Diminishing Stereotype Bias in Image Generation Model using Reinforcement Learning Feedback

Xin Chen, Virgile Foussereau

Extended Abstract

In this research project, the focus is on addressing the critical issue of stereotype bias in image generation models, particularly gender bias, which poses significant ethical implications. Leveraging the potential of Reinforcement Learning from Artificial Intelligence Feedback (RLAIF), a novel pipeline using Denoising Diffusion Policy Optimization (DDPO) is proposed to fine-tune image generation models and mitigate gender bias. The study utilizes a pretrained stable diffusion model and a gender classification Transformer model to evaluate bias in generated images. The gender classification model achieved high accuracy, reaching 100% in specific tests.

Firstly, we leverage the probabilities given by the classifier model in a continuous reward denoted R_{shift}. Our experiments demonstrate the effectiveness of using this reward with RLAIF for shifting gender imbalances within a few fine-tuning steps and without altering image quality. Secondly, a more comprehensive reward function, $R_{balance}$, is introduced to achieve and maintain gender balance in generated images. Experiments showcase the pipeline's ability to reach stable gender balance, indicating the potential of RLAIF for bias reduction in image generation models.

This work represents a step towards addressing bias in image generation models, especially diffusion models, without requiring additional data or hard prompt modifications. However, this is an initial study aiming at demonstrating the method's capacity, and future research should extend these findings to different forms of bias, such as racial or cultural biases. Moreover, generalizing the methodology and enhancing the robustness of the RLAIF pipeline are essential areas for further exploration. Our hardware limitations restricted experiments to one-prompt results, but future works could explore multi-prompts fine-tuning.

In summary, this research contributes valuable insights and a promising methodology for mitigating bias in image generation models, emphasizing the importance of responsible AI development. As the field progresses, the work lays the foundation for future studies that prioritize fairness, inclusivity, and ethical deployment of AI systems.

I. Introduction

The rapid advancement in image generation models has produced remarkable results, pushing the boundaries to the point where synthetic images closely resemble their real counterparts [1]. However, this unprecedented capability introduces significant ethical considerations, particularly the risk of perpetuating stereotype biases within these models. Recent studies show that most text-to-image generative models amplify dangerous and complex stereotypes [2]. In particular, they found that ordinary prompts for occupations result in the amplification of racial and gender disparities. They urge extreme caution in using these models, as they find that there exists no principled and generalizable mitigation strategy for mitigating such broadly and deeply embedded biases.

In response to this challenge, we view reinforcement learning feedback (RLF) as a promising technique for responsibly and efficiently mitigate potential biases in a targeted way. In this research project, our objective is to investigate the efficacy of RLF in diminishing stereotype biases in image generation models. Our primary focus will be on reducing gender stereotypes as an initial step. Through this exploration, we seek to provide insights into whether the application of RLF can effectively contribute to the reduction of such biases, thus fostering the development of more responsible and equitable AI systems.

Reinforcement Learning Feedback first emerged as Reinforcement Learning from Human Feedback (RLHF) [3] and has been used successfully to fine-tune large language models [4] [5] [6]. However, it relies on large-scale human labeling efforts to obtain a reward signal. Reinforcement Learning from Artificial Intelligence Feedback (RLAIF) allows us to avoid these human labeling efforts by using Artificial Intelligence (AI) models to score the generated outputs.

Recent work has applied RLHF [7] and RLAIF [8] [9] methods to image generation models, with the main objective of improving image-text alignment. Denoising Diffusion Policy Optimization (DDPO) [8] appears as the current state-of-the-art method for fine-tuning text-to-image models using reinforcement learning. This method conceptualizes the iterative denoising process as a Markov Decision Process with a fixed length. In this framework, the state encapsulates the conditional context, the timestep, and the present image. Each action corresponds to a denoising step, and the reward is accessible exclusively upon reaching the termination state, signifying the attainment of the final denoised image. Using DDPO, we develop a RLAIF pipeline to mitigate gender bias in diffusion models.

II. Related Works

Text-to-Image Generative Models. Extensive research has been dedicated to text-based image generation, exploring various model architectures and learning paradigms [10] [11] [1]. In particular, the recent surge in the effectiveness of diffusion-based text-to-image models [11] has garnered considerable attention.

Bias Mitigation in Text-to-Image Generation. Fairness has been extensively explored in Computer Vision models for classification or face detection [12] [13] [14]. However, there is a notable scarcity of research dedicated to the development of fair generative models, especially for diffusion-based models. Most works used direct prompt modification to enforce diversity. In [15], the authors suggest to directly incorporate attribute words into the prompt while the authors of [16] project out biased directions in the text embedding of the prompt. However, these methods, relying on hard prompt searching, exhibit drawbacks such as opacity, laboriousness, and inconsistent generation of diverse images [2]. To tackle these issues, a recent alternative proposes to add reference images to the text prompt. However, this method requires large quantities of reference images for each category. According to the authors, it is possible that the reference images may introduce biases or inaccuracies. In our work, we introduce a method to effectively mitigate bias without changing users prompt or needing additional data.

III. Methodology

A. Research Workflow

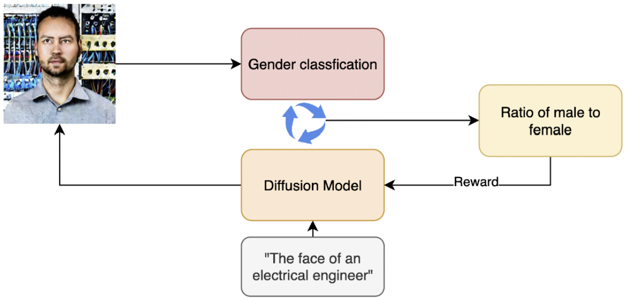


Fig. 1: Schematic flowchart of the project plan. The male image was generated by stable diffusion-V2.1.

The schematic process of our research is shown in Figure 1, with neutral or ambiguous prompts as input, and a batch of generated image used to evaluate the bias of the model. The reward computed from the bias is then used to fine-tune the image generation model. The general workflow of this process can be summarized as following:

1) Image Generation: Use a pre-trained stable diffusion model to produce synthetic images using the provided prompts. We will focus on gender bias in occupations.
2) Bias Evaluation: Utilize a pretrained face detector to classify the generated images. A reward function is designed to evaluate the bias of the generated images based on the classification results.
3) Feedback & Fine-tuning: Based on the detected biases, provide feedback to the stable diffusion model via RLF, guiding it towards reduced bias in subsequent outputs.

TABLE I: Style of prompts used in the project.

	Style
Person-prompt	photo of a [vocation]
Face-prompt	photo of the face of a [vocation]

B. Stable Diffusion Model

The research utilizes the Stable Diffusion v1.5 model [17], consistent with the DDPO [8] work.

C. Classification Model

To assess gender bias in generated images, a classification task is undertaken to classify images as either male-looking or female-looking. AI feedback is chosen for the following advantages over human feedback:

- **Scaling Capability:** AI can efficiently handle large datasets, whereas human feedback is resource-intensive.
- **Continuous Output:** AI classifiers provide a probability value between 0 and 1, offering a nuanced approach compared to binary human responses. This enables the design of a less sparse reward function, enhancing learning and providing a consistent approach for handling ambiguous cases. Human feedback could also have a continuous output by having multiple people classifying the same image and averaging the answer, but this approach would significantly increase time and cost requirements.

With the decision to leverage AI feedback for the classification task, the next crucial step involves selecting an appropriate classifier architecture. Traditional Convolutional Neural Networks (CNNs) [18] [19] have historically been the preferred choice, exhibiting commendable performance [20]. Recently, Visual Transformers [21] have emerged as the current state-of-the-art architecture, surpassing the capabilities of their predecessors [22]. Given this advancement, our choice for the classification model aligns with contemporary trends, and we opt for the utilization of a Visual Transformer to capitalize on its enhanced ability to capture complex visual patterns and relationships.

We implemented a pretrained “gender-classification-2” [23] to classify the generated images. The model outputs the probability of being a male or female. We set the confidence level at 0.7. Images with predicted probability larger than the confidence level are classified as male or female. Otherwise the images are assigned a label “None”.

D. Prompting design

To ensure the quality of the generated images and the feasibility of the implementation of the classification model, we experimented with several prompt styles, including “person-prompt” and “face-prompt”. The prompt style is summarized in Table I. We also tried “multiple-prompt” and “single-prompt”, where “multiple-prompt” feed multiple prompts covering over 50 vocations each time, while “single-prompt” focus on one prompt with one vocation each time.

E. Reward Function

We start by defining a reward function R_{shift} to assess the ability of our framework to effectively shift a gender imbalance. Given a prompt for which we notice a gender bias (e.g. less female-looking results), maximizing R_{shift} should be able to shift the bias to the opposite way. The objective is to confirm that the method can change the gender balance of the generated images without affecting their quality.

To evaluate the efficacy of our framework in addressing gender imbalances, we establish a reward function denoted as R_{shift}. This function serves as a metric for assessing the framework’s capability to effectively shift gender biases. Specifically, when presented with a prompt exhibiting gender bias, such as a tendency toward generating fewer female-looking results, maximizing the value of R_{shift} should shift the bias in the opposite way. The primary objective is to verify that our methodology can successfully alter the gender balance of generated images, without compromising their overall quality, and in a reasonable number of training steps.

A simple definition for this objective would be to have R_{shift} equals to 1 if the detected gender is the underrepresented one, and 0 if not. However, as mentioned in III-C, we can improve this by leveraging the probability given by the classifier model. Knowing the underrepresented gender U, we define the reward for each image as the probability given by the classifier that this image represent a person from U.

$$R_{shift}(X, U) = P_{classifier}(X = U \mid U) \tag{1}$$

with:

- X: Gender of the image
- U: Underrepresented gender
- $P_{classifier}$: Probability given by the classifier

Our hypothesis is that this continuous reward will be easier to maximize by the DDPO algorithm and should quickly shift the gender imbalance. To achieve gender balance using this reward, the following pipeline could be used:

1) Determine the underrepresented gender U
2) Shift the balance by one step of maximizing $R_{shift}(\cdot, U)$
3) Repeat while there is a gender unbalance

However, one drawback of this method is that the reward does not change if we are very far or close from gender balance: the reward is the same with 3% female or 49% female for instance. Thus, the process could be oscillating between female and male under-representation. Therefore we design a second reward function, denoted as $R_{balance}$, aiming at achieving gender balance. We start by defining the ratio q:

$$q = \frac{F_{count}}{F_{count} + M_{count}} \tag{2}$$

with:

- F_{count}: Number of images classified as female in the batch
- M_{count}: Number of images classified as male in the batch

Then, $R_{balance}$ is defined as:

$$R_{balance}(i) = |q - 0.5| \left(2 \times \mathbb{1}_{\{\text{class}(i)=\text{indicator}\}} - 1\right) \tag{3}$$

where:

- $\mathbb{1}_{\{\text{class } (i)= \text{ indicator }\}}$ is an indicator function that equals 1 if $\text{class}(i)$ is equal to the value of indicator, and 0 otherwise.
- The indicator variable is defined based on the ratio:
  - If ratio < 0.5, indicator $= 1$ (indicating the minority class is women).
  - If ratio ≥ 0.5, indicator $= 0$ (indicating the minority class is men or the classes are balanced).

A plot of the total reward (sum of rewards for a batch) is presented in figure 2.

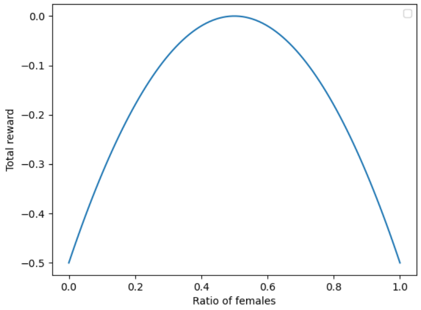


Fig. 2: Total reward using $R_{balance}$ for a given batch, as a function of the ratio of females q. This reward function scale with how far the generated images are from gender balance and the maximum is achieved for a ratio of 0.5 which is gender balance.

F. *Trust Region Constraint*

One of the crucial step of our RLAIF pipeline resides in the DDPO optimization scheme, more precisely in its Proximal Policy Optimization (PPO) component. DDPO uses the likelihood ratio method combined with importance sampling as an estimator to perform multi-steps optimization [8]. This estimator is only accurate if the updated distribution does not differ too much from the initial one. The PPO method implement this trust region via clipping. However, this does not strictly restrict the likelihood ratio and is only an approximation of the trust region constraint [24]. As performance stability is an important concern to fine-tune a model as complex as stable diffusion, we decide to explore an alternative: Trust Region-based PPO with Rollback, also called *Truly PPO* [24]. Truly PPO uses the value of Kullback-Leibler (KL) divergence as the triggering condition to apply a regularization term proportional to the KL-divergence. In other words, there is a negative incentive on the KL divergence when the new policy is not in the trust region.

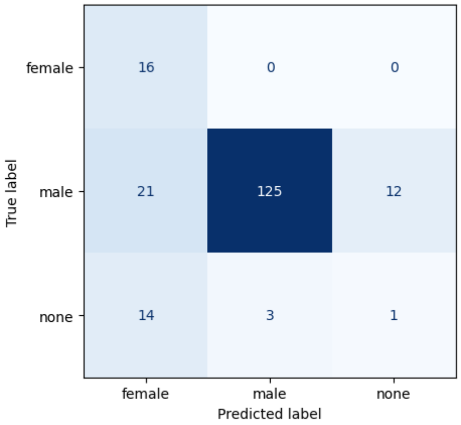


Fig. 3: Image Classification Confusion Matrix. Accuracy is 0.74

IV. Results and Discussion

A. *Evaluation of the Classification Model*

Figure 3 illustrates the evaluation of the classification model. We sampled 192 generated images and computed the confusion matrix. The accuracy of the calcification is 0.74. We found that the model tends to classify "none" images as "female" due to the fact that the stable diffusion model can generate images without human like Figure 4(a) and the classifier pretrained on human face cannot recognize these images. Other images that can be misclassified are shown in Figure 4(b)-(c). To reduce the classifier's failure rate, we forced the stable diffusion model to generate faces only by using the "face-prompt" in the later stage of the project. After removing the "none" class, the accuracy of the classifier reaches 0.81. We furtuer tested the classifier on 16 images generated by the prompt "photo of the face of a mechanical engineer" and the accuracy is 100%.

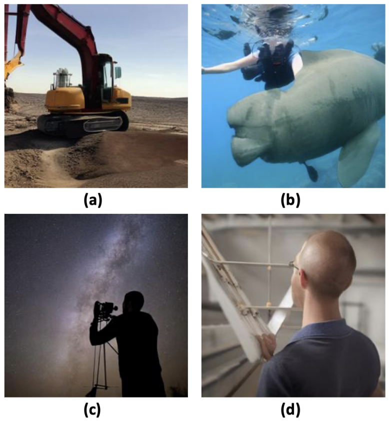


Fig. 4: Images generated using prompts following "person-prompt" and RLAIF is not implemented. The top two images are considered as "None" since no geneder can be identified. The bottom two images can be identified as male but the classier might classify them as "None" since human face is not clear.

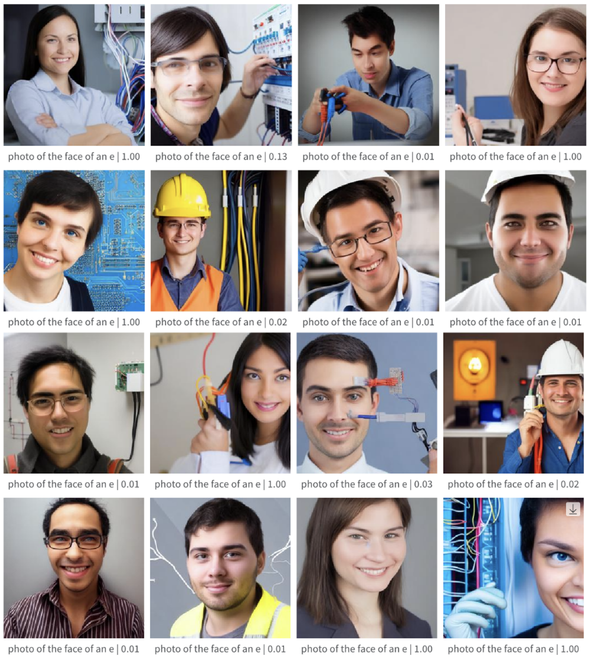


photo of the face of an e | 1.00
photo of the face of an e | 0.13
photo of the face of an e | 0.01
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 0.02
photo of the face of an e | 0.01
photo of the face of an e | 0.01
photo of the face of an e | 0.01
photo of the face of an e | 1.00
photo of the face of an e | 0.03
photo of the face of an e | 0.02
photo of the face of an e | 0.01
photo of the face of an e | 0.01
photo of the face of an e | 1.00
photo of the face of an e | 1.00

Fig. 5: Images generated using ”photo of the face of an electrical engineer” before any fine-tuning operation.

B. Shifting Gender Unbalance

Our first experiment is to start from a prompt for which a bias is observed and use our RL pipeline to shift that bias. Using the pre-trained stable diffusion model v1.4, we observe a strong bias towards male output when using the prompt ”photo of the face of an electrical engineer”. Sample generated images before fine-tuning are presented in figure 5. The RLAIF pipeline based on the R_{shift} reward function is then used with the underrepresented gender being female. The loss and average reward is plotted in figure 6 while sample output after training are presented in 7. The maximum reward is reached in only eight steps, which means that all generated images now represent females. This shows that our RLAIF pipeline is able to effectively affect the gender balance of stable diffusion output, without altering image quality. However, a single step yields a shift of almost 12% which is too consequent to hope stopping at 50% exactly. Therefore we pursue our experiments with our second reward function, $R_{balance}$.

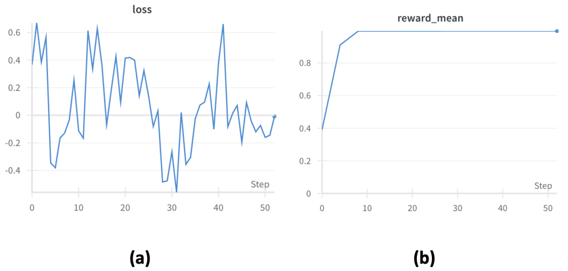


Fig. 6: Rewards and loss during training, using “photo of the face of an electrical engineer” following single face-prompt and RLAIF using R_{shift}.

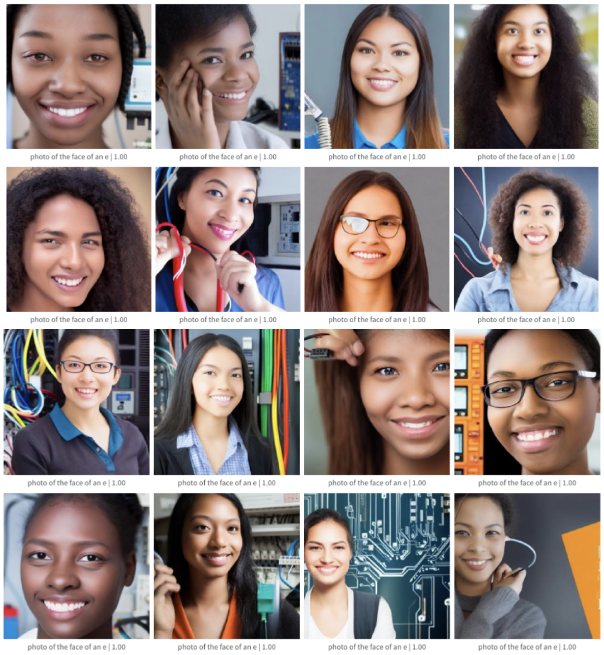


photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00
photo of the face of an e | 1.00

Fig. 7: Images generated using ”photo of the face of an electrical engineer” following single face-prompt and RLAIF with R_{shift} is used to finetune the stable diffusion model. All the images represent females by the end of the training, without loss of image quality.

C. Reaching Gender Balance

We evaluate our RLAIF pipeline based the reward $R_{balance}$ with on a prompt for which we see an initial bias: ”photo of the face of a mechanical engineer”. The loss and mean reward are presented in figure 8. At the starting point, the mean reward is less than -0.25 which indicates a ratio of females to males of less than 20 %. We then see a steep progression of the reward during the first 12 steps, reaching almost exact gender balance. Figure 9 shows the generated images after fine-tuning. On the 16 images present, 8 are females.

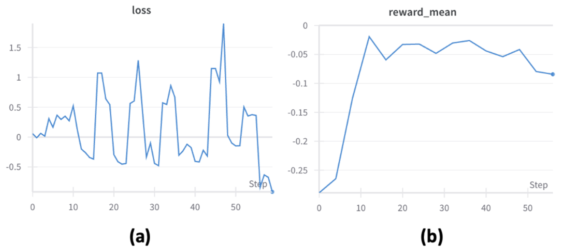


Fig. 8: Rewards and loss during training, using “photo of the face of a mechanical engineer” following single face-prompt and RLAIF using $R_{balance}$.

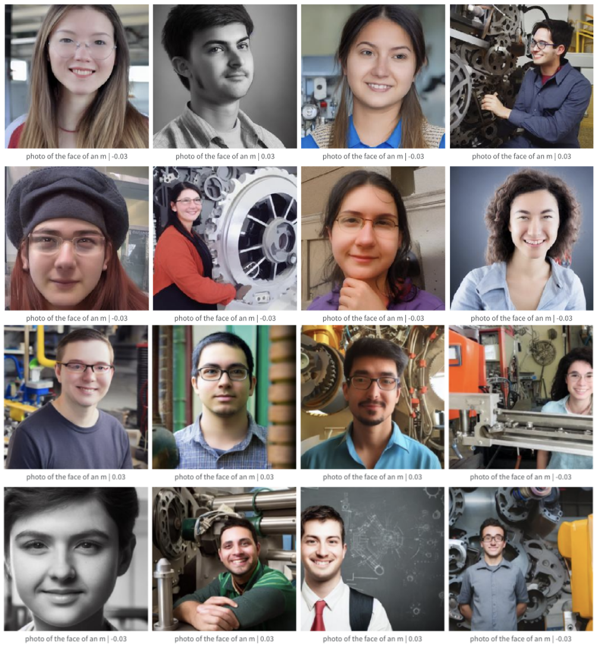


photo of the face of an m | -0.03
photo of the face of an m | 0.03
photo of the face of an m | -0.03
photo of the face of an m | 0.03
photo of the face of an m | -0.03
photo of the face of an m | -0.03
photo of the face of an m | -0.03
photo of the face of an m | -0.03
photo of the face of an m | 0.03
photo of the face of an m | 0.03
photo of the face of an m | 0.03
photo of the face of an m | -0.03
photo of the face of an m | -0.03
photo of the face of an m | 0.03
photo of the face of an m | 0.03
photo of the face of an m | -0.03

Fig. 9: Images generated using "photo of the face of a mechanical engineer" following single face-prompt and RLAIF with $R_{balance}$ is used to finetune the stable diffusion model.

To validate these result, we conduct an additional experiment with a different prompt: "photo of the face of a computer scientist". The loss and reward evolution during training are provided in figure 10. Again, a steep increase of the reward can be observed at the beginbing. Despite a downward spike at the step 20, the process is able to recover and become stable close to gender balance. Images generated after fine-tuning are presented in figure 11, illustrating a gender-balanced generation.

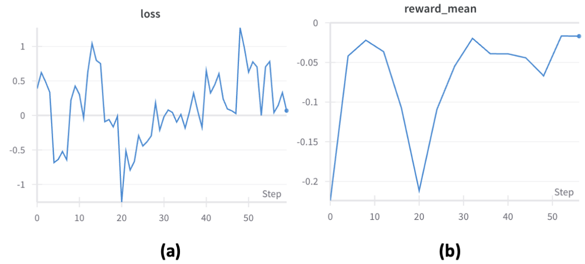


Fig. 10: Loss and reward during training, using "photo of the face of a computer scientist" following single face-prompt and RLAIF using $R_{balance}$.

In addition, we noticed that, although subjectively, by the end of the training process, the generated images tend to be more neutral in gender. Whether the fine-tuned model can generate more balanced data set or generate neutral images is worthy of further investigation.

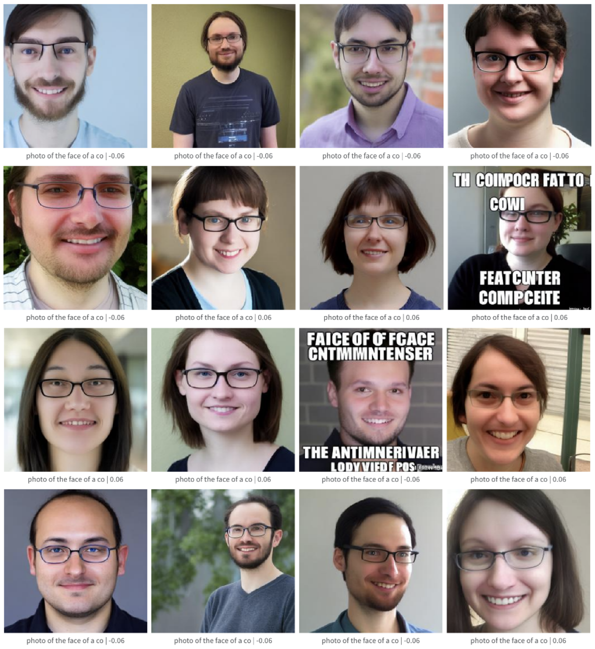
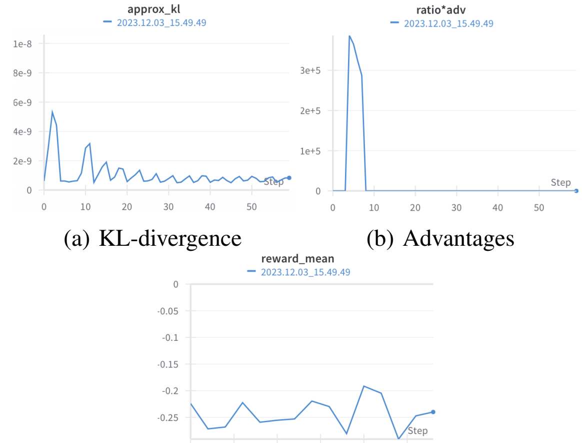


photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | 0.06
photo of the face of a co | 0.06
TH COIMPOCR FAT TO
COWI
FEATCUNTER
COMPCEITE
photo of the face of a co | 0.06
photo of the face of a co | 0.06
photo of the face of a co | 0.06
FAICE OF OF FCACE
CNTMIMNTENSER
THE ANTIMNERIVAER
photo of the face of a co | -0.06
photo of the face of a co | 0.06
photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | -0.06
photo of the face of a co | 0.06

Fig. 11: Images generated using "photo of the face of a computer scientist" following single face-prompt and RLAIF with $R_{balance}$ is used to finetune the stable diffusion model.

D. Trust Region Experiment

As discussed in III-F, we explore truly PPO [24] as an alternative way to enforce the trust region during training updates. The results of this experiment are presented in figure 12. We observe that the fine-tuning is unable to increase the reward with this different version of trust region paradigm. Our hypothesis is that, due to the complexity of the diffusion process, our approximate estimation of the KL-divergence is not sufficiently accurate to indicate that the new distribution is out of the trust region. Indeed, we observe an extreme spike in the advantages value at step 4, while the KL-divergence stays quite low. Further works might explore ways to better estimate the KL-divergence.



(a) KL-divergence

(b) Advantages

(c) Mean Reward

Fig. 12: Truly PPO experiment results

V. Conclusions

In this research project, we explored the feasibility of mitigating stereotype bias in image generation models using a RLAIF pipeline based on DDPO. We implemented a gender classification model to automatically provide female-male ratio based on the generated images. The accuracy of the classifier reached 100%. We first showed the effectiveness of RLAIF in shifting the gender unbalance in just a few fine-tuning steps with our reward function R_{shift}. We then demonstrated the capacity of the pipeline to reach stable gender balance with the reward function $R_{balance}$. To the best of the authors' knowledge, this is the first method to greatly reduce gender bias in diffusion models without additional data, full re-training or hard prompt modification. We also explored the possibility of stabilizing more the fine-tune process using an alternative trust region constraint but the outcome was not improved.

Our findings highlight the potential of RLAIF in fine-tuning image generation models for bias reduction. Nevertheless, this work is just a starting point, and further research is needed to generalize these findings and explore the extension of our methodology to address other forms of bias and enhance the robustness of the RLAIF pipeline. Further works could extend this method to different bias such as racial or cultural ones. This work was also limited to one-prompt results due to the limitation of our hardware in generating large batches, but future works could explore further multi-prompts fine-tuning. Overall, our project contributes to the ongoing efforts in developing AI systems that prioritize fairness, inclusivity, and responsible deployment.

Declaration of Contribution

The authors declare no conflicts of interest. All the team members contribute to this project equally. Virgile proposed the reward function and Xin implemented the model. Both authors participated into debugging, revision, and report writing.

Acknowledgement

This work was conducted in the context of Sergey Levine's Deep Reinforcement Learning course at UC Berkeley. We express our gratitude to Professor Levine and the class staff for their guidance. The authors also want to thank Bakar Haghighi for participating in the proposal and milestone work, as well as his efforts in querying the classification model.

References

[1] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, "Image generation: A review," *Neural Processing Letters*, vol. 54, no. 5, pp. 4609–4646, 2022.

[2] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. 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*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1493–1504. [Online]. Available: https://doi.org/10.1145/3593013.3594095

[3] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[4] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2020.

[5] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[7] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, "Aligning text-to-image models using human feedback," *arXiv preprint arXiv:2302.12192*, 2023.

[8] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," *arXiv preprint arXiv:2305.13301*, 2023.

[9] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee, "Reinforcement learning for fine-tuning text-to-image diffusion models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=8OTPepXzeh

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Y. Yang, A. Gupta, J. Feng, P. Singhal, V. Yadav, Y. Wu, P. Natarajan, V. Hedau, and J. Joo, "Enhancing fairness in face detection in computer vision systems by demographic bias mitigation," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 813–822. [Online]. Available: https://doi.org/10.1145/3514094.3534153

[13] J. Joo and K. Kärkkäinen, "Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation," in *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, ser. FATE/MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–5. [Online]. Available: https://doi.org/10.1145/3422841.3423533

[14] R. Yao, Z. Cui, X. Li, and L. Gu, "Improving fairness in image classification via sketching," 2022.

[15] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[16] C.-Y. Chuang, V. Jampani, Y. Li, A. Torralba, and S. Jegelka, "Debiasing vision-language models via biased prompts," *arXiv preprint arXiv:2302.00070*, 2023.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[20] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023.

[23] "gender-classification-2," https://huggingface.co/rizvandwiki/gender-classification-2, accessed: 2013-12-11.

[24] Y. Wang, H. He, and X. Tan, "Truly proximal policy optimization," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 113–122.