Proceedings of Machine Learning Research 279:119–132, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

Improving Bias Metrics in Vision-Language Models by Addressing Inherent Model Disabilities

Lakshmipathi Balaji Darur LAKSHMIPATHI.BALAJI@RESEARCH.IIIT.AC.IN
IIIT - Hyderabad, India
Shanmukha Sai Keerthi Gouravarapu GOURAVARAPU.SAI@RESEARCH.IIIT.AC.IN
IIIT - Hyderabad, India
Shashwat Goel SHASHWAT.GOEL@RESEARCH.IIIT.AC.IN
IIIT - Hyderabad, India
Ponnurangam Kumaraguru PK.GURU@IIIT.AC.IN
IIIT - Hyderabad, India

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

The integration of Vision-Language Models (VLMs) into various applications has highlighted the importance of evaluating these models for inherent biases, especially along gender and racial lines. Traditional bias assessment methods in VLMs typically rely on accuracy metrics, assessing disparities in performance across different demographic groups. These methods, however, often overlook the impact of the model's disabilities, like lack spatial reasoning, which may skew the bias assessment. In this work, we propose an approach that systematically examines how current bias evaluation metrics account for the model's limitations. We introduce two methods that circumvent these disabilities by integrating spatial guidance from textual and visual modalities. Our experiments aim to refine bias quantification by effectively mitigating the impact of spatial reasoning limitations, offering a more accurate assessment of biases in VLMs.

Keywords: Biases, Vision Language Models

1. Introduction

The advent of Vision Language Models (VLMs) has significantly advanced the field of artificial intelligence by enabling the seamless integration of visual and textual information. Models such as CLIP and BLIP have demonstrated exceptional capabilities across various tasks, including image retrieval (Xue et al., 2022; Bai et al., 2023), captioning (Li et al., 2022, 2023; Liu et al., 2024; Beyer et al., 2024), and visual question answering (Antol et al., 2015; Lin and Byrne, 2022). However, as these models become increasingly integrated into real-world applications, the evaluation of inherent biases, particularly along gender and racial lines is crucial.

Recent evaluations of VLMs have employed diverse methodologies to assess various dimensions of bias, focusing on factors such as gender (Ruggeri and Nozza, 2023; Fraser and Kiritchenko, 2024; Harrison et al., 2023; Janghorbani and De Melo, 2023), and race (Janghorbani and De Melo, 2023; Fraser and Kiritchenko, 2024). These assessments predominantly utilize accuracy as the primary metric, comparing performance across different demographic groups to identify biases. Since these measures are derived from differences in performance across groups, it is necessary to first ensure that these models are fundamentally proficient in underlying tasks. The performance of VLMs on these tasks, however, is heavily influenced by factors such as prompting techniques, and inherent limitations like a lack of spatial reasoning and compositionality.

This work contends that traditional methods of bias evaluation must be complemented by techniques that enhance the spatial reasoning capabilities of VLMs, thereby enabling a more precise and comprehensive quantification of bias. This work systematically investigates the impact of various prompting techniques and impact of spatial reasoning on bias quantification in VLMs. To assess this possibility we conduct experiments using four models CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), BLIP2 (Li et al., 2022, 2023) and PaliGemma-3B (Beyer et al., 2024) using occupational-gender bias becnhmark VisoGender (Hall et al., 2024). We begin by demonstrating that VLMs can perform effectively on gender resolution tasks with simple prompts centered around gendered or occupational terms, using segmentation maps to evaluate performance. We then evaluate these models on more complex tasks that involve captions integrating both occupation and gender. The following sections of this paper explore the textual and visual prompting strategies, highlighting how advancements in spatial reasoning are crucial for a more precise assessment of gender bias. By addressing the limitations of current methodologies, our approach offers a more robust framework for bias assessment, ensuring that (VLMs) are both effective and equitable. Our experiments demonstrate that traditional methods likely overestimate gender bias in CLIP (Radford et al., 2021) compared to our findings, which incorporate spatial guidance. The key contributions of this study are twofold: (i) We demonstrate that while models excel at resolution tasks with simple prompts, they falter with complex prompts due to inadequate spatial reasoning. (ii) We introduce two methods that enhance the spatial reasoning of models and suggest a more accurate approach for measuring biases, minimizing the influence of these limitations.

2. Background

To effectively assess the impact of spatial reasoning on bias evaluation, it is essential to employ a benchmark whose results are influenced by spatial reasoning capabilities. For this purpose, we consider VisoGender (Hall et al., 2024), a benchmark designed for assessing Occupational-Gender bias in VLMs. The VisoGender Dataset (Hall et al., 2024) includes 690 images depicting individuals across 23 distinct occupations, featuring both single-person (**SP**) and two-person scenarios. In the two-person images, one individual, designated as the *main character*, is directly associated with the occupation, while the other, referred to as the *participant*, interacts with or accompanies the main character, forming a *main character-participant* gender pair. These images are further categorized into two-person same-gender (**TPS**) with 5 male-male (MM) and 5 female-female (FF) images per occupation, and two-person different-gender (**TPD**) with 5 male-female (MF) and 5 female-male (FM) images per occupation as shown in Fig. 1.

VisoGender (Hall et al., 2024) introduces a resolution task that assesses the model's ability to correctly associate gender pronouns with given images. For example, given an image accompanied by two captions with differing gender pronouns, as depicted in Fig. 1, the model needs to resolve and pick the correct caption for given image. This is evaluated

Table 1: VisoGender dataset summary, showing the counts of images within each split of the dataset.

	Occ.	Gender Pairs	Img's per Occ.	Overall
SP: Single Person	23	[M, F]	10	230
TPS: Two Person Same Gender	23	[MM, FF]	10	230
TPD: Two Person Different Gender	23	[MF, FM]	10	230

through Resolution Accuracy (**RA**), representing the percentage of correctly resolved captions. Average Resolution Accuracy, RA_{avg} combines the accuracies for male (RA_{m}) and female (RA_{m}) subjects. The gender resolution accuracy gap, **(GG)**, measures the difference between male (RA_{m}) and female (RA_{f}) subjects, indicating potential bias. These are formally described in equations (1) and (2)

$$RA_{\mathrm{avg}} = \frac{RA_{\mathrm{m}} + RA_{\mathrm{f}}}{2} \tag{1}$$

$$GG = RA_{\mathrm{m}} - RA_{\mathrm{f}} \tag{2}$$

A positive GG suggests a bias towards more accurate resolution of male-presenting subjects, and conversely for a negative value. This metric is important for evaluating the fairness and efficacy of VLMs in correctly recognizing diverse occupations from visual inputs.

Given our focus on the impact of a model's limited spatial reasoning on gender biases, we concentrate on images featuring at least two individuals. This approach stems from the observation that performance on *SP images* is already satisfactory, as demonstrated in (Hall et al., 2024). However, in scenarios where both individuals are of the same gender, it becomes ambiguous to identify the main character, which complicates our analysis. Due to the ambiguity in identifying the main character among two-person images, as detailed in Appendix Section A, we face challenges in spatially locating the main character in a given image. Using the perceived gender annotations of the main character in TPD images from VisoGender Dataset (Hall et al., 2024) enabled us to determine their spatial location based on gender. Consequently, our discussions and experimental analyses are confined solely to images of different-gender pairs (TPD).

3. Analyzing Spatial Token Similarities with Segmentation Masks

To assess the spatial reasoning of VLMs in identifying occupational gender with the VisoGender (Hall et al., 2024) benchmark, we conduct an experiment using the CLIP (*ViT-B/32*) (Radford et al., 2021) model. This experiment involves simple, single-word prompts related to gender and occupation to determine how these factors influence model performance. We evaluate the CLIP model with different prompts like common nouns *(man or woman)*, pronouns *(he, she, his, her)*, occupational terms *(doctor, lawyer, ...)*. We introduce an approach that quantifies the model's focus on individuals for a given text prompt. This method helps determine whether the difficulty in resolving tasks in TPD images stems

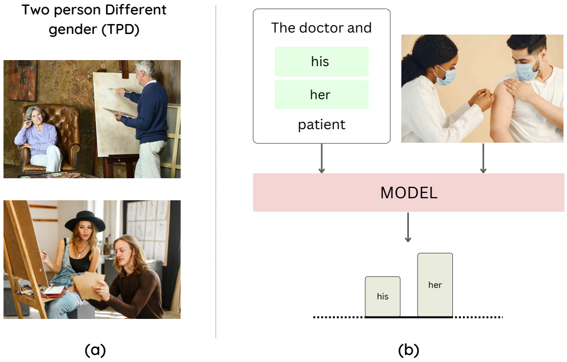

Figure 1: An overview of VisoGender benchmark. **a.** Shows a MF and FM images belonging to painter occupation respectively. **b.** An illustration of a resolution task, as explained in section 4

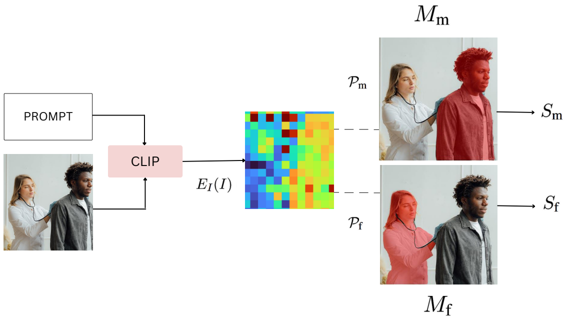

Figure 2: A visualization of similarity scores from the spatial token outputs of the last layer of the CLIP transformer, compared with segmentation masks for a given prompt, to obtain mean similarity scores.

from the model's inability to distinguish between individuals or other complexities hindering its performance.

Here we look in-depth into spatial tokens of the CLIP model, analyzing their similarity scores with given text prompt to understand where the model's focus lies, using segmentation maps. Let us denote the CLIP image encoder by E_I. For an input image I from TPD images, we consider segmentation masks annotated for male and female as M_{m} and M_{f} respectively. The image encoder produces a collection of visual feature tokens, with an adapted implementation from (Dong et al., 2023) as defined in equation (3)

$$E_I(I) = \{f_1, f_2, \ldots, f_N\} \tag{3}$$

where $1, \ldots, N$ are the indices of the spatial tokens from the last transformer layer of E_I. These token features correspond with patches P_i where i ranges from 1 to N in the real image. Given a text prompt T, we compute the similarity scores $S(I, T)$ from $E_I(I)$. Additionally, we have annotated segmentation masks M_{m} for male and M_{f} for female in an image. These masks are used to identify specific regions corresponding to the presence of male and female subjects in the image with respect to the patches as shown in Fig. 2. To determine which individual the model focuses on more, we take the average of similarity scores of the patches P_i that meet specific criteria. For a given mask, each P_i is considered if at least half of its area, $A(P_i)$ is covered by the segmentation masks M_{m} or M_{f}. The sets of patches for each category are defined in equations (4) and (5).

$$\mathcal{P}_{\mathrm{m}} = \{i : \mathrm{A}(P_i \cap M_{\mathrm{m}}) \geq 0.5 \times \mathrm{A}(P_i)\}, \tag{4}$$

$$\mathcal{P}_{\mathrm{f}} = \{i : \mathrm{A}(P_i \cap M_{\mathrm{f}}) \geq 0.5 \times \mathrm{A}(P_i)\}. \tag{5}$$

In above equations $\mathcal{P}_{\mathrm{m}}$ and $\mathcal{P}_{\mathrm{f}}$ represent the patches where the male and female masks, respectively, cover more than half of the patch area. The average similarity scores for the male and female categories are then calculated using equation (6). Finally, we compare S_{m} and S_{f} to ascertain the model's focus based on the given prompt T.

$$S_{\mathrm{m}} = \frac{1}{|\mathcal{P}_{\mathrm{m}}|} \sum_{i \in \mathcal{P}_{\mathrm{m}}} S_i, \quad S_{\mathrm{f}} = \frac{1}{|\mathcal{P}_{\mathrm{f}}|} \sum_{i \in \mathcal{P}_{\mathrm{f}}} S_i, \tag{6}$$

We perform experiments to emphasize and differentiate the model's proficiency in recognizing gender and identifying the main character associated with an occupation. We conduct experiments using gender-specific prompts detailed in Section 3.1 and occupational terms outlined in Section 3.2.

3.1. Gender Identification

In this task, we assess a model's ability to identify gender using TPD images. For each image, we provide a gender-specific prompt (e.g., "*male*") and compare the model's token similarity scores with predefined masks for both genders as shown in Figure 2. A prompt is deemed correctly identified for a given image if S_{m} is greater than S_{f} for a "*masculine*" prompt and reverse for a "*feminine*" prompt. Each of the 230 TPD images contains one male and one female. We define accuracy as the percentage of these images in which the model correctly identifies the gender based on the given prompt. The results of these experiments using various gender-specific prompts are summarized in Table 2.

Table 2: Accuracy scores of the model in distinguishing individuals spatially based on given prompts. The left table shows results for feminine prompts, while the right shows results for masculine prompts.

Feminine		**Masculine**	
Prompt	**Accuracy**	**Prompt**	**Accuracy**
woman	94.35	man	94.35
she	92.17	he	87.83
her	91.30	his	83.91

Given that the accuracy scores for most gender prompts exceed 80%, it indicates that the model effectively focuses on the individual matching the perceived gender of each prompt. However, it is notable that the model shows a slight bias towards correctly identifying female characters based on the given prompts.

3.2. Main Character identification

In this section, we evaluate the model's ability to identify the *main character* within the TPD images using annotated occupational prompts (e.g., "doctor"). If the male is the doctor in a given image, the prompt is considered correctly identified if S_{m} greater than S_{f}. We define gender-specific accuracy using a dataset comprising 5 male and 5 female main character TPD images for each of the 23 occupations. This results in a total of 115 images per gender, where each is depicted as the main character.

Following this approach, we conducted experiments with occupational prompts, presenting results for both genders in Table 3. The accuracies suggest that the model effectively distinguishes individuals based on their occupations. However, it exhibits a male bias when identifying the *main character* from occupational prompts, contrasting the findings from Table 2.

Table 3: Accuracy scores of model in identifying main character spatially for a given occupational prompt.

Gender	**Accuracy**
Masculine	77.39
Feminine	63.48

Given the high accuracy observed in Tables 2 and 3, the model demonstrates a clear ability to differentiate individuals in images using both gender and occupational prompts. Notably, pronouns yield lower accuracies compared to explicit nouns like "man" and "woman", a trend also observed in large language models (Hossain et al., 2023). This suggests that the model's performance is highly sensitive to the specific prompts or metrics used, particularly when assessing gender biases. Given its good performance in this simpler task we continue our experiments with complex prompts that include both gender and occupational terms in them in section 4 inspired by Hall et al. (2024).

4. Spatial guidance with Text Prompting

As mentioned in Section 3, the model effectively recognizes individuals using either gender or occupational prompts alone. However, it struggles when these prompts are combined in TPD images, as shown by the Overall Accuracy scores in Table 4. In this section, we discuss about how the performance changes with more complex prompts that integrate both gender and occupational terms into a single sentence. We evaluate the models CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), BLIP2 (Li et al., 2023) and PaliGemma (Beyer et al., 2024) on this task as shown in Fig. 1b with an example. For clip-like models (CLIP, OpenCLIP) we use similarity score to resolve between the two captions and for

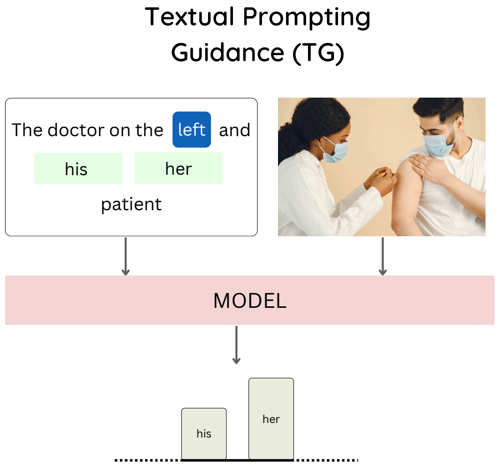

Figure 3: An illustration explaining our approach of adding direction of the main character to textual prompt to provide additional spatial guidance to the model as explained in section 4.

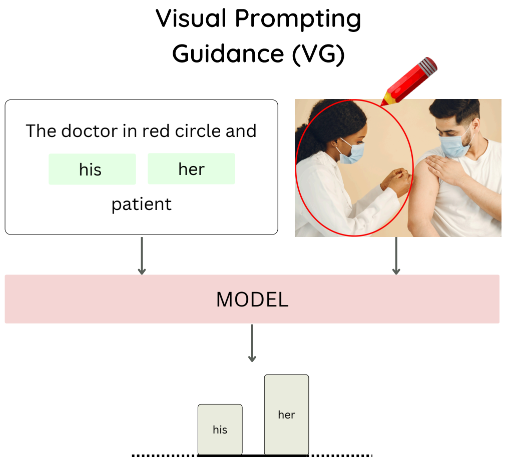

Figure 4: An illustration of our method showing red circles used to emphasize the main character, and improve spatial reasoning for the resolution task as explained in section 5

captioning models (BLIP2, PaliGemma) we compare log-likelihoods of the captions for a given image. This task is relatively complex, because the model needs to understand *who* and *where* is the *main character* and *participant* in the image and determine gender from the image context. Consequently, this experiment is also influenced by other factors such as spatial reasoning, playing a significant role in quantifying gender bias. Thus, we conduct an experiment that modifies the textual prompt to include directional guidance about the main character's position (*left* or *right*) in the image, facilitating the model's spatial reasoning. This method is referred to as *Textual prompting Guidance (TG)*, as depicted in Fig. 3.

In the above Table 4, we assess the performance of various models using complex prompts that integrate both gender and occupational terms. From rows R1, R3, R5 it is evident that the introduction of complex prompts notably diminishes the models' resolution capabilities, as quantified RA_{avg}, compared to the accuracies noted in Tables 2 and 3 in section 3. This reduction in performance suggests that the models struggle with tasks that require simultaneous processing of gender and occupational information. A potential explanation for this issue could be the models' limited spatial reasoning when faced with prompts that combine multiple contextual elements. We believe that inablities like lack of spatial reasoning shouldn't be accounted while quantifying biases. Our approach of *TG* reveals a significant shift in gender gaps and RA_{avg} across all models. This enhanced spatial awareness results in a more accurate estimation of gender bias. Research such as (Kamath et al., 2023) show that VLMs struggle with interpreting simple directional cues in text. We propose that our method of textual prompting guidance could be beneficial for future VLMs designed to better comprehend these textual directions. Considering the limitations

Table 4: Performance comparison of different models with and without Textual prompting Guidance (TG).

	Models	RA_{avg}	RA_m	RA_f	GG
1	CLIP (Radford et al., 2021)	0.38	0.20	0.57	-0.37
2	CLIP *(TG)*	0.48	0.38	0.57	-0.20
3	OpenCLIP$_{400M}$ (Cherti et al., 2023)	0.31	0.22	0.40	-0.18
4	OpenCLIP$_{400M}$ *(TG)*	0.46	0.32	0.61	-0.29
5	OpenCLIP$_{2B}$ (Cherti et al., 2023)	0.41	0.28	0.54	-0.26
6	OpenCLIP$_{2B}$ *(TG)*	0.47	0.63	0.31	-0.32
7	BLIP2 (Li et al., 2023)	0.61	0.47	0.75	-0.28
8	BLIP2 *(TG)*	0.56	0.70	0.42	0.28
9	PaliGemma (Beyer et al., 2024)	0.44	0.45	0.44	0.01
10	PaliGemma *(TG)*	0.36	0.23	0.49	-0.26

in current models' understanding of spatial directions, we suggest an alternative method of visual prompting in the following section, aiming to circumvent these challenges and refine bias quantification.

5. Visual Prompting with Red Circle

In this section, we introduce a prompting technique that offers spatial guidance to better approximate biases. We propose providing spatial guidance to the model by highlighting the main character. For this, we adopt an approach from (Shtedritski et al., 2023), that shows visual prompting images with red circles helps to extract useful behavior from VLMs such as CLIP in a zero-shot manner. This method of *Visual prompting Guidance (VG)* is tested with prompts and images annotated with red circles to provide visual guidance regarding the main character's location as depicted in Fig. 4. We conduct experiments across CLIP (Radford et al., 2021), OpenCLIP$_{400M}$ (Cherti et al., 2023) and OpenCLIP$_{2B}$ (Cherti et al., 2023) for which the results are presented in Table 5.

Table 5: Performance comparison of different models with and without Visual prompting Guidance (VG).

Models	RA_{avg}	RA_m	RA_f	GG
CLIP (Hall et al., 2024)	0.38	0.20	0.57	-0.37
CLIP *(VG)*	**0.58**	0.42	0.75	-0.33
OpenCLIP$_{400M}$ (Hall et al., 2024)	0.31	0.22	0.40	-0.18
OpenCLIP$_{400M}$ *(VG)*	**0.41**	0.29	0.53	-0.24
OpenCLIP$_{2B}$ (Hall et al., 2024)	0.41	0.28	0.54	-0.26
OpenCLIP$_{2B}$ *(VG)*	**0.49**	0.34	0.64	-0.30

With the incorporation of *VG* method, the models exhibit enhanced spatial reasoning, as demonstrated by the performance improvements in Table 5. This method helps mitigate factors such as lack of spatial awareness, providing a more accurate measure of gender bias in the models.

We now evaluate the *TG* and *VG* methods proposed for adding spatial cues that inhibit the effects of inadequate spatial reasoning in bias calculation. For the CLIP model (Radford et al., 2021), both methods consistently reduce the Gender Gap (GG) in magnitude. This reduction indicates that the CLIP model's perceived female bias of 0.37 is is likely an overestimate, once its spatial reasoning shortcomings are addressed. Similarly, for both versions of OpenCLIP, the *GG* consistently increases under both guidance methods as shown in Tables 4 and 5, indicating a stronger female bias than initially apparent without spatial cues. These consistent changes in *GG* values show that our guidance methods effectively measure biases by addressing the models' lack of spatial reasoning.

6. Ethical Considerations

In addressing gender bias evaluation, this paper adheres strictly to binary gender distinctions due to dataset constraints. Similar to the original Visogender dataset, our annotations, whether segmentation masks or red circles are based on the perceived gender presentation of both main characters and participants. We acknowledge that these visual markers may not accurately reflect a subject's self-identified gender, as gender presentation does not necessarily align in a binary manner with an individual's sex, pronouns, or identity. We acknowledge the limitations of this approach, particularly its exclusion of non-binary and LGBTQIA+ perspectives, and the ethical complexities inherent in gender recognition technologies. These technologies, especially when focused on binary gender, risk reinforcing societal biases and may disproportionately impact marginalized communities. Future work should broaden the spectrum of gender inclusivity and critically evaluate the societal implications of enhanced recognition capabilities to mitigate potential harm and ensure equitable advancements in the field.

7. Conclusion

In this study, we introduced enhancements to methodologies for evaluating gender biases in VLMs. Initially, we explored the spatial reasoning abilities of these models through segmentation maps, showing that while models perform well with simple gender and occupational prompts, their effectiveness diminishes when faced with complex prompts combining both elements. A crucial factor for this complexity of the task is the lack of spatial reasoning in these models. Through this work, we demonstrate that traditional methods might not fully consider the impact of limited spatial reasoning when measuring biases in VLMs.

To counter this, we introduced two new prompting strategies—one textual and one visual—to help reduce the effect of these limitations. Observing consistent improvements in Gender Gap and overall model performance with both methods suggests they are reliable. Specifically, our results indicate that the previously estimated bias in the CLIP model in VisoGender benchmark are likely an overestimate due to unaddressed spatial reasoning inabilities. Our work highlights the importance of carefully benchmarking biases in VLMs, by introducing a new dimension to the metrics and evaluation schemes used in the field of Algorithmic Fairness for AI systems.

8. Limitations

Our proposed approach relies on annotations indicating the spatial locations of individuals within each image, a requirement that may not scale effectively for very large datasets. The necessity for detailed annotations could limit the applicability of our methods in expansive, real-world scenarios where such detailed labeling is impractical. Furthermore, while our methods for providing spatial guidance to models are based on prior observations and straightforward techniques, they are not mechanistically validated to enhance model understanding consistently. These techniques presume an improvement in visual context interpretation without strong empirical evidence directly linking the methods to enhanced spatial reasoning capabilities in models.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*, 2023.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel M. Salz, Maxim Neumann, Ibrahim M. Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Martin Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bovsnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiao-Qi Zhai. Paligemma: A versatile 3b vlm for transfer. *ArXiv*, abs/2407.07726, 2024. URL https://api.semanticscholar.org/CorpusID:271088378.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

Kathleen Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In Yvette Graham and

Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.41.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. Run like a girl! sport-related gender bias in language and vision. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14093–14103, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.886. URL https://aclanthology.org/2023.findings-acl.886.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. Misgendered: Limits of large language models in understanding pronouns. *arXiv preprint arXiv:2306.03950*, 2023.

Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.126. URL https://aclanthology.org/2023.eacl-main.126.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Gabriele Ruggeri and Debora Nozza. A multi-dimensional study on bias in vision-language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.403. URL https://aclanthology.org/2023.findings-acl.403.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997, October 2023.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.