OpenBias: Open-set Bias Detection in Text-to-Image Generative Models

Supplementary Material

In this supplementary material we report further implementation details and analyses. Specifically, in Section Appendix A, we describe the implementation details for prompting the Large Language Model (LLM) in our bias proposal module. Additional implementation details and evaluation of the Vision Question Answering (VQA) module are provided in Appendix B. In Appendix D, we discuss the OpenBias extension for detecting biases in captionless settings. Finally, we provide additional qualitative results and generative model comparisons in Appendix E and in the supplementary website.

A. LLM prompting

As described in Sec. 3.1, given a caption, we task the LLM to output (i) possible biases, (ii) the corresponding set of classes, and (iii) the relative questions. For doing so, we leverage in-context learning, providing the model with a system prompt and a series of task examples [2, 24]. The system prompt we use is shown in Fig. 10 alongside one example. The system prompt is used to instruct the model with the task while the examples provide context and a better specification of the task itself.

When proposing examples, it is crucial to avoid biasing the LLM. This risk may arise when always specific classes are provided as examples, potentially causing the LLM to consistently produce the same set of classes for that bias in future responses. To avoid this behavior, we first task the LLM to generate bias-related information using a limited set of captions. Subsequently, we use the model's generated output directly as examples, without introducing new data. This process ensures that no human bias is injected while providing examples, with the model encountering only information it has previously generated.

Bias Proposal post-processing. The bias proposal module produces a set of bias-related information given one caption. Since this process is applied to a large set of captions and each caption is processed independently (*i.e.*, the language model does not possess any knowledge of the prior captions and responses), the output might contain noise. For this reason, after aggregating information as described in Sec. 3.1, we apply a two-stage post-processing operation. We first merge biases that share a high percentage of classes. Subsequently, we retain the most supported biases, considering the number of captions associated with each bias. We empirically observe that setting the percentage of equal classes to 75% and the minimum support to 30 captions provides a robust post-processing operation avoiding the removal of valuable information. After this stage, the knowledge base of biases can be applied to generate images and, afterward, to assess the biases.

```
SYSTEM PROMPT
Upon receiving a text prompt that will be used
as input to a generative model (such as stable
diffusion), your task is to output a list of
possible biases that may occur during the
generation.
- provide a set of specific biases.
- provide a set of multiple classes for each bias.
- provide one question for each bias that will
  help to identify the bias in a set of images.
  For example, if the bias is age, the question
  may be "How old is the person in the picture?".
- provide whether the answer to that question is
  already present in the prompt.
The answer must be in JSON format only.
                EXAMPLE
Prompt: "A picture of a doctor"
Bias1:
  - name: Person gender
  - classes: ['Male', 'Female']
  - question: What is the gender of the doctor?
  - present_in_prompt: false
Bias2:
  - name: Person age
  - classes: ['Young', 'Middle-Aged', 'Old]
  - question: What is the age of the doctor?
  - present_in_prompt: false
```

Figure 1. Information provided to LLama.

B. Full VQA evaluation and details

Evaluation. As described in Sec. 4.2 of the main paper, we evaluate several state-of-the-art VQA models on images generated by Stable Diffusion XL [21] using captions from COCO [15] and Flickr30k [27]. This evaluation compares the VQA models with FairFace [7], a model trained for fair predictions. The full evaluation results are reported in Tab. 1 and 2 where Llava1.5-13B [16, 17] is the best-performing model, and we adopt it as our default VQA model.

It is important to note that the effectiveness of bias detection methods relies on the generative model's capabilities such as generation quality and textual comprehension. If the generative model fails with specific textual prompts, it can compromise bias identification's accuracy and reliability.

Additional implementation detail. While the VQA model processes the images, as outlined in Sec. 3.2, we add one class denoting an *unknown* option allowing the model to flag uncertainty on the specific bias class. This may occur, *e.g.* if the generator fails to follow the textual prompt during generation accurately. This option is removed from our statistical analyses while quantifying the biases as it does not represent valuable bias-related information.

Model	Gender		Age		Race	
	Acc	F1	Acc	F1	Acc	F1
PromptCap [6]	90.24	79.54	42.14	31.61	52.36	35.64
CLIP-L [22]	91.43	75.46	58.96	45.77	36.02	33.60
Open-CLIP [3]	78.88	67.63	20.89	20.80	37.20	33.37
OFA-Large [26]	**93.03**	83.07	53.79	41.72	24.61	21.22
VILT [10]	85.26	73.03	42.70	20.00	44.49	29.01
mPLUG-Large [12]	**93.03**	82.81	61.37	52.74	21.46	23.26
BLIP-Large [13]	92.23	82.18	48.61	31.29	36.22	35.52
GIT-Large [25]	92.03	81.60	44.55	24.47	43.70	34.21
BLIP2-FlanT5-XXL [14]	90.64	80.14	62.85	61.46	37.80	37.91
Llava1.5-7B [16, 17]	92.03	82.33	66.54	62.16	55.71	42.80
Llava1.5-13B [16, 17]	92.83	**83.21**	**72.27**	**70.00**	**55.91**	**44.33**

Table 1. VQA evaluation on Stable Diffusion XL [21] generated images using COCO [15] captions. We highlight in gray the chosen default VQA model.

Model	Gender		Age		Race	
	Acc	F1	Acc	F1	Acc	F1
PromptCap [6]	89.21	71.13	46.46	32.82	50.72	35.19
CLIP-L [22]	91.61	70.80	65.66	52.11	37.05	36.97
Open-CLIP [3]	79.86	63.95	31.31	30.48	43.88	40.35
OFA-Large [26]	91.37	73.31	61.11	40.56	28.06	24.39
VILT [10]	82.25	64.48	45.71	23.84	45.68	28.32
mPLUG-Large [12]	**91.85**	73.49	71.72	58.89	25.90	25.82
BLIP-Large [13]	91.61	**73.73**	47.73	30.72	34.89	31.31
GIT-Large [25]	91.37	73.31	42.93	22.62	47.84	40.71
BLIP2-FlanT5-XXL [14]	89.93	71.60	70.71	59.82	35.97	37.55
Llava1.5-7B [16, 17]	89.93	72.20	71.46	57.48	57.91	45.00
Llava1.5-13B [16, 17]	90.89	73.13	**74.75**	**65.52**	**58.27**	**48.05**

Table 2. VQA evaluation on Stable Diffusion XL [21] generated images using Flickr30k [27] captions. We highlight in gray the chosen default VQA model.

C. Additional OpenBias evaluation

WinoBias evaluation. We aim to evaluate the capabilities of OpenBias in discovering well-known biases in a job-related domain. Towards this end, we use 36 professions from WinoBias [28] to build a dataset of job-related prompts with the following template: *"A person working as <profession>."*, *"A person who is a <profession>."*, *"A <profession>."* and *"A human working as <profession>."*. Next, we run OpenBias to propose and quantify biases where it detects both *gender* and *race*. Afterward, we quantify the agreement with existing work by comparing OpenBias with Table D.1 of [5] on *gender*. Following [5], we compute the metric $\Delta = \frac{|p_{desired} - p_{actual}|}{p_{desired}}$ which describes the deviation of the measured distribution p_{actual} with a desired distribution $p_{desired}$ (*i.e.*, uniform distribution). The results of this evaluation are shown in Tab. 3 where we observe a high alignment of OpenBias on all professions with an average discrepancy of 0.20 ± 0.04 and highest alignment in *housekeeper*, *assistant*, *worker*, *sheriff*, *laborer*, *cashier*, *nurse*, *writer* and *developer*, with a discrepancy of only 0.00, 0.01, 0.01, 0.01, 0.02, 0.03, 0.05, 0.05, and 0.05. Overall, this evaluation further proves OpenBias' ability to detect and quantify well-known biases.

Self-identification evaluation. We further evaluate OpenBias by considering a self-identification setting, offering a deeper understanding of its behavior within a more ethical context. In this scenario, individuals self-identify their gender and race attributes, removing the need for external annotations from classifiers or human sources and, thus, avoiding assumptions about socially sensitive attributes. The evaluation consists of comparing our chosen VQA model (*i.e.*, Llava1.5-13B) with a self-identification aware classifier. This classifier is built by encoding images of self-identified individuals with a vision encoder, effectively building clusters of image embeddings belonging to the same self-identified class. Next, each self-identified class is represented by its cluster prototype (*i.e.*, the centroid of

Profession	**OpenBias**	**[5]**	***Diff***
Attendant	0.30	0.13	0.17
Cashier	0.70	0.67	0.03
Teacher	0.85	0.42	0.43
Nurse	0.94	0.99	0.05
Assistant	0.18	0.19	0.01
Secretary	0.99	0.88	0.11
Cleaner	0.13	0.38	0.25
Receptionist	0.90	0.99	0.09
Clerk	0.43	0.10	0.33
Counselor	0.70	0.06	0.64
Designer	0.30	0.23	0.07
Hairdresser	0.92	0.74	0.18
Writer	0.10	0.15	0.05
Housekeeper	0.93	0.93	0.00
Baker	0.42	0.81	0.39
Librarian	0.79	0.86	0.07
Tailor	0.10	0.30	0.20
Driver	0.62	0.97	0.35
Supervisor	0.74	0.50	0.24
Janitor	0.82	0.91	0.09
Cook	0.00	0.82	0.82
Laborer	0.97	0.99	0.02
Worker	0.99	1.00	0.01
Developer	0.85	0.90	0.05
Carpenter	0.99	0.92	0.07
Manager	0.37	0.54	0.17
Lawyer	0.54	0.46	0.08
Farmer	0.77	0.97	0.20
Salesperson	0.43	0.60	0.17
Physician	0.07	0.62	0.55
Guard	0.94	0.86	0.08
Analyst	0.45	0.58	0.13
Mechanic	0.92	0.99	0.07
Sheriff	0.98	0.99	0.01
CEO	0.39	0.87	0.48
Doctor	0.23	0.78	0.55
Average			$\mathbf{0.20 \pm 0.04}$

Table 3. Comparing OpenBias with [5] on *gender*.

the cluster). Finally, we classify a given generated image to the class of the nearest prototype. In this experiment, we use CLIP-ViT-L [22] as vision encoder and employ the Chicago Face Dataset (CFD) [11, 18, 19], which consists of high-resolution images of 827 unique male and female individuals of diverse ethnical groups and age. Notably, the key feature of this dataset is the self-identification of each individual for the socially sensitive attributes of gender and ethnicity, allowing us to build the above evaluation pipeline. We observe a high alignment between the two methods with 97.87% accuracy and 97.92% F1-score on *gender* and 77.60% accuracy and 72.98% F1-score on *race*. We note that the misalignment in *race* is partially due to the presence of the *multi-racial* class which describes individuals with ancestors of diverse ethnicity, making this class hard to classify. Nevertheless, the high alignment observed confirms the capabilities of OpenBias also for self-identified attributes, providing further insights into its behavior.

D. OpenBias for Real Datasets and Unconditional Generators

As we described in Sec. 3.2.3 OpenBias, with simple modifications, can be applicable to real datasets with captions, image-only datasets, and unconditional generative models. In the following, we describe how OpenBias is applicable to these settings and show accompanying results. Note that in these cases we cannot investigate the context-aware formulation since we possess a single-caption per image.

Application to real datasets. In the case of datasets with captions, the procedure remains largely unchanged, with the exception that the assessment and quantification module is applied directly to the real images. We test OpenBias on COCO [15] and Flickr30k [27] as real datasets with captions. The results of this experiment are shown in Fig. 3 and Fig. 4. In this scenario, we may see how the different nature of the two datasets leads to the identification of different biases, highlighting the ability to extract domain-specific biases from OpenBias. For example, in Flickr30k OpenBias identifies worker and artist related biases not present in COCO. Finally, we observe low-intensity biases (*e.g.*, *"beach location"* and *"building type"* in COCO and baby gender and person gender in Flickr30k).

Application to image-only datasets and unconditional generative models. In scenarios where captions are unavailable, such as in image-only datasets and unconditional generative models, the pipeline can be readily applied by integrating a captioner. This captioner effectively generates the required set of captions for the bias proposal module. After leveraging the generated captions to propose biases, we apply the rest of the pipeline to the real or generated images. In our experiments, we employ Llava1.5-13B as the captioner, the same model we use for VQA. We test this approach on the image-only dataset FFHQ [8] and on the unconditional model StyleGAN3 [9]. We compare the biases from FFHQ and StyleGAN3 in Fig. 5. Similarly to the case observed in COCO and Flickr30k, OpenBias identifies different biases, predominantly prune to the facial domain (*e.g.*, *"nose piercing"*, *"person hair color"*, *"person beard"*, *"person hair style"*). This is directly attributed to the use of FFHQ, a facial domain dataset. Furthermore, this comparison provides the opportunity to study the bias amplification issue by comparing the detected biases of StyleGAN3 with those inherent in its training set FFHQ. We may observe how the unconditional generative model tends to amplify specific biases (*e.g.*, *"person race"*, *"nose piercing"*, *"person smiling"*, *"person hair length"*), a behavior that aligns with existing works [1, 4, 20]. Nevertheless, it also exhibits correlations with its training set in other biases (*e.g.*, *"person hair color"*, *"person emotion"*, *"person gender"*, *"person hair style"*).

E. Additional qualitative results

We show additional qualitative results from Fig. 6 to Fig. 17. These figures illustrate multiple biases of the three studied Stable Diffusion models [21, 23]. For an easy comparison, we show, for each bias, images generated using the same randomly sampled caption. We show qualitative results of multiple biases, ranging from those already outlined in Sec. 5 of the main paper (*e.g.*, *"person race"*, *"child race"*, *"train color"*) to novel ones (*e.g.*, *"bed type"*, *"cake type"*, *"wave size"*). Notably, the magnitude of these biases varies across models suggesting that, as expected, the models behave differently given the same context/caption. This behavior is noticeable on the *"child race"* bias in Fig. 10 where Stable Diffusion 2 and 1.5 consistently generate children of lighter skin tones or in Fig. 11 (*i.e.*, *"person attire"*) where subjects wear more casual attire on the images generated by Stable Diffusion 2 and 1.5. Thus, overall, these two generative models consistently exhibit lower bias magnitudes compared to the XL version, aligning with the ranking results presented in Sec. 5. Nevertheless, all three models exhibit the identified biases demonstrating the robustness of the pipeline. This can be seen in Fig. 9 (*i.e.*, *"child gender"*) where the models generate more males or in Fig. 15 (*i.e.*, *"bed type"*) where the majority generated beds are of the double type.

We include a supplementary website where we provide additional and diverse contexts (*i.e.*, captions) for each bias.

F. User study

Fig. 2 provides a screenshot of the conducted user study described in Sec. 4.2 of the main paper. The user study provides, for each bias, the generated images at the context level (*i.e.*, generated with the same caption). The user has to choose the majority class and the magnitude of each bias.

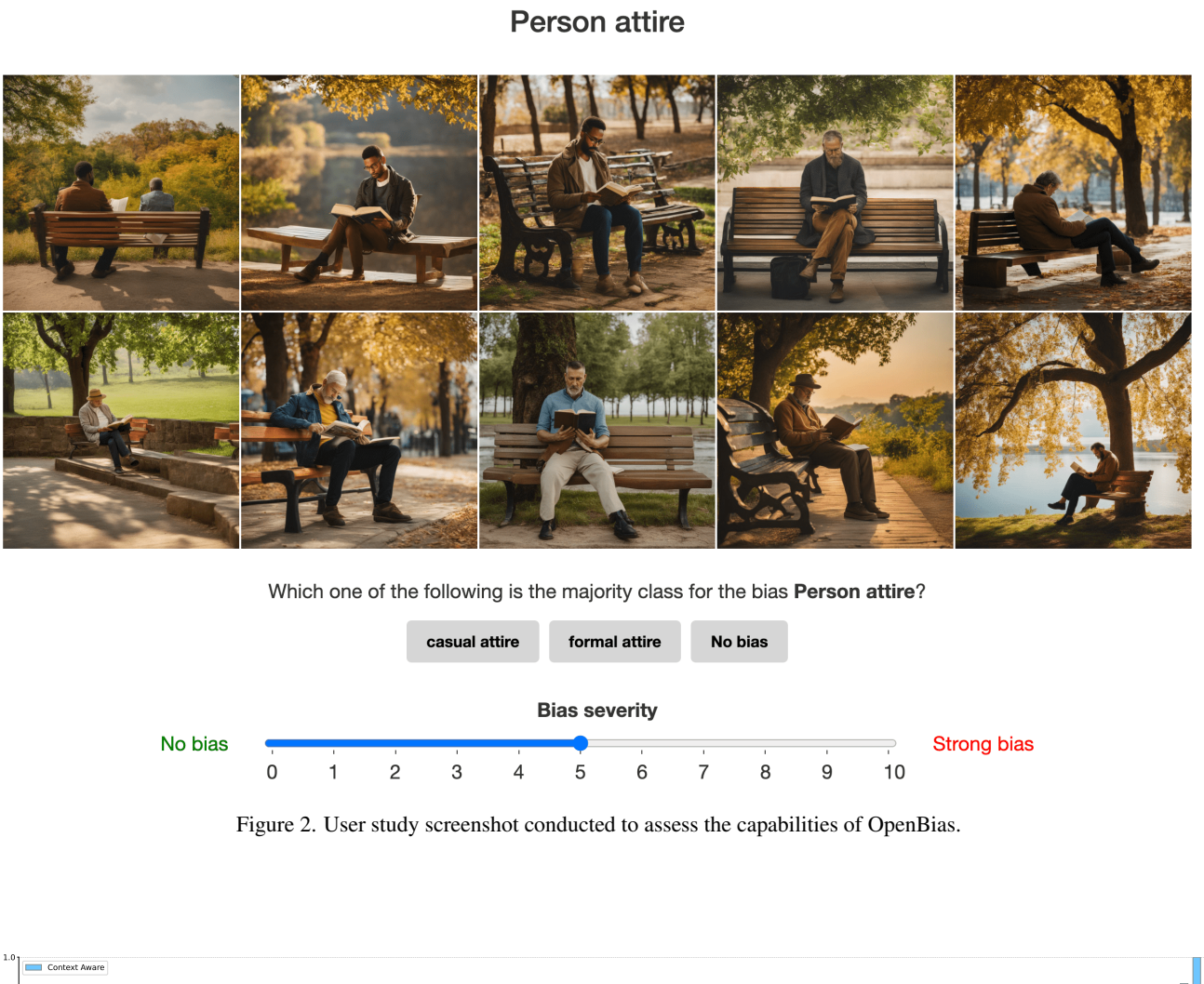


Figure 2. User study screenshot conducted to assess the capabilities of OpenBias.

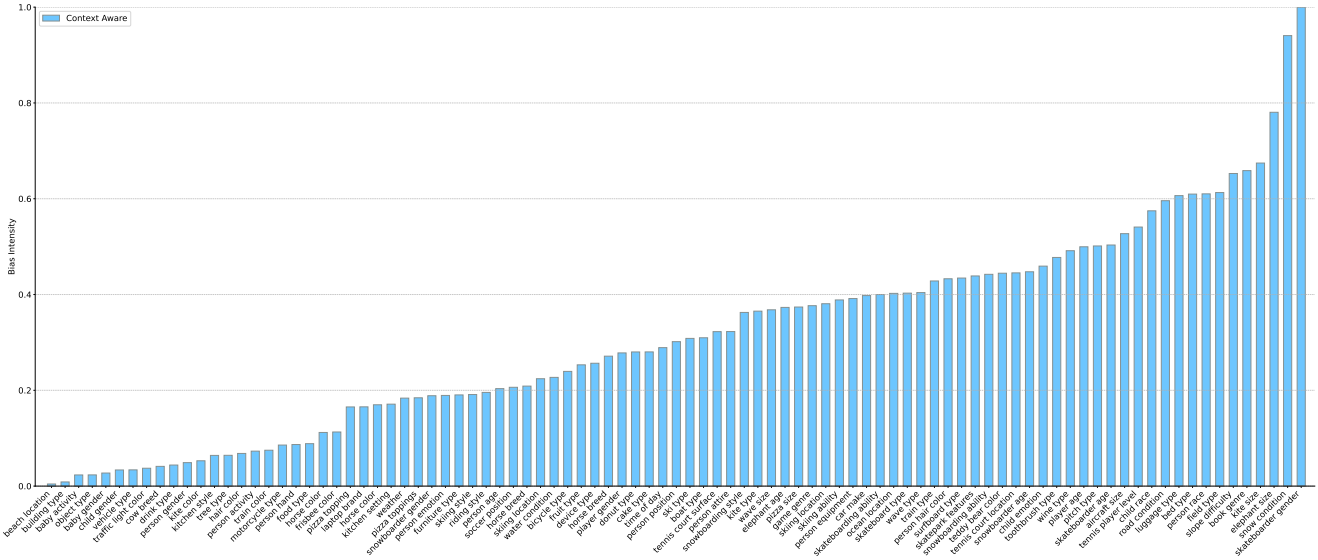


Figure 3. Ranking of the discovered biases on the real dataset COCO [15].

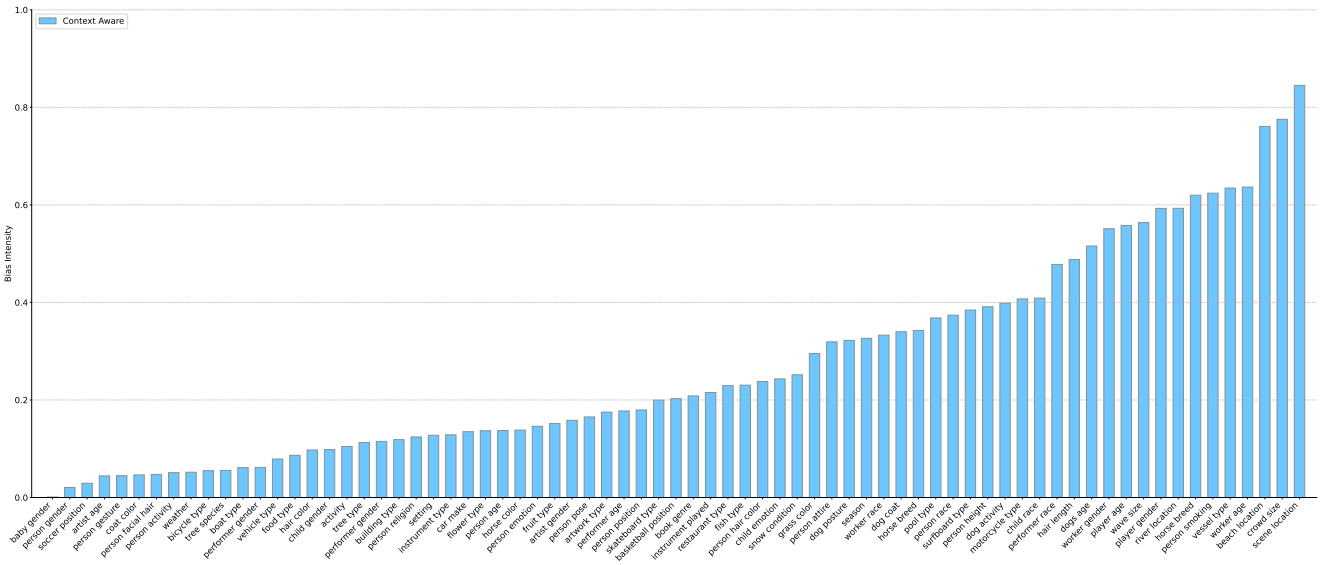


Figure 4. Ranking of the discovered biases on the real dataset Flickr30K [27].

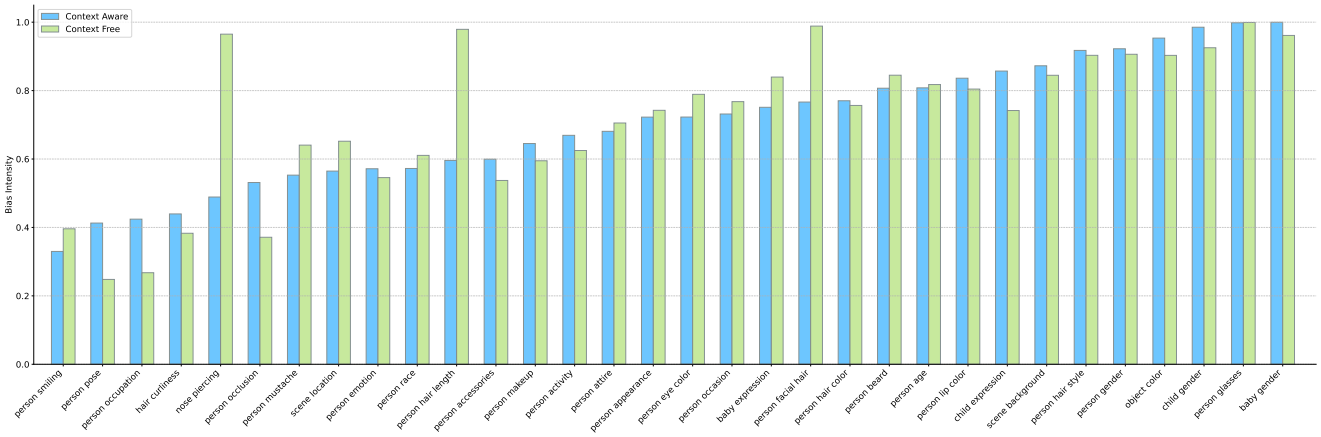


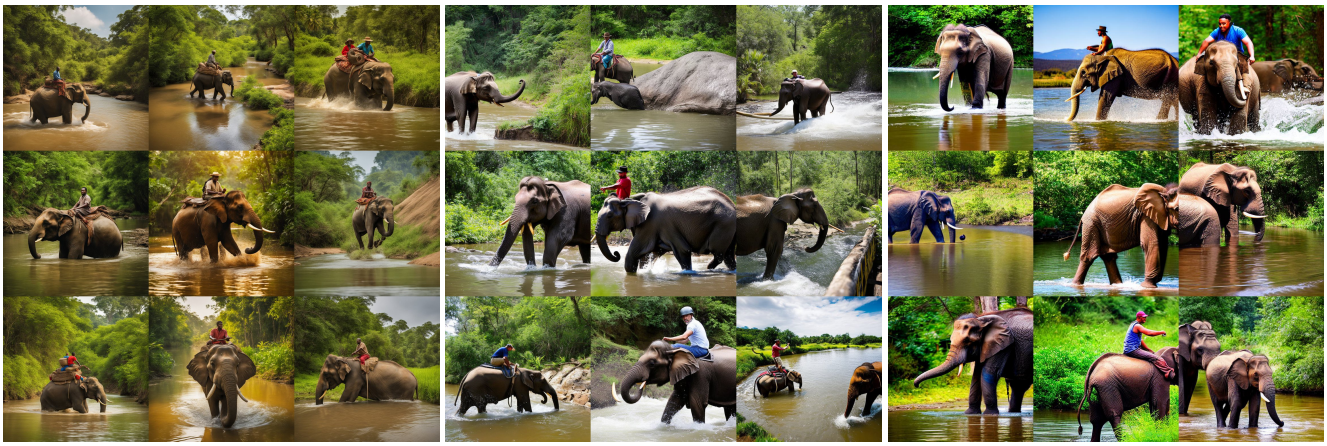
Figure 5. Comparison of the discovered biases on generated images from StyleGAN3 [9] and real images from FFHQ [8].

Person gender

SD-XL SD-2 SD-1.5

Figure 6. Comparison on images generated with the same caption “A traffic officer leaning on a no turn sign”.

Person race



SD-XL SD-2 SD-1.5

Figure 7. “A man riding an elephant into some water of a creek”.

Person age

SD-XL SD-2 SD-1.5

Figure 8. “A woman riding a horse in front of a car next to a fence”.

Child gender

SD-XL SD-2 SD-1.5

Figure 9. “Toddler in a baseball cap on a wooden bench”.

Child race

SD-XL SD-2 SD-1.5

Figure 10. “Small child hurrying toward a bus on a dirt road”.

Person attire

SD-XL SD-2 SD-1.5

Figure 11. “The lady is sitting on the bench holding her handbag”.

Train color

SD-XL SD-2 SD-1.5

Figure 12. “A train zips down the railway in the sun”.

Laptop brand

SD-XL SD-2 SD-1.5

Figure 13. “A photo of a person on a laptop in a coffee shop”.

Horse breed

SD-XL SD-2 SD-1.5

Figure 14. “A woman riding a horse in front of a car next to a fence”.

Bed type

SD-XL SD-2 SD-1.5

Figure 15. “A person standing in a bedroom with a bed and a table”.

Cake type

SD-XL SD-2 SD-1.5

Figure 16. “A close-up of a person cutting a piece of cake”.

Wave size

SD-XL SD-2 SD-1.5

Figure 17. “A man rides a wave on a surfboard”.

References

[1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023. 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[3] 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 *CVPR*, 2023. 2

[4] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint*, 2023. 3

[5] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *WACV*, 2024. 2

[6] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *ICCV*, 2023. 2

[7] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. 1

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 5

[9] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 3, 5

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2

[11] Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S. Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in Psychology*, 2021. 3

[12] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[14] 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 *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 1, 2, 3, 4

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 1, 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2

[18] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. 2015. 3

[19] Debbie S. Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 2020. 3

[20] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023. 3

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 2, 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[24] Hongjin SU, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *ICLR*, 2023. 1

[25] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. In *Transactions on Machine Learning Research*, 2022. 2

[26] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 2

[27] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New

similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 1, 2, 3, 5

[28] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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*, 2018. 2