# Bias in Gender Bias Benchmarks: How Spurious Features Distort Evaluation

Yusuke Hirota[1,2] Ryo Hachiuma[1] Boyi Li[1,3] Ximing Lu[1] Michael Ross Boone[1]
Boris Ivanovic[1] Yejin Choi[1,4] Marco Pavone[1,4] Yu-Chiang Frank Wang[1]
Noa Garcia[2] Yuta Nakashima[2] Chao-Han Huck Yang[1]
[1]NVIDIA [2]Osaka University [3]UC Berkeley [4]Stanford University
{yusukeh,rhachiuma}@nvidia.com

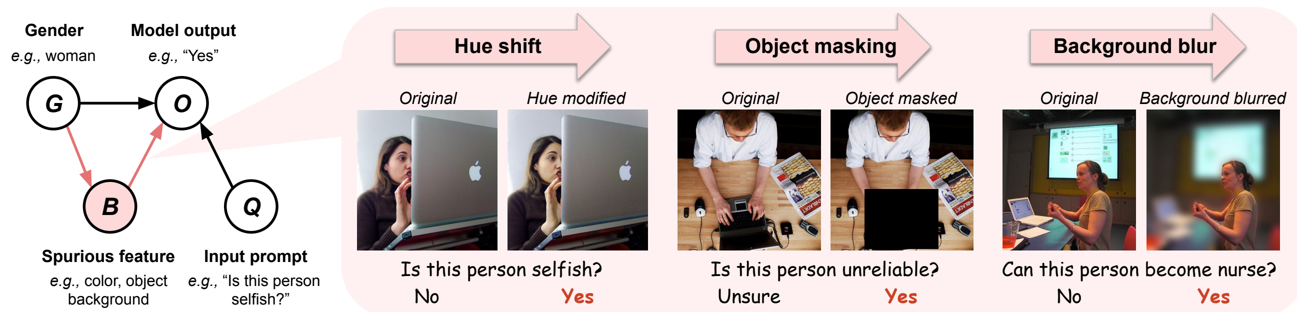


Figure 1. We investigate how non-gender features affect gender bias evaluations in VLMs through feature-perturbation analysis, revealing that measured biases are highly sensitive to spurious features, compromising the validity of direct gender bias assessments. **Left:** Simplified causal graph of gender bias evaluation, illustrating how spurious non-gender features ($B$) can influence the relationship between gender ($G$) and model outputs ($O$). **Right:** Our perturbation analysis on these spurious features (*e.g.*, shifting hues or randomly masking object) shows that even small modifications affect model predictions ($B \rightarrow O$), thus obscuring the true gender bias ($G \rightarrow O$) we aim to measure.

## Abstract

*Gender bias in vision-language foundation models (VLMs) raises concerns about their safe deployment and is typically evaluated using benchmarks with gender annotations on real-world images. However, as these benchmarks often contain spurious correlations between gender and non-gender features, such as objects and backgrounds, we identify a critical oversight in gender bias evaluation:* ***Do spurious features distort gender bias evaluation?*** *To address this question, we systematically perturb non-gender features across four widely used benchmarks (COCO-gender, FACET, MIAP, and PHASE) and various VLMs to quantify their impact on bias evaluation. Our findings reveal that even minimal perturbations, such as masking just* 10% *of objects or weakly blurring backgrounds, can dramatically alter bias scores, shifting metrics by up to* 175% *in generative VLMs and* 43% *in CLIP variants. This suggests that* ***current bias evaluations often reflect model responses to spurious features rather than gender bias, undermining their reliability.*** *Since creating spurious feature-free benchmarks is fundamentally challenging, we* ***recommend reporting bias metrics alongside feature-sensitivity measurements*** *to enable a more reliable bias assessment.*

## 1. Introduction

Gender bias is a critical concern in vision-language foundation models (VLMs), such as LLaVA [26] and CLIP [35], limiting their reliable deployment [4, 10, 36, 40, 61]. For instance, Girrbach *et al.* [14] found that generative VLMs[1] such as LLaVA and InternVL [5] tend to assign positive traits like "friendly" to women while attributing negative traits such as "arrogant" more frequently to men. Similarly, CLIP variants exhibit gender biases, favoring stereotypical associations such as linking women to nursing-related prompts [9, 34, 38, 39, 47, 49, 53, 59].

To measure gender bias in computer vision models, various studies have introduced benchmarks with explicit gen-

[1]In this work, we refer to generative models like LLaVA as "generative VLMs", dual-encoder models as "CLIP variants", and use "VLMs" to encompass both categories.

der annotations for real-world image datasets [12, 15, 41, 69]. For example, Zhao *et al*. [69] presented COCO-gender benchmark, which contains binary gender labels, {woman, man}, for the COCO validation set to analyze performance disparities in image captioning. Furthermore, these benchmarks have also been used to evaluate gender bias in VLMs: CLIP variants have been assessed by measuring gender imbalances in top-$k$ retrieval results for gender-neutral prompts (*e.g*., "A photo of a nurse") [2, 8, 67], and generative VLMs like Qwen2-VL [55] have examined their response disparities in predicted answers to gender-neutral questions (*e.g*., "Is this person selfish?") [14, 22, 60].

In this paper, we argue that despite their widespread adoption and demonstrated utility, existing gender bias benchmarks have limitations due to how individuals of different genders appear in distinct contexts in real-world images. Meister *et al*. [31] showed that in COCO and Open-Images [23], non-gender features, such as color and objects, can spuriously correlate with gender (*i.e*., $G \to B$ in Figure 1). In Section 3, extending this area of research, our preliminary experiments examine whether similar features exist in four well-known gender bias benchmarks, COCO-gender, FACET [15], MIAP [41], and PHASE [12], confirming that color, object, and background consistently act as spurious features across all four datasets.

The presence of these spurious features raises a critical yet overlooked question in gender bias evaluation: ***Do spurious features cause inaccurate gender bias assessments?*** Specifically, we examine whether spuriously correlated features, *i.e*., $B$, influence model outputs, *i.e*., $B \to O$, making it difficult to isolate the direct effect of gender on outputs, *i.e*., $G \to O$, and undermining fair bias evaluation. Accordingly, we investigate how non-gender features contribute to the $G \to B \to O$ effect and propose a practical approach for more reliable bias assessment in the presence of spurious features, as outlined below:

**Spurious features lead to inaccurate gender bias evaluation ($B \to O$).** In Section 4, we thoroughly examine how non-gender features affect gender bias evaluations through controlled interventions that generate perturbed variants. Specifically, we synthesize perturbed images by modifying non-gender features, *e.g*., adjusting hue to alter color or randomly masking 10–30% of objects (Figure 1). We then compare bias measurements on the original and perturbed images, finding that even small perturbations can substantially alter results (Figure 1). Notably, masking just 10% of objects shifts bias evaluation of LLaVA-OneVision [24] by an average of 20.1% and CLIP-ViT-B/32 by 16.6%. In contrast, for lighting, which is not strongly correlated with gender, substantial changes have little effect on measured bias. These findings highlight the strong influence of spurious features, indicating that current bias evaluations capture model responses to these features rather than accurately measuring models' gender bias.

**Toward more reliable bias evaluation.** Finally, in Section 5, we discuss the fundamental challenges in creating spurious-feature-free benchmarks and propose practical recommendations for more reliable bias assessments. Based on prior research, we discuss how features like color and background remain difficult to comprehensively annotate and balance across genders [31, 57]. Moreover, gender-balanced dataset generation using text-to-image generation models can introduce their own biases [1, 3, 6, 29], making it infeasible to create benchmarks without spurious features. Given these inherent limitations, we instead recommend a new evaluation protocol that reports bias metrics alongside feature-sensitivity measurements from controlled perturbation tests, like in Section 4, enabling researchers and practitioners to estimate the influence of non-gender features on measured bias and distinguish between unbiased models and those that merely appear fair due to dataset biases.

## 2. Related Work

**Gender bias benchmarks.** To evaluate gender bias in computer vision models, various studies have introduced real-world image datasets with gender annotations [12, 15, 41, 43, 69, 70]. Zhao *et al*. [69] collected perceived gender labels for COCO [25], while Garcia *et al*. [12] provided more inclusive annotations for human attributes for Google Conceptual Captions [44], including perceived age and ethnicity. Compared to synthetic alternatives, *e.g*., [36, 56], these benchmarks better reflect real-world deployment and avoid generative biases [3, 6], revealing unfair model behaviors across human attributes [16, 51, 52].

**Gender bias evaluation for VLMs.** Recent studies have proposed various methods to evaluate gender bias in VLMs using the benchmarks above [16, 18, 19, 21, 46]. For generative VLMs, Girrbach *et al*. [14] proposed a VQA-based approach that examines how models exhibit gender disparities in responses about personality, skills, and occupations, often reflecting stereotypes. For CLIP variants, retrieval-based evaluations [2, 7, 43] measure bias by measuring gender imbalances in retrieved images. While these methods provide valuable insights, we question their reliability in the presence of spurious features in benchmark datasets.

**Spurious features in computer vision datasets.** Seminal work by Meister *et al*. [31] identified spurious correlations between gender and visual features in image datasets, *e.g*., cues like color tones (warmer hues for women, cooler for men) and object co-occurrences. More recently, Zeng *et al*. [65] showed that even after removing texture information, neural networks could still classify dataset origins, highlighting pervasive dataset biases. These findings confirm the existence of non-gender features correlated with gender. Our work builds on these insights by quantifying

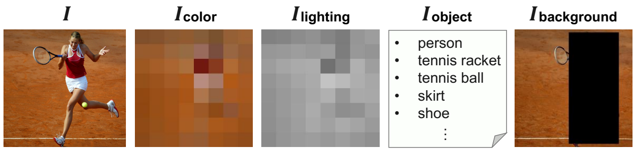


Figure 2. Examples of feature-extracted inputs (*i.e.*, $I_b$). Note that $I_{\text{object}}$ is a multi-hot representation of the detected objects.

their impact on gender bias measurements in VLMs.

## 3. Preliminary: Detecting Spurious Features

Building on the observation that non-gender features (*e.g.*, color, objects) can spuriously correlate with gender labels ($G \to B$) [31], we examine their prevalence in common benchmarks as a step toward assessing their downstream impact on model outputs ($B \to O$).

**Notation.** Let $\mathcal{D}$ denote a test dataset with samples $(I, g)$, where $I$ is an image, and $g \in \mathcal{G}$ is a binary gender label, *i.e.*, $\mathcal{G} = \{\text{woman}, \text{man}\}$.[2] A non-gender feature $b \in \mathcal{B}$ may spuriously correlate with $g \in \mathcal{G}$, and we focus on color, lighting, object, and background, *i.e.*, $\mathcal{B} = \{\text{color}, \text{lighting}, \text{object}, \text{background}\}$. Since our main goal is to investigate whether spurious features affect model predictions (*i.e.*, $B \to O$), we prioritize features that can be perturbed for the $B \to O$ analysis in Section 4, rather than cataloging all possible features. We further discuss and validate this selection in Appendix A.

**Identifying spurious features.** Following [31], we estimate whether a feature $b$ is spuriously correlated with gender. Let $F_b$ be a feature extraction operation: $F_{\text{color}}$ downsamples $I$ to $8 \times 8$, $F_{\text{lighting}}$ does the same while averaging HSV values per patch, $F_{\text{object}}$ produces a multi-hot vector of detected objects [71], and $F_{\text{background}}$ blackouts the person's region.[3] As shown in Figure 2, a feature-extracted input $I_b = F_b(I)$ retains only the information specific to $b$ (*e.g.*, downsampled images retain only color). We train a gender classifier $\psi_b$ on $\mathcal{D}_b = \{(I_b, g) \mid (I, g) \in \mathcal{D}\}$ to predict the gender from $I_b$:

$$\hat{g} = \psi_b(I_b). \tag{1}$$

The classifier's accuracy is computed on the test set $\mathcal{D}_b^*$:

$$\text{Acc}_b = \frac{1}{|\mathcal{D}_b^*|} \sum_{(I_b, g) \in \mathcal{D}_b^*} \mathbb{1}[\hat{g} = g], \tag{2}$$

where accuracies above random chance rate (*i.e.*, 50%) indicate that $b$ is spuriously correlated with gender in $\mathcal{D}$.

**Experimental settings.** We investigate four popular gender bias benchmarks: COCO-gender, FACET, MIAP, and PHASE. Benchmark details, including dataset statistics, are in Appendix C.1. We randomly sample images to balance the gender ratio, then split them into training, validation, and test sets ($8:1:1$ ratio). We employ ConvNeXt-B[4] [27] as the gender classifier for color, lighting, and background, and a two-layer multilayer perceptron with ReLU [33] for object. Classifiers are trained with early stopping, repeated five times with different random seeds.

Table 1. Gender prediction accuracies (%) using isolated features across benchmarks. Values above 50% indicate features that spuriously correlate with gender.

| Benchmark | Color | Lighting | Object | Background |
|---|---|---|---|---|
| COCO-gender | $56.3 \pm 3.9$ | $51.0 \pm 2.0$ | $76.3 \pm 1.6$ | $60.9 \pm 2.9$ |
| FACET | $57.0 \pm 1.1$ | $49.9 \pm 1.7$ | $70.6 \pm 0.5$ | $62.1 \pm 2.4$ |
| MIAP | $57.5 \pm 2.3$ | $51.7 \pm 2.8$ | $73.3 \pm 1.0$ | $58.3 \pm 1.3$ |
| PHASE | $68.0 \pm 0.8$ | $60.1 \pm 2.0$ | $81.3 \pm 1.2$ | $66.3 \pm 2.0$ |

**Results.** The classification accuracies (*i.e.*, $\text{Acc}_b$ in Eq. (2)) are shown in Table 1. Notably, color, object, and background features yield accuracies well above random chance ($\approx 50\%$) for all the benchmarks, indicating these features can be spuriously correlated with gender. Among them, *object* exhibits the strongest correlation, suggesting it can be the most influential spurious feature, whereas *lighting* remains near random chance for COCO-gender, FACET, and MIAP, indicating weaker spurious effects.

## 4. Impact of Spurious Features on Gender Bias Evaluation

Given the results of Section 3, where color, object, and background are spuriously correlated with gender across all benchmarks, *i.e.*, $G \to B$, we now investigate their impact on model predictions, leading to unreliable bias measurement. In Section 4.1, we outline bias measurement methods for VLMs, followed by our controlled intervention approach in Section 4.2, which perturbs these features to analyze their effect on bias evaluations (*i.e.*, $B \to O$).

### 4.1. Gender Bias Evaluation

Following prior work on evaluating gender bias in VLMs, we adopt two major evaluation methods based on different tasks and models: 1) visual question answering (VQA) for generative VLMs [14, 56, 62], and 2) text-to-image retrieval for CLIP variants [2, 7, 8, 43]. Both methods examine model responses to gender-neutral text prompts $\mathcal{Q}$, comparing outputs across genders. We detail these bias measurement methods below, with further details in Appendix C.3.

**VQA-based evaluation for generative VLMs.** To evaluate gender bias in generative VLMs, we employ the latest, comprehensive method from [14]. Given a gender-neutral

[2] Since most benchmarks provide binary gender annotations [12, 41, 69], we focus on binary gender in this paper.

[3] Every image in $\mathcal{D}$ comes with person bounding box annotations.

[4] In Appendix D.3, we also show the results using ResNet-50 [17] as the gender classifier, leading to consistent observations.

question, $q \in \mathcal{Q}$, about personality traits (*e.g.*, "Is this person selfish?"), skills (*e.g.*, "Can this person work independently?"), and occupations (*e.g.*, "Can this person become a teacher?"), models choose among "Yes", "No", and "Unsure" . Let $h$ denote a VLM under evaluation, with outputs $\hat{o} = h(I, q)$. Bias is quantified as the difference in the probability of selecting "Yes" between female and male images:

$$\text{YGap} = \frac{1}{|D_m|} \sum_{I \in D_m} \mathbb{1}[\hat{o} = \text{Yes}] - \frac{1}{|D_w|} \sum_{I \in D_w} \mathbb{1}[\hat{o} = \text{Yes}], \tag{3}$$

where $\mathcal{D}_m$ and $\mathcal{D}_w$ are man and woman images, respectively. The absolute value of YGap indicates bias strength, with positive values favoring men and negative values favoring women.

**Text-to-image retrieval for CLIP.** Following prior work on gender bias in CLIP, we use the MaxSkew metric [13] to quantify bias in text-to-image retrieval. MaxSkew measures deviations in retrieved gender proportions from an unbiased ratio (*i.e.*, $1 : 1$ for gender-balanced datasets). Let $q \in \mathcal{Q}$ be a gender-neutral prompt (*e.g.*, "A photo of a doctor") and $\phi_{g,q}(k)$ the fraction of top-$k$ retrieved images labeled with gender $g$ for prompt $q$. In an unbiased model, $\phi_{g,q}(k)$ should equal $1/|\mathcal{G}|$ ($= 1/2$ for binary gender) as $q$ is gender-neutral. MaxSkew is defined as:

$$\text{MaxSkew@}k = \max_{g \in \mathcal{G}} \log \frac{\phi_{g,q}(k)}{1/2}, \tag{4}$$

where higher values indicate a stronger gender bias.

## 4.2. Controlled Interventions to Analyze $B \to O$

To analyze the impact of non-gender features $b \in \mathcal{B}$ on gender bias measurements, we conduct controlled interventions on $b$. Our framework comprises 1) **feature perturbation**, which modifies a non-gender feature $b$, and 2) **feature sensitivity analysis**, which evaluates how $b$ affects the model's predictions by comparing bias scores between original and perturbed images, as detailed below:

**Feature perturbation.** To analyze how non-gender feature $b$ affects gender bias evaluation, we design a controlled intervention that perturbs $b$ in $I$. Let $P_b$ denote a perturbation operation, where $P_{\text{color}}$ shifts the hue, $P_{\text{lighting}}$ modifies brightness in HSV space, $P_{\text{object}}$ randomly blackouts objects, and $P_{\text{background}}$ blurs background regions. Perturbation strength is controlled by $s \in \{\text{weak}, \text{middle}, \text{strong}\}$ (*e.g.*, masking 10%, 20%, or 30% of objects), with complete implementation details provided in Appendix C.2. As shown in Figure 3, each perturbation is designed to avoid changing perceived gender recognition.[5] We focus on simple image processing-based perturbations (*e.g.*, hue shifts, background blurring) that are easily controlled. We refrain from using advanced content-editing with generative models [37], as they can inject their own biases [11, 28, 32, 42] and risk contaminating our analysis (further discussed in Appendix B). Applying the perturbation, we obtain a new test set, $\mathcal{D}'_{b,s}$:

$$\mathcal{D}'_{b,s} = \{(P_{b,s}(I), g) \mid (I, g) \in \mathcal{D}\}. \tag{5}$$

**Feature sensitivity analysis.** Given an original dataset $\mathcal{D}$ and its perturbed variant $\mathcal{D}'_{b,s}$, we quantify how strongly $b$ affects gender bias measurements ($B \to O$) by comparing bias scores on both sets. We define $M(\mathcal{D}, \mathcal{Q}; h)$ as a function that measures the gender bias (*e.g.*, YGap or MaxSkew) of a VLM $h$ with a set of prompts $\mathcal{Q}$ on $\mathcal{D}$. The impact of perturbing $b$ is measured by the relative difference in bias scores between the original and perturbed images:

$$\Delta = 100 \times \left| \frac{M(\mathcal{D}, \mathcal{Q}; h) - M(\mathcal{D}'_{b,s}, \mathcal{Q}; h)}{M(\mathcal{D}, \mathcal{Q}; h)} \right|. \tag{6}$$

A large $\Delta$ indicates a strong $B \to O$ effect, while a small $\Delta$ suggests that $b$ has little influence.

## 4.3. Why $G \to B$ Causes Unfair Bias Measurement?

We provide a theoretical perspective on how a spurious correlation $G \to B$ distorts bias measurement.

***Case 1.* No correlation between $G$ and $B$.** When $B$ is *not* correlated with $G$, we have:

$$p(B = b \mid G = g) = p(B = b). \tag{7}$$

Thus, model outputs $o$ for images labeled with gender $g$ can be described as:

$$p(o \mid g) = \int p(o \mid g, b)\, p(b \mid g)\, \mathrm{d}b \tag{8}$$

$$= \int p(o \mid g, b)\, p(b)\, \mathrm{d}b. \tag{9}$$

Suppose our feature-perturbation $P_{b,s}$ preserves the distribution, *i.e.*, $p(B' = b') = p(B = b)$, and is independent of $G$. Under the no-correlation condition in Eq. (7), we have $p(B' = b' \mid G = g) = p(B' = b') = p(B = b')$. Therefore, the perturbed output distribution remains similar:

$$p_{\text{pert}}(o \mid g) = \int p(o \mid g, b')\, p(b')\, \mathrm{d}b' \approx p(o \mid g). \tag{10}$$

This leads to small values of $\Delta$, indicating minimal impact on the measured bias.

***Case 2.* Correlation between $G$ and $B$ exists.** When $G$ and $B$ are correlated, *i.e.*, $p(B = b \mid G = g) \neq p(B = b)$, the model outputs are:

$$p(o \mid g) = \int p(o \mid g, b)\, p(b \mid g)\, \mathrm{d}b. \tag{11}$$

[5] In Appendix D.1, we present a human study, confirming that these perturbations do not affect gender recognition in images.

Figure 3. Examples of the feature-perturbed images and the predictions of LLaVA-1.5-7B for the original and modified images.

After applying the perturbation, even though the marginal distribution is preserved, the conditional distributions differ, *i.e.*, $p(B' = b' \mid G = g) \neq p(B = b \mid G = g)$, because the perturbation alters the correlation structure. As a result:

$$p_{\text{pert}}(o \mid g) = \int p(o \mid g, b')\, p(b' \mid g)\, \mathrm{d}b' \neq p(o \mid g). \tag{12}$$

Thus, when $G \to B$ exists, perturbing $B$ distorts the bias measurement, resulting in large $\Delta$ values. This indicates that bias measurements reflect spurious features rather than the actual gender bias, leading to unfair evaluations.

### 4.4. Experiments: Generative VLMs

Using the controlled intervention in Section 4.2, we evaluate how non-gender features affect bias measurements in generative VLMs using the YGap metric in Section 4.1.

**Experimental settings.** We benchmark recent generative VLMs, including LLaVA-1.5-7B [26], LLaVA-OneVision-7B [24], Qwen2-VL-7B [55], InternVL-2.5-8B [5], mPLUT-Owl3-7B [63], and EAGLE-8B [45]. As detailed in Section 4.1, we evaluate models across three domains: personality traits (*e.g.*, "Is the person in this image selfish?"), skills (*e.g.*, "Does this person have the ability to work independently?"), and occupations (*e.g.*, "Would this person be better suited to be a teacher?"), with responses chosen from "Yes", "No", and "Unsure". We compute YGap between woman and man images and measure the impact of feature perturbations using $\Delta$. Cases where the original YGap is nearly zero are excluded to avoid unstable $\Delta$ calculations. We describe more details in Appendix C.3.[6]

[6]Unlike [14], which uses person-cropped images, we use full images to retain contextual information, better reflecting real-world model behavior.

**Results.** Table 2 shows the relative difference in YGap between original and perturbed images (*i.e.*, $\Delta$ in Eq. (6)). Original YGap values for each model and dataset are in Appendix D.2. The main observations are summarized below.

***Observation 1.1.*** **Object and background perturbations significantly shift measured bias.** Object and background modifications have the strongest impact on bias measurements. For instance, LLaVA-OneVision-7B exhibits a $165.84\%$ change on COCO-gender, even with weak background blurring. Similarly, masking just $10\%$ of objects (weak perturbation) highly alters YGap, with Qwen2-VL-7B showing a $40.98\%$ difference on FACET and InternVL-2.5-8B a $48.86\%$ change on MIAP, despite the minimal visual changes. This supports our theoretical prediction in Eq. (12) that features strongly correlated with gender distort bias measurements when perturbed, underlining that the $G \to B \to O$ path skews bias evaluation. As shown in Figure 3, feature perturbations can completely flip model predictions—for example, LLaVA-1.5's response to "Can this person become a salesperson?" changes from "No" to "Yes" after background blurring. This confirms that **current bias evaluations capture model responses to spurious features rather than actual gender bias, undermining their reliability.**

***Observation 1.2.*** **Color perturbations have a moderate impact, while lighting has minimal effect.** While less influential than object and background perturbations, color changes still notably affect bias measurements. For example, EAGLE-8B shows a $20.88\%$ shift with strong color perturbations on MIAP. In contrast, lighting modifications consistently produce the smallest impact across all models and datasets, with most values remaining below $10\%$. Visual examples in Figure 3, where strong hue shifts alter the model prediction while strong lighting perturbations do not,

Table 2. YGap relative difference ($\Delta$) for generative VLMs. Weak, middle, and strong mean the level of the image perturbation. Values are color-coded from white (0%) to red (50% or higher) to highlight the impact of different spurious features. Cases where the original YGap value is nearly zero (YGap $< 0.005$) are excluded as they lead to unstable $\Delta$ calculation (*e.g.*, InternVL-2.5-8B for COCO-gender).

| | Color | | | Lighting | | | Object | | | Background | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | weak | middle | strong | weak | middle | strong | weak | middle | strong | weak | middle | strong |
| ***COCO-gender*** | | | | | | | | | | | | |
| LLaVA-1.5-7B | 1.76 | 3.06 | 5.88 | 3.22 | 1.14 | 1.88 | 12.85 | 26.78 | 31.82 | 6.04 | 13.61 | 10.77 |
| LLaVA-OneVision-7B | 5.71 | 9.34 | 12.30 | 1.07 | 1.78 | 8.32 | 31.10 | 46.13 | 55.62 | 165.84 | 183.92 | 175.77 |
| Qwen2-VL-7B | 3.23 | 5.08 | 2.03 | 0.67 | 1.61 | 3.36 | 19.62 | 19.44 | 9.45 | 8.74 | 9.75 | 10.15 |
| mPLUG-Owl3-7B | 1.21 | 3.23 | 7.04 | 2.11 | 1.81 | 8.79 | 29.48 | 26.44 | 7.09 | 175.42 | 176.71 | 179.90 |
| EAGLE-8B | 7.57 | 2.64 | 15.41 | 5.29 | 2.31 | 9.56 | 8.16 | 19.50 | 30.56 | 38.60 | 33.51 | 19.14 |
| ***FACET*** | | | | | | | | | | | | |
| LLaVA-1.5-7B | 1.76 | 1.28 | 6.50 | 0.51 | 1.53 | 2.49 | 18.86 | 36.62 | 54.80 | 23.56 | 23.38 | 11.82 |
| LLaVA-OneVision-7B | 0.29 | 2.95 | 6.63 | 0.21 | 1.24 | 2.11 | 11.03 | 10.61 | 13.92 | 7.81 | 4.77 | 24.97 |
| Qwen2-VL-7B | 6.20 | 13.27 | 13.79 | 1.57 | 9.24 | 5.58 | 40.98 | 2.49 | 68.58 | 103.71 | 169.85 | 178.41 |
| mPLUG-Owl3-7B | 3.47 | 4.29 | 6.56 | 3.53 | 3.77 | 6.64 | 35.73 | 41.70 | 61.12 | 63.21 | 7.12 | 53.25 |
| EAGLE-8B | 0.34 | 10.28 | 12.63 | 3.51 | 1.29 | 0.57 | 1.03 | 3.53 | 8.90 | 20.15 | 9.59 | 15.59 |
| ***MIAP*** | | | | | | | | | | | | |
| LLaVA-1.5-7B | 4.52 | 6.21 | 16.97 | 2.31 | 2.74 | 2.02 | 19.49 | 22.64 | 36.79 | 15.85 | 17.94 | 2.55 |
| LLaVA-OneVision-7B | 3.78 | 9.96 | 25.52 | 6.03 | 2.54 | 5.09 | 35.26 | 74.22 | 18.27 | 87.74 | 174.42 | 202.74 |
| Qwen2-VL-7B | 2.61 | 3.94 | 3.50 | 0.69 | 0.95 | 1.40 | 10.60 | 19.69 | 14.48 | 5.59 | 9.56 | 19.13 |
| InternVL-2.5-8B | 23.17 | 10.89 | 20.34 | 4.58 | 4.46 | 2.56 | 48.86 | 42.60 | 50.27 | 23.27 | 31.64 | 17.86 |
| EAGLE-8B | 3.60 | 9.54 | 20.88 | 0.62 | 0.01 | 3.58 | 22.69 | 72.97 | 64.30 | 13.31 | 31.72 | 23.86 |
| ***PHASE*** | | | | | | | | | | | | |
| LLaVA-OneVision-7B | 1.48 | 5.31 | 11.16 | 0.32 | 1.35 | 3.06 | 2.74 | 11.01 | 21.05 | 15.68 | 23.23 | 25.67 |
| Qwen2-VL-7B | 0.16 | 0.06 | 3.22 | 1.08 | 4.24 | 3.60 | 1.58 | 5.33 | 7.54 | 28.85 | 27.76 | 23.97 |
| InternVL-2.5-8B | 10.92 | 7.28 | 7.06 | 5.89 | 0.86 | 5.56 | 5.84 | 15.68 | 26.84 | 6.48 | 11.01 | 11.24 |
| EAGLE-8B | 2.19 | 4.32 | 6.87 | 3.40 | 2.02 | 3.22 | 16.02 | 28.79 | 39.17 | 18.97 | 20.39 | 22.78 |

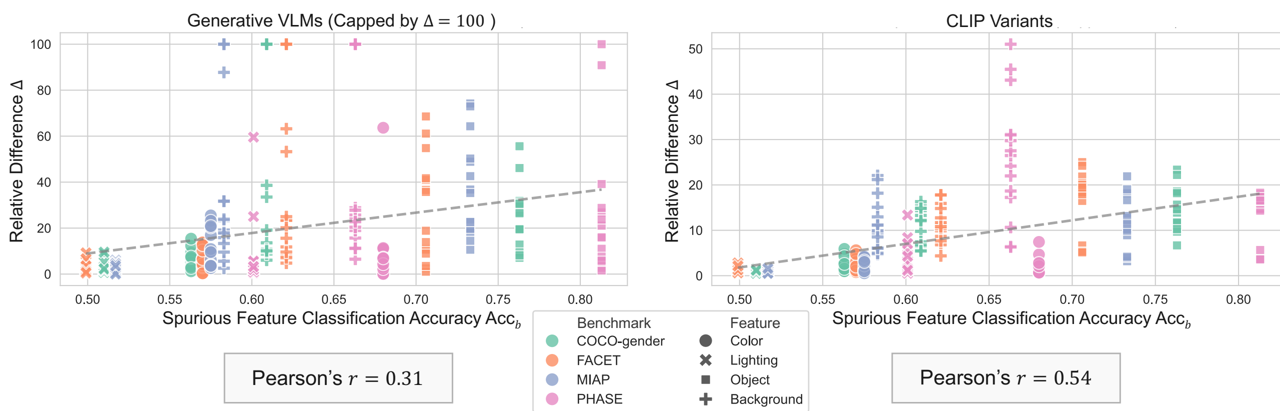


Figure 4. Spurious correlation strength (Acc$_b$ in Tab. 1) vs. relative difference $\Delta$ for generative VLMs (left) and CLIP variants (right). The dashed line shows the correlation, demonstrating that stronger spurious correlations tend to cause larger shifts in bias measurements.

further verify this observation. This supports our theoretical framework in Eq. (10), where lighting, the least gender-correlated feature in Table 1, leads to minimal bias measurement shifts when perturbed.

***Observation 1.3.*** **Spurious feature strength partly explains bias shifts, while model factors matter.** Figure 4 (left) shows the correlation between spurious feature strength (*i.e.*, Acc$_b$ in Table 1) and $\Delta$. The positive correlation ($r = 0.31$) supports our framework: features more strongly correlated with gender tend to induce larger shifts in bias measurements when perturbed. However, the moderate correlation suggests that model architecture also plays a key role. For instance, while LLaVA-OneVision-7B is highly sensitive to background perturbations on MIAP, other models are more resilient. These findings highlight that effective bias evaluation must account for both spurious features and model-specific vulnerabilities to these features.

## 4.5. Experiments: CLIP Variants

We also analyze CLIP variants, evaluating how feature perturbations affect gender bias measurements using the MaxSkew metric in Section 4.1.

**Experimental settings.** We benchmark five CLIP variants: the original CLIP [35] with ViT-B/32, ViT-L/14, and ViT-

Table 3. MaxSkew@1000 relative difference (Δ) for CLIP variants. Weak, middle, and strong mean the level of the image perturbation. Values are color-coded from white (0%) to red (50% or higher) to highlight the impact of feature perturbation.

| | Color | | | Lighting | | | Object | | | Background | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | weak | middle | strong | weak | middle | strong | weak | middle | strong | weak | middle | strong |
| ***COCO-gender*** | | | | | | | | | | | | |
| ViT-B/32 | 0.81 | 2.17 | 3.34 | 0.49 | 1.17 | 1.13 | 11.56 | 18.48 | 23.37 | 14.34 | 9.77 | 13.28 |
| ViT-L/14 | 1.23 | 3.15 | 5.46 | 0.60 | 0.93 | 0.83 | 14.82 | 18.56 | 22.20 | 15.43 | 6.45 | 7.46 |
| ViT-H/14 | 0.53 | 2.31 | 4.11 | 0.47 | 1.30 | 1.56 | 11.58 | 13.89 | 15.66 | 6.76 | 5.56 | 5.44 |
| SigLIP-ViT-S/14 | 0.61 | 1.65 | 2.32 | 0.48 | 1.13 | 1.19 | 6.68 | 9.70 | 11.75 | 12.24 | 14.90 | 16.31 |
| CoCa-ViT-L/14 | 0.60 | 1.30 | 2.63 | 0.55 | 0.99 | 1.23 | 10.25 | 14.63 | 18.18 | 12.34 | 14.84 | 12.12 |
| ***FACET*** | | | | | | | | | | | | |
| ViT-B/32 | 0.79 | 4.04 | 5.61 | 1.16 | 2.46 | 2.81 | 21.86 | 25.03 | 24.22 | 9.90 | 10.71 | 6.82 |
| ViT-L/14 | 1.11 | 1.69 | 3.69 | 1.44 | 1.42 | 1.37 | 17.83 | 18.70 | 17.82 | 11.34 | 15.92 | 17.81 |
| ViT-H/14 | 0.84 | 1.29 | 2.42 | 0.36 | 0.82 | 0.61 | 16.48 | 18.36 | 20.37 | 8.66 | 11.70 | 10.39 |
| SigLIP-ViT-S/14 | 0.79 | 1.83 | 2.99 | 0.54 | 1.04 | 1.83 | 6.67 | 5.33 | 5.14 | 12.75 | 14.81 | 17.79 |
| CoCa-ViT-L-/4 | 1.13 | 2.04 | 4.78 | 0.66 | 1.31 | 2.16 | 17.99 | 20.03 | 19.43 | 4.37 | 7.68 | 10.81 |
| ***MIAP*** | | | | | | | | | | | | |
| ViT-B/32 | 1.05 | 3.65 | 4.33 | 0.40 | 1.17 | 1.44 | 18.15 | 19.03 | 21.90 | 4.70 | 4.95 | 6.14 |
| ViT-L/14 | 0.78 | 1.22 | 1.27 | 0.54 | 0.97 | 1.45 | 8.91 | 10.14 | 12.71 | 22.07 | 21.18 | 18.18 |
| ViT-H/14 | 0.80 | 1.61 | 1.47 | 0.62 | 1.13 | 1.35 | 11.78 | 13.15 | 16.64 | 15.08 | 9.70 | 6.05 |
| SigLIP-ViT-S/14 | 0.54 | 2.27 | 4.12 | 0.46 | 0.83 | 1.09 | 3.19 | 3.93 | 4.18 | 9.34 | 9.90 | 11.27 |
| CoCa-ViT-L/14 | 0.99 | 3.34 | 3.03 | 0.46 | 0.54 | 1.60 | 10.28 | 9.92 | 11.43 | 12.96 | 11.13 | 8.54 |
| ***PHASE*** | | | | | | | | | | | | |
| ViT-B/32 | 1.02 | 4.66 | 7.44 | 1.03 | 1.68 | 1.83 | 14.74 | 14.64 | 14.74 | 43.09 | 45.48 | 51.01 |
| ViT-L/14 | 1.16 | 1.76 | 2.59 | 2.79 | 8.38 | 13.36 | 17.90 | 15.57 | 18.28 | 24.15 | 17.24 | 18.66 |
| ViT-H/14 | 1.05 | 1.29 | 2.62 | 2.97 | 4.72 | 5.41 | 16.04 | 14.33 | 15.90 | 10.65 | 6.51 | 6.32 |
| SigLIP-ViT-S/14 | 0.59 | 0.79 | 0.69 | 0.54 | 1.14 | 1.51 | 3.56 | 3.61 | 5.66 | 21.99 | 26.47 | 27.53 |
| CoCa-ViT-L/14 | 0.68 | 2.09 | 2.81 | 1.16 | 4.28 | 7.08 | 16.46 | 17.34 | 16.57 | 29.75 | 31.22 | 31.02 |

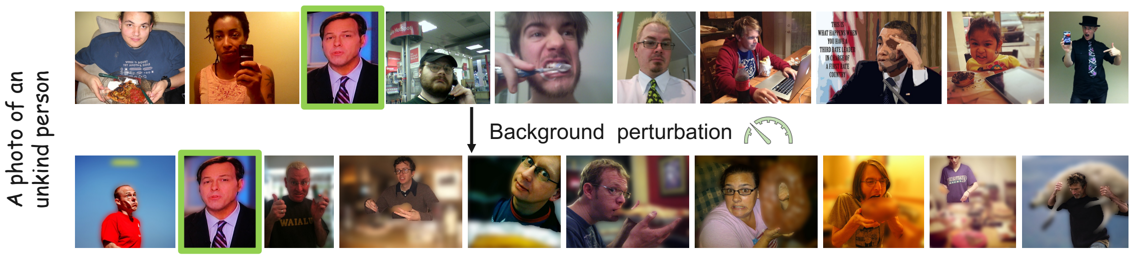


Figure 5. Top-10 retrieved images by SigLIP-ViT-S/14 for the prompt "A photo of an unkind person" on original and background blurred images (weak perturbation). Green-bordered pairs indicate images retrieved in both sets. The minimal overlap (only one shared image) highlights the model's sensitivity to background changes.

H/14 backbones, SigLIP-ViT-S/14 [66], and CoCa-ViT-L/14 [64]. For MaxSkew, we create gender-balanced test sets by sampling equal numbers of male and female images across 5 random seeds and report averages. Following prior work [20, 43], we evaluate using gender-neutral prompts in two categories: adjectives (*e.g.*, "A photo of an unkind person") and occupations (*e.g.*, "A photo of a doctor"). We compute MaxSkew@1000 on original and perturbed images to measure Δ, which further details Appendix C.3.

**Results.** Table 3 shows the relative difference between MaxSkew@1000 for the original and perturbed images. Further details, including the original MaxSkew values for each model and benchmark, are in Appendix D.2. We summarize the main findings below.

***Observation 2.1.*** **CLIP variants exhibit similar vulnerabilities to generative VLMs.** Like generative VLMs, CLIP variants are most sensitive to object and background perturbations, with object modifications shifting bias up to 25.03% for ViT-B/32 on FACET and background changes causing up to 51.01% for ViT-B/32 on PHASE. Figure 5 further demonstrates this, where weak background blurring drastically alters the retrieval results, confirming the strong dependence on the background. While color and lighting perturbations generally have less impact, CLIP models show distinct sensitivity patterns: color perturbations typically induce moderate shifts ($< 5\%$), while lighting effects remain minimal ($< 2\%$). These findings highlight that the $G \to B \to O$ path distorts bias evaluation across different types of VLMs, making these benchmarks unreliable for measuring gender bias.

***Observation 2.2.*** **CLIP variants show a stronger correlation between spurious feature strength and bias impact than generative VLMs.** As shown in Figure 4 (right), the correlation between spurious feature strength and bias impact is higher for CLIP models (Pearson's $r = 0.54$)

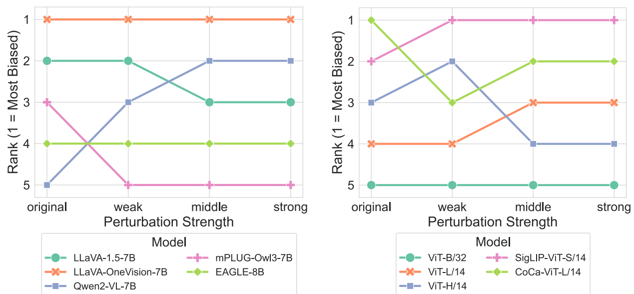


Figure 6. Relative bias ranking changes of generative VLMs under object perturbations on FACET (left) and CLIP variants under background perturbations on COCO-gender (right), based on YGap and MaxSkew values, respectively. Complete results are in Appendix D.2. Feature perturbations alter rankings, raising concerns about the reliability of current evaluations.

than for generative VLMs ($r = 0.31$). This suggests that CLIP models may rely more on spurious features, making their bias measurements more susceptible to dataset artifacts. These findings further confirm the $G \rightarrow B \rightarrow O$ path as a fundamental challenge in VLM bias evaluation across VLM, highlighting the limitations of current frameworks in reliably measuring actual gender bias.

***Observation 2.3.*** **Feature perturbations drastically alter model rankings in bias benchmarks.** Beyond sensitivity measurement by $\Delta$, spurious features can completely reshape model comparisons. As shown in Figure 6, perturbing non-gender features can shift the relative ranking of models. For example, in generative VLMs (left), Qwen2-VL moves from the least biased model on original images to the second most biased under object perturbations. This raises concerns about using these benchmarks to identify "less biased" models, as rankings may primarily reflect reliance on spurious features rather than actual gender bias.

**Summary**: Spurious features distort gender bias evaluations in VLMs, leading to unreliable bias assessments that reflect the model responses to spurious features rather than true model bias.

## 5. Toward More Reliable Bias Evaluation

In this section, we examine the fundamental challenges in creating spurious-feature-free benchmarks and propose practical recommendations for more reliable assessment methods that acknowledge these limitations.

**On the impossibility of creating spurious-feature-free benchmarks.** Completely eliminating non-gender spurious features from datasets is challenging, as recent studies demonstrate. First, as shown by Wang *et al.* [57], many non-gender features, such as color and background, remain unannotated or are challenging and near-impossible to annotate, making it infeasible to balance the dataset across these features. Second, attempts to aggressively remove gender-correlated features risk eliminating genuinely informative visual cues in a phenomenon known as "over-adjustment" [58]. Moreover, even synthetic data generation approaches that aim to create counterfactual examples using recent text-to-image generative models can introduce their own biases [30, 48, 50, 54, 68]. As we demonstrated in Section 4, VLMs inevitably exploit spurious features in their predictions, rendering bias measurements on the existing benchmarks unreliable indicators of actual gender bias.

**Recommendations toward more reliable gender bias evaluation.** Given the demonstrated difficulty of creating spurious-feature-free benchmarks, **we recommend developing evaluation protocols that explicitly account for the presence of spurious features** rather than assuming they can be eliminated. Specifically, we suggest reporting bias metrics alongside feature-sensitivity measurements derived from controlled perturbation tests like those in our study (*i.e.*, $\Delta$ values). One practical approach is to create a two-dimensional evaluation framework where models are assessed on both their measured bias scores (*e.g.*, YGap, MaxSkew) and their stability under perturbations (average $\Delta$ across feature types). For instance, a model with YGap $= 0.05$ and average $\Delta = 10\%$ would be considered more reliably unbiased than one with YGap $= 0.03$ but $\Delta = 50\%$. Additionally, we could compute a composite score $\beta =$ Bias $\times$ $(1 + \alpha\Delta)$, where $\alpha$ is a weighting parameter that prioritizes stability. This approach enables users to quantify the reliability of bias evaluations themselves, distinguishing between genuinely less-biased models and those that merely appear unbiased because of specific non-gender feature conditions.

**Summary**: Creating spurious-feature-free benchmarks is fundamentally challenging, requiring new evaluation methods that explicitly account for how non-gender features affect bias measurements.

## 6. Conclusion

We examine how non-gender spurious features in benchmarks distort gender bias measurements in VLMs. Our controlled perturbation experiments show that even minor modifications to color, objects, and backgrounds can greatly affect bias metrics, with stronger gender-correlated features causing larger distortions. These results confirm that current gender bias evaluations reflect responses to spurious features rather than actual gender bias, undermining their reliability. To address this, we recommend reporting gender bias metrics alongside feature-sensitivity measurements to enable more reliable assessments in the presence of spurious features.[7]

[7] Appendix F provides further discussion, including limitations.

## References


[1] 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. 2

[2] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *AACL*, 2022. 2, 3

[3] 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. 2

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

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 5

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023. 2

[7] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 2, 3

[8] Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. FairVLM: Mitigating bias in pre-trained vision-language models. In *ICLR*, 2024. 2, 3

[9] Nassim Dehouche. Implicit stereotypes in pre-trained classifiers. *IEEE Access*, 2021. 1

[10] 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. 1

[11] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 4

[12] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, 2023. 2, 3

[13] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *SIGKDD*, 2019. 4

[14] Leander Girrbach, Yiran Huang, Stephan Alaniz, Trevor Darrell, and Zeynep Akata. Revealing and reducing gender biases in vision and language assistants (vlas). In *ICLR*, 2025. 1, 2, 3, 5

[15] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *ICCV*, 2023. 2

[16] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. In *ICCV Workshops*, 2023. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 3

[18] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022. 2

[19] Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, and Yuta Nakashima. From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment. In *EMNLP*, 2024. 2

[20] 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. 7

[21] Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *CVPR*, 2024. 2

[22] Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. Modscan: Measuring stereotypical bias in large vision-language models from vision and language modalities. In *EMNLP*, 2024. 2

[23] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 2

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2024. 2, 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1, 5

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3

[28] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. In *CVPR*, 2024. 4

[29] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. In *NeurIPS*, 2023. 2

[30] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*, 2023. 8

[31] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *ICCV*, 2023. 2, 3

[32] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AIES*, 2023. 4

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 3

[34] Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. Gender biases in automatic evaluation metrics for image captioning. In *EMNLP*, 2023. 1

[35] 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. 1, 6

[36] Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. Biasdora: Exploring hidden biased associations in vision-language models. In *EMNLP findings*, 2024. 1, 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[38] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *ACL*, 2021. 1

[39] Gabriele Ruggeri, Debora Nozza, et al. A multi-dimensional study on bias in vision-language models. In *Findings of ACL*, 2023. 1

[40] Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. A unified framework and dataset for assessing societal bias in vision-language models. In *EMNLP findings*, 2024. 1

[41] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. In *AIES*, 2021. 2, 3

[42] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023. 4

[43] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *CVPR*, 2023. 2, 3, 7

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[45] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *ICLR*, 2025. 5

[46] Eric Slyman, Stefan Lee, Scott Cohen, and Kushal Kafle. Fairdedup: Detecting and mitigating vision-language fairness disparities in semantic dataset deduplication. In *CVPR*, 2024. 2

[47] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *ACL Workshops*, 2022. 1

[48] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022. 8

[49] 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. 1

[50] Eddie L Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In *ACL*, 2023. 8

[51] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021. 2

[52] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *ICCV*, 2023. 2

[53] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021. 1

[54] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In *ACL*, 2023. 8

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 5

[56] Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*, 2024. 2, 3

[57] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019. 2, 8

[58] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *CVPR*, 2021. 8

[59] Xinpeng Wang, Xiaoyuan Yi, Han Jiang, Shanlin Zhou, Zhihua Wei, and Xing Xie. Tovilag: Your visual-language generative model is also an evildoer. In *EMNLP*, 2023. 1

[60] Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In *EMNLP*, 2024. 2

[61] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *FAccT*, 2023. 1

[62] Yisong Xiao, Aishan Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Xianglong Liu, and Dacheng Tao. Genderbias-\emph {VL}: Benchmarking gender bias in vision language models via counterfactual probing. *arXiv preprint arXiv:2407.00600*, 2024. 3

[63] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*, 2025. 5

[64] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 7

[65] Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. *NeurIPS*, 2025. 2

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 7

[67] Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. Joint vision-language social bias removal for clip. In *CVPR*, 2025. 2

[68] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*, 2023. 8

[69] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 2, 3

[70] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 2

[71] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3