LLM-Guided Counterfactual Data Generation for Fairer AI

Ashish Mishra
ashish.mishra@hpe.com
Hewlett Packard Labs
Bangalore, India

Gyanaranjan Nayak
gyanaranjan.nayak@hpe.com
Hewlett Packard Labs
Bangalore, India

Suparna Bhattacharya
suparna.bhattacharya@hpe.com
Hewlett Packard Labs
Bangalore, India

Tarun Kumar
tarun.kumar2@hpe.com
Hewlett Packard Labs
Bangalore, India

Arpit Shah
arpit.shah@hpe.com
Hewlett Packard Labs
Bangalore, India

Martin Foltin
martin.foltin@hpe.com
Hewlett Packard Labs
Fort Collins, USA

Reviewed on OpenReview:

Abstract

With the widespread adoption of Deep Learning-based models in practical applications, concerns about their fairness have become increasingly prominent. Existing research indicates that both the model itself and the datasets on which they are trained can contribute to unfair decisions. In this paper, we address the data-related aspect of the problem, aiming to enhance the data to guide the model towards greater trustworthiness. Due to their uncontrolled curation and limited understanding of fairness drivers, real-world datasets pose challenges in eliminating unfairness. Recent findings highlight the potential of Foundation Models in generating substantial datasets. We leverage these foundation models in conjunction with state-of-the-art explainability and fairness platforms to generate counterfactual examples. These examples are used to augment the existing dataset, resulting in a more fair learning model. Our experiments were conducted on the CelebA and UTKface datasets, where we assessed the quality of generated counterfactual data using various bias-related metrics. We observed improvements in bias mitigation across several protected attributes in the fine-tuned model when utilizing counterfactual data.

Keywords: trustworthiness, explainability, counterfactual, fairness

1 Introduction

In the realm of artificial intelligence, the quest for fairness in machine learning algorithms has become paramount, particularly in domains like computer vision, where models directly interact with sensitive attributes such as gender, race, and age. One of the most pervasive challenges in this pursuit is the presence of bias, often ingrained in the data used for training these models, leading to unfair predictions and perpetuating societal disparities.

In facial image analysis, commonly using deep learning neural network models for tasks like gender and age prediction, bias can manifest in various forms. Imbalances in the representation of certain demographic groups or correlations between facial attributes and target labels can introduce biases. These biases not only compromise model accuracy but also pose ethical concerns, potentially reinforcing stereotypes and discrimination.

To address these concerns, researchers have devoted significant effort to developing techniques for bias mitigation in deep learning models. Early approaches focused on algorithmic interventions such as reweighting training samples or modifying loss functions to penalize biased predictions. While these methods showed promise in certain scenarios, they often lacked interpretability and struggled to effectively capture complex interactions between image attributes and target labels.

Recent advancements in the field have spurred the exploration of more sophisticated approaches, including the use of counterfactual generation methods. The idea behind counterfactual generation is to create alternative instances that preserve the semantics of the original data while altering specific attributes that may contribute to bias. By generating counterfactual examples based on textual descriptions of facial attributes such as hairstyle, makeup, smile, and accessories, researchers aim to mitigate bias in the underlying models and promote fairness in predictions.

However, despite the potential of counterfactual generation methods, several challenges and limitations persist. Existing approaches often struggle to generate diverse and realistic counterfactual examples, particularly in complex high-dimensional spaces such as facial images. Moreover, the impact of counterfactual data on model fairness and generalization remains understudied, necessitating further exploration and refinement of these techniques.

In this paper, we present a novel approach for mitigating bias in facial image analysis models. We leverage LLMs and pre trained text-to-image generative models to achieve this. Our methodology involves an iterative process guided by a comprehensive prompt, encapsulating key information about the target task (e.g., age prediction), model performance metrics, bias-related indicators (e.g., disparity impact, equal opportunity difference), and protected attributes (e.g., gray hair). The prompt incorporates importance scores assigned by LIME (local interpretable model-agnostic explanations) Ribeiro et al. (2016) and fairness performance from AIF-360 Bellamy et al. (2018), offering insights into attributes contributing to bias in the model's predictions.

Building upon rich contextual information, we interact with the LLM through queries to unveil the reasons behind observed bias in the model's predictions. Through deep analysis of the prompt and LLM insights, we craft textual descriptions as blueprints for generating counterfactual images. These are created using pre-trained text-to-image generative models, translating textual descriptions into visually realistic images while modifying attributes contributing to bias in the original predictions.

By leveraging LLMs and text-to-image generative models, our approach provides a data-driven, interpretable framework for bias mitigation in facial image analysis. Generating counterfactual examples guided by comprehensive prompts and explainability techniques, we aim not only to identify and understand bias sources in AI models but also to actively intervene, promoting fairness and equity in AI-driven decision-making.

Our contribution is summarized as follows:

- We conduct a multi-metric quantification of bias inherent in deep learning-based classification models, establishing a semantic connection between the observed bias and its origins within the training data.
- To address bias, we propose a pipeline using a pre-trained LLM guided by fairness metrics and image classification explanations for semantic image attributes. We create query templates through iterative interactions with the LLM, obtaining descriptions for counterfactual images and generating corresponding images with a diffusion model.
- We use these synthetic counterfactual images for fine-tuning the original model. Our experimental results show a substantial improvement in the model's trustworthiness without compromising its accuracy.
- Our pipeline integrates state-of-the-art explainability measures and trustworthiness assessment platforms, presenting a flexible solution applicable to any deep-learning-based classifier within the computer vision domain.

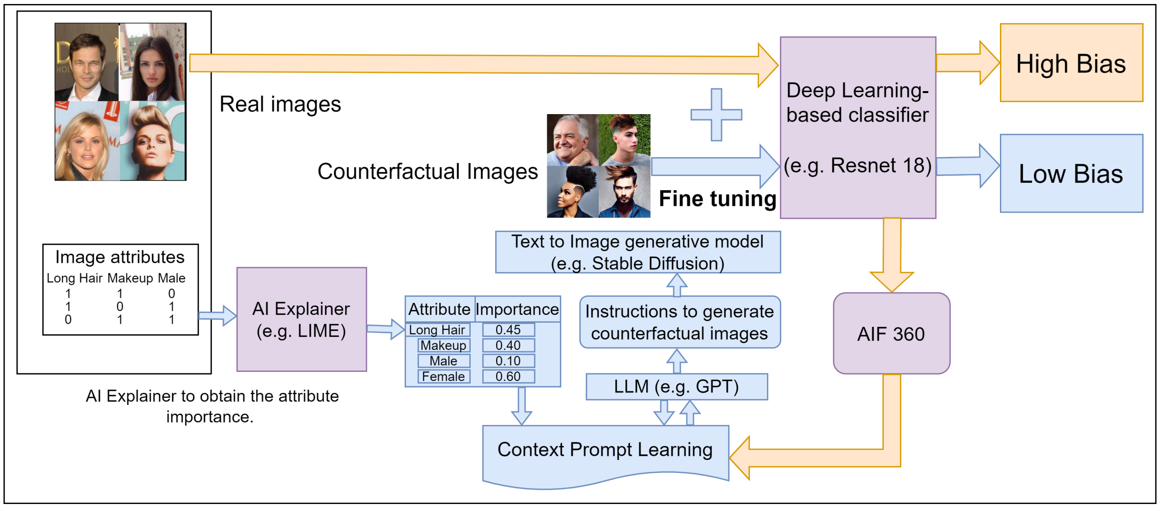

Figure 1: **Overview of task-specific automatic counterfactual example generation:** This diagram shows gender classification using the CelebA image dataset. ResNet18 was trained on these images, and AIF-360 generated bias metrics for model predictions. AI-Explainer provided importance scores for image attributes on the parallel structure dataset. We analyze the bias metrics and explanations to generate counter-data samples using LLM.

2 Related Work

Numerous methods have been suggested to enhance the reliability of model training and alleviate bias in models previously trained solely on real datasets Delaney et al. (2023);

Kim et al. (2023); Rodriguez et al. (2021); Wu et al. (2019); Thiagarajan et al. (2021); Tian et al. (2022). In this section, we provide a brief overview of existing work in this domain.

Adversarial perturbation, a key strategy for counterfactual generation and bias mitigation, involves subtly modifying input data to influence model predictions while preserving visual context Nemirovsky et al. (2022). Recent studies like Yang et al. (2021) explore incorporating adversarial perturbation in the counterfactual generation process for image datasets. These methods strategically perturb input features to prompt the model to generate counterfactual examples that address identified biases.

Text-to-image generation for counterfactual scenarios using models like GANs or diffusion models is a dynamic research area Jeanneret et al. (2024); Wei et al. (2022); Kim et al. (2023). Recent studies, such as Rombach et al. (2022), use these models to generate counterfactual images, translating textual prompts into realistic representations of alternative scenarios. Jeanneret et al. (2024) introduces TIME, a black-box technique using distillation to generate Counterfactual Explanations (CEs) for altering classifier predictions with minimal feature modifications. The Wu et al. (2019) paper enhances counterfactual fairness in classifier construction, validated with experiments on synthetic and real-world data. The Wei et al. (2022) paper introduces the Counterfactual Matching (CFM) framework to improve image-text matching by optimizing for causal effects. The Kim et al. (2023) paper presents CounTEX, a method for generating concept-based explanations for image classifiers without manual annotations, reducing human biases. Several recent bias mitigation approaches aligned with our work have emerged, and tested on the CelebA dataset Dash et al. (2022); Karkkainen and Joo (2021); Savani et al. (2020); Chen et al. (2023).

Our approach differs from existing methods by leveraging the analysis capability of Large Language Models (LLM) based on AI explainability observations and other bias metrics. By utilizing prompts, we aim to generate LLM-guided text that captures sufficient semantic information, such as attributes contributing to model bias. This text is then used to generate counterfactual images, aiding in reducing model bias and fostering fairer learning models.

3 Proposed Approach

Problem Definition

A dataset, denoted as $D = \{(x_1, y1), ..., (x_n, y_n)\}$, consists of pairs where $x_i \in X$ represents feature representations and $y_i \in Y$ represents target labels. A classification model, parameterized by θ, is denoted as F_θ defined as $F_\theta : \mathbb{R}^d \to Y$, where $\mathbb{R}^d$ is the input data space and Y is the model prediction space. It takes an input image x_i and generates class probabilities $\hat{y_i} = F_\theta(x_i)$, where $\hat{y_i}$ is a vector of predicted probabilities for each class.

Further, the dataset D is characterized by n different attributes, represented as $A = \{a_1, a_2, ..., a_n\}$. The classification model F_θ is evaluated using bias metrics M for chosen protected attributes a from A. The set of bias metrics utilized for model evaluation is identified as $M = \{m_1, m_2, ...m_k\}$, such as disparate impact (DI), equalized odds (EOD), and other metrics. Our objective is to train the classification model F_θ for a target task T in a manner that ensures high overall accuracy while minimizing bias across the various attributes. We achieve this using a pre-trained Large language model(LLM), and a text-to-image diffusion-based model, labeled as G_ϕ.

3.1 Solution Flow

Fig 1 and Algorithm 15 summarize our proposed solution. We describe the key steps below.

3.2 Compute fairness metrics and attribute importance for target task

Initially, we analyze the model's performance using feature attribution techniques and bias metrics, guiding the generation of counterfactual examples for mitigation. LIME Tabular explainer Ribeiro et al. (2016) on the parallel structure dataset provides local linear approximations of the model's behavior. Aggregating these explanations yields global importance scores for each image attribute in target class prediction. Bias metrics for the image classification model are computed for each protected variable, establishing connections between bias and responsible attributes. AIF-360 Bellamy et al. (2018) is used to assess bias. Fairness metrics and feature importance scores guide the direction for generating counterfactuals to mitigate bias. These indicators serve as trustworthy anchors in decision-making, steering the entire system toward achieving a fairer model.

3.3 Generate a textual counterfactual description using an LLM as an analyzer

Next, we utilize the text generation abilities of a pre-trained Large Language Model (LLM) by providing it with a carefully constructed prompt derived from detailed observations obtained through LIME and AIF-360. This prompt includes feature importance scores assigned by the classifier, bias metrics related to protected attributes, and overall and class-wise accuracy for the target task. We use these observations as context for the LLM, treating it as an analyzer. Through interactive questioning, our system aims to understand the reasons behind the bias in the classifier and identify neglected features. Based on this analysis, it automatically generates text to produce counterfactual examples (images) that can help mitigate bias within the classifier and guide the training of a fair model. The prompt, denoted as P, encompasses various components crucial for our inquiry, including the target task, model accuracy Acc, bias metrics M, protected attribute P_{att}, dataset attributes A, and corresponding importance scores S. In summary, $P = \{T, Acc, M, P_{\text{att}}, A, S\}$. The LLM analyzer generates textual descriptions to aid in generating counterfactual examples for mitigating bias from model F_θ as follows: $t_i = LLM(P, Q)$, where t_i represents the text for generating a counterfactual for the i^{th} class. Q comprises preset queries utilized to interact with the LLM through the prompt P. The detailed description about the prompt creation and interaction with the LLM is provided in the Appendix A.

3.4 Generate counterfactual images based on generated text

After obtaining textual descriptions from the LLM analyzer, we move on to generate counterfactual examples (images) using a pre-trained text-to-images generative model, such as a diffusion-based generative model. We input the generated text from the LLM into the pre-trained text-to-image stable diffusion model. The model then produces images that serve as counterfactual data, aiding in mitigating bias in our classifier F_θ.

The process can be summarized by the equation: $\hat{x_i} = G_\phi(t_i, t_n)$, where G_ϕ represents the pretrained text-to-image generation model and t_n is the negative prompt. For Example, the LLM generates the following text for generating counterfactual images for the male class:

”Generate an image with minimal changes, reducing the prominence of GrayHair, Double Chin, and Chubby attributes while maintaining the overall appearance as male.”

3.5 Fine-tune simple classifier with counterfactual data

We utilize the counterfactual examples generated in the previous step to fine-tune the classifier F_θ’ for bias mitigation. Following the fine-tuning process, we evaluate our refined classifier $F_{\hat{\theta}}$ in terms of bias metrics for the same protected attributes to make sure that the generated counterfactual examples effectively contribute to mitigating bias in the classifier. Let’s consider $\hat{D} = \{(\hat{x_i}, y_i)\}_{i=1}^{n}$ as the collection of generated counterfactual images utilized to refine the pre-trained model F_θ. Post-fine-tuning the model F_θ with the counterfactual dataset $\hat{D}$, the revised iteration of the model is denoted as $F_{\hat{\theta}}$, exhibiting diminished bias towards image attributes in making conclusive predictions for the target task.

Algorithm 1 Counterfactual images generation using LLM guided text

```
procedure COUNTERFACTUAL IMAGE GENERATION
    $D \leftarrow$ original Image dataset
    model $F_\theta \leftarrow$ train($D$)
    bias $\leftarrow$ fairness metric using model $F_\theta$ in AIF360
    $Exp_i \leftarrow$ Feature importance using LIME tabular ex-plainer
    LLM counterfactual prompt $P \leftarrow$ based on $F_\theta$, bias, $Exp_i$    ▷ To improve fairness
    Initiate Generative model $G$
    while bias $\neq$ Ideal_fair_range do
        $\hat{D} \leftarrow$ generate counterfactual using $G$ and $P$
        $D \leftarrow D + \hat{D}$
        model $F_\theta \leftarrow$ train($D$)
        bias $\leftarrow$ fairness metric using model $F_\theta$ in AIF360
        $Exp_i \leftarrow$ Feature importance using LIME tabular ex-plainer
        LLM counterfactual prompt $P \leftarrow$ based on $F_\theta$, bias, $Exp_i$
    end while
end procedure
```

4 Experimental Results

4.1 Experimental Settings

Datasets and Evaluation Metrics

We evaluated our proposed method on the CelebA Liu et al. (2015) and UTKface Yang and Hairong (2017) datasets, which contain human face images. The CelebA dataset classifies each image into forty different attributes, including gray hair, male, pale skin, etc. The UTKface dataset categorizes images based on three facial attributes: age, race, and gender. The facial attributes serve as protected attributes in assessing the model for potential bias towards these protected attributes in target-level prediction.

	Model: Resnet18		Model: MobileNet		
Prediction attribute:	Male	Young	Male	Young	
Protected attribute:	Gray-hair	Pale-skin	Gray-hair	Pale-skin	
CA (optimal value:1)	**0.9743**	**0.8938**	**0.8367**	**0.894**	Real Data
	0.9634	0.8712	0.785	0.871	Real + Counterfactual Data
BCA (optimal value:1)	0.9723	0.8209	**0.858**	0.8209	Real Data
	0.9744	**0.8217**	0.8125	**0.8217**	Real + Counterfactual Data
DI (optimal value:1)	0.2703	0.5077	0.192	0.508	Real Data
	1.3550	**0.6075**	**0.2745**	**0.608**	Real + Counterfactual Data
SPD (optimal value:0)	-0.4564	-0.0961	-0.367	-0.0961	Real Data
	0.1999	**-0.0932**	**-0.299**	**-0.093**	Real + Counterfactual Data
EOD (optimal value:0)	-0.0118	-0.1191	-0.245	-0.119	Real Data
	0.0063	**-0.1071**	**-0.140**	**-0.107**	Real + Counterfactual Data
AOD (optimal value:0)	**-0.0232**	**-0.0711**	-0.125	**-0.0711**	Real Data
	0.03097	-0.0715	**-0.0669**	-0.0715	Real + Counterfactual Data

Table 1: Comparison of overall performance between predictions for male and young attributes versus gray hair and pale skin as protected attributes, both without and with the utilization of counterfactuals in model training, using the CelebA dataset.

UTK dataset	CA	BCA	DI	SPD	EOD	AOD
Model+Real data	0.8616	0.8601	0.8626	-0.0619	0.09837	0.010967
Model+real+counterfactual data	0.8639	0.86174	0.8888	-0.05167	-0.04817	-0.03698

Table 2: Prediction attribute: Male, protected attribute: Race

CelebA dataset	CA	BCA	DI	SPD	EOD	AOD
Model+Manual text CF	0.8433	0.81491	0.5556	-0.12503	-0.15467	-0.10416
Model+LLM text CF	0.8712	0.8217	0.6075	-0.0932	-0.1071	-0.0715

Table 3: Prediction attribute: Young, protected attribute: Pale-skin. Comparison of counterfactual data generated through manually crafted versus LLM generated text.

4.2 Comparison

To showcase the effectiveness of our counterfactual image generation method based on textual descriptions, we evaluate its performance across various bias-related metrics: Balanced Classification Accuracy (BCA), Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odd Difference (AOD), and Overall Class Accuracy (CA). Additionally, we assess the quality of the generated counterfactual examples by comparing them with models trained solely on real data and those fine-tuned using counterfactual examples. Emphasizing the importance of LLM-generated textual descriptions and capturing key factors for bias mitigation, we create counterfactual images using manually crafted text and compare them with those generated using LLM-generated text 1.

In Table 1, concerning the CelebA dataset, we conducted experiments on two target prediction attributes: male and young. To gauge existing bias, we assessed the classification models, ResNet and MobileNet, for two protected attributes: gray hair and pale skin. It's evident from the table that the fine-tuned model, utilizing generated counterfactual data, effectively mitigates existing bias across all metrics without significantly sacrificing overall accuracy. Likewise, in Table 2, we present observations from the UTKface dataset for the ResNet18 classifier regarding male prediction and race as the protected attribute. The table also illustrates the impact of counterfactual generated data on fine-tuning the model. Table 3 vividly demonstrates the efficacy of counterfactual images generated by the LLM text in mitigating existing bias. In contrast, counterfactual images generated using manually crafted text fail to encompass significant information crucial for bias reduction.

5 Ablation Study

Figure 2: Counterfactual image generation by (l) LLM vs (r) manually crafted text

Figure 2 illustrates the effectiveness of text generated by the LLM for generating counterfactual images based on the prompt. The LLM analyzes the model's performance and potential reasons for bias, ensuring that the generated text encompasses all necessary attributes and information for effective counterfactual images. In contrast, counterfactual images generated with manually curated text lack crucial information about bias attributes, appearing simple and easy to learn. Counterfactual images guided by LLM-generated text serve as more boundary examples between male and female (for gender classification), aiding the model in learning better discriminative features between classes. Figure 3 depicts the part-wise contributions towards the final decision made by the model regarding the target class prediction. Despite the true class being female, the model trained solely on real data erroneously predicts it as male. However, the updated model, trained using counterfactual data, can correctly predict it as female. This discrepancy arises because the counterfactual examples capture sufficient boundary case instances, providing a better understanding and more discriminative features to decide such cases accurately. We made an intriguing observation: fine-tuning the model with counterfactual images sometimes caused the model

(a) W/O fine-tuning: male, With fine-tuning: female, True label: female, CS:0.50701

(b) W/O fine-tuning: male, With fine-tuning: female, True label: male, CS:0.5147

Figure 3: Comparison of flipped examples in real images after fine-tuning using counterfactual data.

to switch its prediction of target labels (Male, Female) in certain examples. However, this phenomenon was most commonly observed in boundary cases. For instance, when a male face exhibits some characteristics traditionally associated with females, such as long hair, makeup, or earrings, and vice versa for a female face, if the fine-tuned model incorrectly predicts the label but its confidence score for the prediction falls near the boundary line, it indicates a wrong prediction with low confidence. Conversely, the model without fine-tuning tends to make incorrect predictions with high confidence scores.

6 Conclusions, Limitations and Future Work

Our early findings demonstrate the promise of using open-source LLMs paired with diffusion models as powerful de-biasers when equipped with existing bias detection and interpretability tools, which allows them to diagnose what to fix and generate counterfactuals based on semantic connections between the observed metrics and their internal knowledge base.

Increasing the complexity of the overall process is evident, encompassing enhancements to the LLM analyzer, observations on model explainability, and text-to-image generation, followed by fine-tuning the model for counterfactual data. However, our aim is to streamline complexity by leveraging the capabilities of pre-trained LLMs and text-to-image generative models. In our ongoing research, we conducted experiments for two distinct tasks: gender classification and age prediction (young or old). However, our proposed approach is more versatile and can be easily applied to other tasks as well, as discussed in the limitations and future work section. The primary risk associated with utilizing pre-trained models lies in the potential for bias stemming from various factors, such as the training data and context. In our research, we strive to mitigate this bias in pre-trained LLMs and text-to-image generation by offering comprehensive observations through bias metrics and importance scores assigned by the model to final predictions. Additionally, we establish a well-defined context for the LLM to analyze data in a fair and unbiased manner. To further address intrinsic biases in image generation, we provide negative prompts to the pre-trained model, which aids in avoiding bias and undesirable examples. We are specifically using it as a reasoning engine to interpret the results of AI bias measurement and explainability tools, but when generating the examples it may add other aspects that could be new biases, but we are still measuring the effect after that iteratively.

Our current implementation uses an interpretability method that relies on the availability of semantic attribute labels (such as grey hair, makeup etc) associated with training and test images. When such annotations are not available in real-world datasets, we are exploring ways to extract them automatically from images or captions. Additionally, vision LLMs may provide opportunities to use image-based interpretability techniques directly in the input passed to the LLM. Another area of exploration is whether the LLM analyzer can also estimate the number of counterfactuals that should be generated and the proportion in which they should be mixed with original images to improve bias without losing accuracy. Finally, an interesting research question is whether these techniques could also be applied to mitigate bias in other types of datasets beyond images.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge the valuable contributions of the members of the Hewlett Packard lab who provided valuable insights and support throughout the course of this research. Their expertise and collaboration greatly enriched our work and contributed to its success. We extend our sincere appreciation to the entire Hewlett Packard lab team for their dedication and assistance.

References

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018. URL `https://arxiv.org/abs/1810.01943`.

R. Chen, J. Yang, H. Xiong, J. Bai, T. Hu, J. Hao, Y. Feng, J. T. Zhou, J. Wu, and Z. Liu. Fast model debias with machine unlearning. *arXiv preprint arXiv:2310.12560*, 2023.

S. Dash, V. N. Balasubramanian, and A. Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 915–924, 2022.

E. Delaney, A. Pakrashi, D. Greene, and M. T. Keane. Counterfactual explanations for misclassified images: How human and machine explanations differ. *Artificial Intelligence*, 324:103995, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.103995. URL `https://www.sciencedirect.com/science/article/pii/S0004370223001418`.

G. Jeanneret, L. Simon, and F. Jurie. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024.

K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

S. Kim, J. Oh, S. Lee, S. Yu, J. Do, and T. Taghavi. Grounding counterfactual explanation of image classifier to textual concept space. In *CVPR 2023*, 2023. URL `https://www.amazon.science/publications/grounding-counterfactual-explanation-of-image-classifier-to-textual-concept-space`.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

D. Nemirovsky, N. Thiebaut, Y. Xu, and A. Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1488–1497. PMLR, 01–05 Aug 2022. URL `https://proceedings.mlr.press/v180/nemirovsky22a.html`.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

P. Rodriguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, and D. Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. *arXiv preprint arXiv:2103.10226*, 2021.

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*, pages 10684–10695, 2022.

Y. Savani, C. White, and N. S. Govindarajulu. Intra-processing methods for debiasing neural networks. *Advances in neural information processing systems*, 33:2798–2810, 2020.

J. Thiagarajan, V. S. Narayanaswamy, D. Rajan, J. Liang, A. Chaudhari, and A. Spanias. Designing counterfactual generators using deep model inversion. *Advances in Neural Information Processing Systems*, 34:16873–16884, 2021.

H. Tian, T. Zhu, W. Liu, and W. Zhou. Image fairness in deep learning: problems, models, and challenges. *Neural Computing and Applications*, 34, 08 2022. doi: 10.1007/s00521-022-07136-1.

H. Wei, S. Wang, X. Han, Z. Xue, B. Ma, X. Wei, and X. Wei. Synthesizing counterfactual samples for effective image-text matching. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4355–4364, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547814. URL `https://doi.org/10.1145/3503161.3547814`.

Y. Wu, L. Zhang, and X. Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1438–1444. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/199. URL `https://doi.org/10.24963/ijcai.2019/199`.

F. Yang, N. Liu, M. Du, and X. Hu. Generative counterfactuals for neural networks via attribute-informed perturbation, 2021.

Z. Z. S. Yang and Q. Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Appendix A. Example of LLM prompt for generating counterfactual Images

You are an excellent analyzer. We will provide you with a task description and the model used. Also, we will provide the results in detail with the importance weights given by the model to predict the target goal. The importance weights are given the dataset attributes. Your job is to analyze the results and find the reasons to show the specific types of observations. For this reason, we aim to mitigate that using the generated counterfactual examples based on the textual description. Please create a short textual description after analyzing the reasons such that we use the textual description to generate the counterexamples. Including these examples to retrain the model can mitigate the bias or show fair learning. The details are given below: Task: The task is to gender classification from an image dataset without any bias towards any attributes. The attributes for the image dataset are also given: 15

Task: gender classification for face image

fairness metrics: Classification accuracy = 0.944444 Test set: Balanced classification accuracy = 0.892716 Test set: Statistical parity difference = 0.055639 Test set: Disparate impact = 1.068707 Test set: Equal opportunity difference = 0.014162 Test set: Average odds difference = -0.066126 Test set: Theil index = 0.027163 Test set: False negative rate difference = -0.014162

LIME Explanations importance scores for attributes: Gray_Hair:0.418, Double_Chin:0.252, Attractive:-0.158, Chubby:0.140, Bald:0.115, Eyeglasses:0.0797, Wearing_Necktie:0.0760, Black_Hair:-0.061,
Big_Nose:0.0577, Bags_Under_Eyes:-0.0572.

LLM prompt for generating counterfactual images:
Male:
Generate an image with minimal changes, reducing the prominence of Gray-Hair, Double-Chin, and Chubby attributes while maintaining the overall appearance as male.
Female:
Create an image with slight adjustments to reduce the emphasis on Attractive, Black-Hair, and Bags-Under-Eyes attributes, ensuring that the modified image remains convincingly female.