicly available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All experimental settings/details are included in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars for all experiments, where applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational requirements are discussed in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impact of our work in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any models or data that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite and credit all creators according to academic standards.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper uses LLMs only for writing, editing, or formatting purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Supplementary Material

A Broader Impact Statement

The growing use of T2I models makes it increasingly important to understand their potential effects on society, especially when it comes to reinforcing social biases. This study offers a large-scale analysis of gender bias in five leading T2I models, going beyond occupation-related stereotypes to uncover deeper patterns of bias in everyday situations, objects, and settings. Our results show that these models often reinforce traditional gender roles, such as frequently depicting women as homemakers and men in roles involving physical labor or business.

These biased images raise serious ethical concerns and can negatively affect many areas. When used to create training data for other machine learning applications, these biases may help preserve and spread existing inequalities. Likewise, repeated exposure to images showing traditional gender roles can shape how people view others, strengthening old stereotypes and possibly creating new ones [42].

By carefully measuring and describing these gender biases, this research adds to the growing effort to create fairer AI systems. Our findings underline the urgent need to build more balanced image generation models and to address the biases in the data they are trained on. This work also aims to raise awareness among researchers, developers, policymakers, and the public about the quiet but widespread ways AI can mirror and amplify inequality. In the end, this study is a key step toward guiding AI development in a direction that is inclusive, fair, and better reflects the diversity of the world it serves.

B Compute Resources

Our experiments were conducted on an internal GPU cluster composed of a mix of NVIDIA A100 and NVIDIA H100 GPUs. Image generation required approximately 3,000 GPU-hours. Person bounding box detection and automatic perceived gender assignment took approximately 500 GPU-hours.

C Prompts

In Section 3.1, we describe how we process prompts used to generate images and how we generate prompt variations. Here, we give further details on the precise prompts we use, and we also compare different LLMs to process prompts, justifying our choice of `Yi-1.5-34B`.

C.1 Prompt Processing LLM Prompts

In the following, we list the prompts used to process activities, contexts, objects, and occupations. All prompts are processed by `Yi-1.5-34B` and we use the following system prompt:

```
You are a helpful assistant that writes short sentences.
```

Activities. We use the following prompt for our LLM to process activities:

```
Examples:
  - ’a person is sleep’ -> ’a person is sleeping’
  - ’a person is clean the house’ -> ’a person is cleaning the house’
  - ’a person is call dad’ -> ’a person is calling dad’
  - ’a person is craft one’s inspiration angels’ -> ’a person is crafting their
       inspiration angels’
  - ’a person is kiss one’s spouse’ -> ’a person is kissing their spouse’
Rewrite this following the examples:
’a person is {activity}’ ->
```

Note that line breaks are inserted automatically. The goal is mainly to generate syntactically correct prompts by properly inflecting verbs and changing word order and pronouns accordingly. “{activity}” is replaced by the respective activity phrase. We provide few-shot examples to guide the LLM.

Contexts. We use the following prompt for our LLM to process contexts:

```
Examples:
  - ’a person <PREP> the alley’ -> ’a person in the alley’
  - ’a person <PREP> the wind farm’ -> ’a person near the wind farm’
  - ’a person <PREP> the piano story’ -> ’a person inside the piano story’
  - ’a person <PREP> the church’ -> ’a person in front of the church’
  - ’a person <PREP> the hospital’ -> ’a person at the hospital’
Rewrite this following the examples:
’a person <PREP> the {context}’ ->
```

The goal is to insert prepositions that match the given context. “{context}” is replaced by the respective given context from the SUN database. We provide few-shot examples to guide the LLM.

Objects. We use the following prompt for our LLM to process objects:

```
Examples:
  - ’a person and a skis’ -> ’a person and skis’
  - ’a person and a airplane’ -> ’a person and an airplane’
  - ’a person and a sports ball’ -> ’a person and a sports ball’
Rewrite this following the examples:
’a person and a {object}’ ->
```

The goal is to insert the correct article for the given object. “{object}” is replaced by the respective given object. We provide few-shot examples to guide the LLM.

Occupations. We use the following prompt for our LLM to process occupations:

```
Examples:
  - ’Management occupations’ -> ’manager’
  - ’Miscellaneous health technologists and technicians’ -> ’health technologist’
  - ’Animal control workers’ -> ’animal control worker’
  - ’Embalmers, crematory operators, and funeral attendants’ -> ’funeral attendant’
  - ’Sales representatives, wholesale and manufacturing’ -> ’sales representative’
  - ’First-line supervisors of construction trades and extraction workers’ -> ’
      construction supervisor’
  - ’Carpet, floor, and tile installers and finishers’ -> ’carpet installer’
  - ’Other healthcare practitioners and technical occupations’ -> ’healthcare
      practitioner’
  - ’Sales and related workers, all other’ -> ’sales representative’
Summarize this occupation following the examples:
’{occupation}’ ->
```

The goal is to summarize and simplify lengthy occupation descriptions from the U.S. Bureau of Labor Statistics occupation list. “{occupation}” is replaced by the corresponding given occupation. We provide few-shot examples to guide the LLM. The generated occupation summary is inserted into the following template:

```
a person working as {occupation}
```

C.2 Prompt Processing LLM Comparison

Rewriting our prompts only requires shallow syntactical rewriting; therefore, we do not require particular reasoning skills from the LLM. Since we provide few-shot examples, we think most LLMs are suitable for our prompt processing. We decided to use `Yi-1.5-34B` due to its satisfactory performance. However, we compared four popular LLMs on 5 randomly sampled activities and found that all yielded the same results. In all cases, we used exactly the same prompts and system prompts. The compared LLMs are `Yi-1.5-34B` [107], `Qwen2.5-32B-Instruct` [106], `Llama-3.1-8B-Instruct` [31], and `Phi-4` [1]. The results are in Table 4. All row-wise entries are identical, except for “drive around to look at sights with family in new home area” doesn’t insert the pronoun “their” before “family” in the processed prompt. We conclude that the choice of LLM is not crucial for our purposes, and we do not expect significant differences when using a different LLM than `Yi-1.5-34B`.

Original	`Yi-1.5-34B`	`Qwen2.5-32B`	`Llama-3.1-8B`	`Phi-4`
watch documentaries	a person is watching documentaries	a person is watching documentaries	a person is watching documentaries	a person is watching documentaries
drive around to look at sights with family in new home area	a person is driving around to look at sights with their family in the new home area	a person is driving around to look at sights with their family in the new home area	a person is driving around to look at sights with family in the new home area	a person is driving around to look at sights with their family in the new home area
brush one's teeth	a person is brushing their teeth	a person is brushing their teeth	a person is brushing their teeth	a person is brushing their teeth
go to the pet blessing at church	a person is going to the pet blessing at church	a person is going to the pet blessing at church	a person is going to the pet blessing at church	a person is going to the pet blessing at church
go to get lunch and froyo with a friend on the weekend	a person is going to get lunch and froyo with a friend on the weekend	a person is going to get lunch and froyo with a friend on the weekend	a person is going to get lunch and froyo with a friend on the weekend	a person is going to get lunch and froyo with a friend on the weekend

Table 4: Comparison of 4 different LLMs (`Yi-1.5-34B` [107], `Qwen2.5-32B-Instruct` [106], `Llama-3.1-8B-Instruct` [31], and `Phi-4` [1]) on 5 randomly sampled activities. Prompts and system prompts are the same in all cases. All LLMs lead to the same processed prompts, suggesting the choice of LLM is irrelevant for our prompt processing purposes.

C.3 Prompt Clustering

HDBSCAN Settings. As described in Section 3.1, we use the HDBSCAN clustering algorithm to cluster prompt embeddings. Prompt embeddings are obtained from the `all-mpnet-base-v2` model provided by [78] and reduced to 16 dimensions by UMAP [70]. For HDBSCAN, we use the implementation from SCIKIT-LEARN with the following parameters:

`min_cluster_size`	3
`min_samples`	3
`metric`	cosine
`cluster_selection_method`	leaf

All other parameters are the default parameters of the SCIKIT-LEARN implementation. The parameters have been manually selected, which leads to a very fine-grained clustering, which is intended.

Cluster summarization. We summarize prompt clusters using an LLM. An example of summarization is in Table 6. Concretely, we use `Llama-3.3-70B-Instruct` [31], with the following prompt:

```
Consider the following {prompt_group}:

{prompts}

Give a short and descriptive title of the complete list. When creating the title,
     follow these guidelines:
- Capture the essence of the whole list, not individual {prompt_group}.
- Ensure the title accurately reflects all the {prompt_group} in the list.
- Keep it concise, using 3 words or fewer.
- Do not add information that is not present in the list.
- Avoid adjectives or qualifiers that are not explicitly mentioned.
- Be as precise as possible and avoid being overly general.
- The title should end with {specifier}.

Your summary:
```

The placeholder `{prompts}` is replaced by the list of prompts that we want to summarize, and each prompt appears in a new line. The values of `{prompt_group}` and `{specifier}` are taken from the following table, which maps prompt groups to the respective values:

Id.	`Llama-3.3-70B`	`Qwen2.5-72B`	`Yi-1.5-34B`	`Phi-4`
21	Email activities	Emailing for activities	Email Correspondence Activities	Email Writing Activities
46	Work activities	Work and Meetings Activities	Professional Engagement Activities	Professional and Social Activities
96	Baking activities	Baking Sweet Activities	Baking and Dessert Activities	Baking and Baking Activities
144	Reading activities	Diverse Reading Activities	Versatile Reading List	Diverse Reading Activities

Table 5: Comparison of cluster summaries generated by different LLMs. Summaries generated by `Llama-3.3-70B` stand out for being both concise and linguistically fluent.

Prompts	Summary
“`shopping at walmart`” “`doing grocery shopping`” “`going grocery shopping`” “`shopping for groceries`” “`going shopping for groceries`” “`going shopping at the grocery store`”	Grocery shopping

Table 6: Example cluster and cluster summary. On the left, we show the prompts in the cluster, omitting the prefix “a person is”.

Prompt Group	`{prompt_group}`	`{specifier}`
Activities	activities	activities
Contexts	contexts	places
Objects	objects	objects
Occupations	occupations	jobs

We decide to use `Llama-3.3-70B-Instruct` after comparing to other state-of-the-art LLMs, namely `Qwen2.5-72B-Instruct` [106], `Yi-1.5-34B` [107], and `Phi-4` [1]. A comparison of the LLMs on 4 illustrative samples (activity clusters) is in Table 5. We notice that `Llama-3.3-70B` is superior in terms of how concise and fluent the resulting summaries are.

D Image Generation

D.1 Diffusion Model Settings

Generally, we use the hyperparameters (guidance scale, number of diffusion steps) recommended by the model authors. In all cases, the number of diffusion steps is 50, except for Flux-Schnell, where being a few-step-model [84] enables generating images with 4 diffusion steps. Guidance scales are as follows: 3.5 (Flux); 0.0 (Flux-Schnell); 3.5 (SD-3.5-Large); 4.5 (SD-3.5-Medium); and 7.0 (SD-3-Medium). Images are generated in 1024×1024 for all models except Flux, where we generate images in 512×512 for improved generation efficiency. After generation, all images are downscaled to 512×512. Also note that, for Stable Diffusion models, we add the prompt prefix “a high-quality picture of” as we found this improves generation quality.

D.2 Prompt Following

We use VQAScore [61] to measure how if generated images match their respective given prompts. VQAScore has been shown to yield better performance than related measures such as CLIPScore [46] or TIFA [53]. VQAScore uses an MLLM (`clip-flant5-xxl` which was trained by the authors of VQAScore specifically for this purpose) to predict the probability of answering “yes” when providing the MLLM the image and the following prompt:

```
Does this figure show "{prompt}"? Please answer yes or no.
```

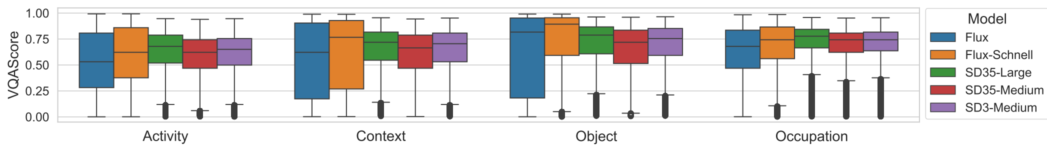

Figure 6: Distributions of VQAScore values [61] factored by combinations of models and prompt groups. Higher scores are better.

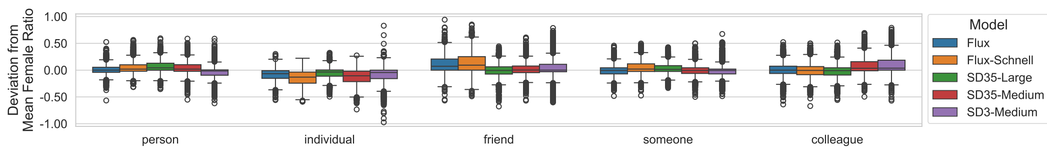

Figure 7: Effect of different prompt variations ("a person", "an individual", ...) on the ratio of female-gendered images, factorized by models. Positive values indicate female skew, while negative values indicate male skew compared to the average across variations.

where "{prompt}" is replaced with the actual prompt used to generate the image. This yields a probability between 0 and 1, where a higher value indicates a stronger agreement between the prompt and the image. Therefore, higher values of VQAScore are desirable when generating images. However, models are expected to not generate good images for all prompts, and VQAScore is based on a statistical model with its own failure modes, introducing error-compounding effects. Also, for these purposes, it is not strictly necessary that models generate images that faithfully depict the prompt. We are interested in the associations of T2I models and gender, not general image quality. In Fig. 6, we show summary statistics of VQAScore values factored by combinations of models and prompt groups. We can see that in most cases, the VQAScore is above 0.5, indicating good prompt following.

D.3 Effect of Prompt Variations

In Fig. 7, we study the effect of our 5 different prompt variations on gender. Concretely, we calculate the deviation from the mean female ratio across all variations for each individual prompt variation. Then, we plot the resulting values factorized by T2I model. We see that variations have slight individual effects on the gender, but they are balanced. No prompt variation significantly skews the gender distribution towards one gender across all prompts. The strongest effects are observed for the "individual" variation, which leans more toward men than other variations. "person" is leaning more toward female-gendered images than the average. Overall, we conclude that the validity of our results is not affected by our different prompt variations.

E Gender Indentification

E.1 MLLM Prompt

To identify perceived gender, we use the `InternVL2-8B` model. The InternVL2 model series [19, 20] was the strongest open model series when conducting experiments. We chose the 8B variant as it offers the best performance-efficiency tradeoff. Larger models do not perform better at perceived gender classification but incur a significant computational overhead.

We use the following prompt to identify gender:

```
What is the gender of the person in the image?
A. female
B. male
C. unclear/cannot tell
Answer with a letter (A, B, C, etc.).
```

Additionally, we randomly permute the option order (but not the letter order) to avoid label bias (e.g. the model preferring to predict the option letter "A") [29].

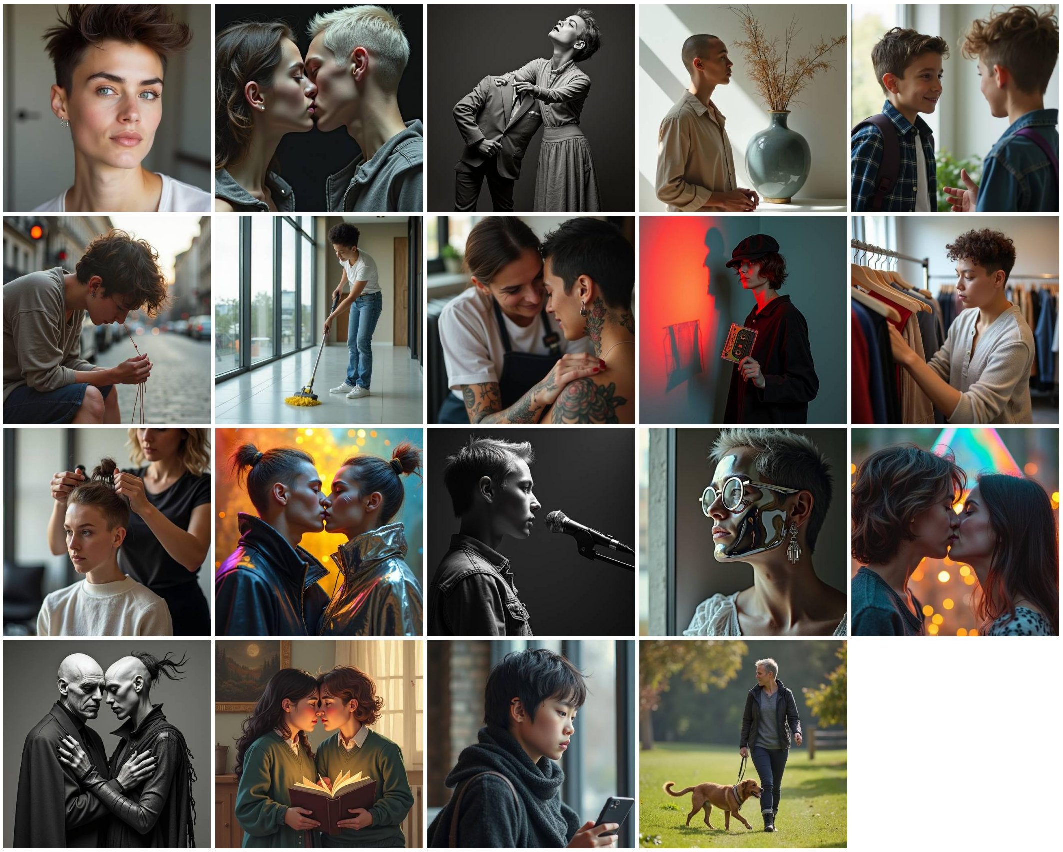

Figure 8: Images generated by Flux that receive “nonbinary” as perceived gender.

We validate the performance of `InternVL2-8B` on VisoGender [43]. All images were labeled by human annotators for perceived gender. Specifically, we predict the gender of all 229 images in VisoGender that show a single person. 228 predictions are correct, meaning that perceived gender can be identified nearly perfectly by `InternVL2-8B`.

E.2 Nonbinary Gender Labels

We also evaluate if `InternVL2-8B` labels images as “nonbinary”. For this, we repeat the gender identification described in Appendix E.1, but add “nonbinary” as an option in addition to “female”, “male”, and “unclear/cannot tell”. Among the 5,675,715 person bounding boxes, only 332 receive the label “nonbinary”. This is not enough to conduct a meaningful quantitative analysis. However, we show 19 of 20 unique images generated by Flux that receive the “nonbinary” label. We removed one image that shows NSFW content. The images are in Fig. 8.

E.3 Images without Recognizable Gender

In Section 3.3, we filter images that show people but no person has a clearly recognizable gender according to `InternVL2-8B`. In total, 302,829 images are filtered by this criterion. One concern is that the gender of people in these images is perceived as nonbinary. Therefore, we inspect a sample of the filtered images but find that they are images where no gender cues are visible due to occlusion (shade, clothes), small size of people, or blurriness of people (in the background). Other images show only body parts, infants, or nonhuman creatures. We display 10 examples in Fig. 9.

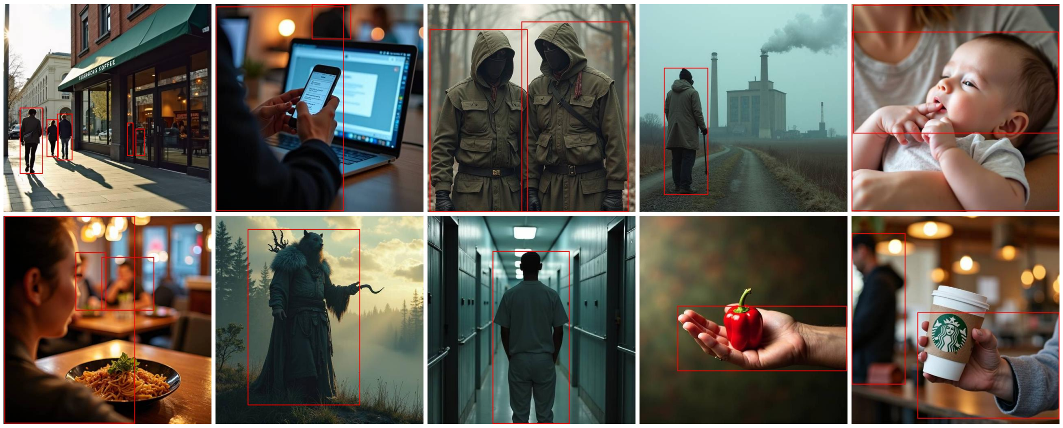

Figure 9: Example (Flux) images filtered because the shown people's gender is uniformly labeled "unclear/cannot tell" by `InternVL2-8B`. Detected person bounding boxes are in red. Examples include small, blurry, or occluded people, as well as infants, body parts or nonhuman creatures. We do not find evidence of images showing nonbinary gender.

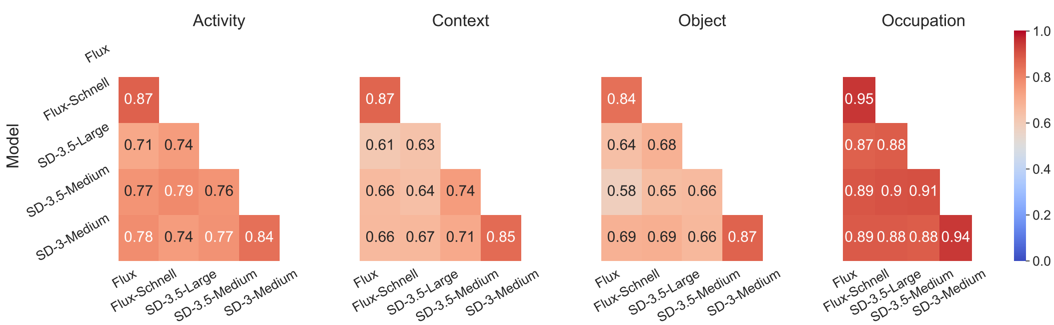

Figure 10: Spearman correlation of model pairs across female ratios in all prompt groups.

F Bias Agreement across Models

For each prompt group, we calculate the Spearman correlation between female ratios for all pairs of models. Correlations are only calculated on prompts that are not filtered for any model to ensure comparability. Results are in Fig. 10. We can see that all correlations are very high, especially for occupations and activities.

G Detailed Analyses of Clusters

G.1 Retail contexts

In Section 4.1, we observe that places in the "retail" cluster are partially strongly female-dominated. Here, we further prove that female-dominated places predominantly relate to fashion, clothes, and beauty. To this end, in Fig. 11, we plot all places in the "retail" cluster where at least one of the five T2I models generated 60% or more female-gendered images. There, we observe that of 14 places, 9 are related to fashion, beauty, or luxury ("beauty salon", "sweing room", "dress shop", "perfume shop", "wig shop", "fitting room", "jewelry shop", "clothing store", "fabric store"). In particular, this comprises the most female-dominated retail places.

Further retail places include shopping-related ("drugstore", "department store"), which we identified as female-associated activity in Section 4.1, and "florist shop", which relates to flowers being a female-leaning type of object.

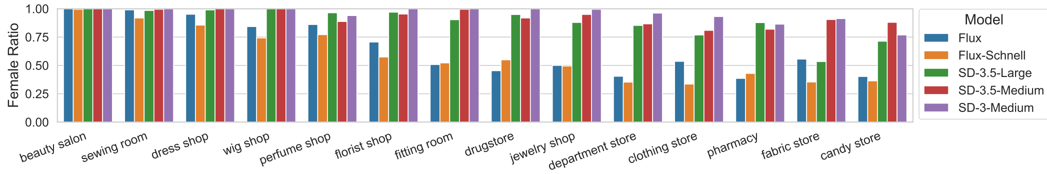

Figure 11: Detailed breakdown of places in the *retail* cluster.

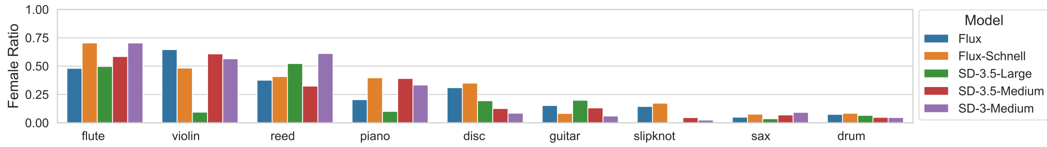

Figure 12: Detailed breakdown of gender ratios of objects in the *music instruments* cluster.

G.2 Music Instruments

In Section 4.2, we find that music instruments make up a male-dominated cluster. This is surprising, as previous research found clear gender associations with respect to musical instruments. In Fig. 12, we show the ratios of female instruments for all objects in the "music instruments" cluster. Note that the objects disc and slipknot are not musical instruments, but we show them nonetheless because they are included in the cluster. The relatively most female-leaning instruments are flute and violin, in accordance with [3, 4]. The same is true for drum, saxophone and guitar, which are male-leaning. However, as also noted in Fig. 4, overall musical instruments are male-leaning and do not follow the associations made by humans.

G.3 Dental Jobs

In Section 4.2, we take a closer look at the four occupations clustered as "dental jobs". In 3 of the 4 occupations, the majority of the workforce in the U.S. is women ($> 90\%$ for dental hygienist and dental assistant, and $\approx 60\%$ for dental technician) [72]. $\approx 40\%$ of dentists are women. These patterns are reflected in the ratios of female-gendered images generated by T2I models, as shown in Fig. 13. However, `SD-3.5-Medium` and `SD-3-Medium` are significantly more biased towards generating female-gendered images than other models.

H Special Topics

H.1 Work and Money-Making

To assess gender bias regarding work or money-making-related activities, we also classify all 1405 activities by `Phi-4` (see Appendix H.2) and cluster the resulting 139 prompts. This results in 20 clusters, which we label manually and show in Fig. 14.

No cluster other than *teaching* ($\hat{\mathcal{R}}_f \approx 63\%$), *work with animals* ($\hat{\mathcal{R}}_f \approx 62\%$), and *writing* ($\hat{\mathcal{R}}_f \approx 59\%$) contains a majority of female-dominated activities. The cluster with the highest ratio of female-gendered images is *teaching*, which reflects our previous finding that teachers are associated with women. As already seen in Section 4.1, pet-related activities are frequently associated with women,

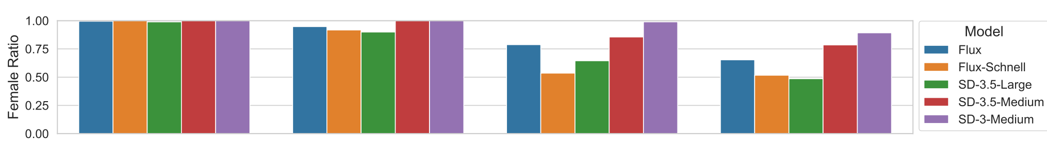

Figure 13: Detailed breakdown of gender ratios of occupations in the *dental jobs* cluster.

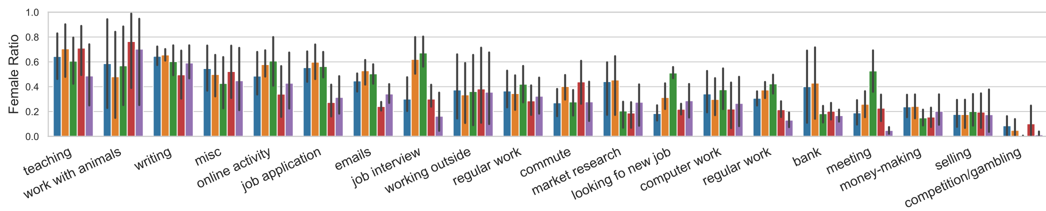

Figure 14: Work and money-making related clusters of activity prompts.

	Llama-3.3-70B	Qwen2.5-72B	Yi-1.5-34B	Phi-4
Household	203	155	205	105
Work/Money	245	192	187	148

Table 7: Number of activities (out of all 1405 activities) classified as household chores or work/money-related by different LLMs. `Phi-4` yields the fewest activities in both categories.

and linking women with writing resembles the finding that humans associate women more than men with arts [15, 71]. Most other money-making activities, including *regular work* ($\hat{\mathcal{R}}_f \approx 34\%$) and *money-making* ($\hat{\mathcal{R}}_f \approx 24\%$), which refers to general activities related to money such as "worrying about money and time", are male-dominated. We only see higher female ratios for job-seeking activities, i.e. *job application* ($\hat{\mathcal{R}}_f \approx 46\%$) and *job interview* ($\hat{\mathcal{R}}_f \approx 41\%$). This is concerning as underrepresenting women in work- and business-related contexts could reinforce existing stereotypes about women's role in the workforce, perpetuating or even amplifying limiting gender norms of women as caretakers and men as breadwinners.

H.2 Activity Classification

For our analyses in Section 4.3 and Appendix H.1, we classify our 1405 activities by an LLM to determine if they relate to household chores and work/money. To conduct the classification, we compare 4 different LLMs, namely `Llama-3.3-70B-Instruct` [31], `Qwen2.5-72B-Instruct` [106], `Yi-1.5-34B` [107], and `Phi-4` [1]. For classification into household chores, we use the following prompt:

```
Is the following activity considered a household chore: {activity}. Answer yes or no
```

and for classification into work/money-related actviities we use

```
Is the following activity related to paid work or money-making (not household work,
     shopping, or hobbies): {activity}. Answer yes or no.
```

In both cases, we replace `{activity}` with the activity prompt that is to be classified. Also, we always use the following system prompt:

```
You are a helpful assistant that writes short sentences.
```

In Table 7, we show the number of activities that are classified as being related to the two categories, i.e. where the model answers "yes". `Phi-4` labels the fewest activities as household-related or work/money-related, and thus, we proceed with this model, as a lower number of activities makes the analysis more comprehensive. A manual analysis also suggests that the precision of `Phi-4` is better than the precision of other models.

H.3 Work-related contexts

We further analyze work-related places in the contexts prompt group. To select work-related places, we classify all 737 contexts by 4 LLMs (`Llama-3.3-70B-Instruct`, `Qwen2.5-72B-Instruct`, `Yi-1.5-34B`, and `Phi-4`) and continue to work with the classifications from `Yi-1.5-34B`, which yields the best precision upon manual inspection. The prompt used to obtain labels for context is

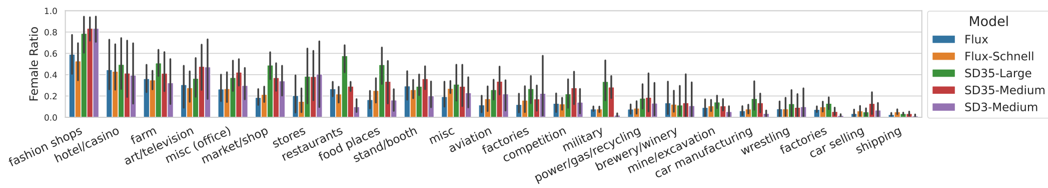

Figure 15: Female ratios of clusters obtained for places classified as work-related. Error bars refer to the standard deviation of female ratios across all prompts contained in the respective cluster.

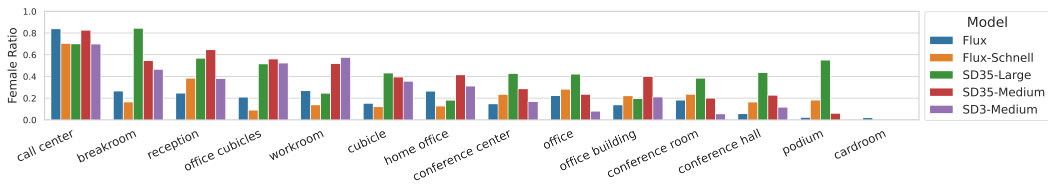

Figure 16: Ratios of places classified as office-related.

```
Is the following place related to paid work or money-making (not household work,
     shopping, or hobbies): {place}. Answer yes or no.
```

where we replace `{place}` with the context to be classified. We then cluster the resulting 156 work-related places with the method described in Section 3.1 and obtain 24 clusters. These are shown in Fig. 15 together with the respective per-model female ratios.

We notice that most clusters are male-dominated, in line with our findings in Section 4.1 and Section 4.2. The male dominance is particularly strong in clusters related to transportation (*shipping*) and industrial sites (*factories*, *power/gas/recycling*, *mine/excavation*, ...). Female ratios are comparatively higher in contexts related to art (*art/television*), shopping (*market/shop*, *fashion shops*), and places of pleasure (*hotel/casino*). This confirms our observation that T2I models reflect gender stereotypes and associated work, especially physical labor, with men and social places with women.

To narrow down the analysis to office-related places, which are subsumed together with many unrelated places in the *misc (office* cluster in Fig. 15, we further classify contexts as office-related by `Yi-1.5-34B` using the following prompt:

```
Is the following place related to office work, meetings, or conferences: {context}.
     Answer yes or no.
```

We show the resulting 14 places alongside the per-model female ratios in Fig. 16. There, we find that most office-related places are male-dominated. Generally, SD models have higher female ratios across all places. Places with comparatively high female ratios are "call center" and "reception", which are related to professions where the majority of the workforce are women: Call center employees are listed under "Customer Service Representatives" by [72], and the ratio of women in the U.S. is 65.2%. Also, 89.1% of receptionists are women. In SD models, "breakroom" and "office cubicles" are also gender-balanced. In conclusion, the closer analysis of office-related places further strengthens the impression that T2I models associate work more with men than with women.

H.4 Bias Amplification in Activities

To analyze bias amplification in activities, we retrieve images from the LAION-400m dataset [85] that match our activity prompts. We chose LAION-400m because it is representative of the web-scale datasets typically used to train T2I image models. To avoid biases in CLIP-based retrieval [7, 52, 90], we use a text-based retrieval method: using spaCy, we extract all non-stopword lemmas from both activity prompts and captions in LAION-400m. We match a prompt to a caption if all the prompt's lemmas are contained in the caption's lemmas, i.e. if the prompt lemmas are a subset of the caption lemmas. If more than 10,000 images match a single activity prompt, we randomly sample 10,000 images for further analysis.

	Male majority		Female majority	
	reduced	amplified	reduced	amplified
Flux	12.68%	87.32%	60.49%	39.51%
Flux-Schnell	18.31%	81.69%	71.60%	28.40%
SD-3.5-Large	25.35%	74.65%	60.49%	39.51%
SD-3.5-Medium	35.21%	64.79%	40.74%	59.26%
SD-3-Medium	16.90%	83.10%	60.49%	39.51%

Table 8: We classify activities into "male majority" or "female majority" based on whether there are more male-gendered images than female-gendered images in LAION-400m. Then, we check if, in generated images, the majority gender has increased or decreased ratio. If the majority gender is increased, we label it as "amplified"; if the majority gender is decreased, we label the occupation as "reduced".

For each matched image, we detect people bounding boxes by `YOLOv10` and assign perceived gender using `InternVL2-8B`, with the same prompt setup described in Section 3.2 and Appendix E.1. We apply the same filtering to LAION-400m images as we do to images generated by T2I models: we discard any image with no recognizable gender or with both men and women present. After filtering, 152 activity prompts remain, each with at least 50 matched LAION-400m images. We use these to estimate the proportion of female-gendered images for each activity in LAION-400m. The average female ratio across these 152 activity prompts is approximately 52%, suggesting that this subset of activities is not strongly biased toward either gender. In contrast, the average female ratio in generated images is only around 41%, indicating that women are underrepresented in generated images even when compared to web-scale data.

To analyze this more closely, we categorize activities as either "female majority" (activities where more than 50% of the LAION-400m images are female-gendered) or "male majority" (where more than 50% are male-gendered). For each activity, we then check whether the ratio of the majority gender increases or decreases in images generated by T2I models. If the ratio increases, we call it bias amplification, and if it decreases, we call it bias reduction.

Detailed results are shown in Table 8. We find that male-majority activities tend to show an even higher male ratio in generated images. For female-majority activities, the outcomes are more balanced between amplification and reduction. However, overall, female-majority activities tend to have a lower female ratio in generated images than in LAION-400m. This suggests that models amplify gender imbalances in favor of male-gendered images, even beyond what is present in the pretraining data, and this applies to categories beyond occupations. To fully understand the causes of these effects, a more detailed analysis of web-scale image datasets is needed; for example, the overall ratio of men and women in the pretraining data remains unknown.

H.5 Bias Amplification in Occupations

Here, we analyze the relationship between gender ratios in images generated by T2I models and the actual representation of women in the U.S. workforce, as reported by [72]. Of the 575 occupations in our study, [72] provides the percentage of women for 365 occupations. For each occupation prompt p, we compute

$$\Delta(p) = \mathcal{R}_f^{\text{bls}}(p) - \mathcal{R}_f(p) \in [-1, 1] \tag{4}$$

where $\mathcal{R}_f^{\text{bls}}(p)$ represents the proportion of women in the U.S. workforce. A positive Δ indicates that T2I models generate fewer women than the actual workforce proportion, while a negative Δ indicates that they generate more women than expected. In Fig. 17, we present the distributions of Δ values for all five T2I models. Overall, the distributions tend to be centered above zero, indicating that, on average, T2I models depict a higher proportion of men compared to actual workforce statistics.

To further explore this perspective, we analyze bias amplification in occupations based on whether the majority of the workforce is male or female. This analysis is presented in Table 9. First, we classify each occupation as either "male majority" or "female majority" based on the actual proportion of women in that occupation. If more than 50% of the workforce is female, the occupation is labeled as "female majority"; otherwise, it is labeled as "male majority". Next, we examine whether the

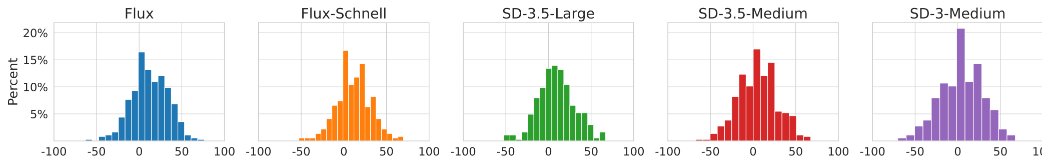

Figure 17: Distribution of differences Δ between female ratios in occupation images generated by T2I models and real-world (U.S.) ratio of women in the workforce for the respective occupation. Positive values indicate more men in generated images than in the workforce, and negative values indicate more women in generated images than in the workforce.

	Male majority		Female majority	
	reduced	amplified	reduced	amplified
Flux	6.85%	44.38%	19.18%	29.59%
Flux-Schnell	6.30%	44.93%	20.00%	28.77%
SD-3.5-Large	7.67%	43.56%	23.56%	25.21%
SD-3.5-Medium	10.96%	40.27%	28.77%	20.00%
SD-3-Medium	8.77%	42.47%	29.32%	19.45%

Table 9: We classify occupations into "male majority" or "female majority" based on whether there are more men than women in the workforce (actual U.S. statistics). Then, we check if, in generated images, the majority gender has increased or decreased ratio. If the majority gender is increased, we label it as "amplified"; if the majority gender is decreased, we label the occupation as "reduced".

proportion of men or women increases or decreases in images generated by T2I models. If the ratio of the majority group increases, we say the bias is amplified, whereas if it decreases, we say the bias is reduced.

From Table 9, we observe that bias in male-majority occupations is almost always amplified. For Flux models, bias in female-majority occupations is more often amplified than reduced. In contrast, for Stable Diffusion models, bias in female-majority occupations is more often reduced than amplified. Overall, these findings confirm our observation that T2I models tend to increase the proportion of men in generated images, while also showing numerous cases where female-majority bias is amplified.

I Detailed Discussion of Limitations

While our study makes a valuable contribution to understanding gender bias in current T2I models and extends insights from previous work, there are several important areas that we do not address. These include gender identities beyond the binary, social categories beyond gender, intersectional biases, and debiasing techniques. Below, we explain why these topics cannot currently be properly analyzed using the methods applied in our study. Furthermore, we justify our use of automatic methods for labeling perceived gender.

Non-binary gender identities. In the generated images, we do not find clear evidence of images that unambiguously depict non-binary gender identities. We believe that such an analysis should involve judgments or annotations from people who identify as non-binary, similar to [97]. Without this input, it is unclear how to identify relevant images or analyze stereotypes within them. This is also supported by our findings in Appendix E.2. Currently, automatic methods do not label images with "nonbinary", and as mentioned above, we are not aware of any other techniques that enable automatic analysis of images that may depict non-binary gender identities.

Automatic gender labeling. Using automatic methods to assign sensitive attributes such as gender (as well as race or age) can be problematic because models may introduce errors, carry their own biases, and in doing so, undermine the validity of analyses based on automatic labels. Even worse, if models are biased, they may reinforce those biases throughout the analysis. At the same time, using automatic tools is essential for conducting large-scale studies like ours. Therefore, we take steps to ensure our results are as valid as possible by addressing issues that arise from automatic methods. First, we filter images based on detected people, using state-of-the-art object detectors

White		Black		East Asian		Latino-Hispanic		Middle Eastern		Indian		SE Asian		Other	
a trout	0.99	going to a reggae concert	0.87	studying mandarin chinese.	0.99	eating tacos.	0.36	praying the obligatory 5 daily prayers.	0.59	in the slum	0.33	near the rice paddy	0.67	playing skyrim	0.34
going to a bass pro elite competition.	0.99	a basketball	0.67	watching anime	0.95	buying an awesome burrito.	0.32	in the medina	0.49	inside the kasbah	0.31	in the slum	0.34	playing world of warcraft.	0.33
by the fjord	0.99	on the savanna	0.61	a china	0.95	fast food worker	0.28	at the caravansary	0.49	at the temple	0.25	agricultural worker	0.30	playing the computer game lords of the fallen.	0.11
reading the new anthology with christine feehan in it.	0.99	tutoring their basketball players before their history exam.	0.59	brewing tea gong-fu style.	0.94	making guacamole	0.27	outside the mosque	0.47	in the village	0.24	in the village	0.28	a spear	0.07
flying to the adirondacks with their girlfriend or boyfriend.	0.99	usher	0.59	in the japanese garden	0.91	a bikini	0.25	inside the kasbah	0.44	inside the fort	0.24	at the temple	0.27	the hoodoo	0.06
trimming their beard.	0.98	at the basketball court	0.58	going out to dinner with their family to enjoy delicious chinese food.	0.90	eating a burrito bowl at chipotle.	0.24	going to a far away place for religious reasons.	0.38	laborer	0.20	farming or fishing worker	0.27	watching game of thrones	0.04
at the hunting lodge	0.98	playing basketball	0.56	a japan	0.90	licensed practical nurse	0.24	religious worker	0.37	in the medina	0.19	near the garbage dump	0.25	playing mario	0.03
installing a new door sweep.	0.98	meeting up with a friend and playing basketball for the entire afternoon.	0.54	going to hunan garden with their girlfriend/boyfriend.	0.85	registered nurse	0.24	chef	0.32	an elephant	0.18	rice	0.22	watching all of the lord of the rings movies.	0.03
hiking with their dog.	0.98	preparing for the upcoming fantasy football draft.	0.51	reading a book called "the taker" by alma katsu.	0.84	meat processing worker	0.23	near the mastaba	0.30	near the garbage dump	0.16	at the bazaar	0.19	a squid	0.03
hiking	0.98	working on their fantasy football lineup.	0.47	in the zen garden	0.84	physician assistant	0.23	baker	0.25	at the bazaar	0.15	within the rainforest	0.18	within the rainforest	0.03

Table 10: Top 10 prompts with highest avg. ratio of generated people for each race across T2I models.

[99]. Then, we crop person bounding boxes to reduce bias from background or contextual elements. Most importantly, we evaluate whether gender assignments from `InternVL2-8B` align with human annotations of perceived gender. As shown in Appendix E.1, this is indeed the case. Given the near-perfect alignment between human labels and automatically determined labels, we do not expect automatic methods to introduce significantly more errors or reinforce stereotypes beyond what human annotators might. While gender bias remains a concern in MLLMs [40], it is less pronounced in discriminative tasks that aim specifically to label gender.

Debiasing methods. The aim of our study is to provide a detailed, in-depth analysis of gender bias in current T2I models across everyday scenarios. In addition to understanding the societal issues related to T2I models, exploring ways to address these problems is also an important area of research. However, as models continue to be used without explicit steering mechanisms [11, 24, 26, 108], it becomes crucial to develop a clear understanding of their underlying issues. Determining how and when to apply steering or other debiasing techniques is another complex challenge, which lies beyond the scope of this study. For instance, it remains an open question whether solutions to these identified problems should be implemented by model providers or users. One possible approach is "ambiguity in, diversity out" [56], although this too raises concerns, such as maintaining contextual appropriateness. Given these challenges, detailed insights into model biases, like those provided in our study, are essential for making informed decisions about modifying or restricting model outputs. For the same reason, we do not aim to develop a benchmark. The fact that models exhibit bias has been shown before, and benchmarks typically construct one or a few measures of bias that help guide researchers and developers toward creating less biased models. However, such benchmarks can only indicate the degree of bias, not the specific manifestations of bias that we provide in this study.

Social categories beyond gender. We find that T2I models show strong biases in other social categories, such as race and age, when generating images from the underspecified prompts used in our study. To illustrate this, we detect perceived race and age for all identified people in the generated images using `InternVL2-8B`. The prompts used are similar to those employed for detecting perceived gender. For detecting race and age, we use the following prompts:

```
What is the race of the person in the image?
A. black
B. east asian
C. indian
D. middle eastern
E. latino-hispanic
F. southeast asian
```

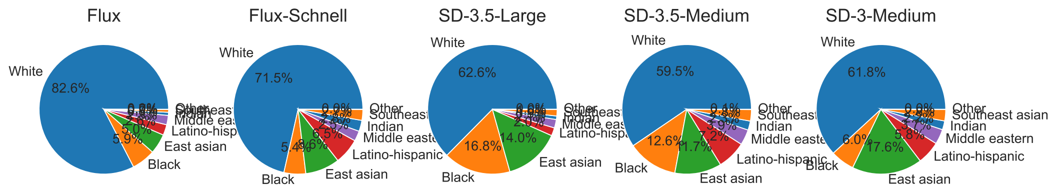

Figure 18: Race ratios for people in all images generated by T2I models, as assigned by `InternVL2-8B`.

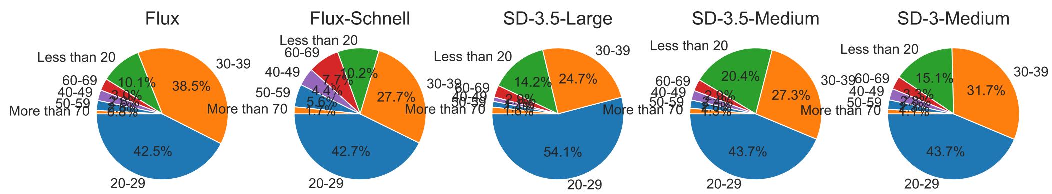

Figure 19: Age ratios for people in all images generated by T2I models, as assigned by `InternVL2-8B`.

```
G. white
H. other
I. unclear/cannot tell
Answer with a letter (A, B, C, etc.).
```

And for age, we use the following prompt:

```
What is the age of the person in the image?
A. less than 20
B. 20-29
C. 30-39
D. 40-49
E. 50-59
F. 60-69
G. more than 70
H. unclear/cannot tell
Answer with a letter (A, B, C, etc.).
```

In both cases, we randomly permute the option order (but not the option letters) to avoid label bias. Race and age categories are from FairFace [55]. However, we truncate the underage age categories to a single label ("less than 20").

We then calculate the overall ratios of people assigned to each race and age category for all 5 models in this study. Before calculating ratios, we drop all people who receive the "unclear/cannot tell" label. Results for race are in Fig. 18 and for age in Fig. 19. From these results, it is clear that models predominantly generate white and young (age 20-29 or 30-39) people, confirming results in [39, 102].

In Table 10, we also show the top 10 prompts with the highest average ratio of generated people across models for each race. There, we find that White and East Asian individuals have a notable number of prompts that, consistently across models, generate predominantly images of the respective race in all T2I models. Moreover, only prompts associated with White people tend to be fairly general, while prompts linked to other races are mostly tied to cultural or national stereotypes. For example, East Asian-looking people are generated from prompts mentioning East Asian cultural elements, such as "anime" or "mandarin chinese", while Latino-looking people appear in images generated from prompts like "tacos" or "burrito". An analysis of such cultural stereotypes in T2I models has been conducted in [27, 54].

Based on these findings, we conclude that the dominance of young, White individuals in generated images makes it difficult to perform intersectional analysis under the current experimental settings. To properly study race and age biases, as well as their intersection, it is necessary to explicitly prompt T2I models for these attributes and analyze the resulting stereotypes.

	Precision	Recall	F1-Score	Support
Black	0.83	0.91	0.87	1556
East Asian	0.67	0.86	0.75	1550
Indian	0.83	0.65	0.73	1516
Latino-Hispanic	0.56	0.53	0.55	1623
Middle Eastern	0.62	0.54	0.57	1209
Southeast Asian	0.58	0.48	0.53	1415
White	0.75	0.74	0.74	2085

(a) Detailed race labeling results by `InternVL2-8B` wrt. human annotations on the FairFace validation set. Overall accuracy is 68%.

	Precision	Recall	F1-Score	Support
20-29	0.66	0.60	0.63	3300
30-39	0.48	0.45	0.47	2330
40-49	0.48	0.22	0.30	1353
50-59	0.42	0.37	0.39	796
60-69	0.30	0.56	0.39	321
less than 20	0.81	0.82	0.81	2736
more than 70	0.34	0.60	0.43	118

(b) Detailed age labeling results by `InternVL2-8B` wrt. human annotations on the FairFace validation set. Overall ccuracy is 56%.

Table 11: Race and age classification results on the FairFace validation set.

Lastly, we validate the performance of MLLM race and age detection using human annotations from the FairFace dataset. Using the prompts described above, we assign race and age labels to all images in the FairFace validation set. Detailed results are shown in Table 11. Overall, the accuracy is 68% for race and 56% for age. While these values are lower than the reported accuracies for gender, they are still significantly better than random chance, considering the larger number of categories. Therefore, we conclude that our observations about race and age stereotypes are approximately accurate, although a fine-grained analysis remains difficult due to the lower agreement between automatic methods and human labels.

J Detailed Comparison to Previous Work

In this section, we provide a detailed comparison between our work and previous studies on analyzing gender bias in T2I models. We focus on works that use gender-neutral prompts, as this matches the experimental setup in our study. The comparison is shown in Table 12. For each paper, we include the number of gender-neutral prompts, the total number of images generated per evaluated model, a brief summary of the main findings, and a short note on how our study differs from that work.

In comparison to previous work, our study significantly improves the understanding of gender bias in T2I models by offering a detailed analysis across a wide range of everyday activities, places, objects, and occupations. As noted by [98] and clearly shown in Table 12, most prior studies have focused mainly on occupational prompts to highlight bias. While this focus is valuable, examining gender bias beyond occupations is also essential for a more complete understanding of how such bias manifests in T2I models.

Another aspect is the typically very small scale of studies, as also shown in Table 12. While this allows us to conclude that models are biased, gaining concrete insights into these biases requires a broader analysis like ours. Two other studies also generate a large number of images: [102] generated images from 200,000 distinct prompts, but used them not to analyze gender distributions for individual prompts or prompt groups, but to examine representational similarities between images from gender-neutral and gendered prompts. This setup is well-suited to reveal an overall male bias in the evaluated models, but does not support a detailed analysis of the specific stereotypes replicated by the models. Similarly, [64, 65] (these two papers have significant textual overlap) generated images for 2,123,200 prompts, about 70% of which focus on occupations. This study uses the images to compute holistic bias scores for comparing and ranking models, whereas our goal is to document biases in detail.

	# Images	# Prompts	Main Findings	Novelty of our work
[8]	2000	20	This study includes 20 gender-neutral prompts (in addition to prompts that either explicitly specify gender and race or focus on objects from diverse cultural contexts). Of these, 10 prompts describe people (e.g., "an exotic person", "a terrorist"), and 10 describe occupations. The paper reports on how gender and racial stereotypes are reflected and perpetuated in these 20 analyzed cases.	Our study enables a more thorough analysis of gender bias in T2I image models by including a larger set of prompts and generated images. This makes it possible to automatically evaluate broader trends, such as associations with household chores or workplaces, beyond just a few manually examined examples.
[21]	1050	105	Generated 10 images each for 105 different occupations. After collecting gender and race annotations from human labelers, a filtered set of 67 occupations was compared with respect to race and gender ratios in the U.S. workforce and the generated images. The study finds strong bias amplification, i.e., the images often depict only men or only women for a given occupation.	Our study examines gender bias not only in occupations but also in related categories such as activities, places, and objects. Within occupations, we include the complete set from the U.S. BLS list. Our method produces more reliable estimates of gender ratios by sampling a larger number of images and filtering out unsuitable prompts and images.
[23]	747	83	Generated 9 images based on 83 gender-neutral occupation prompts (excluding variants that explicitly specify gender). Gender, skin tone, and 15 other attributes were automatically detected. The results show that T2I models generally generate more men than women. Additionally, skirts appear only on women, while suits are more commonly shown on men.	Our study analyzes gender bias not only in occupation-related prompts but also in everyday activities and locations. In addition, our evaluation protocol provides a more reliable estimate of gender ratios by sampling more images and filtering out unsuitable ones. Lastly, we reduce contextual bias in automatic gender detection by cropping the images to focus on person bounding boxes.
[63]	4380	146	Generated 30 images using 146 gender-neutral occupation prompts. Gender and race distributions were analyzed with a non-parametric method that does not rely on explicit gender or race labels. A comparison with U.S. BLS statistics shows that women — especially Black women — are underrepresented.	We analyze gender biases beyond just occupations while also including a larger set of occupations. This broader analysis helps us identify bias trends on a wider scale. At the same time, we ensure our results are reliable by using large-scale sampling and filtering.

[64, 65]	2 123 200	2654	Generated 2,654 prompts related to occupations, social relationships, and attributes. A significant portion of the prompts include explicit gender or race identifiers, and about 70% of the prompts focus on occupations. The study evaluates different models using various bias scores to determine which ones are the most or least biased.	While this study also provides a large-scale evaluation of bias in T2I models, our work offers two key contributions: First, instead of presenting overall bias scores to compare models, we closely examine which specific biases (e.g., those related to household chores) the models exhibit. Second, our study goes beyond occupations, which are the main focus of the FaintBench benchmark, and explores a broader range of categories.
[66]	2000	100	This study evaluates the reliability and validity of gender bias analysis pipelines, identifying several issues, such as images featuring people of different genders or no people at all. The prompts used in the analysis cover all the categories included in this study, but on a much smaller scale (10 to 40 prompts).	Our study focuses on a detailed analysis of gender bias in T2I models at a large scale. To achieve this, we include a significantly higher number of prompts and analyses, comparing our results to those related to human stereotypes. However, the insights from this study shaped our experimental design, particularly emphasizing the need for careful and rigorous filtering.
[89]	31000	62	This study investigates bias amplification using 62 occupation prompts and concludes that bias amplification is largely explained by distribution shifts between the training and probing distributions.	Our study thoroughly documents the gender bias in recent T2I models, including observations of bias amplification. However, we do not explore the causes behind this bias amplification. Instead, we analyze a broad range of activities, places, objects, and occupations to provide in-depth insights.
[97]	924	231	Generated 4 images for each of 321 prompts centered on non-binary identities. The key findings are that non-binary identities are poorly represented by T2I models, often resulting in the creation of NSFW or degrading content.	Our study focuses specifically on binary gender. We also note that, without explicit instructions, models do not produce images that clearly represent non-binary identities. As a result, it is currently impossible to quantitatively explore biases related to non-binary identities using the models and methods applied in this study. However, we believe that addressing this issue is an important direction for future research.

[102]	800 000	200 000	This study examines how gender-neutral prompts are represented across different T2I models (text, latent noise, and images). The key finding is that when gender (or other image characteristics) is not specified in the prompt, the generated images tend to resemble those created from masculine prompts.	While this study analyzes a large number of prompts, it does not estimate gender ratios for prompts or prompt groups. Instead, it focuses on examining representational similarities. In contrast, our method allows for a deeper exploration of biases across a wide range of activities, places, objects, and occupations. This approach enables us to make precise statements about whether models display specific types of gender bias.

Table 12: Detailed comparison to previous work. We show the number of images in the study for each evaluated model, the number of prompts, a summary of the study's findings, and a comment on how our study contributes beyond the respective prior work.