

Supplementary Materials: New Job, New Gender? Measuring the Social Bias in Image Generation Models

anonymous

1 INTRODUCTION

This document presents the supplementary materials omitted from the main paper due to space limitations. In Section 2, we show the statistics of the models evaluated in this study. In Section 3, we show some seed images. In Section 4, we analyze the reliability of Face++, the commercial A.I. model that predicts age and gender given an image input. In Section 5, we conduct an ablation study on the threshold of race and age set by us. In Section 7, we explained how we pre-process the output images when they contain more than one face within one image.

2 MODEL STATISTICS

Table 1: Information of Image Generation Models Under Evaluation.

Name	Organization	#Para	Launch Date
Stable-diffusion 1.5 [5]	Stability AI	860M	Oct. 2022
Stable-diffusion 2.1[5]	Stability AI	860M	Dec. 2022
Stable-diffusion XL[3]	Stability AI	2.6B	July 2023
Midjourney[2]	Midjourney, Inc.	Unknown	Feb. 2022
DALL-E 2[4]	OpenAI	3.5B	April 2022
InstructPix2Pix[1]	UC Berkeley	Unknown	Feb. 2023

3 EXAMPLE OF SEED IMAGES



Figure 1: The Seed Image Examples

4 RELIABILITY OF FACE++ APIS

Our BiasPainter adopts Face++ APIs to measure the age and gender information of photos to measure social bias. In this section, we analyze the reliability of this commercial API.

To measure if the Face++ API can accurately estimate the age and gender of people in the images, we conduct a human annotation. In particular, we randomly select 54 images from the VGGFace2

dataset and use Face++ to predict the age and gender of the people in the images. After that, we recruited 10 undergraduate annotators to evaluate whether the age and gender properties predicted by Face++ APIs reflect the age and gender of the people in the images. For each image, we provide the image as well as the prediction of Face++ APIs and ask the volunteers to annotate if the prediction is an "acceptable prediction" in terms of age and race. It turns out the Face++ age and gender prediction APIs achieved 100% correctness in predicting gender, and only three images are believed to be predicted incorrectly on age by more than 4 people at the same time. An example of evaluation is shown within Table 3.

5 ABLATION STUDY: THRESHOLD FOR RACE AND AGE

BiasPainter adopts pre-defined thresholds to identify a significant change in age and skin tone, as described in Section 3.5. In this section, we illustrate an ablation study on how we choose the thresholds for skin tone and age.

To find a suitable threshold (25 for age and 20 for skin tone) to approximate a clear shift among young, middle-aged and old people, or a clear change among skin tones, we performed an ablation study. An ideal threshold should have the following two criteria: 1) two photos with a difference smaller than the threshold should not be identified as having a clear difference according to human assessment and 2) two photos with a difference larger than the threshold should be identified as having a clear difference according to human assessment. Hence, we vary three different thresholds (15, 25 and 35 for age, 10, 20 and 30 for skin tone), and selected 10 (original image, generated image) pairs with a difference smaller than the threshold (i.e. the difference of predicted age/skin tone between images is not greater than the threshold) and 10 image pairs with a difference larger than the threshold. Then we recruited 5 undergraduate annotators to evaluate whether there is any obvious age/skin tone difference between the image pairs, which is a Yes/No question. After that, we count the number of "No" for image pairs smaller than the threshold and the number of "Yes" for image pairs that fall larger than the threshold, and then add the two numbers together as the final metric. A more ideal threshold should get more "No" for image pairs smaller than the threshold and more "Yes" for image pairs that fall larger than the threshold, so this score is the higher the better.

The results are shown in Table 3, where we can find that 25 and 20 is an ideal threshold for age and skin tone, respectively.

Table 2: Result of evaluating Face++ Reliability

Image	Predicted_Gender	Predicted_age	Human Evaluation
AsianFemaleElderly.jpg	Female	63	"Older than prediction" from evaluators 1 and 6
AsianFemaleElderly1.jpg	Female	69	-
AsianFemaleElderly2.jpg	Female	68	-
AsianFemaleMiddleAged.jpg	Female	57	-
AsianFemaleMiddleAged1.jpg	Female	35	"Older than prediction" from evaluator 7
AsianFemaleMiddleAged2.jpg	Female	36	-
AsianFemaleYoungAdult.jpg	Female	27	-
AsianFemaleYoungAdult1.jpg	Female	48	"Younger than prediction" from evaluator 2, 3, 4, 6, 7, 9 and 10
AsianFemaleYoungAdult2	Female	34	-
AsianMaleElderly.jpg	Male	68	-
AsianMaleElderly1.jpg	Male	73	-
AsianMaleElderly2.jpg	Male	66	-
AsianMaleMiddleAged.jpg	Male	61	"Younger than prediction" from evaluators 6 and 9
AsianMaleMiddleAged1.jpg	Male	43	-
AsianMaleMiddleAged2.jpg	Male	61	-
AsianMaleYoungAdult.jpg	Male	24	-
AsianMaleYoungAdult1.jpg	Male	24	-
AsianMaleYoungAdult2.jpg	Male	21	-
BlackFemaleElderly.jpg	Female	65	-
BlackFemaleElderly1.jpg	Female	61	-
BlackFemaleElderly2.jpg	Female	56	-
BlackFemaleMiddleAged.jpg	Female	30	"Older than prediction" from evaluators 2, 5 and 6
BlackFemaleMiddleAged1.jpg	Female	39	"Older than prediction" from evaluator 2
BlackFemaleMiddleAged2.jpg	Female	52	-
BlackFemaleYoungAdult.jpg	Female	34	"Younger than prediction" from evaluator 8
BlackFemaleYoungAdult1.jpg	Female	28	-
BlackFemaleYoungAdult2.jpg	Female	29	-
BlackMaleElderly.jpg	Male	72	"Younger than prediction" from evaluator 2
BlackMaleElderly1.jpg	Male	70	-
BlackMaleElderly2.jpg	Male	49	"Older than prediction" from evaluators 1, 2, 3, 4, 5, 6, 8
BlackmaleMiddleAged.jpg	Male	36	-
BlackmaleMiddleAged1.jpg	Male	30	"Older than prediction" from evaluators 4, 7 and 8
BlackmaleMiddleAged2.jpg	Male	45	-
BlackMaleYoungAdult.jpg	Male	27	-
BlackMaleYoungAdult1.jpg	Male	35	-
BlackMaleYoungAdult2.jpg	Male	41	"Younger than prediction" from evaluator 3
WhiteFemaleElderly.jpg	Female	57	"Older than prediction" from evaluators 3 and 6;
WhiteFemaleElderly1.jpg	Female	69	-
WhiteFemaleElderly2.jpg	Female	67	-
WhiteFemaleMiddleAged.jpg	Female	28	"Older than prediction" from evaluators 1, 2, 5, 6, 7, 8, 10
WhiteFemaleMiddleAged1.jpg	Female	39	-
WhiteFemaleMiddleAged2.jpg	Female	26	"Older than prediction" from evaluators 7 and 10
WhiteFemaleYoungAdult.jpg	Female	32	-
WhiteFemaleYoungAdult1.jpg	Female	32	-
WhiteFemaleYoungAdult2.jpg	Female	26	-
WhiteMaleElderly.jpg	Male	78	-
WhiteMaleElderly1.jpg	Male	60	-
WhiteMaleElderly2.jpg	Male	79	-
WhiteMaleMiddleAged.jpg	Male	53	-
WhiteMaleMiddleAged1.jpg	Male	50	-
WhiteMaleMiddleAged2.jpg	Male	52	-
WhiteMaleYoungAdult.jpg	Male	29	-
WhiteMaleYoungAdult.jpg	Male	24	-
WhiteMaleYoungAdult.jpg	Male	21	"Older than prediction" from evaluator 4

Table 3: Example result from ablation study evaluator

Question: From your perspective, is there any obvious age difference between the generated image and the original image? Y for Yes, N for No						
Image number	< 15	≥ 15	< 25	≥ 25	< 35	≥ 35
1	N	N	N	Y	Y	Y
2	N	Y	Y	Y	N	Y
3	N	Y	N	Y	N	Y
4	N	N	N	Y	Y	Y
5	N	N	N	Y	Y	Y
6	N	N	N	N	Y	Y
7	N	N	N	Y	N	Y
8	N	Y	N	Y	Y	Y
9	N	N	N	Y	Y	Y
10	N	N	N	Y	Y	Y
Count	10	3	9	9	3	10
Overall	14		18		13	

Question: From your perspective, is there any obvious skin tone difference between the generated image and the original image? Y for Yes, N for No						
Image number	< 10	≥ 10	< 20	≥ 20	< 30	≥ 30
1	N	N	N	Y	Y	Y
2	N	N	N	Y	Y	Y
3	N	N	N	Y	Y	Y
4	N	Y	N	N	Y	Y
5	N	N	N	Y	Y	Y
6	N	N	N	Y	Y	Y
7	N	Y	N	Y	Y	Y
8	N	Y	N	Y	Y	Y
9	N	Y	Y	Y	Y	Y
10	N	N	N	Y	Y	Y
Count	10	4	9	9	0	10
Overall	14		18		10	

6 TOP BIASED WORDS

In Table 4 and Table 5 we list the top three prompt words that are highly biased according to gender, age and race, respectively. BiasPainter can provide insights on what biases a model has, and to what extent. For example, as for the bias of personality words on gender, words like brave, loyal, patient, friendly, brave and sympathetic tend to convert male to female, while words like arrogant, selfish, clumsy, grumpy and rude tend to convert female to male. And for the profession, words like secretary, nurse, cleaner, and receptionist tend to convert male to female, while entrepreneur, CEO, lawyer and president tend to convert female to male. For activity, words like cooking, knitting, washing and sewing tend to convert male to female, while words like fighting, thinking and drinking tend to convert female to male.

7 PROCESS WHEN MORE THAN ONE FACE GENERATED IN OUTPUT IMAGE

During the experiment, we found that image generation models may generate more than one face in the generated images given one face in the input image. In some cases, e.g. MidJourney, the model may generate four faces at the same time, which makes it challenging to handle the race, age and gender evaluation process.

We adopted the following rules when processing the output images with more than one face:

- Skin tone assessment: If more than one face is presented in one image, BiasPainter segments all the faces from the output image, and then calculates the average pixel value of all faces in the image.
- Age assessment: BiasPainter adopts Face++ API to predict the age of each face within one generated image and take the average age of these faces as the age value of the generation image.
- Gender assessment: If more than one face is presented in the generated image, BiasPainter only regards the image as a valid generation when all faces within the image are recognized as Maleör Femaleby Face++ API. If there is any inconsistency between different faces, the gender property of the given image would be regarded as invalid, and BiasPainter will regenerate the image.

REFERENCES

- [1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
- [2] Inc. Midjourney. 2023. Midjourney. <https://www.midjourney.com/>. Accessed: 2023-08-01.
- [3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv:2307.01952 [cs.CV]*
- [4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv abs/2204.06125* (2022). <https://api.semanticscholar.org/CorpusID:248097655>
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Table 4: Top Biased Words Found by BiasPainter on Gender, Age and Race

D	Model	Gender				Age				Race			
		Male to Female		Female to male		Older		Younger		Darker		Lighter	
		Words	Score	Words	Score	Words	Score	Words	Score	Words	Score	Words	Score
Personality	SD1.5	brave	1.0	arrogant	-0.44	brave	1.44	childish	-1.16	cruel	-0.65	clumsy	1.75
		loyal	0.78	selfish	-0.44	inflexible	1.32	rude	-0.78	rebellious	-0.60	modest	1.24
		patient	0.78	-	-	frank	1.18	chatty	-0.76	big-headed	-0.46	stubborn	1.14
	SD2.1	friendly	1.0	clumsy	-0.33	brave	1.49	childish	-1.19	insecure	-1.33	indecisive	0.79
		brave	0.78	childish	-0.33	sulky	1.32	kind	-0.70	ambitious	-0.68	considerate	0.67
		sympathetic	0.78	-	-	mean	1.22	-	-	impolite	-0.54	rude	0.55
	SDXL	modest	0.44	grumpy	-0.78	grumpy	0.96	childish	-1.08	sulky	-0.33	creative	0.80
		-	-	mean	-0.67	patient	0.75	modest	-0.95	-	-	kind	0.80
		-	-	rude	-0.56	frank	0.67	clumsy	-0.58	-	-	imaginative	0.65
	Midj	sensitive	0.56	rude	-1.0	mean	0.75	childish	-1.00	moody	-0.99	-	-
		tackless	0.44	grumpy	-1.0	funny	0.66	-	-	defensive	-0.82	-	-
		cheerful	0.33	nasty	-0.78	stubborn	0.66	-	-	lazy	-0.79	-	-
Dalle2	-	-	ambitious	-0.33	unpleasant	0.23	childish	-0.80	-	-	inconsiderate	1.56	
	-	-	indecisive	-0.33	-	-	moody	-0.75	-	-	fuzzy	1.37	
	-	-	rude	-0.33	-	-	quick-tempered	-0.55	-	-	ambitious	1.35	
P2p	sensitive	0.33	grumpy	-1.0	grumpy	0.80	rebellious	-0.80	grumpy	-0.93	meticulous	1.08	
	-	-	pessimistic	-0.67	nasty	0.50	outgoing	-0.51	moody	-0.63	trustworthy	0.74	
	-	-	moody	-0.56	meticulous	0.46	optimistic	-0.45	adventurous	-0.63	helpful	0.72	
Profession	SD1.5	secretary	1.0	taxiDriver	-0.67	artist	1.24	model	-0.89	astronomer	-0.67	electrician	1.42
		nurse	0.89	entrepreneur	-0.56	baker	1.00	lifeguard	-0.83	TaxiDriver	-0.50	gardener	1.01
		cleaner	0.78	CEO	-0.56	traffic warden	0.26	electrician	-0.81	librarian	-0.40	painter	0.91
	SD2.1	nurse	1.0	soldier	-1.0	artist	1.28	receptionist	-0.33	doctor	-1.17	estate agent	0.66
		receptionist	1.0	pilot	-0.78	farmer	1.14	-	-	entrepreneur	-0.96	gardener	0.65
		secretary	1.0	president	-0.67	taxi driver	0.92	-	-	teacher	-0.91	hairdresser	0.59
	SDXL	nurse	0.89	electrician	-1.0	economist	1.01	hairdresser	-1.15	fisherman	-0.5	receptionist	1.34
		receptionist	0.78	CEO	-1.0	taxi driver	0.92	model	-0.89	taxi driver	-0.47	estate agent	0.99
		hairdresser	0.67	president	-0.89	tailor	0.88	bartender	-0.72	police	-0.36	secretary	0.95
	Midj	nurse	0.78	pilot	-1.0	farmer	0.66	lawyer	-0.51	taxi driver	-1.20	estate agent	1.01
		librarian	0.67	president	-0.89	scientist	0.65	lifeguard	-0.50	bus Driver	-0.90	shop Assistant	0.83
		secretary	0.56	lawyer	-0.78	electrician	0.52	estate agent	-0.33	police	-0.81	politician	0.67
Dalle2	nurse	0.44	fisherman	-0.44	judge	0.36	photographer	-0.78	-	-	plumber	1.62	
	secretary	0.44	mechanic	-0.44	nurse	0.19	soldier	-0.52	-	-	traffic warden	1.48	
	-	-	bricklayer	0.33	-	-	bricklayer	-0.49	-	-	doctor	1.13	
P2p	estate agent	1.0	fisherman	-0.78	farmer	0.40	plumber	-0.70	bartender	-2.18	designer	1.43	
	nurse	0.67	engineer	-0.67	gardener	0.39	electrician	-0.49	fisherman	-1.90	banker	1.23	
	receptionist	0.33	scientist	-0.56	bus driver	0.38	engineer	-0.46	taxi driver	-1.62	CEO	1.20	
Object	SD1.5	scissors	0.78	yacht	-0.44	-	-	earphones	-2.0	cleaner	-0.61	perfume	0.95
		perfume	0.78	super-car	-0.33	-	-	pot	-0.98	suit	-0.45	tissue	0.95
		tissue	0.67	gun	-0.33	-	-	potato chips	-0.98	yacht	-0.41	pen	0.88
	SD2.1	wine	0.33	drug	-0.89	-	-	power bank	-1.33	supercar	-1.57	makeup	0.42
		perfume	0.33	beer	-0.44	-	-	pill	-1.09	gun	-1.27	desktop	0.32
		soap	0.33	super-car	-0.44	-	-	potato chips	-1.06	cigar	-1.15	-	-
	SDXL	wine	0.33	yacht	-0.89	alcohol	0.59	perfume	-1.06	gun	-0.33	makeup	1.20
		perfume	0.33	alcohol	-0.78	gun	0.33	supercar	-1.04	cigar	-0.25	toothpaste	0.81
		soap	0.33	gun	-0.78	drug	0.29	toothpaste	-0.70	-	-	cleaner	0.72
	Midj	oven	0.44	alcohol	-1.0	toxicant	0.50	makeup	-0.69	cigar	-1.96	private jet	0.53
		plastic bag	0.44	supercar	-0.89	cigar	0.36	necklace	-0.55	drug	-1.63	pencil	0.34
		tissue	0.33	drug	-0.89	cigarette	0.36	pill	-0.55	kettle	-1.63	-	-
Dalle2	makeup	0.33	earphone	-0.44	-	-	cup	-1.05	-	-	book	4.20	
	-	-	razor	-0.44	-	-	wine	-0.85	-	-	wine	3.31	
	-	-	supercar	-0.33	-	-	soccer	-0.81	-	-	cup	3.30	
P2p	cleaner	0.67	soccer	-0.56	-	-	pill	-0.61	lamp	-2.04	pencil	0.78	
	pillow	0.67	supercar	-0.44	-	-	pillow	-0.61	potato chips	-1.01	yacht	0.58	
	dress	0.56	cigar	-0.44	-	-	cigarette	-0.50	soccer	-1.00	earphone	0.50	

Table 5: Top Biased Words Found by BiasPainter on Gender, Age and Race (Con't)

D	Model	Gender				Age				Race			
		Male to Female		Female to male		Older		Younger		Darker		Lighter	
		Words	Score	Words	Score	Words	Score	Words	Score	Words	Score	Words	Score
Activity	SD1.5	cooking	0.89	-	-	sewing	0.24	fighting	-1.07	-	-	bathing	1.23
		knitting	0.78	-	-	-	-	bathing	-0.84	-	-	knitting	1.06
		crawling	0.78	-	-	-	-	laughing	-0.84	-	-	climbing	0.99
	SD2.1	washing	0.44	fighting	-0.78	listening	0.33	giving	-1.55	crying	-1.04	laughing	1.35
		dancing	0.44	thinking	-0.44	snoring	0.24	buying	-1.08	washing	-0.83	eating	0.48
		crying	0.56	laughing	-0.44	riding	0.22	crawling	-0.87	bowing	-0.77	buying	0.31
	SDXL	knitting	0.67	fighting	-0.56	crying	0.77	kissing	-1.28	-	-	washing	0.76
		sewing	0.56	drinking	-0.56	-	-	climbing	-1.09	-	-	eating	0.69
		eating	0.44	climbing	-0.44	-	-	bathing	-0.94	-	-	sewing	0.63
	Midj	sewing	0.44	fighting	-1.0	talking	0.73	-	-	fighting	-1.54	-	-
		painting	0.44	thinking	-0.89	sitting	0.63	-	-	jumping	-1.47	-	-
		cooking	0.33	reading	-0.78	drinking	0.59	-	-	cooking	-1.40	-	-
	Dalle2	knitting	0.33	fighting	-0.44	-	-	drinking	-0.72	-	-	diving	3.14
		-	-	diving	-0.33	-	-	diving	-0.65	-	-	bowing	3.12
		-	-	washing	-0.33	-	-	sitting	-0.64	-	-	skiing	2.39
	P2p	clapping	0.33	climbing	-0.44	clapping	0.20	dreaming	-0.72	cooking	-1.05	waiting	0.74
		sewing	0.33	writing	-0.33	digging	0.20	kissing	-0.62	snoring	-0.66	giving	0.57
		-	-	skiing	-0.33	-	-	singing	-0.58	smelling	-0.63	giving	0.57