GAMMA-FACE: GAussian Mixture Models Amend Diffusion Models for Bias Mitigation in Face Images

JOHNS HOPKINS UNIVERSITY — THE UNIVERSITY OF TEXAS AT DALLAS

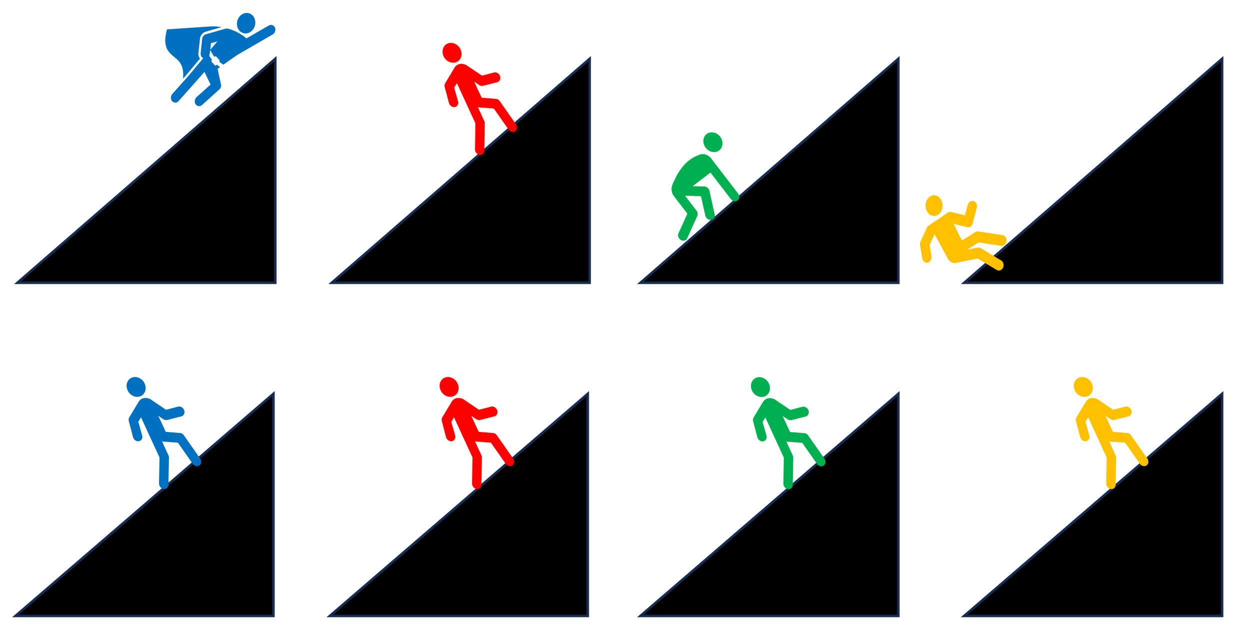
Basudha Pal[1*], Arunkumar Kannan[1*], Ram Prabhakar[1], Alice J.O'Toole[2], and Rama Chellappa[1]

[1] Johns Hopkins University [2] The University of Texas at Dallas

* Indicates equal contribution

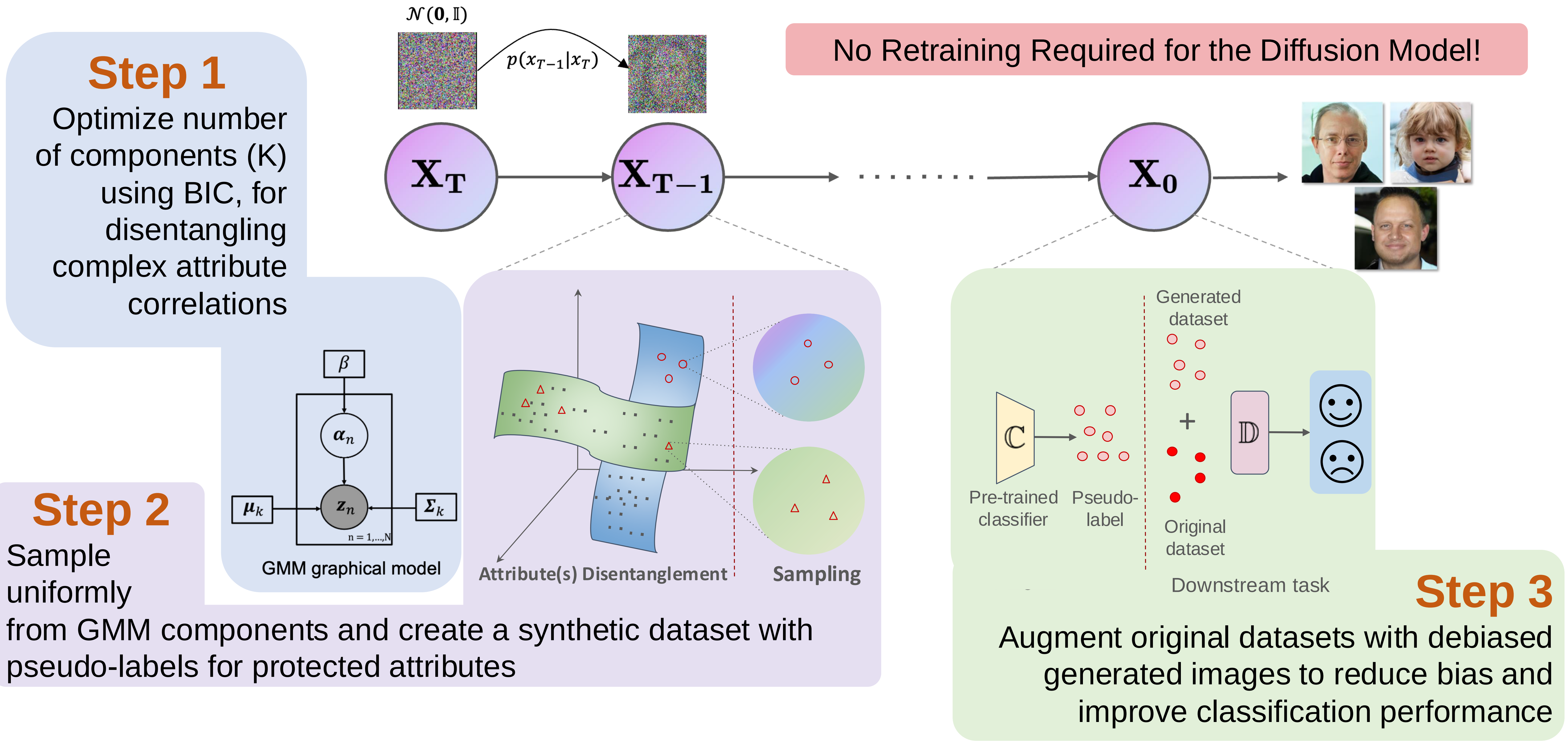
Motivation : Bias in Unconditional Diffusion Model

- Deep learning models trained on sensitive data often show demographic biases, raising fairness concerns, particularly with limited datasets.
- Diffusion models amplify bias[1] - excel in image generation but challenging to use generated images in downstream tasks due to amplified biases.
- Proposed solution - a novel yet simple technique, GAMMA-FACE to debias the attributes in the images generated by unconditional diffusion models.
- Utilized Gaussian Mixture Models (GMMs) to disentangle the attributes in the latent space of diffusion models.



Pictorial analogy depicting bias in protected attributes for a same target downstream task

Our Approach: GAMMA-Face



Step 1

Optimize number of components (K) using BIC, for disentangling complex attribute correlations

Step 2

Sample uniformly from GMM components and create a synthetic dataset with pseudo-labels for protected attributes

Step 3

Augment original datasets with debiased generated images to reduce bias and improve classification performance

Quantitative Results

	FairFace								
	$A_t = g \mid A_p = a, r$			$A_t = r \mid A_p = a, g$			$A_t = a \mid A_p = r, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.187	1.36	81.67	0.237	**1.27**	78.10	0.112	1.502	78.62
[50]	0.142	1.38	84.14	0.163	1.393	79.3	**0.097**	**1.43**	80.13
[8]	0.169	1.53	82.28	0.218	1.781	75.5	0.130	1.62	76.51
Ours	**0.088**	**1.29**	**86.5**	**0.102**	1.36	**80.23**	0.128	1.510	**81.00**
	FFHQ								
	$A_t = s \mid A_p = a, g$			$A_t = h \mid A_p = a, g$			$A_t = gl \mid A_p = a, g$		
Method	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)	B (↓)	BA (↓)	Acc. (↑)
[34]	0.015	**1.48**	91.68	0.221	1.787	84.03	0.028	0.995	96.50
[50]	0.0064	1.61	93.16	0.153	1.798	**88.87**	0.031	1.008	97.29
[8]	0.019	1.77	91.58	0.192	1.84	82.11	0.040	1.156	96.10
Ours	**0.0056**	1.52	**94.84**	**0.146**	**1.756**	82.81	**0.0208**	**0.987**	**98.70**

	FairFace					
	$A_t = g \mid A_p = a, r$		$A_t = r \mid A_p = a, g$		$A_t = a \mid A_p = r, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.886	0	0.798	0	**0.769**
Ours	**0.085**	**0.801**	**0.118**	**0.740**	**0.454**	0.783
	FFHQ					
	$A_t = s \mid A_p = a, g$		$A_t = h \mid A_p = a, g$		$A_t = gl \mid A_p = a, g$	
Method	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)	BPC (↑)	KL (↓)
Baseline	0	0.782	0	0.95	0	1.814
Ours	**0.673**	**0.698**	**0.4244**	**0.912**	**0.128**	**0.918**

Bias evaluation metrics: Bias (B), Bias Amplification (BA), Overall accuracy (Acc.), Bias Performance Coefficient (BPC) and KL divergence (KL)

	FairFace						FFHQ					
	$A_t = g \mid A_p = a, r$			$A_t = r \mid A_p = a, g$			$A_t = s \mid A_p = a, g$			$A_t = h \mid A_p = a, g$		
%Gen+%Org	B	BA	Acc.	B	BA	Acc.	B	BA	Acc.	B	BA	Acc.
100	0.117	1.59	83.25	0.125	1.72	72.83	0.0204	1.928	93.16	0.181	1.98	76.58
70+30	0.1098	1.41	82.25	**0.0956**	1.541	71.30	0.119	1.75	92.18	0.216	1.824	77.98
30+70	**0.089**	**1.33**	**86.21**	0.104	**1.354**	**77.93**	**0.0044**	**1.513**	**93.85**	**0.152**	**1.76**	**81.19**

	FairFace				FFHQ			
	$A_t = g \mid A_p = a, r$		$A_t = r \mid A_p = a, g$		$A_t = s \mid A_p = a, g$		$A_t = h \mid A_p = a, g$	
%Gen+%Org	BPC	KL	BPC	KL	BPC	KL	BPC	KL
100	-0.266	1.01	-0.093	0.989	-0.243	0.95	-0.176	1.27
70+30	-0.204	1.23	-0.031	**0.85**	-0.201	0.852	-0.376	1.14
30+70	**0.117**	**0.978**	**0.055**	0.913	**0.0056**	**0.787**	**-0.0023**	**1.05**

The effect of different mixing ratios (Generated + Original) on FairFace and FFHQ

Qualitative Results

Face images generated by GAMMA-Face after localizing the image attributes in the latent space of the DDPM for *Left*: FFHQ and *Right*: FairFace datasets.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Perera, M.V. *et.al.*, Analyzing bias in diffusion-based face generation models. *IEEE IJCB* 2023

[34] Ramaswamy, V.V *et.al.* Fair attribute classification through latent space de-biasing. *IEEE/CVF CVPR* 2021

[50] Zhang, F. *et. al* Distributionally Generative Augmentation for Fair Facial Attribute Classification. *IEEE/CVF* CVPR 2024

[8] Dhar, P. *et. al.* Pass: protected attribute suppression system for mitigating bias in face recognition. *IEEE/CVF* ICCV *2021*

Webpage: https://bas-2k.github.io/gamma-face/