The APA Benchmark: A People-centric Benchmark for Testing Vision-Language Models

Yuri Ishitoya[1⋆], **Veronica Flores**[2⋆], **Ziyan Yang**[4], **Paola Cascante-Bonilla**[3], **Vicente Ordonez**[4]
[1]Ochanomizu University, [2]Santa Clara University, [3]University of Maryland, [4]Rice University

Abstract

We introduce the APA Benchmark, consisting of images of Actors, Politicians, and Athletes paired with a series of text prompts. Our benchmark serves as a tool for practitioners and researchers who are considering VLMs for people-centric tasks. We demonstrate the usefulness of our benchmark by systematically probing a large variety of modern VLMs for their associative abilities in this domain. We discuss the implications of these experiments and examine how model scale affects their basic associative abilities, influence from societal biases and capacity for identity recognition. APA Benchmark is publicly available at our group repository.[1]

1 Introduction

Vision-Language Models (VLMs) have made significant progress in reasoning over images and text. Image-text association models such as CLIP (Radford et al., 2021) achieve impressive classification performance, while recent autoregressive VLMs (Bai et al., 2025) are able to answer open questions about an image in a wide variety of domains. These models are trained at scale with open-vocabulary supervision from web images, making it difficult to assess the full extent of their capabilities. As a result, important benchmarks have been designed to probe VLM capabilities in complex tasks such as compositionality (Thrush et al., 2022; Zhao et al., 2022), text-chart understanding (Fu et al., 2024; Masry et al., 2022), or science and math (Lu et al., 2022, 2024; Yue et al., 2024).

However, we still lack a systematic evaluation benchmark that probes how VLMs perceive people, which is also the most common and socially consequential entity in natural images. We therefore aim to ask three main questions that cover VLM capacities of association to entity recognition, role association, bias assessment, and identity recognition, which we consider crucial for filling the specific gap of people-centric VLM evaluation.

In this work, we introduce APA, a people-centric benchmark consisting of pictures of public figures including Actors, Politicians and Athletes. We first demonstrate the utility of APA by testing the capability of VLMs to associate the basic level category `person` with our images, compared to subordinate categories such as `actor`, `politician`, `athlete`, and more specific subordinate categories such as `Academy Award winner`, `Senator`, or `Marathon runner`. On a positive note, our benchmark shows that generally, as models become more robust and capable, they rely less on stereotypical associations related to gender. However, our experiments also uncover models where this might not be the case. As a whole, our evaluation framework serves a dual purpose: it can help assess a VLM capacity for linking names with faces for applications where this is required, and it can flag VLMs that use identity recognition for predictions, which is crucial for applications where this is undesirable.

2 Dataset

The APA benchmark comprises high-quality portrait images of actors, politicians, and action shots of athletes, each annotated with basic demographic information such as name, gender, and category-specific attributes. The actor set includes film appearances and prioritizes those with *Academy Award* history, comprising 30 women and 60 men. The politician set consists of United States politicians, including all 100 Senators, 436 House Representatives, some delegates, and 100 Mayors from the largest cities, with additional metadata such as party affiliation and region; all politicians held public office on June 2022. The athlete set contains 109 pictures of prominent athletes, comprising 30 women and 79 men, each labeled with their respec-

[1]https://github.com/uvavision/apa-bench

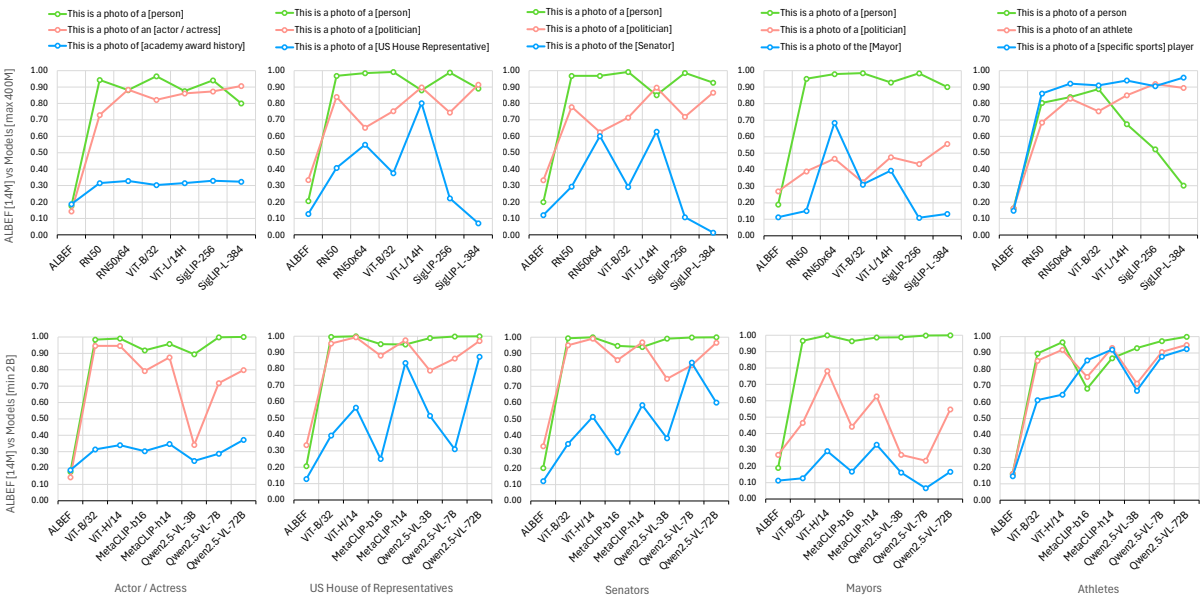

Figure 1: Classification accuracy across prompt levels (■ green: category, ■ pink: occupation, ■ blue: specialty). Top: models trained on ≤400M samples; bottom: >400M. Models are ordered by parameter size within each type. When only the visual backbone is specified, the model is CLIP. The top row shows models trained at the 400 million image-text pairs scale, and the bottom shows models trained at a larger scale.

tive sport. All images were sourced from Wikipedia and have a Wikimedia Commons compatible license. The dataset, code, and framework will be released under the MIT License and is included with this submission as supplementary material.

3 Evaluation Framework

Our benchmark proposes to uncover the capabilities of modern VLMs through association tests. Fig 2 shows some of the prompts that can be generated from our metadata for two images. Our prompts are of the form "This is a photo of [C]". For instance, at the most basic level, all of our images depict the category `person`. In this case, we also pair the images with the following six distractor categories: `dog`, `giraffe`, `plant`, `tree`, `bed`, and `chair`. These categories are included as a sanity check, as they are frequent categories in image datasets; a reliable model should consistently assign high scores to `person` for all benchmark images. Additionally, we design similar tests at two additional levels of granularity: General occupation and specific occupation (e.g., "This is a photo of a `politician`" and "This is a photo of a `Senator`"). A different set of negative prompts is selected manually for each type of occupation. A full list of positive and negative prompts for all levels of granularity can be found in our Appendix.

Our evaluation tests whether models can assign the highest matching score to the positive prompt.

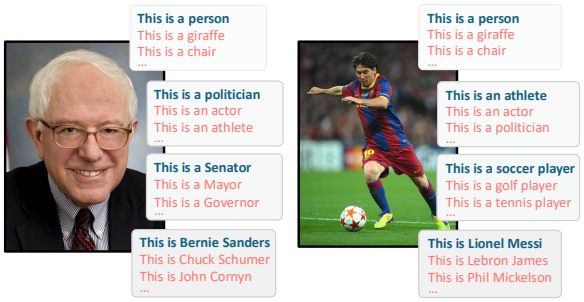

Figure 2: APA Benchmark images with associated prompts. In blue is the positive prompt and in red are negative prompts at various levels of granularity from more general to more specific prompts.

CLIP-style VLMs directly output a matching score between prompts and images. Generative VLMs are prompted to answer either true or false as in VQAScore (Lin et al., 2024), and we use the logit score of `true` as the matching score. We evaluate 14 VLMs in total, and find that: 1) It is generally easier for almost all VLMs to categorize things at the most general level of granularity i.e. `person`. 2) Models with larger capacity tend to be more capable in their associations, but with several exceptions. Results are shown in Fig 1.

Particularly, models such as SigLIP-256 and SigLIP-L-384 seem to struggle with scoring `person` above negative prompts for the images of athletes, yet they reliably recognize the people in this picture with the category `athlete`. This shows that using this model might affect downstream applications if relying on predictions at this superor-

Data	Model	Gender	Classes								Bias score (b)
			Sci.	*Pol.*	*Athl.*	*Teach.*	*Recep.*	*Asst.*	*Sales*	*Actor*	
Actor / Actress	RN101	woman	11.03	9.44	7.53	12.78	40.43	5.66	3.41	78.61	1.004
		man	18.80	17.24	7.43	13.51	3.72	7.92	18.86	78.32	
	DS-VL2	woman	0.79	0.36	0.84	0.94	0.38	0.51	0.42	47.80	0.549
		man	2.13	1.54	1.87	2.33	0.37	1.01	1.76	87.01	
Politicians	RN50	woman	4.10	76.32	0.30	8.03	5.51	1.98	3.61	0.17	1.008
		man	5.87	75.73	0.64	5.04	0.49	1.65	10.38	0.19	
	SigLIP	woman	0.48	27.76	0.57	41.06	6.03	3.50	0.08	18.08	0.317
		man	1.12	87.70	1.36	4.40	0.43	0.39	0.26	4.34	
Athletes	ViT-B/16	woman	0.01	0.01	74.61	0.02	0.11	25.22	0.03	0.01	1.001
		man	0.82	1.19	74.56	1.55	0.00	16.56	4.84	0.44	
	MetaCLIP	woman	0.19	0.31	75.54	1.43	0.13	2.12	0.18	18.93	0.927
		man	0.37	2.40	81.45	1.81	0.07	0.93	0.95	10.27	

Table 1: Gender bias in occupational classification. Bias = score for correct category. DS-VL2 = DeepSeek-VL2-small; ViT-B/16 = ViT-B/16 with L4M; MetaCLIP = MetaCLIP-b32-400M.

dinate level of categorization. On the other hand, SigLIP is able to categorize images of politicians as `person` but clearly can not recognize the pictures of politicians at more detailed levels of granularity. For completeness, our Appendix shows an evaluation on a total of 39 VLMs across 10 families of VLM models including ALBEF (Li et al., 2021), BLIP (Li et al., 2022), CLIP (Radford et al., 2021), OpenCLIP (Ilharco et al., 2021), SigLIP (Zhai et al., 2023), MetaCLIP (Xu et al., 2023), EVA-CLIP (Sun et al., 2023), Gemma3, Qwen2.5 (Bai et al., 2025), and DeepSeek-VL2 (Wu et al., 2024).

Bias Assessment. We assume that correct predictions for occupations such as `politician` stem from VLM knowledge of individual public figures. However there is also a risk that these associations take place due to biased predictions with respect to demographic variables such as gender. We evaluate this by measuring category score disparities across gender, a variable in our dataset sourced from Wikipedia biographies. Let s_m and s_w denote the average scores for the positive prompt for both images of `men` and `women`, since these are the only two categories present in our data. We define the bias score as the ratio $b = s_w/s_m$, where values significantly larger than 1 would indicate bias toward women, and values significantly less than 1 would indicate a bias toward men.

Table 1 reports gender bias score per dataset, highlighting the most balanced model (ratio ≈ 1.0) and the lowest-scoring one, indicating lower accuracy for women. A ratio greater than 1 would indicate difficulty in recognizing men; however, as most ratios fall below 1, we focus on cases with substantially lower values. Smaller models seem to rely more on stereotypical associations such as (`woman`, `receptionist`) and (`man`, `salesperson`), but are also less accurate overall, but even the best performing models have some disparities. On the actor and actress data, models trained on under 400M samples show lower accuracy, often relying on gendered cues—e.g., assigning `receptionist` to women and `salesperson` to men. Some generative models also show reduced accuracy specifically for actresses. Although they do not explicitly assign stereotypical roles. In the case of politicians, SigLIP models—particularly the smallest one—exhibits a high degree of bias, frequently misclassifying female politicians as teachers. This may reflect the limited exposure of SigLIP to political figures, resulting in biased outputs. In contrast, none of the models exhibit noticeably biased scores on the athlete dataset, but biases are evident in misclassifications—such as labeling men as politicians and women as assistants. On a positive note, our full set of results in the Appendix show that when picking most of the best performing models, bias ratios are the closest to 1 with a few notable exceptions, e.g. Gemma3-27B-it.

Identity Recognition. At the finest level of granularity, our prompts probe VLMs for their ability to associate a specific person's name with their image, rather than with the names of other individuals in their respective group. For instance, we issue prompts of the format "This is a picture of `Bernie Sanders`", and use the names of all other senators as negative prompts. Similar to Winoground (Thrush et al., 2022), we have the model score all images against the prompt using the VLM, and then given an image, we have the VLM score all prompts. We define this as the Text Score, and the Image Score.

	US House of Rep.		Senators		Mayors		Actor/Actress		Athletes	
	Text	Image	Text	Image	Text	Image	Text	Image	Text	Image
ALBEF	0.25	0.25	1.06	1.08	1.12	1.10	1.09	1.12	1.10	1.20
ViT-B/32	27.98	30.69	83.79	85.06	26.42	25.15	88.62	88.92	79.35	83.79
ViT-L/14	42.80	39.92	94.92	93.51	34.23	32.67	96.00	97.71	87.84	88.82
SigLIP-224	6.70	2.66	12.79	10.67	5.08	4.10	58.84	60.11	73.83	75.44
SigLIP-L-384	10.10	4.53	18.38	16.75	6.06	6.86	71.58	70.56	86.43	87.16
ViT-B/32 w/ L2B	33.81	34.13	93.07	93.65	31.25	30.86	94.29	92.33	79.79	81.74
ViT-L/14 w/ L2B	41.02	38.94	95.95	95.36	34.47	36.04	97.66	97.02	87.45	88.87
Qwen2.5-3B	1.00	1.10	7.30	9.40	3.50	3.50	7.60	8.70	37.30	36.60
Qwen2.5-7B	0.70	0.90	7.80	11.50	2.90	3.30	41.30	40.50	60.80	64.60

Table 2: Identity recognition. Can VLMs associate images of a person with their name?

Table 2 shows that larger models tend to rely more heavily on individual-specific information, whereas smaller models such as ALBEF appear to lack substantial knowledge about individual identities, as this model was trained on a smaller dataset. Models trained on at least 400M samples generally show strong recall for all categories, except SigLIP, which performs well only on actors and athletes. Models trained with more than 2B samples show similar performance as models trained with 400M samples, but generative models show higher accuracy as the number of parameters increases, and the values are not significantly higher.

4 Discussion on Data Representation

We analyze the LAION-400M dataset (Schuhmann et al., 2021) which is typically used for training CLIP and other models. We counted the number of captions that overlap with categories or names of people included in the APA Benchmark. The actors consisted of the most frequently mentioned names, without a significant difference in frequency across gender. However, the high rate of classification due to spurious statistical associations and the appearance of bias in the generative VLMs suggest that they may learn from characters played by the actors. Aside from the actors, there was a large gap between mentions of men and women, which suggests that models with larger number of parameters might avoid learning biases through more robust visual representations. The significantly higher accuracy for senators compared to other politician categories in *Identity recognition* experiments is likely due to their over representation in online media. In comparison, the overall low ability to discriminate between politicians of SigLIP suggests that there might not be enough of this data in the training set. If that is the case, it is worth noting that this model may exhibit stronger bias than other models by making biased assignments in data-sparse areas. On the other hand, a different pattern emerges with the athlete set, where the difference in ratios between mean and women is particularly wide compared to other categories. This imbalance suggests that VLMs may be learning to encode features about specific individuals rather than abstract athlete characteristics. As a result, while the accuracy of *Identity recognition* is higher, it also tends to be biased. This reliance on individual-specific information rather than additional categorical cues (e.g., sports equipment such as balls), may explain the higher accuracy for athletes on the subordinate category experiments.

5 Related work

Our work is in the spirit of other benchmark tests that have been designed in the past for Large Language Models. For instance, the WinoBias (Zhao et al., 2018) and WinoGender (Rudinger et al., 2018) benchmarks designed to test models for biases in co-reference resolution. StereoSet (Nadeem et al., 2021) was designed to measure biases across various sensitive protected variables. Honnavalli et al. also proposes a benchmark for language models to probe implicit gender and seniority biases when referring to politicians and academics. In contrast, our benchmark provides a comprehensive VLM evaluation of basic association capabilities, biases and recognition of public figures through image-text association.

6 Conclusion

We introduce the APA benchmark to evaluate VLMs in terms of their people-centric knowledge. Our work uncovers that, while large VLMs generally demonstrate strong classification performance, they also tend to rely more heavily on individual-specific information. Our findings highlight that,

despite recent advancements, such models risk overlooking essential features or reinforcing gender stereotypes, which likely depend on the data they are exposed to during training.

Limitations Our benchmark is not intended to test a wide range of capabilities in vision-language models. Our benchmark is also limited to the English language although there is little that ties its construction and general framework to one language. Additionally, APA Bench should be used together with other benchmarks such as 1) The original downstream benchmarks proposed by (Radford et al., 2021) which evaluate zero-shot learning on standard image classification tasks such as Imagenet-1k, 2) The VL-Checklist benchmark (Zhao et al., 2022) which probes for individual capabilities such as objects, attribute, and relations, and 3) the Winoground benchmark (Thrush et al., 2022) which probes for compositionality. Additionally, favorable performance in our bias assessment tests under our metric does not guarantee that a model is insulated by other biases or that even biases with respect to gender have been mitigated as there are many concurrent factors that can lead to a problematic bias score. The demographics of the people depicted in the benchmark data do not necessarily follow that of a target population in a different context or even necessarily that of the United States which is nevertheless overrepresented as a source of public figures in our dataset. Evaluation in our dataset provides a useful metric during development but responsible deployments should consider a benchmark more targeted toward their individual particular use case.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2024. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning.

Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang. 2022. Towards understanding gender-seniority compound bias in natural language generation. Conference on Learning Resources and Evaluation (LREC).

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating Text-to-Visual Generation with Image-to-text Generation. In *European Conference on Computer Vision (ECCV)*, pages 366–384. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2507–2521.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models

from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL-HLT*, pages 8–14.

Cristoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Proceedings of Neurips Data-Centric AI Workshop*.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2023. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying CLIP Data.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

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

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

A Detailed Dataset Statistics

Image Statistics. In the actor and actress portraits, there are 100 portraits overall, 60 of them male and 40 female. 38 of them have won and been nominated for an Academy Award. 37 of them have been nominated for an Academy Award but have not won. 25 of them have never been nominated for or won an Academy Award In the politician portraits, there are 100 senators (76 of them male and 24 female), and the 436 U.S. House of Representatives (310 of them male and 126 female), and 100 mayors (67 of them male and 33 female). Among them, there are 343 Democratic and 291 Republican politicians, which are annotated in our data but not used in our analysis. The athlete images are generally not portraits but often action shots of the athletes in context. There are 109 images, 79 of them male and 30 female. Each portrait is labeled with a specific sport, and there are 18 football players, 20 tennis players, 14 basketball players, 11 soccer players, 4 racing players, 1 martial arts player, 2 snowboarding players, 1 softball player, 14 baseball players, 2 track athletes, 11 golf players, 6 hockey players, 1 swimmer, 2 boxers, 1 biker, and 1 gymnast.

Prompt Statistics. We create prompts of the form "This is a photo of C" at four levels, where C can make the text-image pair a positive or a negative pair depending on whether the prompt entails the image. Table 3 shows a list of all the categories we consider. The ■ green-level and ■ pink-level prompts are applied to all images, while special prompts are created for athletes, politicians and benchmarks for ■ blue-level prompts. A total of 845 images $\times$ 7 prompts, lead to 5,915 ■ green-level image-prompt pairs, and 845 $\times$ 9 prompts $=$ 7,605 ■ pink-level image-prompt pairs. For actors we consider three ■ blue-level prompts but we also consider the words "actor" or "actress", leading to 109 images $\times$ 6 $=$ 600 image-prompt pairs, for politicians the number is 636 images $\times$ 7 prompts $=$ 4,452 image-prompt pairs and for athletes the number is 109 images $\times$ 14 sports $=$ 1,744 pairs. The total number of image-prompt pairs that will require evaluation across all these tests is 20,316. Additionally, we create prompts to evaluate *Identity recognition* which requires significant computation since we pair every image with the names of every person in the group. As a result, we obtain an additional 100 images $\times$ 100 prompts $=$ 10,000 image-prompt pairs for actors, 100 images $\times$ 100 images $=$ 10,000 image-prompt pairs for US Senators, 436 images $\times$ 436 prompts $=$ 190,096 image-text prompts, 100 images $\times$ 100 prompts $=$ 10,000 image-prompt pairs for mayors, and 109 images $\times$ 109 prompts $=$ 11,881 image-prompt pairs for athletes. Leading to a total of 1,942,777 image-text prompts. Despite the large number of image-text prompt pairs especially for identity recognition, it should be noted that the image-text prompt pairs can be created on-the-fly and taking advantage of

Type	Categories
■ green: category	person dog giraffe plant tree bed chair
■ pink: occupation	politician scientist athlete teacher receptionist assistant salesperson actor actress
■ blue: Actors	won and nominated for Academy Award nominated but not won Academy Award never nominated or won Academy Award
■ blue: Politicians	President Senator Vice President Mayor Governor US House Representative Attorney General
■ blue: Athletes	Football player Tennis player Basketball player Soccer player Racer Martial Artist Snowboarder Softball player Baseball player Track Athlete Golfer Hockey player Swimmer Boxer Biker Gymnast

Table 3: List of prompts used in the experiments, grouped by difficulty and role type. The phrase "This is a photo of" is omitted for brevity.

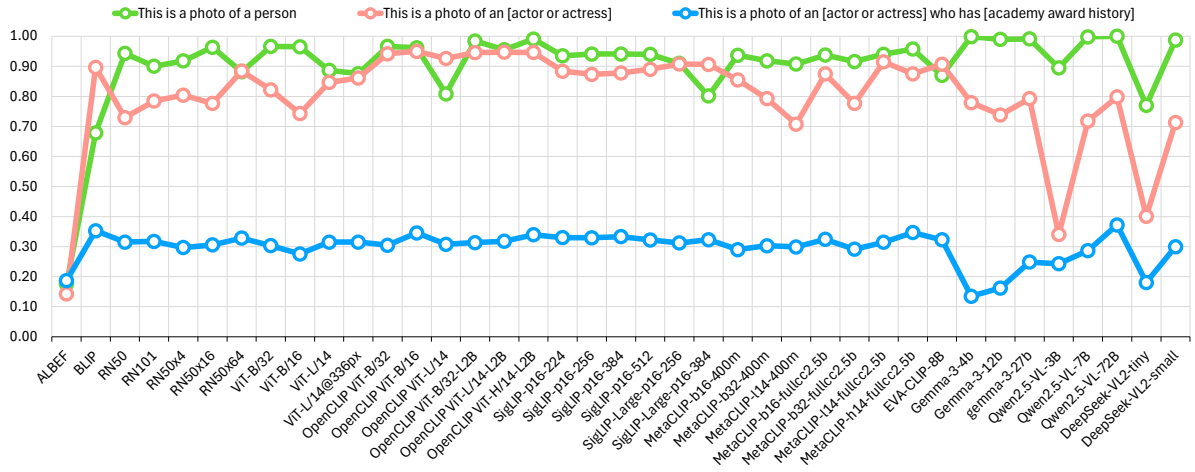

Figure 3: Classification accuracy of all vision-language and generative models toward three levels of prompts by using actor and actress portraits.

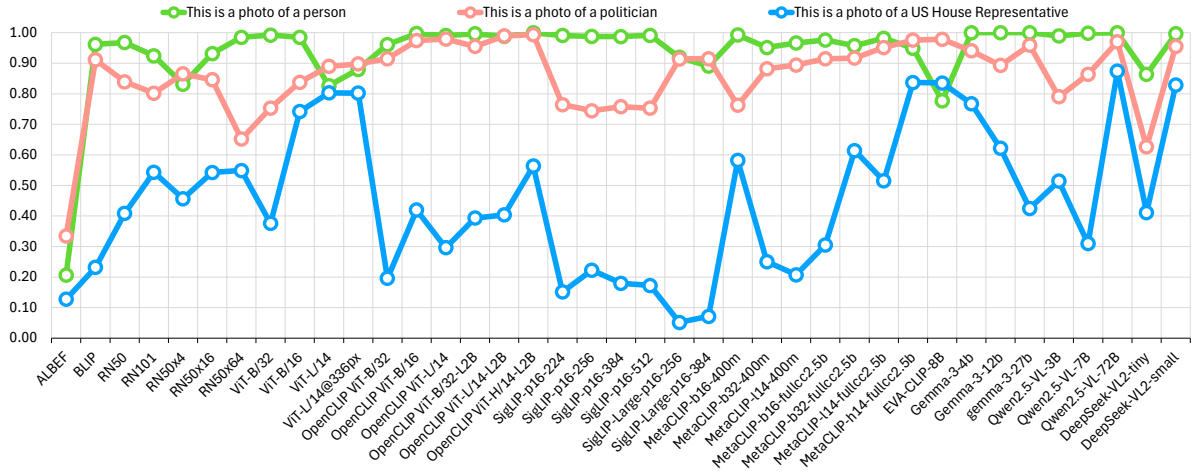

Figure 4: Classification accuracy of all vision-language and generative models toward three levels of prompts by using US House of Rep. portraits.

batch processing in modern VLMs and model parallelism. Table 4 is a Data Card that includes our intended license for release and stated intended use.

B Prompt Format

For the generative vision language models, we employed following form:

```
Indicate if the following
statement is true or false, and
please only output 'true' or
'false': [statement]
```

This format was used to obtain the logit score for "true", and verified classification accuracy by comparing relative to other prompts

C Full set of Results

Our full set of results is presented here for 39 VLMs including powerful variants of recent generative VLMs such as Gemma3 by Google, DeepSeek-VL2, and Qwen2.5-VL. Figures 3 4, 5, 6, and 7 show a battery of tests for basic associations across all VLMs, significantly expanding on our summarized results. Our goal is to provide an easy tool to evaluate future models. Similarly, we present bias assessment for the full roster of VLMs in Tables 5, 6, and 7. Finally Table 8 presents the full set of results for our task of *identity recognition*, showcasing across a wide variety of model sizes and training settings, the subtle nuances across models in their capabilities to memorize how individual public figures look like. Surprisingly here generative VLMs even with their larger capacity fail to associate names to the same extent as the CLIP-based models. Note: In all our tables and figures if only the model backbone is mentioned, the model is the basic CLIP model from OpenAI.

Data Card: APA Benchmark	
Dataset Name	APA Benchmark
Summary	Benchmark to evaluate people-centric capabilities in VLMs. Includes images of public figures (actors, athletes, politicians) and text prompts of the form "This is an image of a {C}".
Supported Tasks	Evaluation only — not for model training.
Languages	English
License	MIT License
Data Content	
Example Instance	`{'image': 'actors/01.jpg', 'prompt': 'This is an image of an actor', 'type': 1}`
Fields	• `image_name`: Path to image • `prompt`: Descriptive text • `type_prompt`: Positive/Negative label
Data Splits	Evaluation only
Curation Rationale	Designed to evaluate recognition of people-related concepts in modern VLMs.
Source and Annotation	
Source	Images and metadata from Wikipedia (Creative Commons or Public Domain).
Annotation Process	Manual annotations derived from Wikipedia content.
Annotators	Dataset creators and Wikipedia editors.
Sensitive Content	Includes information about real people who are public figures with Wikipedia entries written about them.
Use and Limitations	
Social Impact	Helps evaluate people-centric reasoning in VLMs. Should be used alongside other compositional/bias-probing datasets.
Biases	Dataset is not gender-balanced; gender annotations are included. Other demographic skews are likely.
Known Limitations	*[Needs More Information]*
Additional Metadata	
Homepage	*[To be Announced]*
Repository	*[To be Released]*
Paper	*[Not Published]*
Leaderboard	*[Not Public]*
Point of Contact	*[Anonymous Authors]*
Citation	*[This manuscript when published]*

Table 4: Preliminary Data Card for our proposed APA Benchmark

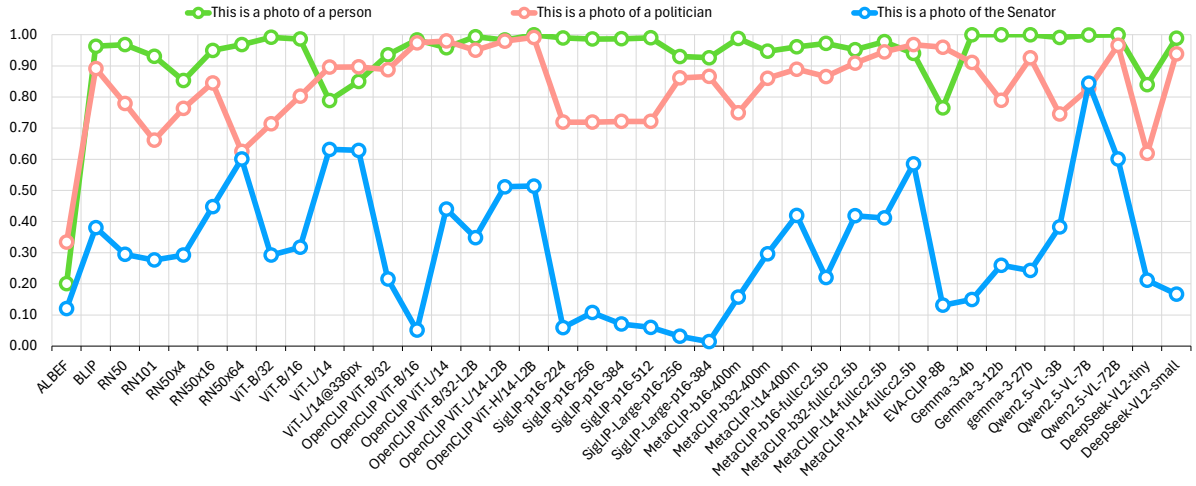

Figure 5: Classification accuracy of all vision-language and generative models toward three levels of prompts by using senator portraits.

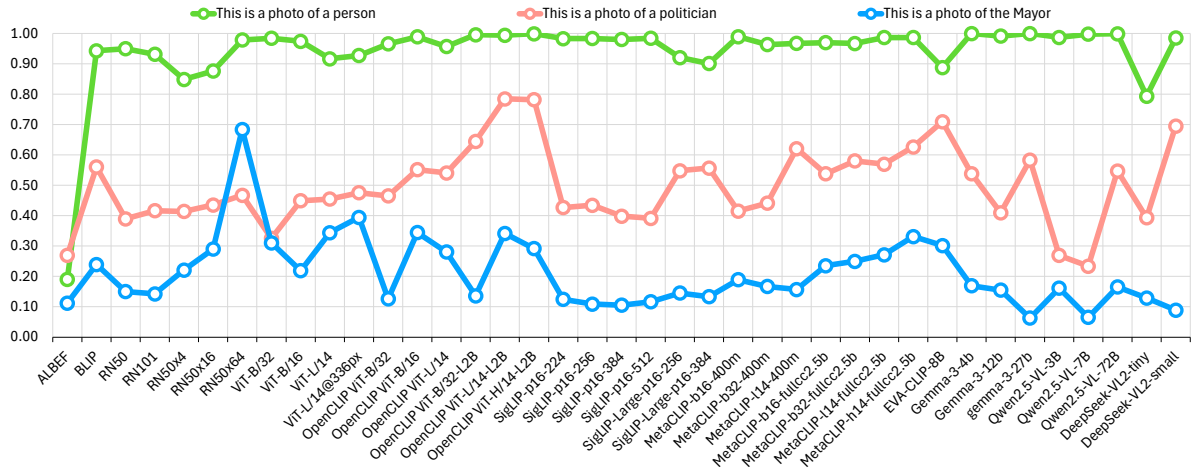

Figure 6: Classification accuracy of all vision-language and generative models toward three levels of prompts by using mayor portraits.

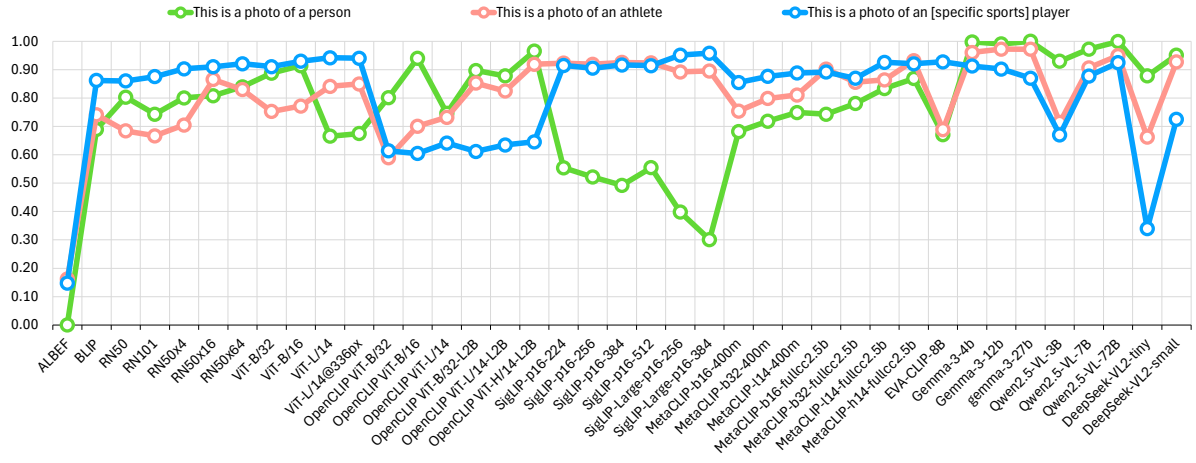

Figure 7: Classification accuracy of all vision-language and generative models toward three levels of prompts by using athlete portraits.

Model	Gender	Classes								Ratio
		Sci.	*Pol.*	*Athl.*	*Teach.*	*Recep.*	*Asst.*	*Sales*	*Actor*	
ALBEF	woman	8.42	13.78	8.97	8.46	11.69	9.85	11.4	15.47	1.147
	man	9.03	18.21	8.1	7.9	11.38	9.85	11.38	13.49	
BLIP	woman	0.02	0.3	0.35	2.01	0.28	0.41	0.4	88.67	0.981
	man	0.15	4.39	1.16	0.66	0.06	0.23	0.82	90.38	
RN50	woman	7.35	9.81	6.99	27.44	25.66	7.78	8.52	79.39	1.156
	man	16.7	11.89	6.7	24.55	6.32	7.6	19.36	68.7	
RN101	woman	11.03	9.44	7.53	12.78	40.43	5.66	3.41	78.61	1.004
	man	18.8	17.24	7.43	13.51	3.72	7.92	18.86	78.32	
RN50x4	woman	19.31	12.23	5.86	17.04	21.04	5.67	12.18	82.67	1.047
	man	23.82	19.49	5.45	16.41	1.92	8.79	15.89	78.96	
RN50x16	woman	12.71	11.06	2.58	26.9	31.96	1.17	4.35	65.97	0.772
	man	21.96	12.44	6.29	23.68	8.42	2.78	14.49	85.4	
RN50x64	woman	19.09	6.1	3.73	5.91	11.35	28.15	8.72	82.03	0.885
	man	25.54	4.4	7.93	10.14	1.8	17.79	15.61	92.72	
ViT-B/32	woman	10.77	7.4	8.22	14.28	17.35	14.22	2.39	83.06	1.017
	man	19.85	5.91	6.23	9.41	1.85	16.41	10.17	81.64	
ViT-B/16	woman	14.17	5.7	4.04	12.36	31.93	14.75	6.2	65.28	0.813
	man	20.26	16.56	3.93	18.68	1.74	10.45	14.34	80.32	
ViT-L/14	woman	4.86	1.04	1.89	4.27	28.86	24.94	12.03	80.91	0.928
	man	8.29	1.97	1.86	6.83	6.28	24.71	26.27	87.21	
ViT-L/14@336px	woman	3.94	1.08	1.52	3.94	32.45	24.56	11.56	83.4	0.948
	man	8.07	2.34	1.98	5.99	8.73	23.44	25.59	87.94	
OpenCLIP ViT-B/32	woman	2.15	2.66	2.43	4.44	18.07	65.28	3.02	93.95	0.995
	man	6.29	40.72	6.35	5.11	0.77	22.16	10.06	94.38	
OpenCLIP ViT-B/16	woman	0.55	3.82	11.68	2.94	2.23	67.97	6.48	92.92	0.965
	man	2.03	28.2	10.32	3.3	0.23	26.27	19.29	96.34	
OpenCLIP ViT-L/14	woman	3.13	12.16	3.27	1.88	30.88	42.77	3.44	93.65	1.019
	man	4.35	18.49	4.41	6.19	0.71	39.75	16.0	91.89	
OpenCLIP ViT-B/32-L2B	woman	6.87	7.72	11.53	6.67	0.87	35.99	2.86	94.48	0.996
	man	4.81	28.83	10.4	3.78	0.02	13.23	4.34	94.87	
OpenCLIP ViT-L/14-L2B	woman	7.06	5.69	11.79	2.62	18.2	6.42	2.34	97.17	1.046
	man	17.81	20.98	7.91	1.75	0.93	2.29	4.25	92.92	
OpenCLIP ViT-H/14-L2B	woman	0.73	0.62	12.62	2.71	1.61	1.99	0.42	94.34	0.996
	man	2.52	1.56	10.58	0.69	0.13	0.87	2.74	94.73	
SigLIP-p16-224	woman	0.08	0.01	0.24	1.18	0.17	0.24	0.02	82.86	0.9
	man	0.9	1.66	1.52	1.78	0.08	0.26	0.08	92.04	
SigLIP-p16-256	woman	0.04	0.02	0.23	1.69	0.23	0.27	0.02	84.28	0.942
	man	1.0	3.15	1.73	2.56	0.07	0.33	0.15	89.45	
SigLIP-p16-384	woman	0.21	0.01	0.3	2.3	0.57	0.47	0.05	83.15	0.915
	man	0.91	1.99	1.63	2.64	0.08	0.3	0.09	90.92	
SigLIP-p16-512	woman	0.15	0.01	0.24	1.3	0.33	0.44	0.04	84.28	0.915
	man	1.07	1.56	1.28	1.92	0.07	0.25	0.06	92.14	
SigLIP-Large-p16-256	woman	0.05	0.17	0.03	1.02	0.15	2.06	0.05	84.28	0.886
	man	0.29	0.82	0.19	0.62	0.03	1.42	0.04	95.12	
SigLIP-Large-p16-384	woman	0.09	0.16	0.09	0.68	0.25	3.01	0.11	85.11	0.903
	man	0.37	0.73	0.35	0.89	0.03	1.63	0.02	94.29	
MetaCLIP-b16-400m	woman	0.36	0.23	0.34	0.82	0.51	0.44	0.63	75.0	0.811
	man	0.69	2.44	0.19	0.37	0.01	0.09	1.27	92.53	
MetaCLIP-b32-400m	woman	0.16	0.35	0.5	0.82	0.4	1.72	0.31	89.06	1.224
	man	0.65	3.91	2.28	1.75	0.19	2.66	1.85	72.75	
MetaCLIP-l14-400m	woman	1.05	1.94	4.61	4.19	1.31	2.18	0.44	73.63	1.07
	man	2.28	9.73	4.3	5.04	0.61	2.07	2.6	68.85	
MetaCLIP-b16-fullcc2.5b	woman	0.07	0.2	0.13	0.64	0.2	0.55	0.15	83.5	0.927
	man	0.17	2.57	0.35	0.7	0.07	0.86	0.42	90.09	
MetaCLIP-b32-fullcc2.5b	woman	0.3	0.43	0.11	4.48	0.07	1.1	0.05	79.83	1.048
	man	1.03	2.24	0.31	13.0	0.06	1.91	0.3	76.17	
MetaCLIP-l14-fullcc2.5b	woman	0.04	0.06	0.11	0.38	0.12	0.32	0.13	88.18	0.941
	man	0.04	0.37	0.14	0.58	0.03	0.92	0.1	93.7	
MetaCLIP-h14-fullcc2.5b	woman	0.03	0.04	0.15	0.53	0.07	0.26	0.14	84.86	0.951
	man	0.2	0.33	0.4	1.13	0.05	0.45	0.4	89.26	
EVA-CLIP-8B	woman	0.17	0.1	0.1	0.49	0.06	5.32	0.15	85.5	0.908
	man	0.38	0.75	0.15	0.55	0.04	2.6	0.31	94.14	
Gemma-3-4b-it	woman	0.04	0.35	0.03	0.44	0.0	0.0	2.24	71.53	0.87
	man	4.42	5.35	2.08	0.31	0.01	0.0	4.05	82.23	
Gemma-3-12b-it	woman	0.09	2.63	0.0	0.13	0.0	0.0	0.95	63.57	0.788
	man	8.88	2.68	6.62	0.0	0.0	0.0	0.03	80.66	
gemma-3-27b-it	woman	0.0	2.49	2.48	0.0	0.0	0.0	0.0	62.26	0.687
	man	0.0	2.8	5.0	0.0	0.0	0.0	0.02	90.58	
Qwen2.5-VL-3B-Instruct	woman	5.66	5.05	5.97	6.48	6.05	15.17	8.36	26.93	0.695
	man	6.14	6.05	7.2	7.73	4.8	16.5	8.83	38.75	
Qwen2.5-VL-7B-Instruct	woman	1.92	2.85	2.33	1.69	3.05	2.37	3.22	58.25	0.72
	man	2.4	1.81	2.42	1.84	3.27	2.26	3.2	80.86	
Qwen2.5-VL-72B-Instruct	woman	2.02	1.22	2.69	3.5	3.86	2.7	3.42	69.19	0.796
	man	1.57	1.12	1.5	2.01	1.61	1.41	1.5	86.96	
DeepSeek-VL2-tiny	woman	2.31	3.19	5.15	3.8	10.09	2.26	5.47	39.45	0.973
	man	4.58	6.67	7.2	4.62	8.83	3.05	9.02	40.55	
DeepSeek-VL2-small	woman	0.79	0.36	0.84	0.94	0.38	0.51	0.42	47.8	0.549
	man	2.13	1.54	1.87	2.33	0.37	1.01	1.76	87.01	

Table 5: Results by using actor and actress dataset for all models that showcase disparities in the association of different occupations with people of different genders.

Model	Gender	Classes								Ratio
		Sci.	*Pol.*	*Athl.*	*Teach.*	*Recep.*	*Asst.*	*Sales*	*Actor*	
ALBEF	woman	7.16	28.05	7.94	8.39	11.22	7.79	10.13	10.35	0.82
	man	6.62	34.23	7.93	7.24	11.1	7.07	10.66	8.91	
BLIP	woman	0.17	65.19	0.78	4.59	8.49	1.85	2.97	15.04	0.697
	man	0.12	93.46	0.67	0.43	0.46	0.16	1.65	3.07	
RN50	woman	4.1	76.32	0.3	8.03	5.51	1.98	3.61	0.17	1.008
	man	5.87	75.73	0.64	5.04	0.49	1.65	10.38	0.19	
RN101	woman	1.97	64.65	0.16	3.68	24.27	1.17	4.02	0.07	0.864
	man	2.17	74.85	0.36	0.79	0.61	0.77	20.39	0.06	
RN50x4	woman	8.42	70.07	0.08	5.58	7.51	0.78	7.52	0.08	0.866
	man	5.24	80.96	0.15	1.9	0.21	0.75	10.74	0.06	
RN50x16	woman	3.86	71.19	0.18	3.31	17.37	0.3	3.56	0.21	0.879
	man	4.56	80.96	0.91	2.41	1.55	0.36	9.11	0.14	
RN50x64	woman	1.88	68.75	0.56	0.59	3.59	3.18	21.36	0.1	1.164
	man	1.75	59.08	1.64	1.06	0.18	1.79	34.33	0.17	
ViT-B/32	woman	6.87	67.04	0.37	13.73	2.48	5.91	2.93	0.68	0.98
	man	9.88	68.41	0.58	1.92	0.18	5.51	12.87	0.65	
ViT-B/16	woman	5.89	67.97	0.15	4.17	13.62	1.59	6.28	0.32	0.841
	man	4.63	80.81	0.34	1.9	0.3	0.82	10.97	0.22	
ViT-L/14	woman	0.61	81.84	0.02	1.49	4.23	5.7	6.06	0.07	0.993
	man	0.81	82.42	0.2	0.84	0.73	5.13	9.81	0.07	
ViT-L/14@336px	woman	0.74	81.35	0.03	1.48	4.23	5.96	6.11	0.12	0.97
	man	1.11	83.84	0.21	0.91	0.77	4.6	8.45	0.1	
OpenCLIP ViT-B/32	woman	0.17	65.58	0.0	4.34	7.34	19.69	2.86	0.02	0.717
	man	0.52	91.46	0.11	0.37	0.07	1.93	5.53	0.02	
OpenCLIP ViT-B/16	woman	0.31	84.13	0.09	1.22	1.79	5.1	7.32	0.03	0.9
	man	0.17	93.46	0.27	0.22	0.0	0.66	5.24	0.01	
OpenCLIP ViT-L/14	woman	0.47	87.89	0.0	0.56	2.23	6.79	2.03	0.02	0.952
	man	0.59	92.33	0.24	0.45	0.02	2.36	3.99	0.01	
OpenCLIP ViT-B/32-L2B	woman	1.95	81.25	0.1	4.48	0.53	6.87	4.73	0.08	0.861
	man	1.22	94.34	0.53	0.39	0.0	0.95	2.46	0.14	
OpenCLIP ViT-L/14-L2B	woman	1.43	93.12	0.04	1.99	1.06	0.1	1.31	0.97	0.965
	man	0.82	96.53	0.31	0.2	0.0	0.09	1.94	0.09	
OpenCLIP ViT-H/14-L2B	woman	0.08	95.17	0.2	0.92	0.13	0.23	2.12	1.17	0.988
	man	0.13	96.34	1.35	0.18	0.0	0.05	0.88	1.06	
SigLIP-p16-224	woman	0.48	27.76	0.57	41.06	6.03	3.5	0.08	18.08	0.317
	man	1.12	87.7	1.36	4.4	0.43	0.39	0.26	4.34	
SigLIP-p16-256	woman	0.46	20.11	0.62	52.54	4.82	2.93	0.08	16.7	0.226
	man	1.04	89.01	1.28	4.86	0.29	0.26	0.36	2.89	
SigLIP-p16-384	woman	0.71	23.95	0.55	41.06	5.69	3.1	0.05	22.05	0.272
	man	1.4	87.99	1.47	3.95	0.4	0.37	0.24	4.17	
SigLIP-p16-512	woman	0.57	23.86	0.53	41.94	4.28	3.17	0.05	22.69	0.273
	man	1.37	87.4	1.5	4.47	0.41	0.36	0.27	4.21	
SigLIP-Large-p16-256	woman	0.51	65.87	0.27	6.92	0.88	10.42	0.05	13.35	0.712
	man	0.74	92.48	0.89	0.67	0.17	1.37	0.16	3.52	
SigLIP-Large-p16-384	woman	0.33	67.19	0.2	5.33	1.06	9.29	0.03	14.51	0.728
	man	0.63	92.33	1.13	0.53	0.2	0.99	0.09	4.07	
MetaCLIP-b16-400m	woman	6.8	48.27	0.5	13.56	5.24	6.81	4.42	8.79	0.606
	man	5.64	79.59	1.02	2.5	0.05	0.98	4.14	6.03	
MetaCLIP-b32-400m	woman	3.1	66.46	2.23	4.69	2.44	12.48	2.28	5.19	0.766
	man	1.99	86.77	1.84	1.09	0.18	3.27	3.64	1.17	
MetaCLIP-l14-400m	woman	2.71	76.95	0.86	8.16	0.92	2.15	0.95	5.97	0.872
	man	2.07	88.28	1.69	3.43	0.11	0.84	1.14	2.35	
MetaCLIP-b16-fullcc2.5b	woman	2.07	70.61	0.7	3.77	4.33	4.09	5.08	6.86	0.78
	man	1.13	90.48	0.77	0.52	0.18	1.03	2.95	2.86	
MetaCLIP-b32-fullcc2.5b	woman	1.13	86.18	0.18	7.75	0.41	1.8	0.41	1.62	0.998
	man	1.74	86.33	0.84	8.3	0.03	1.16	0.74	0.87	
MetaCLIP-l14-fullcc2.5b	woman	1.05	81.15	0.38	2.31	1.84	8.83	1.06	2.86	0.88
	man	1.01	92.24	0.64	0.83	0.11	3.05	1.17	0.9	
MetaCLIP-h14-fullcc2.5b	woman	0.99	86.52	0.33	6.13	0.78	2.65	1.0	1.11	0.919
	man	0.71	94.14	0.69	1.83	0.06	0.91	1.03	0.62	
EVA-CLIP-8B	woman	0.8	89.55	0.14	2.24	0.87	0.77	1.75	3.44	0.944
	man	0.61	94.87	0.4	0.97	0.11	0.4	1.26	1.35	
Gemma-3-4b-it	woman	4.86	81.49	0.04	8.76	0.0	0.0	4.69	0.05	0.909
	man	3.84	89.65	0.42	2.34	0.0	0.2	3.33	0.22	
Gemma-3-12b-it	woman	10.42	73.29	0.82	1.01	0.0	0.0	7.09	2.8	0.885
	man	5.49	82.81	3.95	0.38	0.0	0.01	5.98	1.39	
gemma-3-27b-it	woman	4.26	87.01	0.0	2.08	0.0	0.0	3.69	1.11	0.962
	man	0.94	90.48	1.03	0.81	0.0	0.03	5.6	1.13	
Qwen2.5-VL-3B-Instruct	woman	2.47	63.13	1.2	3.11	2.12	19.43	3.09	2.56	0.865
	man	1.63	73.0	1.41	2.45	0.97	15.0	2.73	2.06	
Qwen2.5-VL-7B-Instruct	woman	0.05	71.58	0.05	0.04	0.07	25.93	0.09	0.6	0.922
	man	0.05	77.64	0.66	0.06	0.06	20.74	0.15	0.46	
Qwen2.5-VL-72B-Instruct	woman	0.57	85.35	0.2	5.53	0.9	1.83	1.49	2.27	0.924
	man	0.3	92.38	0.63	2.02	0.17	0.68	2.58	1.01	
DeepSeek-VL2-tiny	woman	2.5	49.73	3.52	5.1	9.5	2.49	5.87	11.09	0.795
	man	2.73	62.55	4.43	3.74	4.65	1.77	5.18	10.96	
DeepSeek-VL2-small	woman	1.47	87.55	0.57	2.16	1.4	1.02	1.74	2.36	0.945
	man	1.08	92.68	1.13	0.97	0.39	0.78	1.62	1.11	

Table 6: Results by using politicians dataset for all models that showcase disparities in the association of different occupations with people of different genders.

Model	Gender	Classes								Ratio
		Sci.	*Pol.*	*Athl.*	*Teach.*	*Recep.*	*Asst.*	*Sales*	*Actor*	
ALBEF	woman	9.03	13.12	17.79	9.2	10.28	11.19	10.46	10.53	1.117
	man	9.19	16.15	15.92	8.49	10.53	11.02	10.51	9.53	
BLIP	woman	0.13	1.08	75.24	0.2	0.05	0.6	0.23	21.45	1.03
	man	0.85	5.54	73.05	0.43	0.03	0.29	0.8	17.03	
RN50	woman	10.22	4.03	69.73	2.88	2.15	7.7	2.13	1.17	1.029
	man	6.84	5.92	67.77	4.64	0.55	4.36	8.13	1.78	
RN101	woman	5.37	4.64	76.86	1.33	1.46	8.28	1.46	0.62	1.183
	man	5.62	8.03	64.94	2.01	0.64	6.31	11.4	1.05	
RN50x4	woman	10.96	5.56	71.04	2.01	1.88	2.29	3.02	3.25	1.052
	man	9.2	8.86	67.53	1.8	0.51	1.38	7.03	3.7	
RN50x16	woman	2.47	2.07	88.72	2.76	1.7	0.86	0.38	1.05	1.014
	man	1.73	2.15	87.5	1.96	0.56	0.42	4.92	0.79	
RN50x64	woman	1.55	2.61	91.46	0.49	0.14	2.46	0.58	0.73	1.05
	man	4.92	2.21	87.06	1.04	0.27	0.73	2.67	1.08	
ViT-B/32	woman	2.02	0.73	82.13	1.34	1.2	9.51	0.69	2.38	1.133
	man	4.08	2.56	72.51	1.91	0.35	7.41	6.2	4.95	
ViT-B/16	woman	2.21	2.36	84.18	1.5	4.32	2.41	1.27	1.77	1.071
	man	1.93	3.54	78.56	4.01	0.3	1.54	9.07	1.02	
ViT-L/14	woman	0.1	1.11	90.58	0.38	1.41	4.75	1.11	0.58	1.088
	man	0.65	2.58	83.25	1.22	1.19	4.84	5.25	1.03	
ViT-L/14@336px	woman	0.11	0.82	92.68	0.3	0.8	3.79	0.85	0.65	1.077
	man	0.67	2.27	86.04	1.02	0.88	4.31	3.49	1.32	
OpenCLIP ViT-B/32	woman	0.03	0.31	59.38	0.18	0.19	39.67	0.21	0.03	0.931
	man	0.76	4.8	63.77	1.9	0.4	24.82	2.56	0.98	
OpenCLIP ViT-B/16	woman	0.01	0.01	74.61	0.02	0.11	25.22	0.03	0.01	1.001
	man	0.82	1.19	74.56	1.55	0.0	16.56	4.84	0.44	
OpenCLIP ViT-L/14	woman	0.01	0.06	87.16	0.06	0.22	12.02	0.12	0.35	1.118
	man	0.14	1.2	77.93	0.86	0.05	11.05	6.59	2.17	
OpenCLIP ViT-B/32-L2B	woman	0.05	0.07	88.53	0.08	0.01	10.83	0.2	0.26	1.044
	man	0.52	0.91	84.81	0.1	0.01	9.36	2.69	1.58	
OpenCLIP ViT-L/14-L2B	woman	0.14	0.04	96.78	0.13	1.28	1.12	0.22	0.26	1.198
	man	0.75	1.91	80.76	0.72	0.02	3.59	10.46	1.78	
OpenCLIP ViT-H/14-L2B	woman	0.02	0.07	96.09	0.02	0.05	2.38	0.06	1.32	1.059
	man	0.11	0.3	90.77	0.35	0.0	1.16	1.49	5.82	
SigLIP-p16-224	woman	0.05	0.02	93.65	0.06	1.14	2.34	0.1	2.49	1.021
	man	0.18	0.53	91.7	0.37	0.03	0.24	0.07	6.73	
SigLIP-p16-256	woman	0.06	0.02	92.48	0.12	0.81	2.69	0.04	3.68	1.009
	man	0.2	1.01	91.65	0.43	0.02	0.28	0.11	6.16	
SigLIP-p16-384	woman	0.05	0.01	93.95	0.06	0.76	1.87	0.04	3.14	1.024
	man	0.21	0.54	91.75	0.2	0.02	0.17	0.05	6.9	
SigLIP-p16-512	woman	0.05	0.01	93.12	0.05	0.82	1.57	0.02	4.16	1.012
	man	0.18	0.91	91.99	0.19	0.02	0.12	0.04	6.39	
SigLIP-Large-p16-256	woman	0.04	0.05	88.82	0.02	0.04	2.22	0.01	8.64	0.987
	man	0.23	0.31	89.99	0.1	0.04	0.89	0.06	8.26	
SigLIP-Large-p16-384	woman	0.02	0.02	86.57	0.01	0.05	2.25	0.01	10.99	0.953
	man	0.18	0.69	90.82	0.03	0.02	0.55	0.03	7.61	
MetaCLIP-b16-400m	woman	0.15	0.18	77.25	0.28	0.78	3.45	0.15	14.89	1.035
	man	0.92	4.1	74.66	1.37	0.02	0.67	1.58	16.49	
MetaCLIP-b32-400m	woman	0.22	0.05	93.75	0.2	0.14	2.86	0.15	2.46	1.055
	man	0.43	1.28	88.87	0.52	0.1	2.34	2.69	3.13	
MetaCLIP-l14-400m	woman	0.19	0.31	75.54	1.43	0.13	2.12	0.18	18.93	0.927
	man	0.37	2.4	81.45	1.81	0.07	0.93	0.95	10.27	
MetaCLIP-b16-fullcc2.5b	woman	0.24	0.14	87.45	0.44	0.4	2.21	0.38	8.0	1.031
	man	0.54	2.06	84.81	0.48	0.06	1.85	1.73	7.87	
MetaCLIP-b32-fullcc2.5b	woman	0.15	0.12	86.47	2.55	0.02	1.96	0.16	7.64	1.095
	man	1.13	0.92	79.0	11.17	0.06	1.84	1.82	3.89	
MetaCLIP-l14-fullcc2.5b	woman	0.06	0.07	88.87	0.11	0.11	3.05	0.07	6.71	1.039
	man	0.22	1.11	85.5	0.47	0.03	2.02	0.88	9.46	
MetaCLIP-h14-fullcc2.5b	woman	0.12	0.1	95.95	0.13	0.1	0.68	0.11	2.19	1.041
	man	0.22	1.05	92.14	0.56	0.04	0.38	0.77	4.6	
EVA-CLIP-8B	woman	0.45	0.41	81.05	0.48	0.04	3.61	0.32	13.01	1.03
	man	1.18	1.93	78.71	0.98	0.03	2.54	2.12	12.3	
Gemma-3-4b-it	woman	0.01	0.03	96.97	0.01	0.01	0.01	0.01	0.36	1.014
	man	0.12	0.27	95.65	0.02	0.06	0.01	1.89	1.94	
Gemma-3-12b-it	woman	0.0	0.0	96.68	0.0	0.0	0.0	0.0	3.33	0.991
	man	0.0	0.0	97.51	0.0	0.0	0.0	0.0	1.36	
gemma-3-27b-it	woman	0.0	0.0	96.92	0.0	0.0	0.0	0.0	0.09	0.995
	man	0.0	0.0	97.36	0.0	0.0	0.0	0.0	2.64	
Qwen2.5-VL-3B-Instruct	woman	0.73	0.8	85.16	1.26	0.88	6.26	1.37	1.77	1.033
	man	0.86	1.15	82.42	1.53	0.97	6.79	1.99	3.23	
Qwen2.5-VL-7B-Instruct	woman	0.01	0.0	99.02	0.01	0.01	0.81	0.01	0.09	1.044
	man	0.01	0.01	94.82	0.01	0.01	2.86	0.01	2.16	
Qwen2.5-VL-72B-Instruct	woman	0.03	0.05	99.37	0.05	0.06	0.05	0.16	0.15	1.04
	man	0.2	0.22	95.56	0.29	0.36	0.32	0.61	1.91	
DeepSeek-VL2-tiny	woman	0.71	1.92	68.75	1.47	3.89	1.68	3.81	7.96	1.055
	man	0.73	3.52	65.19	1.84	4.33	2.65	4.17	15.22	
DeepSeek-VL2-small	woman	0.1	0.1	98.29	0.06	0.03	0.14	0.07	0.72	1.031
	man	0.09	0.32	95.36	0.11	0.03	0.36	0.13	3.42	

Table 7: Results by using athletes dataset for all models that showcase disparities in the association of different occupations with people of different genders.

	US House of Rep.		Senators		Mayors		Actor/Actress		Athletes	
	Text	Image	Text	Image	Text	Image	Text	Image	Text	Image
ALBEF	0.25	0.25	1.06	1.08	1.12	1.10	1.09	1.12	1.10	1.20
BLIP	1.81	2.27	12.82	13.64	6.31	6.56	38.26	42.70	31.96	42.31
RN50	31.86	33.03	86.13	85.21	28.74	28.15	88.18	89.31	78.96	82.37
RN101	30.03	31.59	84.18	84.96	27.93	26.83	90.82	90.97	81.74	83.64
RN50x4	38.87	37.26	89.06	89.94	32.52	32.54	95.46	94.82	86.67	87.01
RN50x16	47.02	44.07	91.06	88.23	36.18	36.50	95.95	95.95	87.84	88.62
RN50x64	48.49	44.19	93.12	92.72	39.67	37.21	96.53	97.80	88.77	89.75
ViT-B/32	27.98	30.69	83.79	85.06	26.42	25.15	88.62	88.92	79.35	83.79
ViT-B/16	29.98	30.54	82.32	82.08	26.39	25.39	94.48	93.07	81.05	85.16
ViT-L/14	42.80	39.92	94.92	93.51	34.23	32.67	96.00	97.71	87.84	88.82
ViT-L/14@336px	41.11	39.60	94.73	93.07	35.13	33.79	96.39	97.51	88.48	88.92
OpenCLIP ViT-B/32	27.34	26.90	89.16	84.33	26.25	26.64	90.72	90.48	74.80	77.93
OpenCLIP ViT-B/16	29.98	28.96	88.72	86.13	29.64	24.45	92.72	93.80	79.49	83.30
OpenCLIP ViT-L/14	38.84	36.23	93.99	91.60	34.45	33.06	96.73	96.53	85.74	86.04
OpenCLIP ViT-B/32-L2B	33.81	34.13	93.07	93.65	31.25	30.86	94.29	92.33	79.79	81.74
OpenCLIP ViT-L/14-L2B	41.02	38.94	95.95	95.36	34.47	36.04	97.66	97.02	87.45	88.87
OpenCLIP ViT-H/14-L2B	51.66	50.15	98.68	99.17	41.50	38.13	98.24	98.14	89.26	90.48
SigLIP-p16-224	6.70	2.66	12.79	10.67	5.08	4.10	58.84	60.11	73.83	75.44
SigLIP-p16-256	6.71	2.38	11.66	9.14	5.59	4.41	58.30	57.81	71.19	74.12
SigLIP-p16-384	7.05	2.74	12.29	9.81	5.28	4.79	61.72	61.67	74.27	77.20
SigLIP-p16-512	7.14	2.76	12.31	9.42	5.40	4.57	64.65	64.31	73.29	79.79
SigLIP-Large-p16-256	9.45	3.77	15.23	14.55	5.95	6.26	67.24	67.19	82.47	83.50
SigLIP-Large-p16-384	10.10	4.53	18.38	16.75	6.06	6.86	71.58	70.56	86.43	87.16
MetaCLIP-b16-400m	37.30	35.30	76.95	75.34	22.12	23.72	84.18	85.16	81.25	80.18
MetaCLIP-b32-400m	26.51	30.22	75.00	75.68	22.89	22.58	75.78	78.22	73.83	75.98
MetaCLIP-l14-400m	42.55	45.12	89.45	90.04	28.93	30.54	94.58	94.58	84.03	87.99
MetaCLIP-b16-fullcc2.5b	38.21	38.87	87.94	88.13	28.20	29.44	95.17	93.65	84.18	84.18
MetaCLIP-b32-fullcc2.5b	29.08	33.62	82.28	84.33	22.42	23.69	83.50	81.79	76.12	77.69
MetaCLIP-l14-fullcc2.5b	48.29	46.14	94.29	95.31	35.06	35.45	97.85	96.63	90.43	91.89
MetaCLIP-h14-fullcc2.5b	66.50	60.99	98.44	97.75	39.50	40.31	96.68	98.19	91.65	92.24
EVA-CLIP-8B	43.31	42.97	95.12	96.88	34.20	34.35	98.88	98.68	88.38	89.40
Gemma-3-4b-it	0.52	1.27	4.10	4.70	1.30	1.50	12.90	29.70	49.20	46.90
Gemma-3-12b-it	1.27	1.23	6.80	10.80	0.90	0.60	24.80	35.20	61.30	58.80
Gemma-3-27b-it	2.17	1.72	16.39	15.88	2.55	3.28	35.67	49.56	71.48	76.17
Qwen2.5-VL-3B-Instruct	1.00	1.10	7.30	9.40	3.50	3.50	7.60	8.70	37.30	36.60
Qwen2.5-VL-7B-Instruct	0.70	0.90	7.80	11.50	2.90	3.30	41.30	40.50	60.80	64.60
Qwen2.5-VL-72B-Instruct	7.79	7.43	45.80	45.65	10.42	7.62	63.38	61.72	70.51	70.46
DeepSeek-VL2-tiny	0.40	0.40	1.40	1.70	1.80	1.90	2.80	3.50	3.40	3.80
DeepSeek-VL2-small	0.70	0.80	4.00	4.50	3.40	3.30	10.60	14.80	36.00	37.50

Table 8: Results of *identity recognition* experiment for all models.