# The P-FI Benchmark: Probing Vision-Language Models for Societal Bias and Retention across Domains

**Veronica Flores (vflores@scu.edu)**


## Abstract

Vision-language models such as CLIP that are able to compute matching scores between images and text have become extremely capable. We propose the Public Figure Benchmark (P-FI) to probe the capabilities of the associations computed by these models, the societal biases induced by these associations, and the capacity of these models to retain knowledge. The P-FI Benchmark consists of public domain portraits of significant politicians, athletes, and actors in the United States including elected officials from the Senate, the House of Representatives, and mayors of the most populated cities. We discuss some of the implications of our results and discover the role of scale in each of the properties targeted in our study. Similar to the pure textual domain, there are capabilities in vision-language models that seem to emerge only in the largest models. As more variants of vision-language models are trained on publicly available data, we expect that our benchmark will be an easy test to replicate. Our code and data are included with this submission and will be released upon publication.


## 1 Introduction

There has been a lot of progress in recent years in training large-scale models that can reason about images and text. Particularly, general-purpose models trained to learn associations between images and text have become incredibly powerful. One prominent example is the CLIP model by OpenAI (Radford et al., 2021). Although this model is trained on a web dataset of images with freely associated text, it can be used at evaluation time with prompts of the type: *This is a photo of [$X$]* in order to work as a zero-shot classifier for class $X$, rivaling performance with current models trained on the challenging Imagenet-1k classification task (Russakovsky et al., 2015). However, these models are trained with an open vocabulary and have been exposed to a much larger number of object categories, concepts and image types than what Imagenet can capture. In this work, we propose a complementary benchmark that aims to explore human-centric capabilities, biases and retention capacity of these models through a database of portrait photographs of political figures.

Our benchmark aims to first test the basic capabilities of these models to associate the basic level category `person` with these images, compared to subordinate categories such as `politician`, `athlete`, `actor` and more specific subordinate categories e.g. `Senator`, `(sport) player`, `leading actor`. Large language models (LLMs) have been shown to have some emergent properties that only are exhibited at certain scale (Wei et al., 2022). Similarly, for vision-language models, while most models can easily predict the first two types of categories, we show that a combination of training data size and model size seems to be required for them to recall more specific subordinate categories. Our experiments also combine demographic information to estimate the amount of societal bias with respect to the gender of the people in these pictures, and the disparities that different models make with respect to occupations. Finally, we estimate the capacity for these models to retain knowledge of specific people in our benchmark by prompting them to recognize the names of the individual political figures depicted in each picture. We define this capacity to recall specific facts as *retention*.

The proposed P-FI benchmark consists of 845 high-quality portrait pictures of several groups of Public Figures which include politicians, athletes, and actors. These figures also include basic demographic information depending on the Public Figure such as gender, political affiliation, type of athlete, movie role, and district, state, or city. This benchmark includes 636 politicians, 109 actors, and 100 athletes. The politicians in these pictures correspond to public figures in the United States who were elected members of the House of Repre-

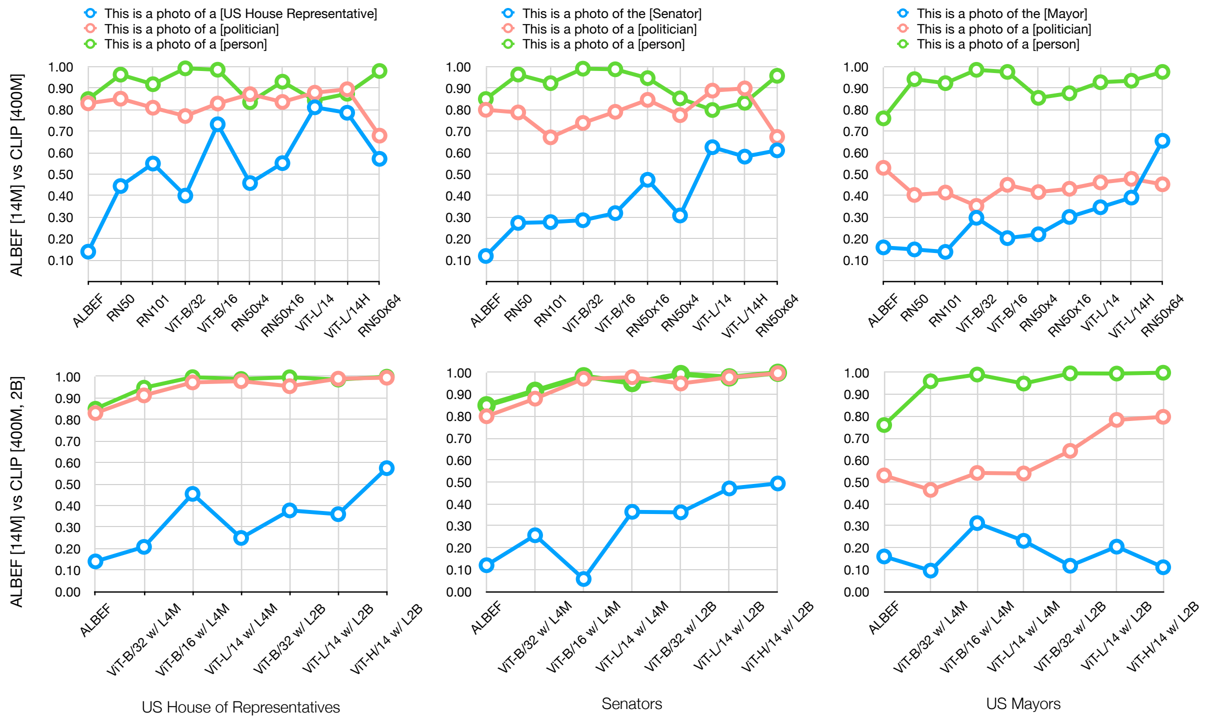


Figure 1: Results of testing the basic capabilities of vision-language models, from their ability to distinguish the basic level category `person`, which most models are able to do, to the more specific category `politician` which these models are still able to perform. However identifying the specific role of the person in the picture can only be recalled by the largest vision-language models even when smaller models – except ALBEF [14M images+text] which is included as a baseline – have access to the same large scale web training data as some of the other larger models [400M and 2B images+text].

sentatives, the Senate, and the Mayor of the most populous cities during the summer of 2022. These pictures correspond to 183 women and 453 men, and depict the subjects in a relatively similar manner. The actor portraits were selected from the top 100 actors working in Hollywood today and the athlete portraits were from the Top 100 athletes in sports history. The actor portraits contain 40 women and 60 men, the athlete portraits contain 30 women and 79 men. Information was obtained from public and official sources and was manually curated to correct inaccuracies.

## 2 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 were designed to test models for societal biases in the downstream task of coreference resolution. StereoSet (Nadeem et al., 2021) was designed to measure stereotypical biases across various sensitive protected variables. Honnavalli et al. also propose a benchmark for language generation models that involves US politicians by probing models for their implicit associations with respect to gender and seniority for members of congress and academia. More recently, Wei et al. (2022) designed a test to probe emergent and often surprising abilities that arise in large language models after certain model scale such as their ability to perform basic arithmetic and instruction following tasks.

In the vision-language domain, the Winoground benchmark was proposed by Thrush et al. (2022) to probe the capability of these models to perform compositional reasoning. Their benchmark aims to test models for their ability to distinguish syntactically similar but semantically different prompts such as "there are three people and two windows" and "there are two windows and three people" and corresponding images. VL-Checklist (Zhao et al., 2022) is another systematic benchmark designed to test various individual capabilities in vision-language models. Our P-FI Benchmark is complementary to these tests, and probes for capabilities across different categorical levels that are likely to be challenging for models with smaller capacity

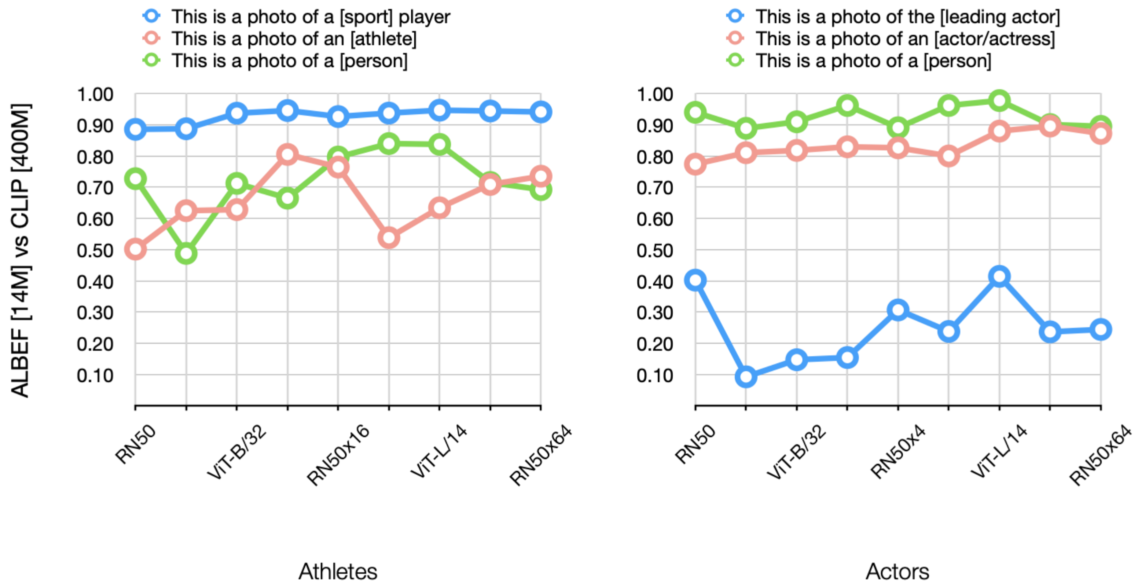


Figure 2: These results demonstrate the averages from CLIP 400M in correspondence to the portrait pictures of athletes and actors. Again these tests depict the basic capabilities of vision-language models, from their ability to distinguish the basic level category `person`, which most models are able to do, to the more specific category `athlete` or `actor` which these models are still able to perform.

and training datasets, and introduces a test for probing the capacity for *retention* of facts possibly seen during training.

## 3 The P-FI Benchmark

The P-FI Benchmark data consists of high-quality portrait pictures of United States figures who were star athletes, award-winning actors, and elected politicians. These include all the 100 senators, 436 House of Representatives including some delegates, 100 Mayors from the top 100 largest cities, 100 top actors, and 100 top athletes. We also compiled information from these political figures such as gender, political affiliation, type of athlete, movie role, and district, state, or city, respectively. We downloaded all the politician images from Wikipedia. The athlete images were downloaded from a Bleacher Report and Forbes List, then the actor images from an IMDB list. In the majority of cases for political figures, the images are the official pictures that are made available to the public by the official office of each representative. There are 253 women across these sets and 592 men. For the politician portraits, in terms of political affiliation, there are 291 members of the Republican party and 343 members of the Democratic party. The tests ran in our benchmark were calculated based on the images in each folder of 'senator', 'mayors', 'representatives', 'athletes', and 'actors'. Our framework, code, and data will be released under the MIT License.

Our benchmark tests consist of three types of evaluations that test for basic-level and subordinate-level categorization, societal bias estimation, and retention. Next, we define each of these evaluations and the motivation behind each:

**Capabilities: Politicians** Our first test aims to test the capabilities of each vision-language model to perform categorization starting from a basic-level category: `person`. In this test we prompt the model with a template in the format "This is a photo of a [$C$]", where $C$ is `person`, or one of the following six distractor categories: `dog`, `giraffe`, `plant`, `tree`, `lamp`, and `chair`. These categories are chosen so that they cover animals, vegetation, and objects, and are meant to provide a basic sanity check for the capability of the vision-language model. A reasonably good model should assign a matching score close to 1 for the category `person` for any of the images in the benchmark data. Our second capability test, involves the subordinate category `politician`, and six distractor categories corresponding to other occupations: `scientist`, `athlete`, `teacher`, `receptionist`, `assistant`, and `salesperson`. The dataset contains all pictures of politicians so the expectation is that most

| Model | Gender | Classes | | | | | | | Ratio |
|---|---|---|---|---|---|---|---|---|---|
| | | *scientist* | *politician* | *athlete* | *teacher* | *receptionist* | *assistant* | *salesperson* | |
| ALBEF | woman<br>man | 1.11<br>0.81 | 70.13<br>81.85 | 0.56<br>0.60 | 2.31<br>0.85 | 7.60<br>1.31 | 9.45<br>3.26 | 8.85<br>11.32 | 0.8568 |
| CLIP ViT-B/16 | woman<br>man | 5.96<br>4.66 | 68.19<br>80.92 | 0.15<br>0.35 | 4.17<br>1.89 | 13.60<br>0.30 | 1.59<br>0.82 | 6.34<br>11.06 | 0.8427 |
| CLIP ViT-L/H | woman<br>man | 0.74<br>1.10 | 81.41<br>83.91 | 0.03<br>0.22 | 1.50<br>0.92 | 4.25<br>0.77 | 5.92<br>4.58 | 6.15<br>8.49 | 0.9702 |
| OpenCLIP ViT-L/400M | woman<br>man | 0.46<br>0.59 | 87.81<br>92.18 | 0.00<br>0.24 | 0.55<br>0.49 | 2.35<br>0.02 | 6.79<br>2.32 | 2.04<br>4.15 | 0.9526 |
| OpenCLIP ViT-H/2B | woman<br>man | 0.13<br>0.17 | 95.56<br>96.71 | 0.32<br>1.72 | 1.49<br>0.30 | 0.15<br>0.00 | 0.28<br>0.08 | 2.01<br>1.02 | 0.9881 |

Table 1: Results for vision-language models that showcase disparities in the association of different occupations with people of different genders. We can see that in general even the less gender biased models under this test tend to associate `men` with *politician* more than they do for `women`.

models would assign a high matching score to this category for any of the images in the dataset since all these individuals are or have been politicians. Finally, our last test probes whether these models can assign with the highest score the specific role of the politicians as either a `US House Representative`, a `senator`, or a `mayor`. In addition to these three prompts, in this last test we also include the four distractor categories `president`, `vicepresident`, `governor`, and `attorney general`.

**Capabilities: Athletes** This is our prompts for our first test on athletes. In this test we prompt the model with a template in the format "This is a photo of a [$C$]", where $C$ is `person`, or one of the following six distractor categories: `dog`, `monkey`, `plant`, `tree`, `lamp`, and `chair`. These categories are chosen so that they cover animals, vegetation, and objects, and are meant to provide a basic sanity check for the capability of the vision-language model. A reasonably good model should assign a matching score close to 1 for the category `person` for any of the images in the benchmark data. Our second capability test involves the subordinate category `athlete`, and six distractor categories corresponding to other occupations: `artist`, `coach`, `teacher`, `receptionist`, `athletic trainer`, and `assistant coach`. The dataset contains all pictures of athletes so the expectation is that most models would assign a high matching score to this category for any of the images. Finally, our last test probes whether these models can assign with the highest score the specific sport of the athletes as either a `basketball player`, a `tennis player`, a `soccer player`, a `hockey player`, a `golf player`, `softball player`, and a `baseball player`.

**Societal Bias.** We do not expect that vision-language models would be able to predict a person's occupation based on facial features as there is no scientific basis for this assumption but rather due to two other factors: Either the model has seen enough images of the specific individual in our benchmark data, or even the precise specific image in our data and has enough text associations to recall this knowledge, or the model is making a prediction based entirely on spurious associations based on stereotypes. We measure to what extent this might be happening in our benchmark for various models by computing the disparity in the scores for the category `politician` compared to other distractor categories for both a `men` and `women` split of the data. Assuming that the score for `men` is $s_m$ and the score for women is $s_w$, then our bias score is defined as the ratio $b = s_w/s_m$, which means the closer this number is to 1.0, the more neutral the model, and the smallest the score, the more it is biased negatively toward women, as they are seen as less associated as politicians than their male counterparts. While this use of vision-language models would be problematic, once a model is deployed as part of larger and more general system for retrieval or captioning, these biases will emerge in these downstream applications.

**Retention.** We define retention as the ability of vision-language models to recall facts that they were likely exposed to during their training. We probe models for their ability to recall the names of each individual in each of our three groups: representatives, senators and mayors. For this purpose we issue prompts of the format "This is a picture of `Bernie Sanders`", and the names of all other senators as similarly formatted distractor prompts. We conduct this experiment in two directions, first given one prompt, have the model score all im-

| | US House of Rep. | | Senators | | Mayors | |
|---|---|---|---|---|---|---|
| | Text Score | Image Score | Text Score | Image Score | Text Score | Image Score |
| ALBEF | 0.32 | 0.32 | 1.42 | 1.56 | 1.63 | 1.71 |
| ViT-B-32 | 30.76 | 28.28 | 85.37 | 84.53 | 25.17 | 25.94 |
| ViT-B-16 | 30.66 | 30.05 | 82.11 | 82.23 | 25.48 | 26.88 |
| ViT-L-14 | 39.99 | 42.61 | 93.61 | 94.97 | 32.73 | 33.94 |
| RN50x64 | 44.42 | 48.84 | 92.67 | 92.83 | 37.87 | 39.24 |
| ViT-B-16 w/ L4M | 28.86 | 29.96 | 86.20 | 88.65 | 24.56 | 29.80 |
| ViT-L-14 w/ L4M | 36.17 | 38.88 | 91.64 | 93.93 | 32.98 | 34.26 |
| ViT-L-14 w/ L2B | 38.81 | 40.94 | 95.38 | 96.15 | 35.57 | 33.69 |
| ViT-H-14 w/ L2B | 50.23 | 51.85 | 99.09 | 98.87 | 38.03 | 41.51 |

Table 2: Results when evaluating large vision-language models to assess their prior knowledge about the politicians based on the picture and its corresponding name. The aim is to evaluate to which extent the model is familiar with a specific person by testing its ability to identify their unique name (i.e., Text Score) and the model's ability to identify their unique image (i.e., Image Score).

ages against the prompt using the vision-language model, and then given an image, have the vision-language score all prompts. We define this as the Text Score, and the Image Score under this test.

## 4 Results

Fig. 2 presents detailed plots for each subset of our benchmark data, probing the capabilities for 16 pretrained models, 9 versions of the official CLIP model by OpenAI in increasing order of model size (top three plots), and 6 versions of the Open-CLIP (Ilharco et al., 2021) in increasing order of model size (bottom three plots). Additionally, we report on each plot as baseline performance, the scores obtained by ALBEF (Li et al., 2021) which is a model trained with considerably less training data, 14 million as opposed to 400 million image text pairs, or 2 billion image text pairs as is the case in some of the OpenCLIP models. The main observation we have is that all models are relatively equivalent in matching the category `person` and `politician` but only the models trained at a considerably larger scale (400M and 2B) have seen enough data to infer the specific branch of government of the politicians. Moreover, from the models trained on the same dataset of 400M images, the models with the largest number of parameters are more consistently associating branches of government despite having been exposed to the same training data as the other smaller models. Table 1 shows the bias ratios $b$ across the `man` and `woman` splits of the benchmark. Smaller models seem to rely more on stereotypical associations such as (`woman`, `receptionist`) and (`woman`, `assistant`) but even the highest performing models have some disparities in this regard. Finally, Table 2 shows our *retention* capacity experiment results where the goal is to probe how much knowledge does each model have about the specific individuals in each of the portraits in our benchmark data by its capacity to assign a high matching score to the correct prompt or the correct image out of all other possibilities within each set as distractors. We observe that a model trained on a smaller scale web dataset such as ALBEF (Conceptual Captions (Changpinyo et al., 2021)) does not contain much knowledge of the people in these pictures while models trained on larger and more general data e.g. LAION-400M (Schuhmann et al., 2021) are able to recall most of the senators, although it does not seem to provide as much information about politicians in the other two groups.

## 5 Conclusion

Our work presents a benchmark that demonstrates for vision-language models their human-centric capabilities, societal biases, and capacity for retention of facts in their training. The Public Figures Benchmark (P-FI) represents a relatively homogeneous set of inputs corresponding to politicians who often have to legislate and regulate matters related to the use of technology in their own societal context.

## References

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

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, 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.

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, et al. 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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

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*, pages 5238–5248.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

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 Limitations

Our benchmark is not intended to test a wide range of capabilities in vision-language models. Ideally it 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 attributes, and 3) the Winground benchmark (Thrush et al., 2022) which probes for compositionality. Additionally, high performance in our societal biases test 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 high bias score ratio. 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.