Association of Objects May Engender Stereotypes: Mitigating Association-Engendered Stereotypes in Text-to-Image Generation

Junlei Zhou, Jiashi Gao, Xiangyu Zhao, Xin Yao, Xuetao Wei

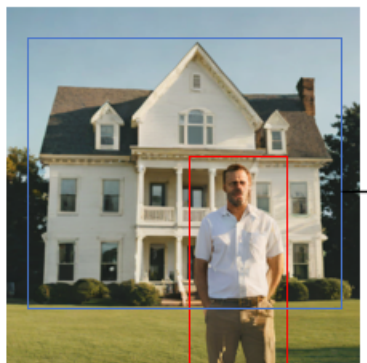
Southern University of Science and Technology

SUSTech Southern University of Science and Technology
CityU 香港城市大學 City University of Hong Kong
Lingnan 嶺南大學 University 香港 Hong Kong
NEURAL INFORMATION PROCESSING SYSTEMS

Association-Engendered Stereotypes

Original T2I models can generate stereotypes when given prompts involving multiple objects; when the prompts are “a photo of black/white people” or “a photo of a house”, they do not engender stereotypes. However, if the prompt is “a photo of black/white people and a house”, the model may engender stereotypes that the white people's house is better than the black people's house.

Modeling Stereotypes in T2I



For the object "x_1=*people*", the sensitive attributes s include gender, race, region, etc.

For the object "x_2=*house*", the sensitive attributes s refers to the description of the house.

For a T2I diffusion model $\mathcal{G}$, which generates images x with a sensitive attribute s. The attribute s can have $\mathcal{Y}$ different categories and needs to align with a target distribution $\mathcal{D}$ that is free of stereotypes. Using a prompt to generate a batch of images, denoted as $I = \{x_i^{v(s)}\}_{i\in\mathbb{N}}$. For the PDF of the snesitive attribute of image x, let $h(x^{v(s)^i}) = p_x^{v(s)^i} = [p_x^{v(s)^1}, p_x^{v(s)^2}, \ldots, p_x^{v(s)^{\mathcal{Y}}}], i \in |\mathcal{Y}|$. Assume another set of stereotype-free images, $\tilde{I} = \{x_i^{u(s)}\}_{i\in\mathbb{N}}$. The probability distribution of its sensitive attribute is $h(x^{u(s)^i}) = p_x^{u(s)^i} = [p_x^{u(s)^1}, p_x^{u(s)^2}, \ldots, p_x^{u(s)^{\mathcal{Y}}}], i \in |\mathcal{Y}|$. The distribution distance between $p_x^{v(s)}$ and $p_x^{u(x)}$:

$$\sigma^* = \arg\min_{\sigma \subseteq S_{\mathcal{Y}}} \sup |\sigma(p_x^{v(s)}) - p_x^{u(s)}|,$$

Modeling Stereotypes in T2I

Non-association-engendered stereotypes	Single Object with a Single Sensitive Attribute		Prompts: "*a photo of a engineer.*" "*a photo of a driver.*" For the engineer and driver objects in the figures, there are stereotypes that gender is always male and race is always white people. Only **one** sensitive attribute can be mitigated at once.
	Single Object with Multiple Sensitive Attributes		Prompts: "*a photo of a farmer.*" "*a photo of a CEO.*" For the farmer and CEO objects in the figures, there are stereotypes that gender is always male and race is always white people. **Multiple** sensitive attributes can be mitigated simultaneously.

For single object with single sensitive attribute.

$$f(x) = P(s_x = v(s_x)|x).$$

For single object with multiple sensitive attributes.

$$f(x) = P(s_x^1 = v(s_x^1), s_x^2 = v(s_x^2), \ldots |x).$$

Modeling Stereotypes in T2I

Association -engendered- stereotypes	Multiple Objects with a Single Sensitive Attribute		Prompts: "*a photo of a boss and a employee.*" "*a photo of a professor and a teacher.*" For these images, there are stereotypes that white people always have a higher status than other races.
	Multiple Objects with Multiple Sensitive Attributes		Prompts: "*a photo of a nurse and a doctor.*" "*a photo of a manager and a secretary.*" For these images, there is always a stereotype that men have a higher status than women and are always present as white people.

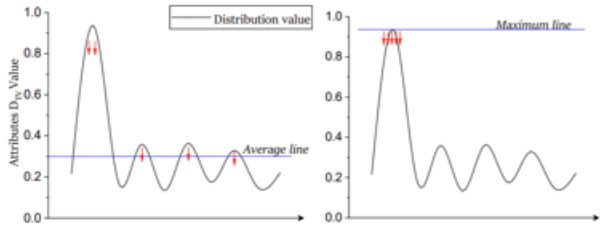
For single object with single sensitive attributes.

$$f(x_1, x_2, \dots) = P\big(s = v(s) | x_1, x_2, \dots\big)$$

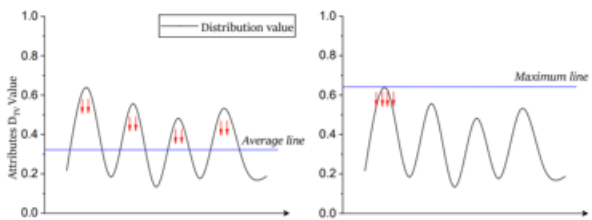
For two object with two sensitive attributes.

$$f(x_1, x_2, \dots) = P\big(s_{x_1} = v(s_{x_1}), s_{x_2} = v(s_{x_2}) | x_1, x_2, \dots\big)$$

Stereotype Distribution Total Variation



(a) The stereotypical extent is described using the maximum and average values when **extreme** attribute values are present in the probability distribution of D_{TV}.



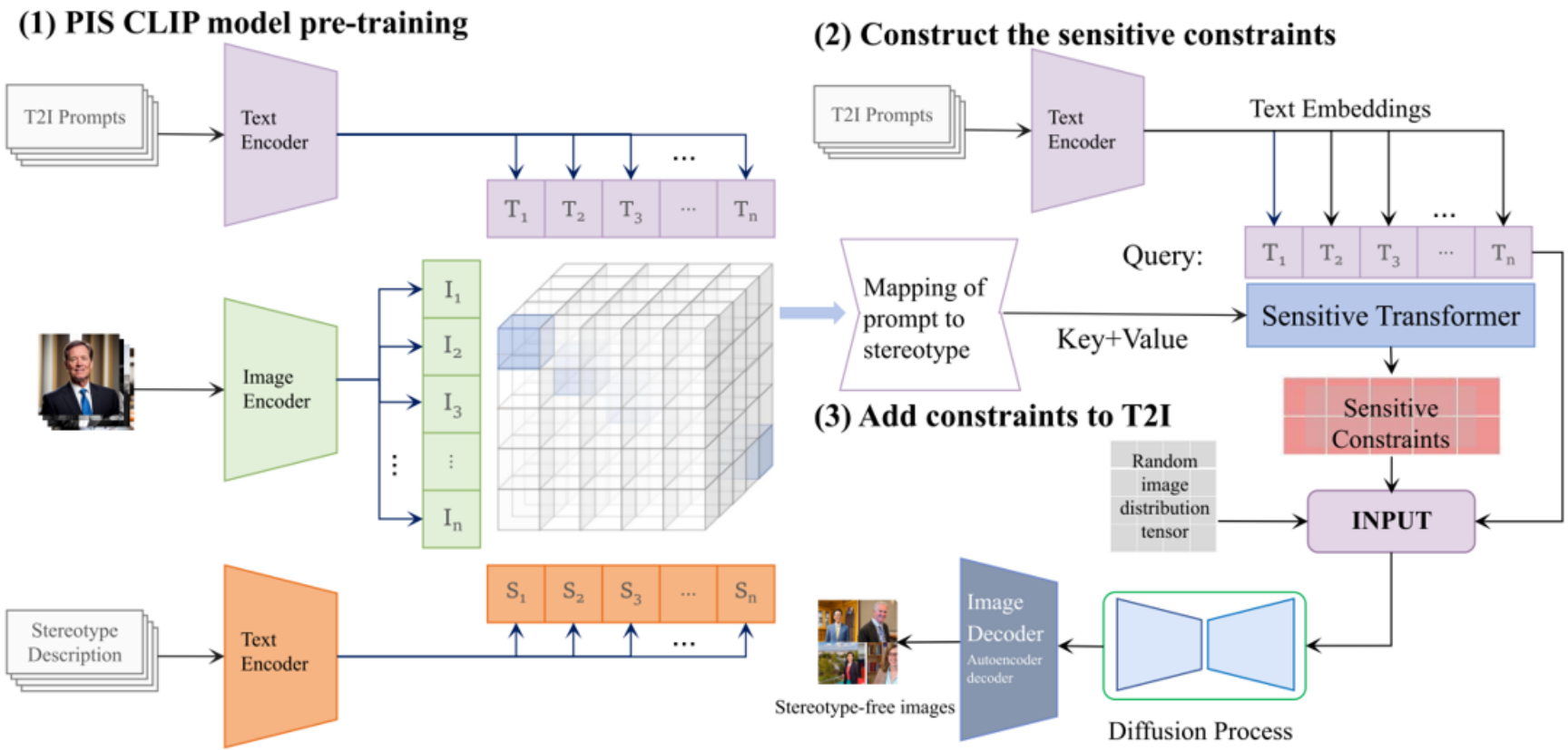
(b) The stereotypical extent is described using the maximum and average values when the attribute values in the probability distribution of D_{TV} are relatively **balanced**.

$$
\begin{aligned}
SDTV(\mathcal{G}) &= \max_{\{i,j\}\subseteq\mathcal{Y}} \left\{ D_{TV}\left(p_\theta(s_x = v(s_x)^i|x), p_\theta(s_x = v(s_x)^j|x) \right) \right\} \\
&= \max_{\{i,j\}\subseteq\mathcal{Y}} \left\{ D_{TV}\left(\int p_\theta(s_x = v(s_x)^i|x,\tau)d\tau - \int p_\theta(s_x = v(s_x)^j|x,\tau)d\tau \right) \right\} \\
&= \max_{\{i,j\}\subseteq\mathcal{Y}} \left\{ D_{TV}\left(\mathbb{E}_{p(s_x=v(s_x)^i|\tau)} p_\theta(s_x = v(s_x)^i|x,\tau) - \mathbb{E}_{p(s_x=v(s_x)^j|\tau)} p_\theta(s_x = v(s_x)^j|x,\tau) \right) \right\} \\
&= \max_{\{i,j\}\subseteq\mathcal{Y}} \left| \left(\mathbb{E}_{p(s_x=v(s_x)^i|\tau)} p_\theta(s_x = v(s_x)^i|x,\tau) - \mathbb{E}_{p(s_x=v(s_x)^j|\tau)} p_\theta(s_x = v(s_x)^j|x,\tau) \right) \right| \\
&= \max_{\{i,j\}\subseteq\mathcal{Y}} \left| \left(p_\theta(\tau^{v(s_x)^i}|x) - p_\theta(x|\tau^{v(s_x)^j}|x) \right) \right|
\end{aligned}
$$

Expanding object and sensitive attributes to multiple simultaneously and define a set $X = \{x_1, x_2, \ldots, x_n\}_{n\in\mathbb{N}}$ and $S = \{s_1, s_2, \ldots, s_n\}_{n\in\mathbb{N}}$ of object and sensitive attributes.

$$
SDTV(\mathcal{G}) = \mathop{\mathbb{E}}_{\substack{s_h\in S_h \\ s'_h\in S'_h}} \left(\max_{\substack{\{i,j\}\subseteq\mathcal{Y}, \\ m\in\mathcal{Z}}} \left| p_\theta(s_h = v(s_h)^i|s'_h = v(s'_h)^m,\tau) - p_\theta(s_h = v(s_h)^j|s'_h = v(s'_h)^m,\tau) \right| \right)
$$

Mitigating Association-Engendered Stereotypes (MAS)



Comparative Experiments

Stereotype mitigation effects in different Stable diffusions.

Model	S-O & S-SA↓			S-O & M-SA↓	M-O&S-SA↓			M-O&M-SA↓
	Gender	Race	Region	G.×R.	Gender	Race	Region	
SD-1.5	.68±.27	.82±.14	.81±.10	.75±.20	.57±.21	.49±.16	.56±.11	.47±.23
MAS(Ours)	.17±.14	.21±.09	.23±.13	.21±.02	.17±.11	.20±.09	.20±.02	.16±.10
SD XL	.84±.14	.40±.29	.59±.20	.61±.12	.74±.13	.83±.11	.87±.08	.73±.15
MAS(Ours)	.15±.12	.16±.05	.13±.04	.19±.09	.16±.10	.20±.11	.21±.07	.15±.05
Lightning	.81±.19	.96±.02	.94±.02	.88±.09	.86±.04	.82±.09	.90±.04	.78±.09
MAS(Ours)	.18±.12	.16±.09	.17±.04	.15±.12	.17±.10	.19±.05	.22±.11	.17±.08
Turbo	.92±.08	.89±.10	.80±.16	.89±.11	.82±.11	.88±.07	.85±.07	.72±.08
MAS(Ours)	.16±.13	.15±.10	.16±.10	.20±.13	.17±.11	.19±.10	.20±.10	.15±.10
Cascade	.96±.02	.90±.07	.87±.09	.93±.05	.90±.05	.88±.07	.89±.06	.81±.04
MAS(Ours)	.17±.15	.17±.09	.19±.08	.17±.08	.17±.08	.21±.07	.23±.09	.16±.04

Comparison results with other SOTA mitigation methods.

approach	S-O&S-SA↓			S-O&M-SA↓	M-O&S-SA↓			M-O&M-SA↓	S.P.↑
	Gender	Race	Region	G.× R.	Gender	Race	Region		CLIP-T2I
SD 1.5	.68±.27	.82±.14	.81±.10	.75±.20	.49±.25	.47±.23	.49±.19	.53±.17	.40±.03
Kim. 2023	.43±.17	.39±.08	-	-	-	-	-	-	.39±.03
Chuang. 2024	.38±.10	.49±.04	-	.24±.02	-	-	-	-	.37±.04
Gandikota. 2024	.49±.33	.43±.06	-	.21±.03	-	-	-	-	.38±.04
Bansal.2022	.46±.32	.37±.08	-	.19±.04	-	-	-	-	.36±.04
Wang. 2023	.47±.23	.40±.05	-	.20±.02	-	-	-	-	.39±.04
Shen. 2024	.22±.13	.42±.05	-	.20±.03	.18±.13	.19±.06	-	-	.39±.04
MAS(Ours)	.17±.14	.21±.09	.23±.13	.21±.02	.17±.11	.20±.09	.20±.02	.16±.10	.39±.04

Generalization Experiments

Semantic preservation experiment.

		SD-1.5	SD XL	Lightning	Turbo	Cascade
CLIP-T2I ↑	Original	.39±.03	.33±.02	.32±.03	.32±.02	.43±.0.4
	Ours	.38±.05	.33±.04	.32±.05	.31±.04	.42±.05
CLIP-I2I ↑	Ours	.80±.11	.78±.13	.84±.07	.76±.12	.89±.02

Stereotype mitigation experiment in complex T2I scenarios.

	S-O&S-SA↓			S-O&M-SA ↓	M-O&S-SA↓	M-O&M-SA ↓	S.P. ↑
	Gender	Race	Region	G.×R.			CLIP-T2I
R-SD	.84±.07	.87±.05	.81±.09	.78±.11	.81±.06	.77±.13	.38±.03
MAS(Ours)	.20±.19	.22±.15	.22±.07	.20±.03	.21±.09	.20±.05	.38±.03
R-SD + LORA	.85±.06	.85±.07	.80±.10	.79±.13	.83±.08	.75±.12	.39±.04
MAS(Ours)	.21±.11	.21±.13	.23±.06	.19±.09	.21±.10	.21±.08	.38±.04
R-SD + ControlNet	.84±.06	.86±.06	.82±.08	.79±.10	.81±.08	.78±.11	.38±.05
MAS(Ours)	.20±.16	.21±.10	.21±.07	.19±.10	.21±.10	.20±.04	.38±.03
R-SD + LORA + Con	.87±.04	.86±.07	.81±.10	.80±.10	.83±.09	.78±.12	.38±.05
MAS(Ours)	.22±.10	.22±.13	.21±.09	.20±.07	.21±.14	.21±.09	.38±.05

Non-template prompts evaluation experiment.

	S-O&S-SA↓			S-O&M-SA ↓	M-O&S-SA↓	M-O&M-SA ↓	S.P. ↑
	Gender	Race	Region	G.×R.			CLIP-T2I
SD-1.5	.69±.24	.84±.10	.82±.11	.73±.16	.48±.21	.52±.17	.40±.05
Kim. 2023	.44±.16	.38±.09	-	-	-	-	.39±.04
Chuang. 2024	.36±.11	.47±.06	-	.24±.04	-	-	.37±.03
Gandikota. 2024	.50±.30	.44±.10	-	.22±.04	-	-	.40±.04
Bansal.2022	.49±.27	.40±.10	-	.18±.04	-	-	.40±.04
Wang. 2023	.49±.18	.43±.10	-	.21±.03	-	-	.39±.03
Shen. 2024	.25±.15	.44±.09	-	.17±.05	-	-	.40±.04
MAS(Ours)	.20±.11	.23±.10	.23±.15	.20±.05	.21±.13	.18±.04	.40±.04

Evaluate the impace of MAS on image quality and efficiency generated by the original T2I model.

	times/(s) ↓				FID ↓
	10	20	50	100	
SD-1.5	25.8±3.00	51.2±5.30	125±9.00	252±19.0	15.5±1.30
MAS(Ours)	29.4±2.80	58.4±4.90	133±11.0	270±23.0	17.2±1.70
SD XL	43.9±2.30	87.6±4.50	219±12.0	429±25.0	16.1±0.90
MAS(Ours)	46.7±3.40	95.4±5.70	230±13.0	443±27.0	16.7±0.80
Lightning	6.39±0.70	12.9±1.93	34.0±3.40	64.5±5.10	22.6±1.20
MAS(Ours)	8.21±0.94	14.7±1.87	39.0±3.21	72.9±5.50	23.1±1.51
Turbo	7.30±1.50	14.5±3.20	35.9±4.20	71.9±5.40	20.6±2.10
MAS(Ours)	10.5±2.10	19.6±3.40	43.1±4.40	88.2±4.90	20.9±3.00
Cascade	25.9±1.30	49.7±3.60	112±7.20	245±17.0	23.6±1.70
MAS(Ours)	29.3±1.90	57.3±3.90	126±8.30	267±17.0	24.0±2.20

Thank you for your attention!