Supplementary Material for Bias in Gender Bias Benchmarks: How Spurious Features Distort Evaluation

Appendix

This appendix includes:

- Candidate feature selection (Appendix A).
- Further discussion on the perturbation strategy selection (Appendix B).
- Implementation details (Appendix C).
- Additional experimental results (Appendix D).
- Additional visual examples (Appendix E).
- Discussion (Appendix F).

A. Candidate Feature Selection

In this study, we focus on four non-gender features—color, lighting, object, and background—as potential spurious features in gender bias evaluation. These are selected based on prior research [9, 16] and practical considerations for our perturbation-based methodology.

Meister *et al*. [9] identified several visual cues correlated with gender in computer vision datasets. They showed that low-level features, such as color distributions, differ by gender, with images of women often having warmer hues and those of men having cooler tones. Additionally, they found that contextual elements, including objects and backgrounds, can serve as gender predictors even when the person is masked.

We prioritize these four features due to:

- **Established gender correlation**: Prior studies [9, 16] have shown that these features strongly correlate with gender in datasets like COCO and OpenImages. Our results in Section 3 further confirm that color, object, and background enable above-chance gender prediction across all examined benchmarks.
- **Controlled perturbability**: They can be modified in isolation while preserving other image characteristics, allowing precise analysis of their impact.
- **Diversity of feature types**: Our selection covers both low-level properties (color, lighting) and high-level contextual elements (objects, background).

While other potential spurious features exist (*e.g*., image composition, text overlays, photographic style), **our selection provides a strong foundation for studying spurious features effects while maintaining experimental tractability.**

B. Selection of Perturbation Strategies

We opt for straightforward, simple image processing-based feature perturbations (*e.g*., hue shifts, background blurring) rather than text-to-image (T2I) generative model-based feature editing as:

- **T2I models introduce additional biases**. State-of-the-art models like Stable Diffusion [12] are known to encode gender, racial, and cultural stereotypes [3, 8, 10, 14], raising fairness concerns. **Using T2I models for perturbations would risk contaminating our study with these biases.**
- **T2I models lack precise control over feature isolation**. While they can generate image variations, maintaining all other aspects constant is difficult. Even with careful prompting, generative models may subtly alter multiple attributes simultaneously, making it challenging to attribute model output changes to specific perturbations.

In contrast, our selected perturbation techniques ensure controlled and consistent transformations across the dataset. They offer parametric control over perturbation strength (weak, middle, strong) while preserving individual identity and gender-relevant attributes. This allows us to isolate spurious effects with greater precision while avoiding the introduction of additional biases.

C. Implementation Details

C.1. Details of Gender Bias Benchmarks

In this work, we focus on the four well-known gender bias benchmarks: COCO-gender [18], FACET [6], MIAP [13], and PHASE [4]. These benchmarks provide binary gender annotations for real-world images such as COCO. We detail each benchmark below:

- **COCO-gender** provides human attribute annotations, including gender and skin tone, for the COCO validation set.
- **FACET** presents more inclusive human attribute annotations, such as gender, age, and hair color, for a subset of Segment Anything 1 Billion.

Table 1. Statistics of gender bias benchmarks used in our experiments. Woman/Man indicates the number of woman/man images. For all datasets, we focus on images that do not contain multiple people for robust analysis.

Benchmark	Woman	Man	Total
COCO-gender	1,568	3,156	4,724
FACET	7,009	21,776	28,785
MIAP	1,459	4,501	5,960
PAHSE	4,031	6,831	10,862

- **MIAP** presents gender and age annotations for a subset of OpenImages.
- **PHASE** provides comprehensive attribute annotations (*e.g*., gender, ethnicity, emotion) for a subset of Google Conceptual Captions.

For all benchmarks, we select images that do not contain multiple people for robust analysis. The statistics of the datasets are shown in Table 1.

C.2. Details of Feature Perturbation

Color perturbation. For color perturbation, we alter the Hue component in HSV color space while preserving saturation and value (brightness). We implement different perturbation strengths through controlled random Hue shifts:

- Weak perturbation: Random Hue shifts within ±10 units in the HSV Hue channel ($0 - 255$ scale)
- Middle perturbation: Random Hue shifts within ±20 units, excluding the central range of ±10
- Strong perturbation: Random Hue shifts within ±30 units, excluding the central range of ±10

Lighting perturbation. For lighting perturbation, we modify the Value (brightness) component in HSV color space while preserving hue and saturation. We implement different perturbation strengths by controlling the range of random Value shifts:

- Weak perturbation: Random Value shifts within ±10 units in the HSV Value channel ($0 - 255$ scale)
- Middle perturbation: Random Value shifts within ±20 units, excluding the central range of ±10
- Strong perturbation: Random Value shifts within ±30 units, excluding the central range of ±10

Object perturbation. For object perturbation, we randomly mask 10%, 20%, or 30% objects in images. We do not remove persons to keep gender recognition. The objects are detected by the object detector [19].

Background perturbation. For background perturbation, we selectively apply Gaussian blur to only the background regions of images while preserving the person in focus. We implement different perturbation strengths by varying the blur radius:

- Weak perturbation: Gaussian blur with radius= 10 applied to background areas
- Middle perturbation: Gaussian blur with radius= 25 applied to background areas
- Strong perturbation: Gaussian blur with radius= 40 applied to background areas

The implementation uses person masks to ensure that only non-person regions are affected. This approach maintains all gender-relevant visual information about the person while modifying the contextual background, allowing us to isolate the impact of background features on model predictions.

C.3. Details of Gender Bias Evaluation

This section provides additional details on the gender bias evaluation methods used in our study, expanding on the information presented in Section 3.1.

VQA-based evaluation for generative VLMs. Following the methodology in Girrbach *et al*. [5], we evaluated generative VLMs using a comprehensive set of questions designed to assess gender bias across multiple domains:

- **Personality traits:** We use 20 personality traits (*e.g*., “friendly”, “honest”, “humble”) and a specific question format, “Is the person in this image [personality trait]”.
- **Skills:** We included 21 skill-related questions (*e.g*., “work independently”, “interact with individuals”, “work in team”) formatted as “Does this person have the ability to [skill]?”.
- **Occupations:** We employed 40 occupation-related questions (*e.g*., “laborer”, “accountant”, “teacher”) using the format “Would this person be better suited to be a [occupation]?”

For each question, an instruction “A. Yes, B. No, C. Unsure. Indicate the correct option as your answer.” is added to prompt models to select one of the options.

For calculating YGap values, we processed model outputs to extract the probability of a “Yes” response for each question and image. The YGap was then computed as the difference between the average probability of “Yes” responses for men versus women across all images in the dataset.

Text-to-image retrieval for CLIP variants. Following the previous works [1, 2, 7, 15, 17], we used a diverse set of gender-neutral prompts across two categories:

- **Adjectives (85 prompts):** Covering various personality traits, both positive (*e.g*., “good”, “kind”, “smart”) and negative (*e.g*., ”evil”, ”criminal”, ”violent”). These were formatted using templates such as “This is a photo of a [trait] person” or “This person is [trait]”.
- **Occupations (97 prompts):** Spanning diverse professional roles, including both stereotypically male-dominated fields (*e.g*., “computer programmer”, “engineer”, “CEO”) and stereotypically female-dominated ones (*e.g*., “nurse”, “childcare worker”, “social worker”).

Table 2. Consistency between human-identified gender and original gender labels for perturbed images. Results show percentage agreement across 200 randomly sampled images from COCO-gender, demonstrating that our feature perturbations preserve gender recognition.

Benchmark	Color	Lighting	Object	Background
COCO-gender	100.0	100.0	99.6	100.0
FACET	100.0	100.0	99.8	100.0
MIAP	100.0	100.0	100.0	100.0
PHASE	100.0	100.0	99.5	100.0

These used templates like "This is a photo of a [occupation]" or "A [occupation]".

For each prompt, we calculated the cosine similarity between the text embedding and all image embeddings in the gender-balanced test set. We then ranked the images based on these similarity scores and computed the MaxSkew@1000 metric as described in Section 3.1. To ensure robustness, we performed each experiment 5 times with different random seeds for sampling gender-balanced test sets and reported the average MaxSkew values.

D. Additional Experimental Results

D.1. Human Study on Gender Recognition Robustness After Perturbations

To ensure that our feature perturbations do not affect gender recognition, we conducted a human evaluation study. Specifically, we show randomly sampled 200 feature-perturbed (strong perturbations) images from COCO-gender, asking about the gender of individuals in these images. We then compute the consistency between the answered gender and the gender label of the original images. The results are shown in Table 2. While object perturbations occasionally obscure facial features, slightly reducing consistency, overall agreement remains high for all the features.

D.2. Complete Results of YGap and MaxSkew

In Tables 3 and 4, we show the complete results of YGap and MaxSkew@1000, respectively. The results verify that feature perturbations highly affect the measured bias scores for both types of VLMs.

D.3. Using ResNet50 for Gender Classifier

In Section 3.1, we employ ConvNeXt-B for the gender classifier. To check whether the insights are consistent across the backbone selection of the gender classifier, we conduct the same analysis using ResNet-50 for the gender classifier. Table 5 shows the results of Acc_b, and Figure 1 presents the correlation between Acc_b and Δ values, leading to the consistent observations across backbone selections.

E. Additional Visual Examples

In Figure 2, we provide additional visual examples of feature-perturbed images and generative model predictions, further confirming that spurious features (*i.e.*, color, object, and background) influence model outputs. For CLIP variants, while the main paper presented visual examples for background perturbations, Figures 3 to 5 provide additional examples for color, lighting, and object perturbations, showing the top-10 retrieved images for both original and perturbed inputs. These results further support our main findings: strong spurious features (*i.e.*, objects) distort model outputs more significantly than weaker ones (*i.e.*, color, lighting). Additionally, color perturbations moderately impact model predictions, whereas lighting perturbations have minimal effects.

F. Discussion

F.1. Additional Recommendations for Fairer Evaluation

While in the main paper, we recommend reporting bias metrics alongside feature-sensitivity measurements, here we expand on our second recommendation. Specifically, we also **recommend being more intentional in dataset curation**. The metrics for bias evaluation hinge upon the training and evaluation data selected for a model. This paper reminds readers of the criticality of data. Researchers, data scientists, and developers will be wise to pay close attention to identify potential spurious features, perform similar testing methodology, and use discerning judgment. In our paper, we recognize specific and seemingly-ambiguous spurious features may be in plain sight and surmise there are likely more to be identified. Our analysis encourages teams to review and reconsider their data sourcing strategies.

F.2. Cropping Person to Remove Background Features

Unlike the original YGap study [5] that uses cropped images focusing only on persons within bounding boxes, we intentionally use full images to preserve the contextual information that models encounter in real-world applications. While cropping may reduce the influence of background spurious features, it eliminates valuable contextual cues that VLMs typically utilize in practical deployments, where images are rarely presented as isolated subjects. Our approach allows for a more realistic assessment of how these models process and respond to complete visual scenes, better reflecting their behavior in actual use cases.

F.3. Limitations

Larger Models for Generative VLMs While we evaluate the latest, state-of-the-art generative VLMs, we focus on 7B-8B parameter variants due to computational resource

Table 3. YGap results (scaled by 100) of the generative VLMs. Weak, middle, and strong mean the level of the image perturbation, and original means the results for the original images. Gray cells indicate cases where the original YGap value is nearly 0 (YGap < 0.005 before scaled by 100), leading to unstable Δ computation, which we exclude from the analysis in the main paper.

	Color				Lighting				Object				Background			
Model	original	weak	middle	strong	original	weak	middle	strong	original	weak	middle	strong	original	weak	middle	strong
COCO-gender																
LLaVA-1.5-7B	-3.69	-3.63	-3.58	-3.47	-3.69	-3.57	-3.65	-3.62	-3.69	-3.22	-2.70	-2.52	-3.69	-3.88	-4.15	-4.05
LLaVA-OneVision-7B	-1.13	-1.07	-1.03	-0.99	-1.13	-1.14	-1.11	-1.04	-1.13	-1.48	-1.66	-1.76	-1.13	-3.01	-3.22	-3.12
Qwen2-VL-7B	2.80	2.89	2.94	2.86	2.80	2.78	2.85	2.90	2.80	2.25	2.26	2.54	2.80	2.56	2.53	2.52
InternVL-2.5-8B	-0.09	-0.10	-0.11	0.11	-0.09	0.05	0.14	0.04	-0.09	-0.15	0.11	0.03	-0.09	0.09	0.23	0.14
mPLUG-Owl3-7B	-0.92	-0.93	-0.95	-0.98	-0.92	-0.90	-0.93	-0.84	-0.92	-1.19	-1.16	-0.98	-0.92	-2.53	-2.54	-2.57
EAGLE-8B	0.57	0.61	0.58	0.65	0.57	0.60	0.58	0.62	0.57	0.52	0.46	0.39	0.57	0.35	0.38	0.46
FACET																
LLaVA-1.5-7B	-1.62	-1.59	-1.60	-1.52	-1.62	-1.61	-1.60	-1.58	-1.62	-1.32	-1.04	-0.73	-1.62	-2.01	-2.00	-1.81
LLaVA-OneVision-7B	-1.70	-1.71	-1.66	-1.63	-1.70	-1.67	-1.69	-1.67	-1.70	-1.52	-1.52	-1.47	-1.70	-1.57	-1.79	-2.13
Qwen2-VL-7B	0.65	0.69	0.74	0.74	0.65	0.66	0.71	0.69	0.65	0.92	1.20	1.10	0.65	1.32	1.75	1.81
InternVL-2.5-8B	0.00	-0.07	-0.12	-0.20	0.00	-0.08	0.01	-0.05	0.00	-0.00	-0.01	-0.00	0.00	0.26	0.28	0.34
mPLUG-Owl3-7B	-0.72	-0.74	-0.75	-0.67	-0.72	-0.74	-0.74	-0.75	-0.72	-0.46	-0.42	-0.28	-0.72	-0.26	-0.77	-1.10
EAGLE-8B	0.68	0.68	0.61	0.60	0.68	0.71	0.69	0.68	0.68	0.69	0.66	0.63	0.68	0.82	0.75	0.79
MIAP																
LLaVA-1.5-7B	-2.28	-2.17	-2.14	-1.89	-2.28	-2.22	-2.21	-2.23	-2.28	-1.83	-1.76	-1.44	-2.28	-2.64	-2.69	-2.33
LLaVA-OneVision-7B	-0.65	-0.68	-0.59	-0.49	-0.65	-0.69	-0.67	-0.62	-0.65	-0.88	-1.14	-0.77	-0.65	-1.22	-1.79	-1.97
Qwen2-VL-7B	2.29	2.23	2.20	2.21	2.29	2.27	2.26	2.32	2.29	2.04	1.84	1.95	2.29	2.41	2.50	2.72
InternVL-2.5-8B	0.83	0.63	0.74	0.66	0.83	0.86	0.86	0.85	0.83	0.42	0.47	0.41	0.83	0.63	0.56	0.68
mPLUG-Owl3-7B	0.02	-0.02	0.11	0.66	0.02	0.01	0.07	0.08	0.02	0.06	-0.10	-0.06	0.02	-0.15	-0.52	-0.64
EAGLE-8B	0.75	0.78	0.68	0.59	0.75	0.75	0.75	0.78	0.75	0.58	0.20	0.27	0.75	0.65	0.51	0.57
PHASE																
LLaVA-1.5-7B	-0.04	-0.04	-0.02	0.14	-0.04	-0.05	-0.13	-0.19	-0.04	0.44	0.35	0.37	-0.04	-0.19	-0.28	-0.39
LLaVA-OneVision-7B	-2.33	-2.29	-2.20	-2.07	-2.33	-2.34	-2.29	-2.25	-2.33	-2.26	-2.58	-2.82	-2.33	-2.69	-2.87	-2.92
Qwen2-VL-7B	3.82	3.82	3.82	3.94	3.82	3.78	3.66	3.68	3.82	3.88	3.61	3.53	3.82	4.92	4.88	4.73
InternVL-2.5-8B	2.43	2.16	2.25	2.25	2.43	2.28	2.40	2.29	2.43	2.28	1.86	1.77	2.43	2.27	2.16	2.15
mPLUG-Owl3-7B	0.21	0.18	0.08	0.19	0.21	0.20	0.16	0.08	0.21	0.26	0.02	-0.12	-0.21	-0.66	-0.78	-0.82
EAGLE-8B	3.03	2.97	2.90	2.83	3.03	2.93	2.97	2.94	3.03	2.55	2.16	1.85	3.03	2.46	2.42	2.34

Table 4. MaxSkew@1000 results (scaled by 100) of the CLIP variants. Weak, middle, and strong mean the level of the image perturbation, and original means the results for the original images.

Model	Color				Lighting				Object				Background			
	original	weak	middle	strong	original	weak	middle	strong	original	weak	middle	strong	original	weak	middle	strong
COCO-gender																
ViT-B/32	12.57	12.54	12.35	12.15	12.57	12.58	12.60	12.59	12.57	11.18	10.37	9.80	12.57	14.20	13.82	14.15
ViT-L/14	12.72	12.57	12.36	12.21	12.72	12.69	12.78	12.75	12.72	11.05	10.57	10.12	12.72	14.55	13.09	12.88
ViT-H/14	14.29	14.36	14.21	14.21	14.29	14.25	14.13	14.09	14.29	12.72	12.29	12.09	14.29	14.94	14.84	14.67
SigLIP-ViT-S/14	15.33	15.30	15.01	14.73	15.33	15.29	15.29	15.35	15.33	14.33	13.84	13.50	15.33	17.22	16.59	16.59
CoCa-ViT-L/14	13.17	13.24	13.20	13.22	13.17	13.18	13.21	13.18	13.17	11.96	11.39	11.10	13.17	14.73	15.03	14.74
FACET																
ViT-B/32	15.98	15.95	15.39	15.08	15.98	16.01	16.16	16.18	15.98	12.73	12.22	12.23	15.98	17.16	17.34	16.82
ViT-L/14	16.45	16.36	16.27	15.90	16.45	16.36	16.39	16.45	16.45	13.81	13.49	13.75	16.45	18.13	18.79	19.04
ViT-H/14	17.24	17.18	17.08	16.93	17.24	17.24	17.24	17.17	17.24	14.56	14.24	13.87	17.24	18.44	17.93	18.18
SigLIP-ViT-S/14	17.55	17.60	17.59	17.55	17.55	17.52	17.64	17.72	17.55	17.28	17.08	17.07	17.55	19.44	19.67	20.13
CoCa-ViT-L/14	17.68	17.61	17.27	16.80	17.68	17.62	17.47	17.28	17.68	14.84	14.23	14.21	17.68	18.23	18.84	19.21
MIAP																
ViT-B/32	20.40	20.21	19.72	19.57	20.40	20.42	20.34	20.31	20.40	17.09	16.89	16.34	20.40	21.30	21.32	21.61
ViT-L/14	19.89	19.84	19.73	19.73	19.89	19.84	19.89	19.82	19.89	18.20	17.92	17.37	19.89	23.35	23.17	22.70
ViT-H/14	19.96	19.91	19.78	20.01	19.96	19.96	19.94	19.94	19.96	17.76	17.46	16.71	19.96	22.62	21.55	19.37
SigLIP-ViT-S/14	24.46	24.53	24.63	24.81	24.46	24.53	24.61	24.65	24.46	23.70	23.65	23.45	24.46	26.18	27.04	27.41
CoCa-ViT-L-14	20.75	20.60	20.33	20.41	20.75	20.75	20.65	20.55	20.75	18.81	18.77	18.45	20.75	23.06	22.84	22.43
PHASE																
ViT-B/32	18.03	18.05	18.03	18.03	18.03	18.00	18.17	18.32	18.03	15.81	15.64	15.55	18.03	24.95	25.35	26.25
ViT-L/14	18.47	18.60	18.53	18.38	18.47	18.99	20.01	20.05	18.47	15.31	15.62	15.14	18.47	22.83	21.63	21.84
ViT-H/14	20.50	20.67	20.45	20.08	20.50	21.10	21.05	21.59	20.50	17.43	17.68	17.26	20.50	22.75	21.85	21.79
SigLIP-ViT-S/14	20.60	20.62	20.69	20.64	20.60	20.56	20.53	20.61	20.60	20.03	20.55	20.80	20.60	24.74	24.78	24.85
CoCa-ViT-L/14	20.01	20.09	20.33	20.27	20.01	20.11	20.48	20.58	20.01	17.00	16.75	16.89	20.01	20.92	21.45	21.64

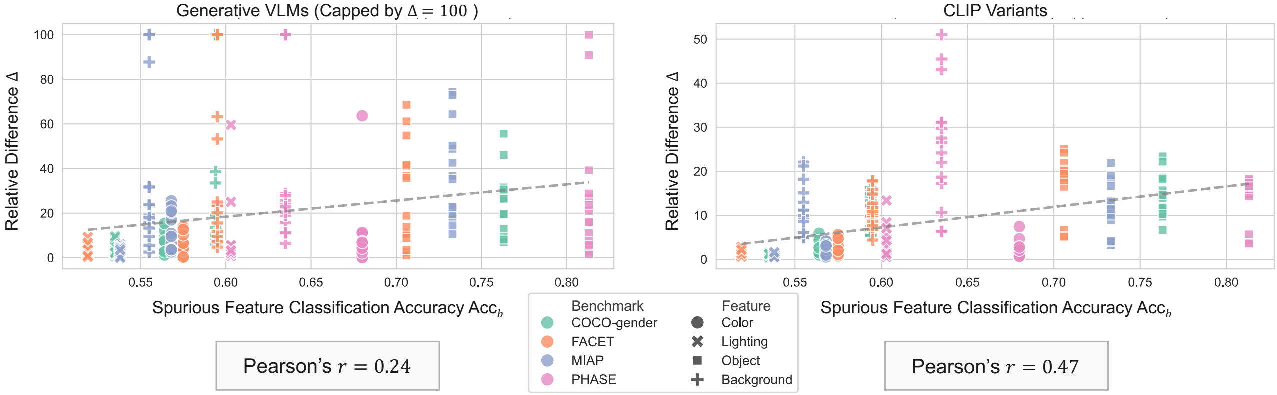

Figure 1. Relationship between spurious feature strength (Acc$_b$ in Table 1) and relative difference Δ for generative VLMs (left) and CLIP variants (right). The dashed line shows the correlation, demonstrating that stronger spurious features tend to cause larger shifts in bias measurements.

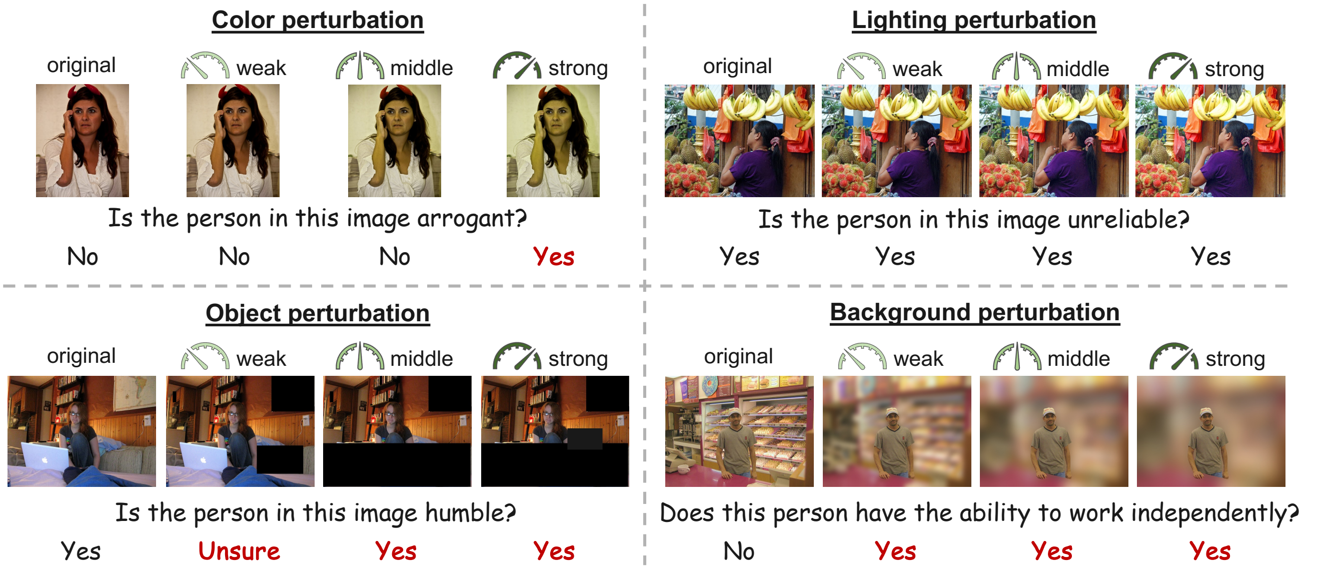

Figure 2. Examples of the feature-perturbed images and the predictions of LLaVA-1.5-7B (color perturbation), LLaVA-OneVision-7B (lighting perturbation), and Qwen2-VL-7B (object and background perturbations) for the original and modified images.

Table 5. Gender prediction accuracies (%) using isolated features across benchmarks. Values above 50% indicate features that correlate with gender, acting as confounders.

Benchmark	Color	Lighting	Object	Background
COCO-gender	56.4 ± 1.7	53.5 ± 2.7	76.3 ± 1.6	59.4 ± 1.1
FACET	57.5 ± 1.0	51.9 ± 1.5	70.6 ± 0.5	59.5 ± 0.4
MIAP	56.8 ± 0.9	53.8 ± 1.7	73.3 ± 1.0	55.5 ± 1.7
PHASE	68.0 ± 2.6	60.3 ± 1.8	81.3 ± 1.2	63.5 ± 1.9

constraints. Larger variants (*e.g.*, LLaVA-1.5-13B) and closed source models like GPT-4V [11] may exhibit different sensitivity to spurious factors due to their enhanced representational capacity. We leave this exploration as future work.

Other Evaluation Metrics While YGap and MaxSkew are widely adopted metrics, there are other evaluation metrics for VLMs (*e.g.*, NDKL [1, 17] for CLIP variants), which might yield different insights. Although our findings with YGap and MaxSkew provide strong evidence that spurious features affect gender bias evaluations, extending our analysis to other metrics would further strengthen these conclusions and potentially reveal additional insights into how spurious factors influence different aspects of bias measurement.

Potential Bias in Object Detector The object detector used in our analysis [19], while state-of-the-art, necessarily introduces its own detection patterns, which could interact with our analysis. We selected the detector for its

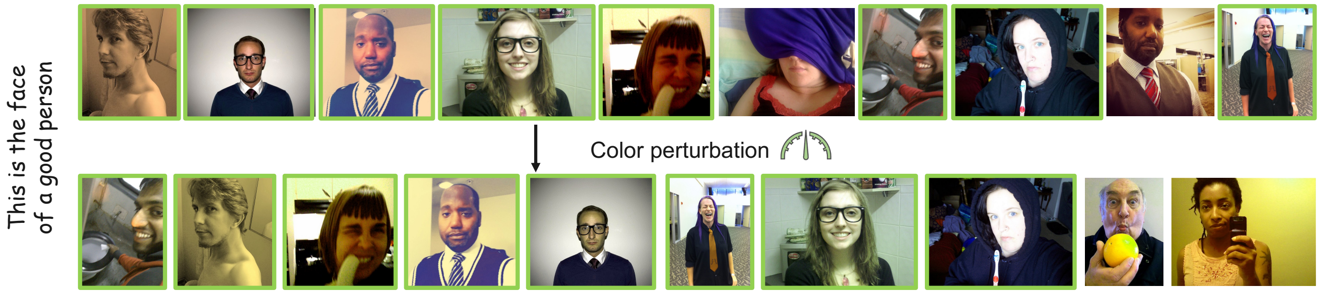

Figure 3. Top-10 retrieved images by ViT-B/32 for the prompt "This is the face of a good person" on original and hue-modified images (middle perturbation). Green-bordered pairs indicate images retrieved in both sets.

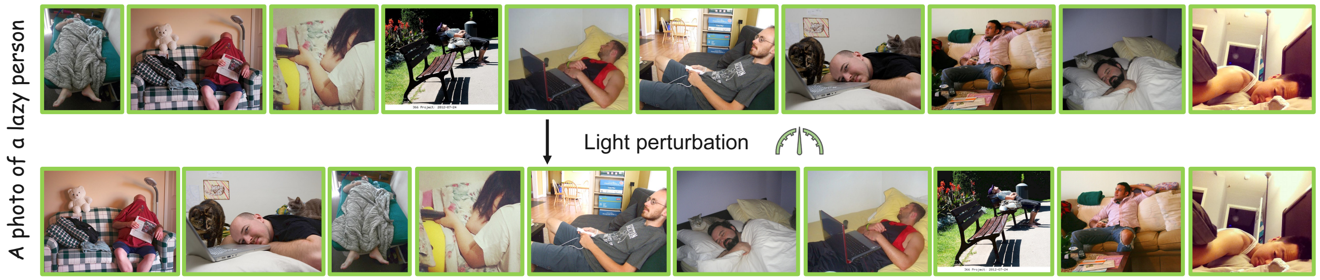

Figure 4. Top-10 retrieved images by ViT-L/14 for the prompt "A photo of a lazy person" on original and brightness-modified images (middle perturbation). Green-bordered pairs indicate images retrieved in both sets.

widespread adoption and demonstrated reliability across diverse datasets. While potential detector biases are minimized through our controlled experimental design, fully disentangling detector characteristics from VLM bias measurements presents a fundamental challenge in this research domain. Future work incorporating ensemble detection approaches or self-supervised methods could further isolate the effects of detector choice from the underlying bias phenomena we aim to measure.

References

[1] 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, 5

[2] 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] 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. 1

[4] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, 2023. 1

[5] 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. 2, 3

[6] 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. 1

[7] 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. 2

[8] 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. 1

[9] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *ICCV*, 2023. 1

[10] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AIES*, 2023. 1

[11] OpenAI. Gpt-4v(ision) system card, 2023. 5

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[13] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. In *AIES*, 2021. 1

[14] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023. 1

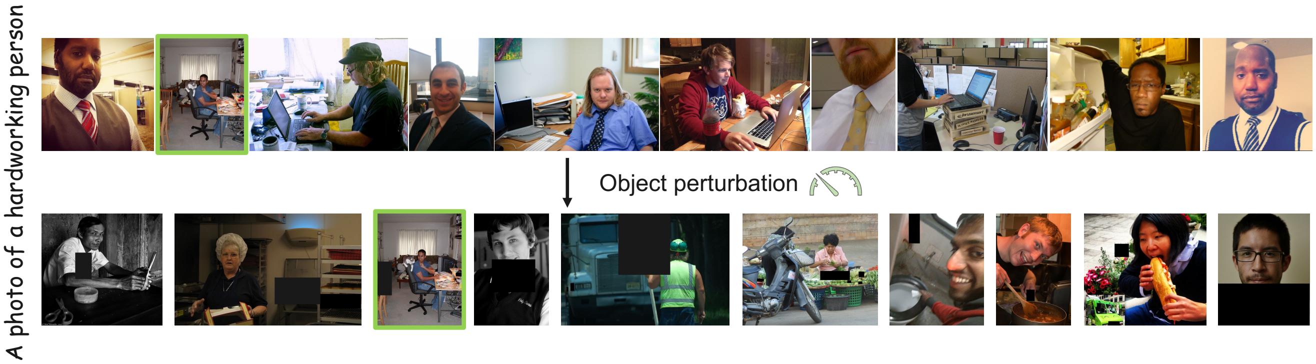

Figure 5. Top-10 retrieved images by CoCa-ViT-L/14 for the prompt “A photo of a hardworking person” on original and object-masked images (weak perturbation). Green-bordered pairs indicate images retrieved in both sets.

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

[16] Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. *NeurIPS*, 2025. 1

[17] Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. Joint vision-language social bias removal for clip. In *CVPR*, 2025. 2, 5

[18] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 1

[19] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 5