Latent Directions: A Simple Pathway to Bias Mitigation in Generative AI

Carolina López Olmos, Alexandros Neophytou, Sunando Sengupta, Dim P. Papadopoulos

CVPR SEATTLE, WA JUNE 17-21, 2024

Problem: lack of diversity + biased generations

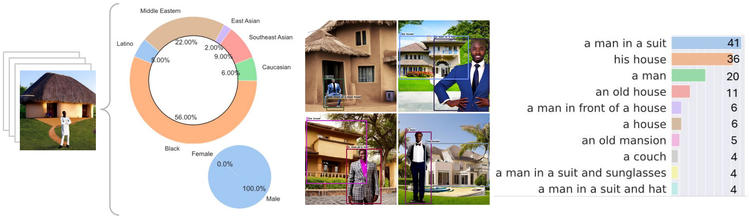
According to Stable Diffusion …

Today's **text-to-image models** leverage stereotypes.
In this work we propose a simple novel technique to achieve debiased generations without prompt modification or embedding alteration.

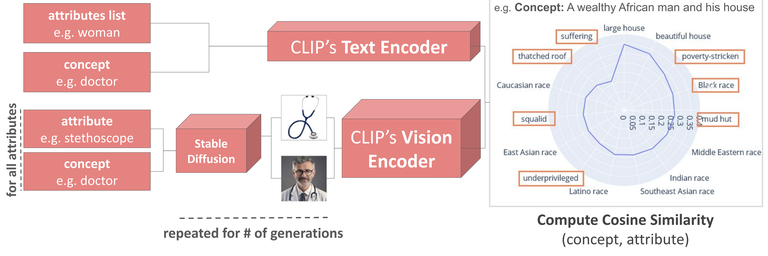
Understanding biases

To mitigate certain biases we first need to understand **if** and **why** they are present in our image generations.

IF: Detecting social characteristics and objects in the images.

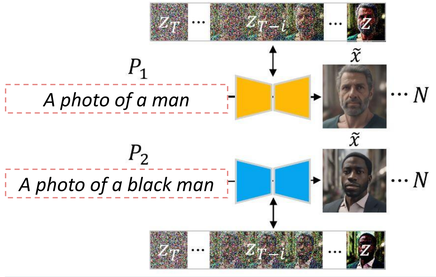


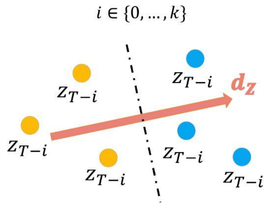
WHY: Comprehending the connections between embeddings and generations.



Method: Learning and applying latent directions

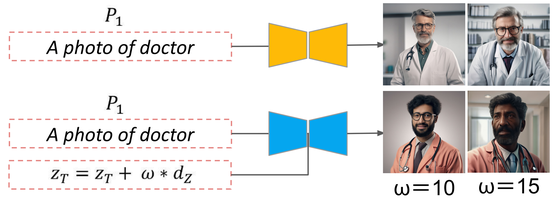
The prompts generate N images & their latents are saved for training





latent direction is learned

The latent direction is applied to the latent noise with a weight



We use a SVM classifier to linearly separate the latents across our labeled dataset. From it we obtain a latent direction, which we apply together with a neutral prompt to obtain debiased generations.

Tuning: Selecting the optimal weight and latent

The **choice of weight has a higher impact** on the debiasing than the choice of latent. Two approaches to find the optimal configuration:

➔ *clean-fid:* similarity between debiased images and small subset for every configuration.

➔ CLIP as a zero-shot classifier within the configurations.

Results

d_Z / P_1	Skin Tone		Gender			Landbird	Indian	Wealth
	Man	*Woman*	*Doctor*	*Firefighter*	*Cleaner*	*Waterbird*	*Wedding*	*African man*
(SD XL, ours)	0.87	0.78	**0.52**	**0.08**	0.09	-	0.33	0.28
(SD 2.1, PD [1])	0.91	0.90	0.14	0.06	0.01	0.29	0.79	0.47
(SD 2.1, ours + PD [1])	**0.94**	**1.00**	0.29	0.04	**0.22**	**0.68**	**1.00**	**0.96**

Table 1. Quantitative results with Statistical Parity Difference after debiasing.
PD[1][Prompt Debiasing] 1 - Chuang et al. "Debiasing Vision-Language Models via Biased Prompts." arXiv (2023).

- It is possible to alter biased relations such as those in cultural events while maintaining **unaltered embeddings** and a **neutral prompt**.
- The application of latent directions achieves **successful results debiasing diverse and complex scenarios.**

Combining latent directions

$$z_T = z_T + \sum_{i=1}^{\infty} \omega_i \cdot d_{Zi}$$

Combining debiasing methods

Latent directions can leverage embedding alteration (PD) to obtain more effective debiasing.