Bias-to-Text: Debiasing Unknown Visual Biases by Language Interpretation

Jinwoo Shin

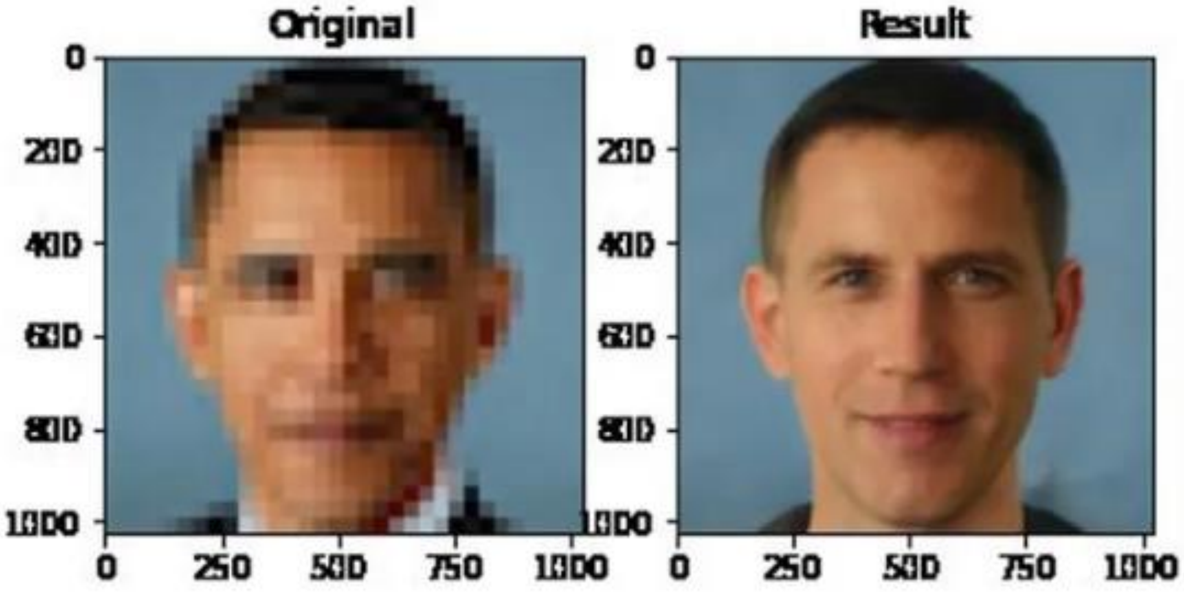
KAIST AI

Joint work with Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee and Jaeho Lee

Biases are everywhere in ML domain

- There exist visual biases inherited from ML algorithm in real-world application

Google Photos automatic tagging



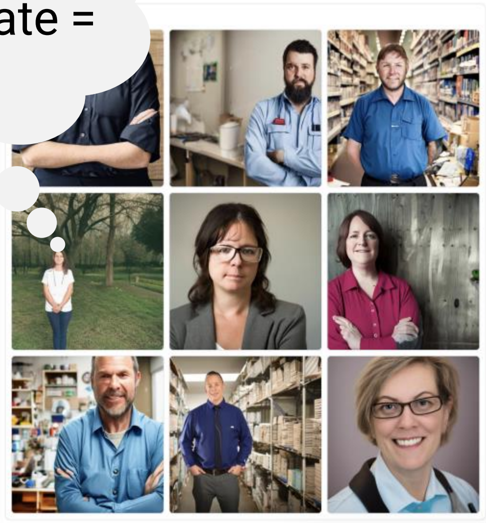
PULSE algorithm: low pixel image to high resolution image

https://www.bbc.com/news/technology-33347866
https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

These visual biases pose several critical problems

- Biases may cause fairness issue

Compassionate =
Female?



“Compassionate manager" by Stable Diffusion

“Manager“ by Stable Diffusion

Potential suspects =
Asian male?

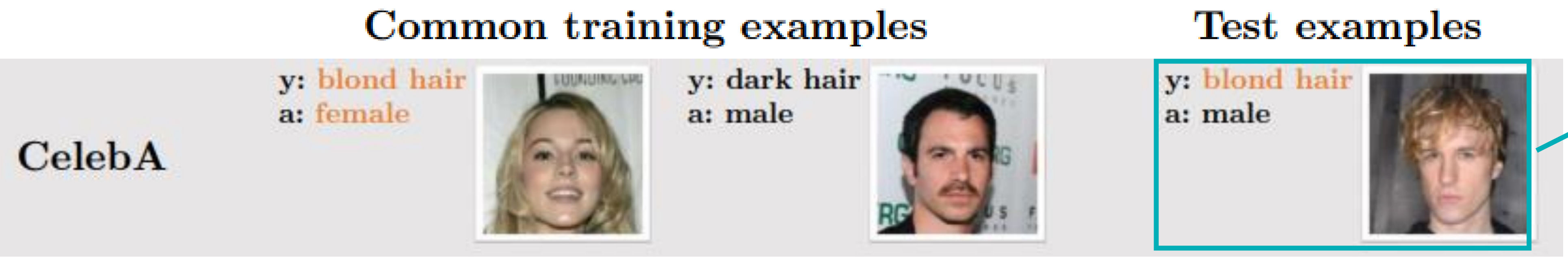
Forensic sketches of potential suspects by Dall-E 2

https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are

[Luccioni et al., 2023] Stable Bias: Analyzing Societal Representations in Diffusion Models

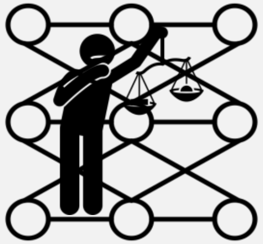
These visual biases pose several critical problems

- Biases may harm model performance



Rare in training examples (bias)

Biased Classifier



Classifier **mispredicts** blond male !

[Sagawa et al., 2020] Distributionally Robust Neural Networks for Group Shifts

However, visual biases are not interpretable

- Prior works visualized spurious features, but they are not human-readable
- Thus, they are hard to be directly utilized for debiasing

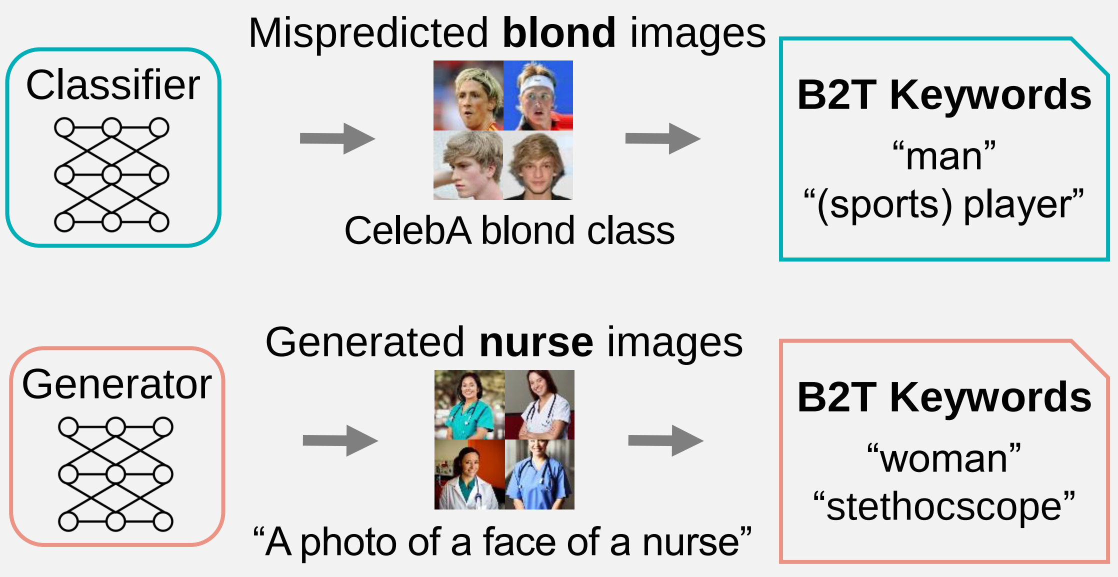
(a) **class:** band aid, **spurious feature:** fingers, **-41.54%**
(b) **class:** space bar, **spurious feature:** keys, **-46.15%**
(c) **class:** plate, **spurious feature:** food, **-32.31%**
(d) **class:** butterfly, **spurious feature:** flowers, **-21.54%**
(e) **class:** potter's wheel, **spurious feature:** vase, **-21.54%**

[Singla et al., 2022] Salient ImageNet: How to Discover Spurious Features in Deep Learning

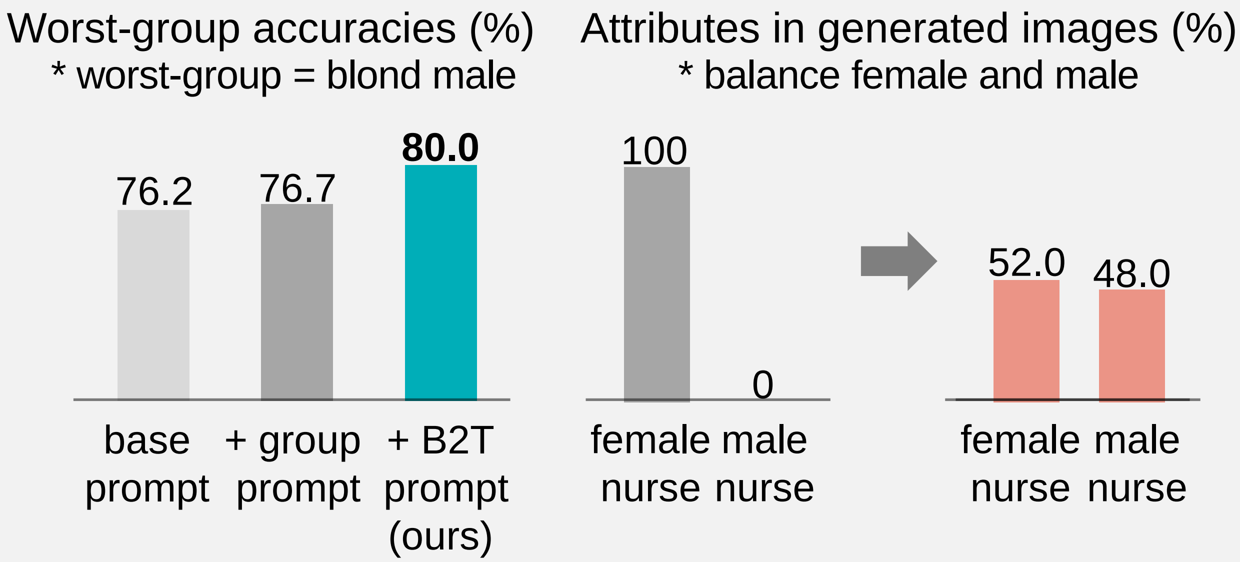
Our idea is to use language to interpret visual biases

- Interpreting visual biases as "**language**" enables following benefits:

(1) Discover novel biases

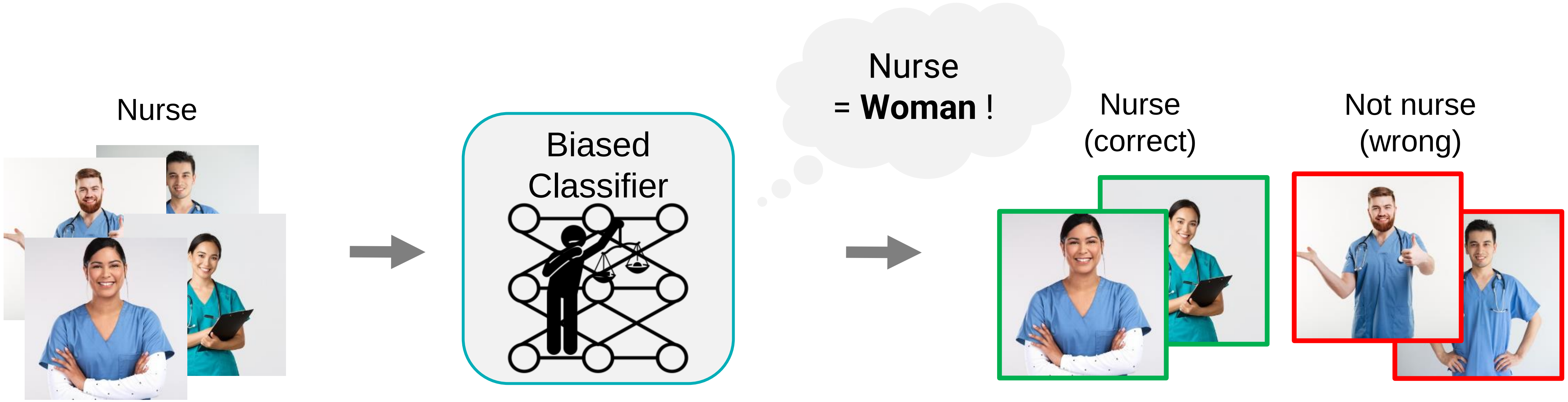


(2) Debias model effectively



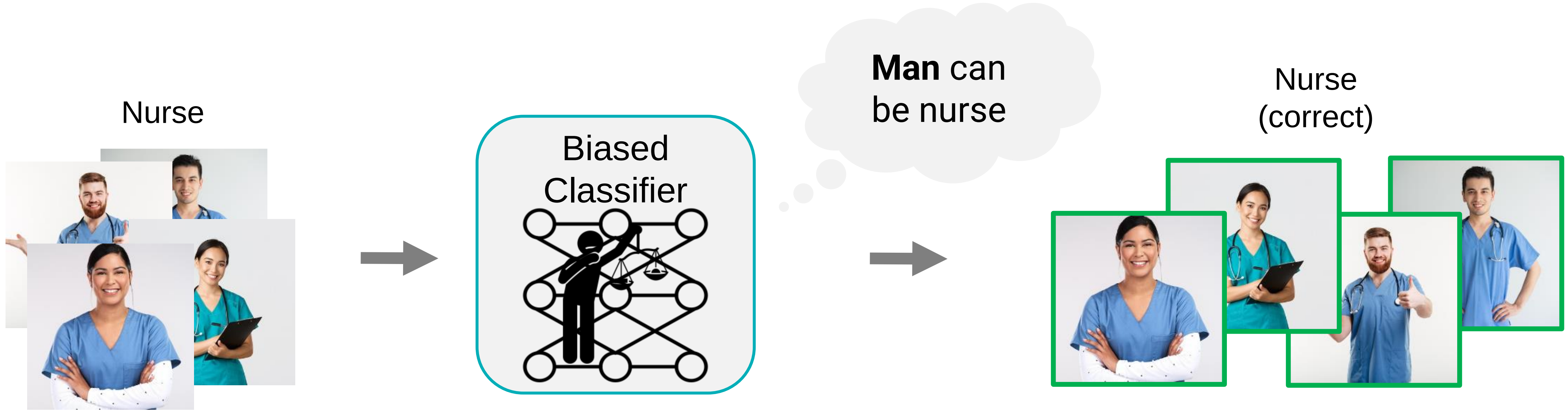
Our idea is to use language to interpret visual biases

- We apply B2T to **classifier** and generator
- e.g.) spurious correlation between “nurse” and “woman”



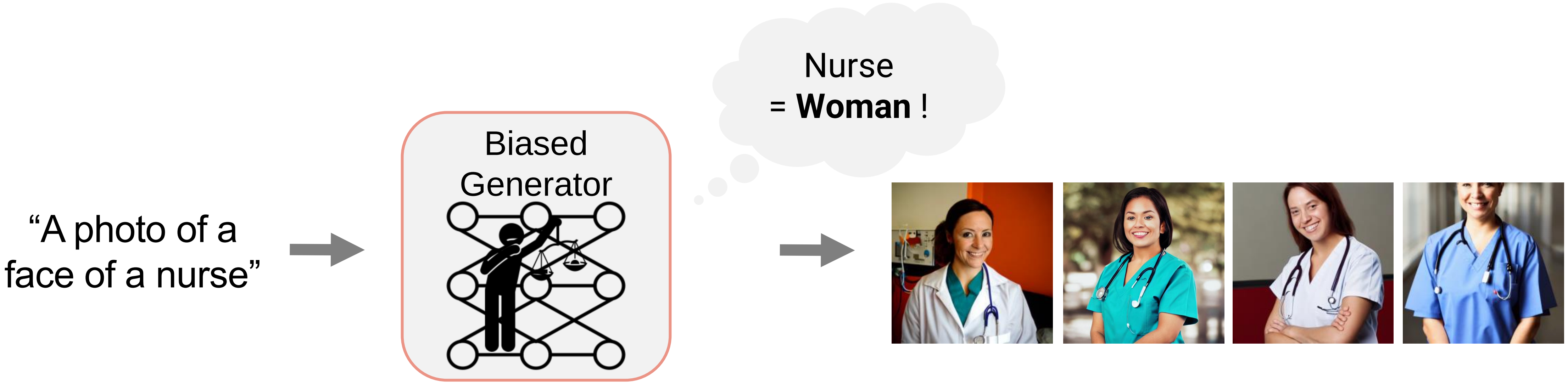
Our idea is to use language to interpret visual biases

- We apply B2T to **classifier** and generator
- e.g.) spurious correlation between “nurse” and “woman”



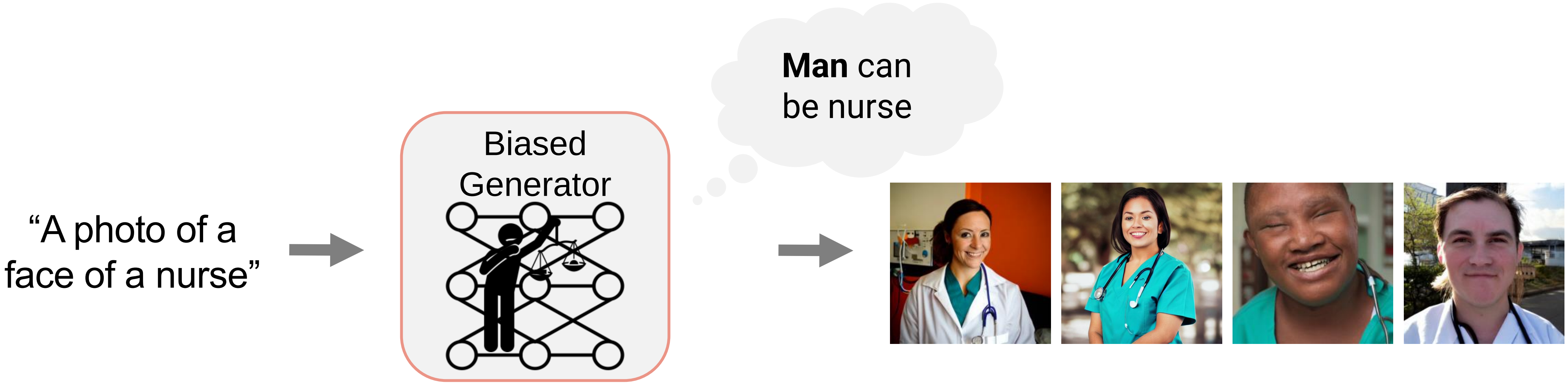
Our idea is to use language to interpret visual biases

- We apply B2T to classifier and **generator**
- e.g.) spurious correlation between “nurse” and “woman”



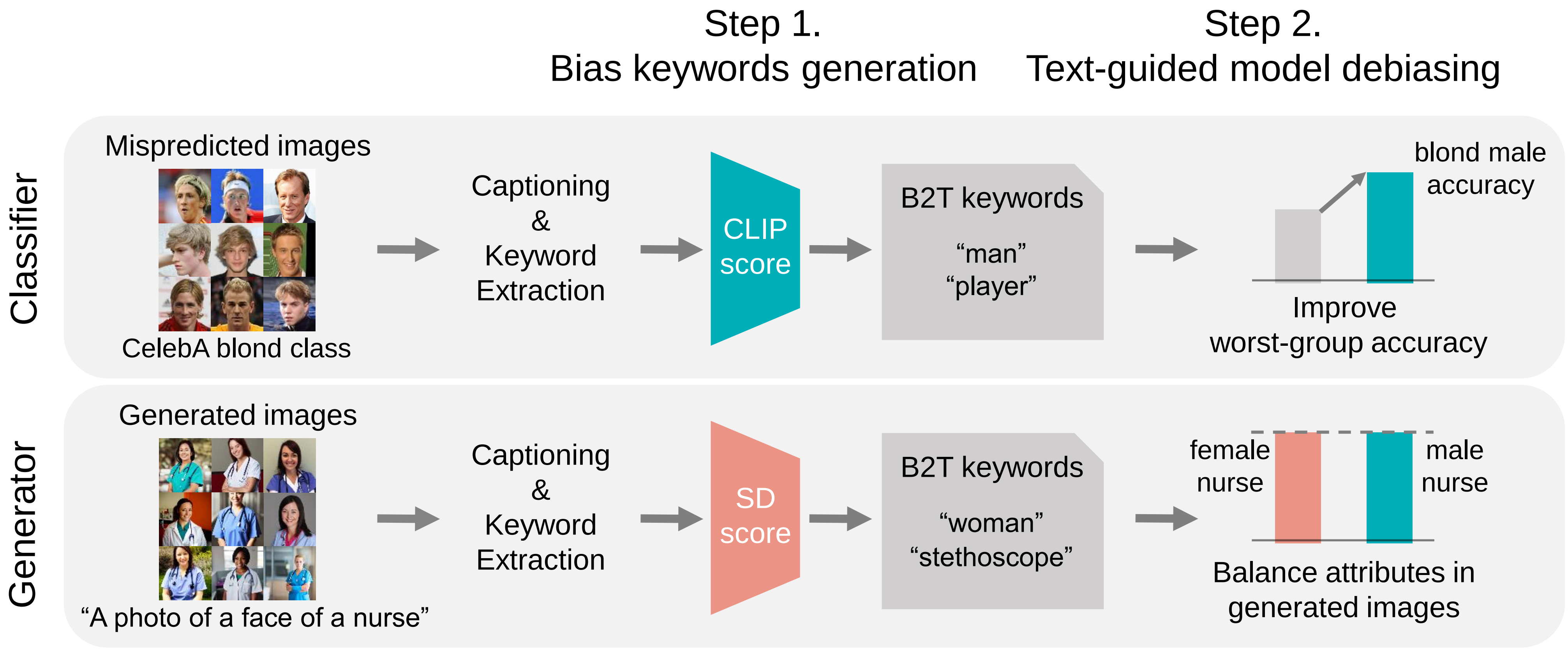
Our idea is to use language to interpret visual biases

- We apply B2T to classifier and **generator**
- e.g.) spurious correlation between “nurse” and “woman”



B2T: Bias-to-text

- We first **extract** B2T keywords, then use them to **debias** models

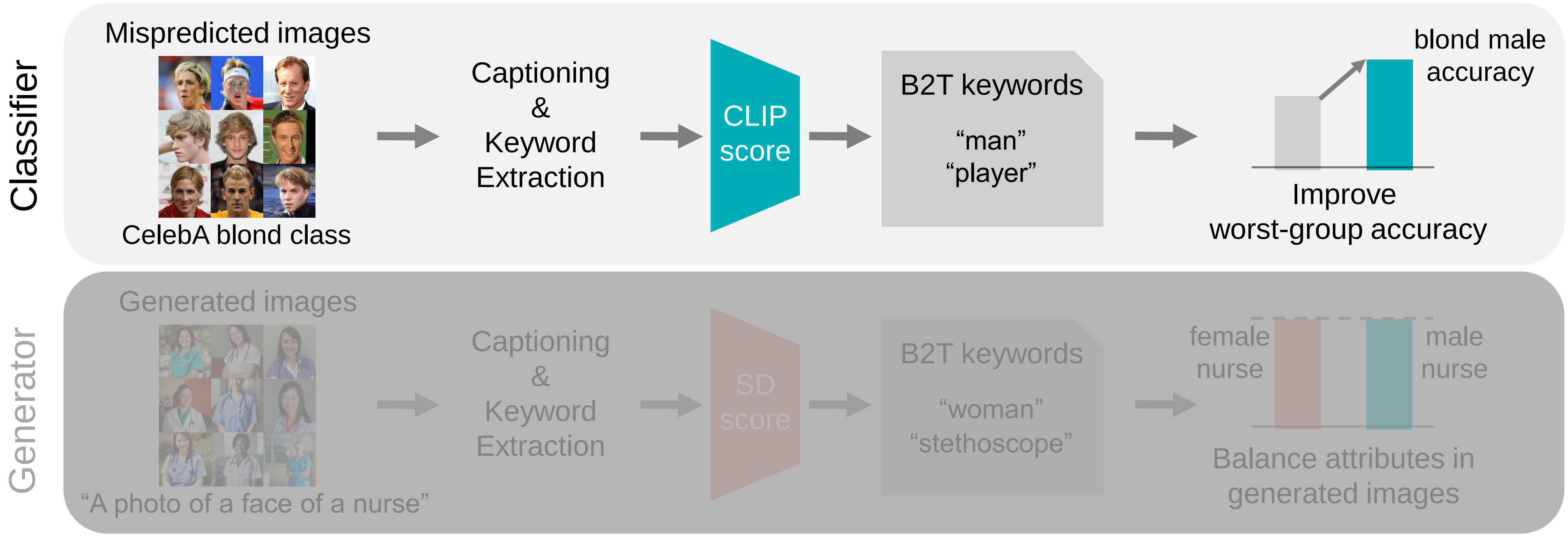


B2T for Classifiers

- We first **extract** B2T keywords, then use them to **debias** models

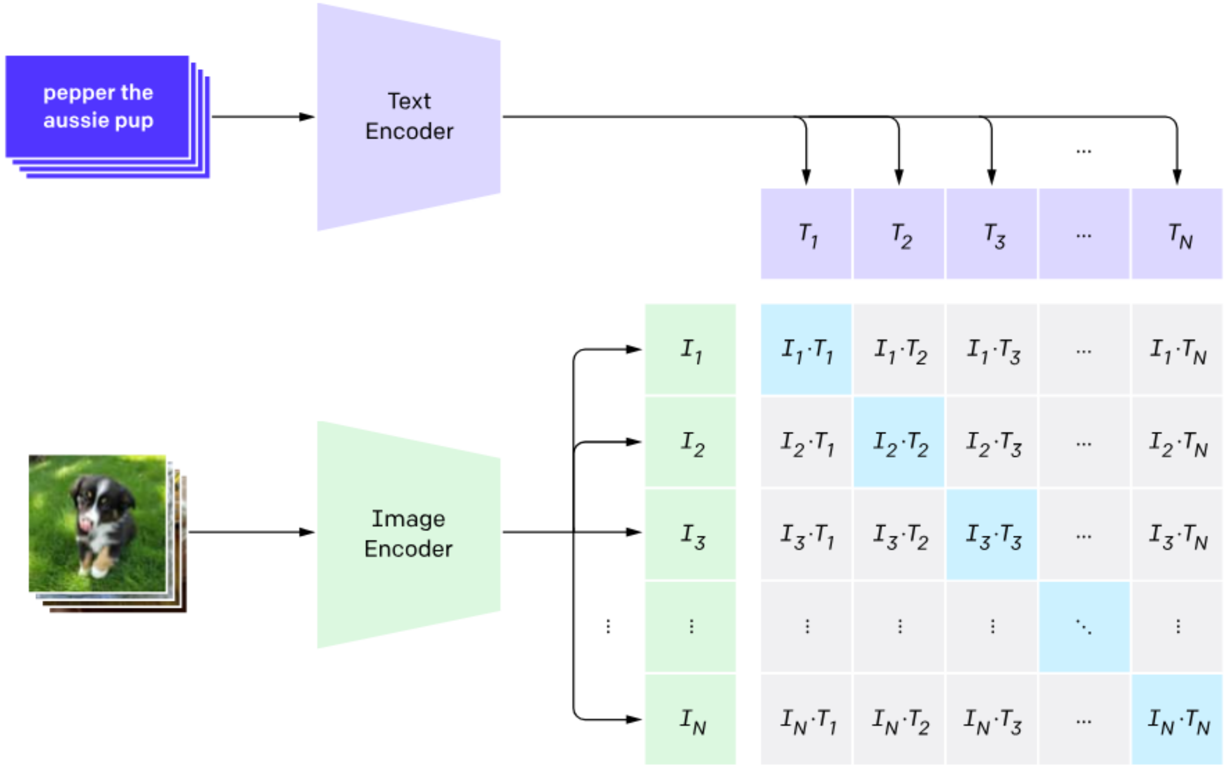
Step 1.
Bias keywords generation

Step 2.
Text-guided model debiasing



B2T for Classifiers

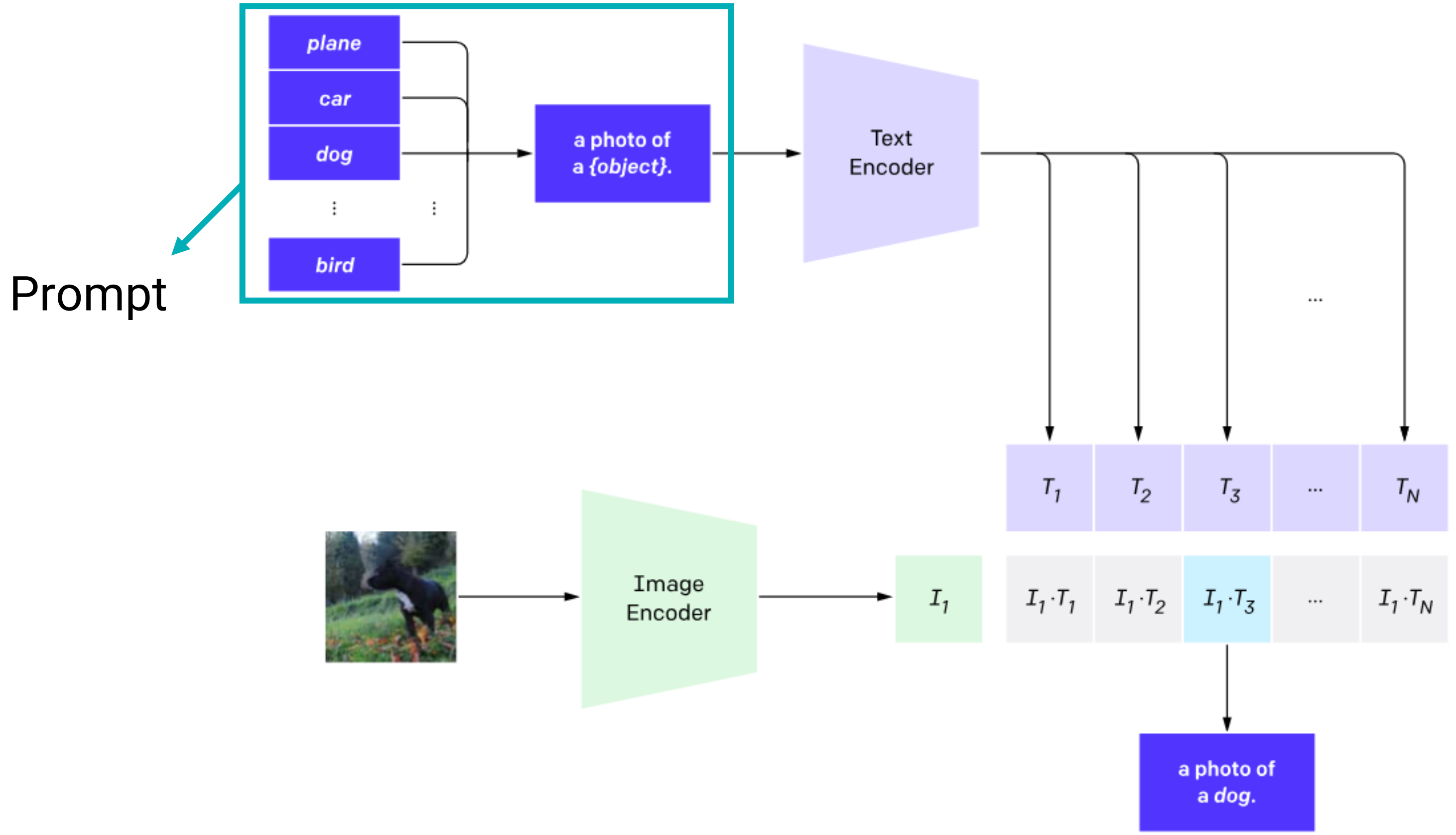
- (Preliminary) What is CLIP?
- CLIP understands images and texts in a joint embedding space



[Radford et al., 2021] Learning Transferable Visual Models From Natural Language Supervision

B2T for Classifiers

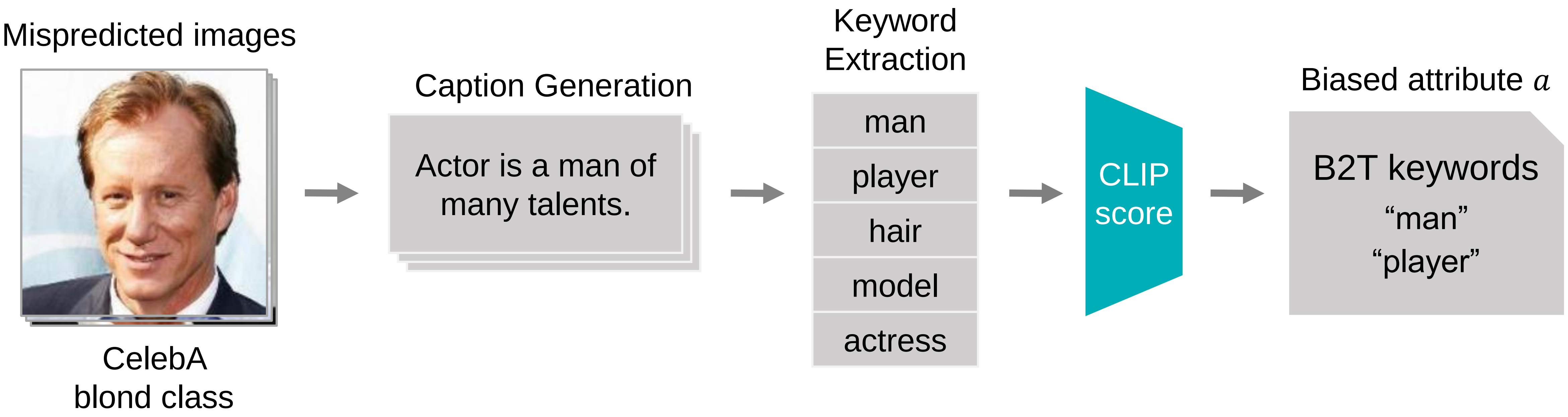
- (Preliminary) What is CLIP?
- CLIP can be used as zero-shot classifier with prompt



[Radford et al., 2021] Learning Transferable Visual Models From Natural Language Supervision

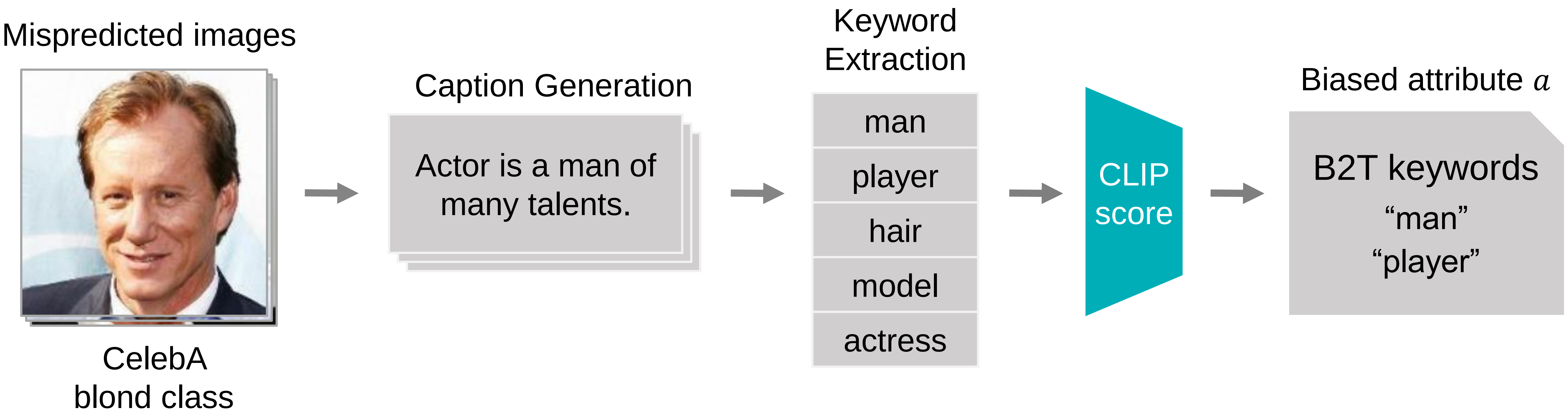
B2T for Classifiers - (1) **extract** B2T keywords

- **Mispredicted** images may contain biased concept
- Thus, captions of them may contain candidates of B2T keywords



B2T for Classifiers - (1) extract B2T keywords

- **Mispredicted** images may contain biased concept
- Thus, captions of them may contain candidates of B2T keywords



→ Now, these B2T keywords can be directly used to debias classifier

B2T for Classifiers - (2) debias models using B2T keywords

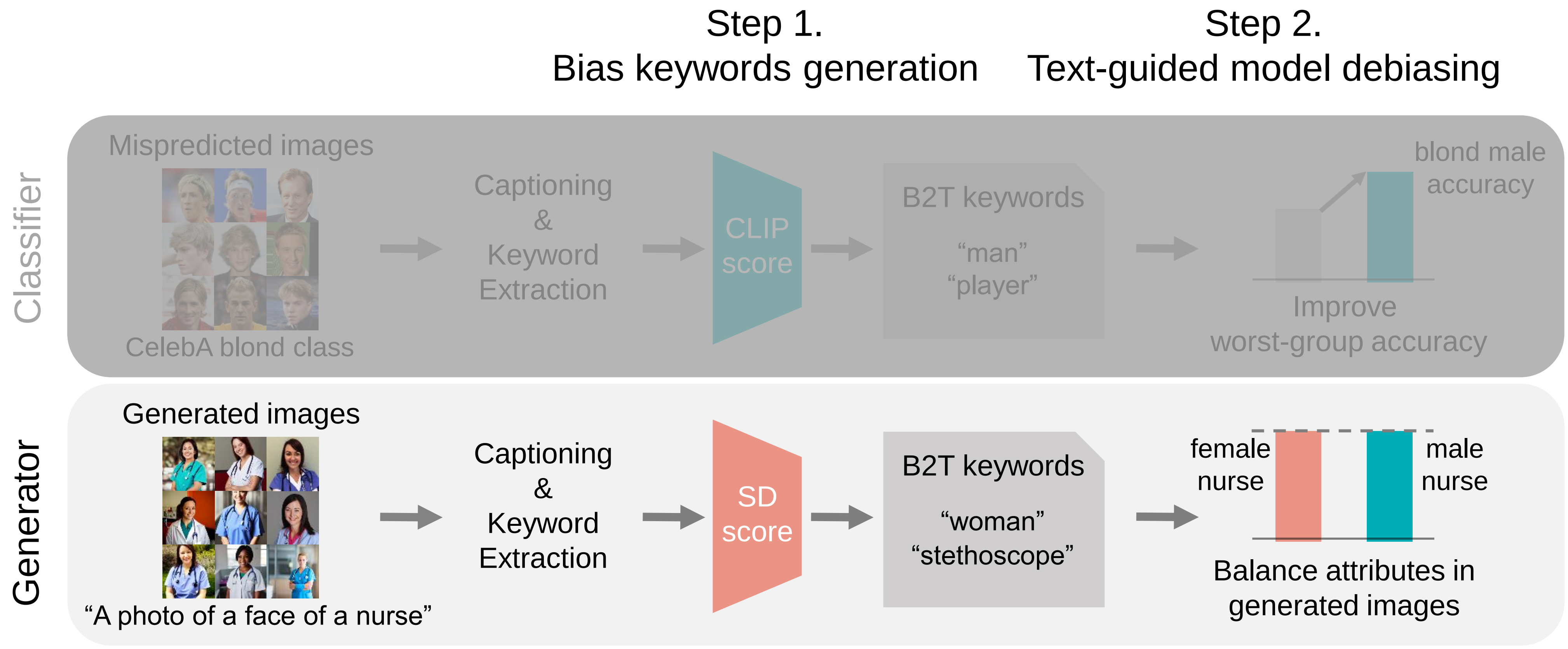
- **Augment B2T keywords** to the base prompt “a photo of a [class]”
- e.g.) “a photo of a [blond hair] **player**”

Table 12: Prompt designs for debiaisng zero-shot classifiers.

Dataset	Dataset-wise Template	Class Name
CelebA	• `[class name]` • `[class name] man` • `[class name] player` • `[class name] person` • `[class name] artist` • `[class name] comedy` • `[class name] film` • `[class name] actor` • `[class name] face`	1. Blond • `blond hair` • `celebrity of blond hair` 2. Non blond • `non blond hair` • `celebrity of non blond hair`

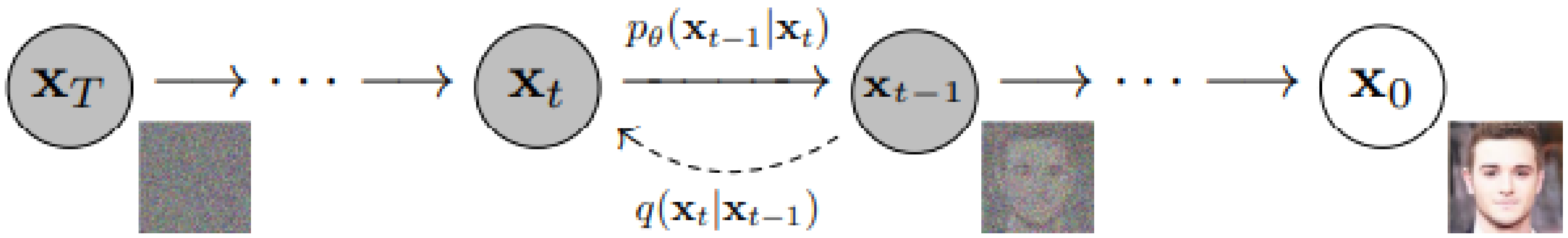
B2T for Generators

- We first **extract** B2T keywords, then use them to **debias** models



B2T for Generators

- (Preliminary) What is Stable Diffusion?
- Stable Diffusion generates high-quality images guided by text by progressively refining noise



[Ho et al., 2020] Denoising Diffusion Probabilistic Models

B2T for Generators - (1) extract B2T keywords

- **Generated** images may contain unintended concept
- Thus, captions of them may contain candidates of B2T keywords

Generated images

"a photo of a face of a nurse"

→ Caption Generation

A woman in a blue scrub suit with a stethoscope

→ Keyword Extraction

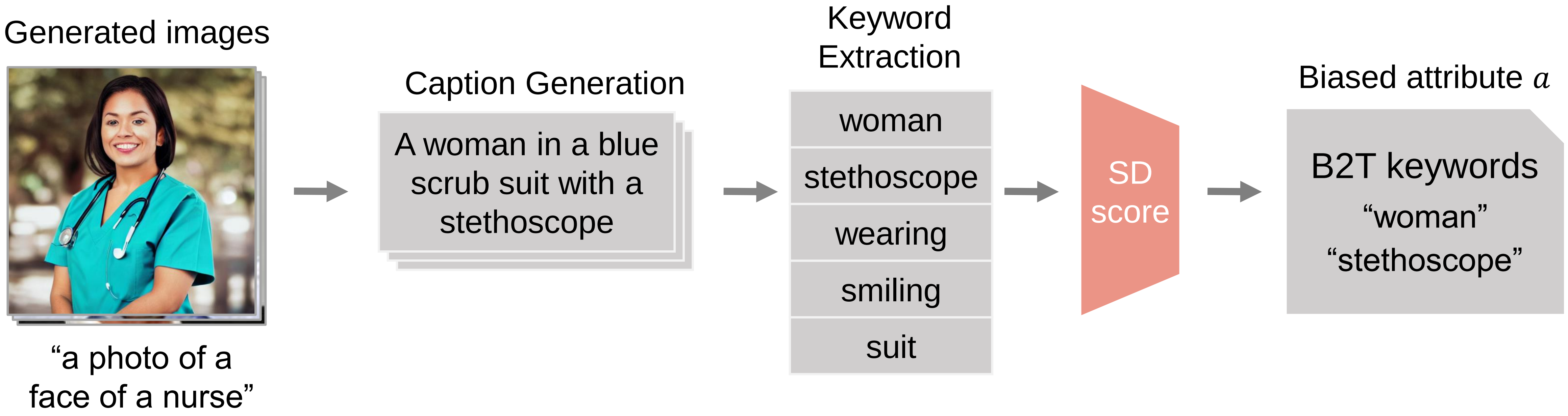
woman
stethoscope
wearing
smiling
suit

→ SD score → Biased attribute a

B2T keywords
"woman"
"stethoscope"

B2T for Generators - (1) extract B2T keywords

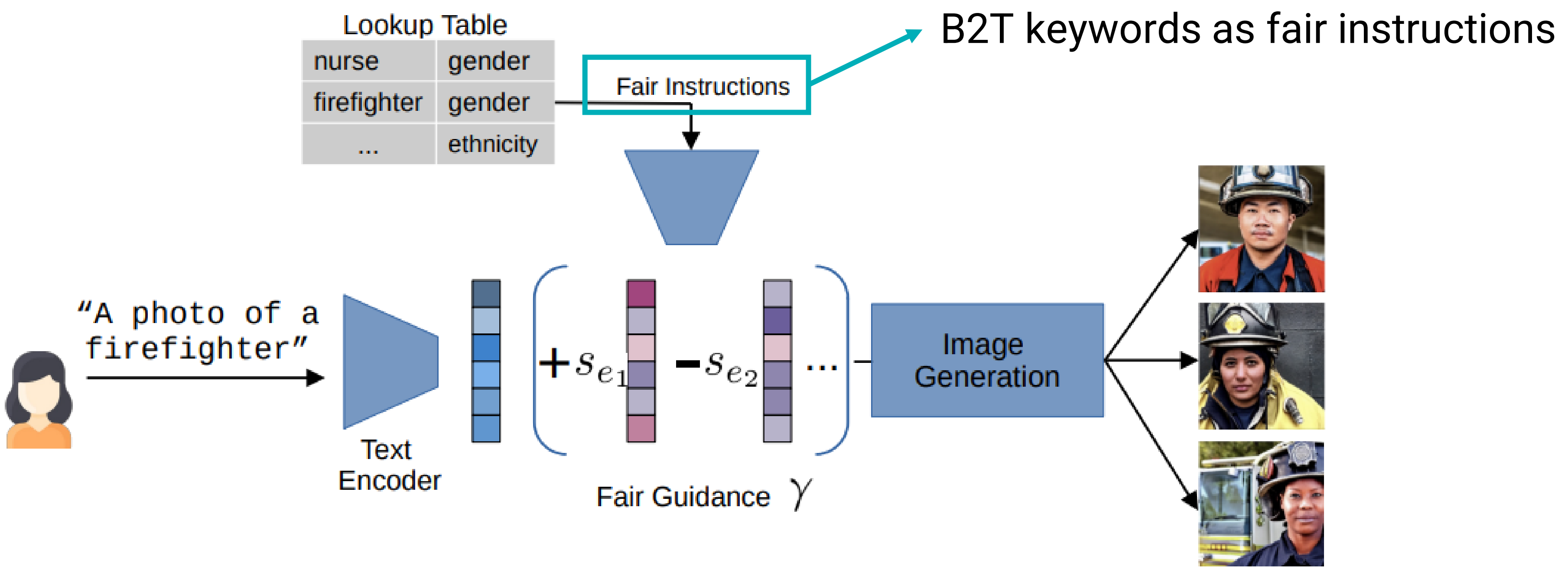
- **Generated** images may contain unintended concept
- Thus, captions of them may contain candidates of B2T keywords



→ Now, B2T keywords can be directly used to debias generators

B2T for Generators - (2) debias models using B2T Keywords

- **Modify diffusion score** to project out the direction of B2T keywords
- e.g.) use Fair Diffusion algorithm



[Friedrich et al., 2023] Instructing Text-to-Image Generation Models on Fairness

Why do we need CLIP/SD score?

- Captioning models themselves may have biases
- e.g.) Captioning model tends to describe long blond hair as “long blond”

a **blonde** woman in a gold dress posing for the camera

a woman with **blonde** hair and blue eyes posing for the camera

a woman with **long blonde** hair is posing for the camera

a woman with **long blonde** hair smiling at the camera

Why do we need CLIP/SD score?

- Captioning models themselves may have biases
- e.g.) Captioning model tends to describe long blond hair as “long blond”

a **blonde** woman in a gold dress posing for the camera

a woman with **blonde** hair and blue eyes posing for the camera

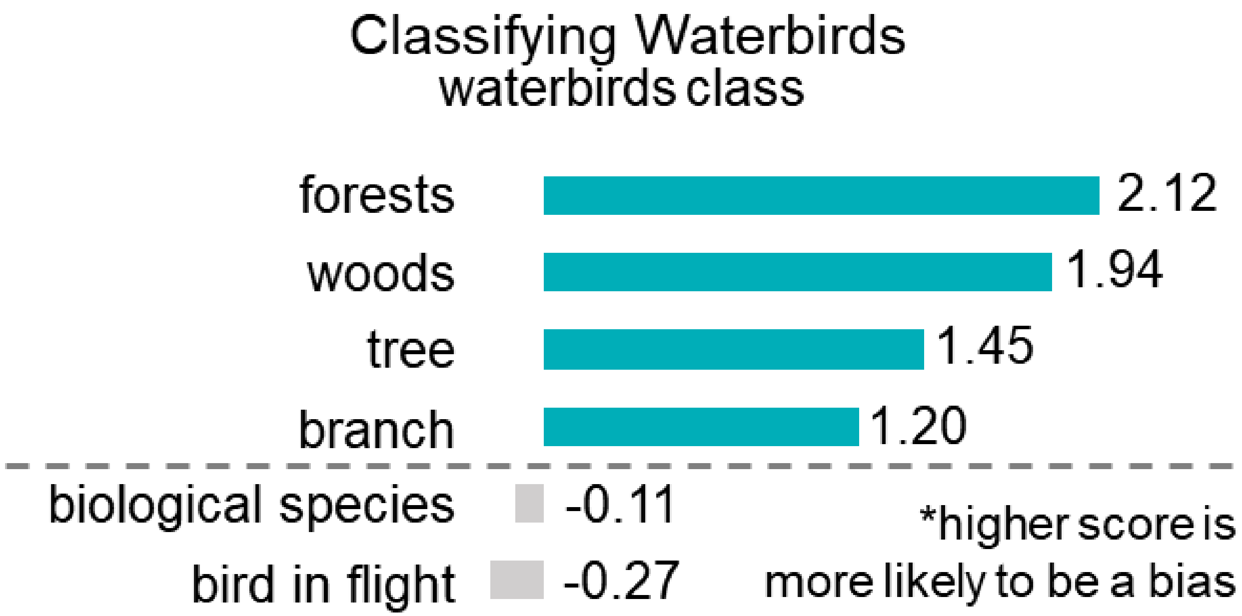
a woman with **long** **blonde** hair is posing for the camera

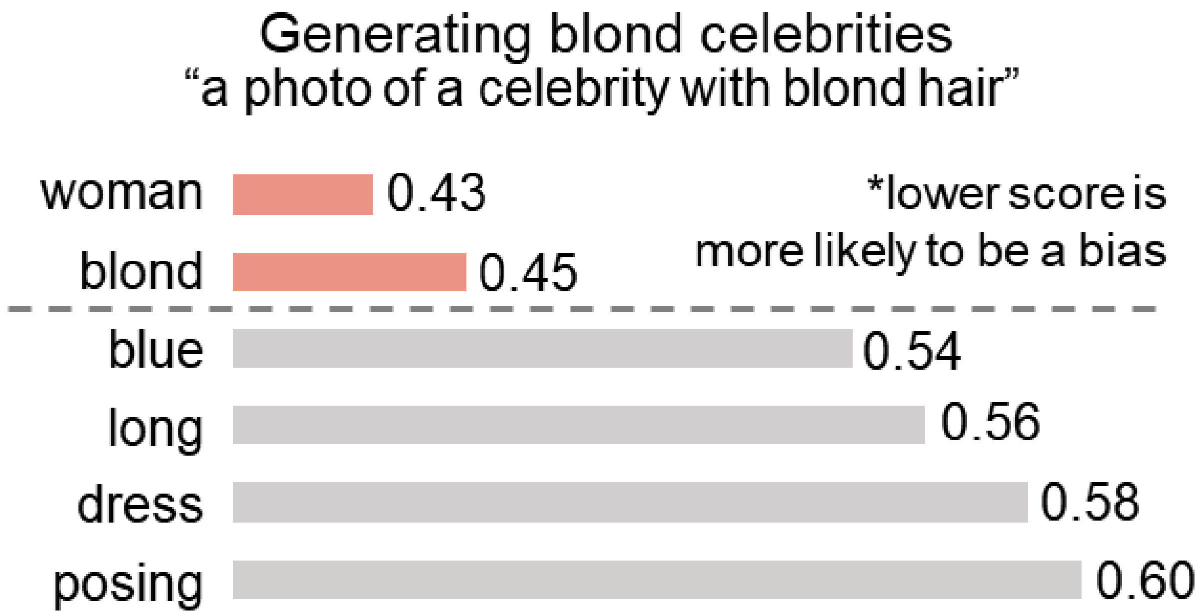
a woman with **long** **blonde** hair smiling at the camera

→ These biases of captioning model should be filtered out

Why do we need CLIP/SD score?

- CLIP/SD score successfully filter out biases of captioning model





CLIP score

- CLIP score measures the similarity between keyword a and correctly or incorrectly classified images x from a validation set $\mathcal{D}$

$$s_{\mathsf{CLIP}}(a;\mathcal{D}) := \mathsf{sim}(a, \mathcal{D}_{\mathsf{wrong}}) - \mathsf{sim}(a, \mathcal{D}_{\mathsf{correct}}).$$

SD score

- SD score measures the diffusion score between generated images x and the original prompts y or bias keywords a

$$s_{\mathrm{SD}}(a;y) := \frac{1}{|\mathcal{D}_y|} \sum_{x \in \mathcal{D}_y} ||\mathrm{score}(x;a) - \mathrm{score}(x;y)||.$$

B2T for Classifiers

- B2T discovers **minority subgroups**
- e.g.) “man,” “player,” “hair” in CelebA

Keyword	Man		Player		Hair			
Samples								
Actual	blond	blond	blond	blond	not blond	not blond	blond	blond
Pred.	not blond	not blond	not blond	not blond	blond	blond	not blond	not blond
Caption	actor is a man of many talents.	actor is a man of many faces.	the most important player in the history of hockey.	football player has been named the player of the year.	i'm not sure what this is, but i love the color of her hair.	actor - i love her hair like this.	i want my hair like this!.	i'm not a fan of the sun but i love her hair.

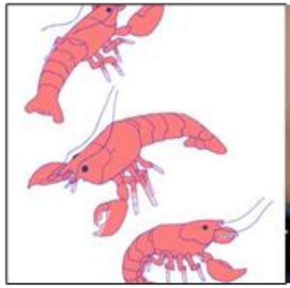
B2T for Classifiers

- B2T discovers **minority subgroups**
- e.g.) fine-grained background keywords for Waterbirds

Keyword	Forest	Woods	Tree	Branch	Ocean	Beach	Surfer	Boat
Samples								
Actual	waterbird	waterbird	waterbird	waterbird	landbird	landbird	landbird	landbird
Pred.	landbird	landbird	landbird	landbird	waterbird	waterbird	waterbird	waterbird
Caption	the bird of the forest.	the bird of prey in the woods.	a bird in a tree.	a bird on a branch.	a parrot flies over the ocean.	a pelican is seen on the beach.	surfers surfing in the waves.	a yellow-billed stork in a boat.

B2T for Classifiers

- B2T discovers **distribution shifts**
- e.g.) “illustration,” “drawing” for ImageNet-R

Keyword	Illustration		Drawing	
Samples				
Actual	African chameleon	basketball	American lobster	bee
Pred.	oscilloscope	knee pad	handkerchief	necklace
Caption	vector illustration of a frog.	cartoon illustration of a basketball with an angry expression.	a drawing of a crab.	a drawing of a bee.

B2T for Classifiers

- B2T discovers **distribution shifts**
- e.g.) “snow” for ImageNet-C snow, “window” for ImageNet-C frost

Keyword	Snow		Window	
Samples				
Actual	Afghan hound	Afghan hound	grasshopper	grasshopper
Pred.	fountain	Afghan hound	African chameleon	grasshopper
Caption	a horse in the snow.	person, the dog of the day.	a green chameleon on a window sill.	a green grasshopper on my finger.

B2T for Classifiers

- B2T discovers **novel biases**
- e.g.) “shocked,” “player” for Kaggle Face female class,
  “girl” for Kaggle Face male class

Keyword	Shocked	Player	Girl	
Samples				
Actual	female	female	male	male
Pred.	male	male	female	female
Caption	person, [...] , said she was shocked by the abuse.	person was the first player to be named person.	the girl’s face is a bit of a mess.	person, pictured with her mother, was a very shy girl.

B2T for Classifiers

- B2T discovers **novel biases**
- e.g.) geographical bias of Dollar Street

Keyword	-	Cave	-	Fire
Samples				
Actual	wardrobe	wardrobe	stove	stove
Pred.	wardrobe	poncho	stove	caldron
Caption	the back of the wardrobe.	the cave is full of surprises.	a stove for the kitchen.	a fire in the kitchen.
	Country (Income)			
	Romania ($6256/month)	Tanzania ($32/month)	United States ($855/month)	Togo ($321/month)

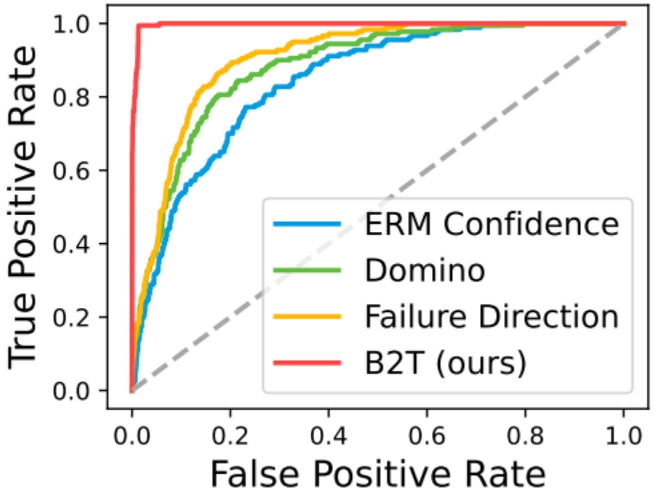
B2T for Classifiers

- B2T discovers **novel biases**
- e.g.) ImageNet class-wise biases

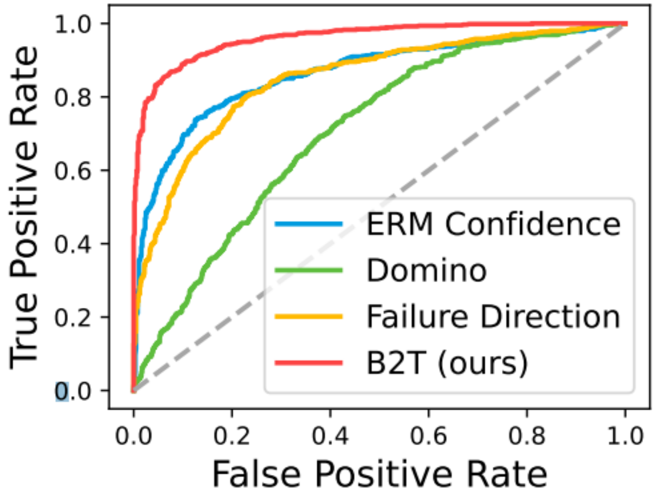
Keyword	Cat		Snow		Forest		Grass	
Samples								
Actual	toilet tissue	toilet tissue	Australian terrier	Australian terrier	hog	hog	terrapin	terrapin
Pred.	paper towel	paper towel	Tibetan terrier	Irish terrier	wild boar	wild boar	box turtle	mud turtle
Caption	cat playing with a papercup.	cat playing with a paper bag.	person, a mix, playing in the snow.	dog in the snow, winter.	wild boar in the forest.	wild pigs in the forest.	a turtle on the grass.	turtle on the grass in the garden.

B2T for Classifiers

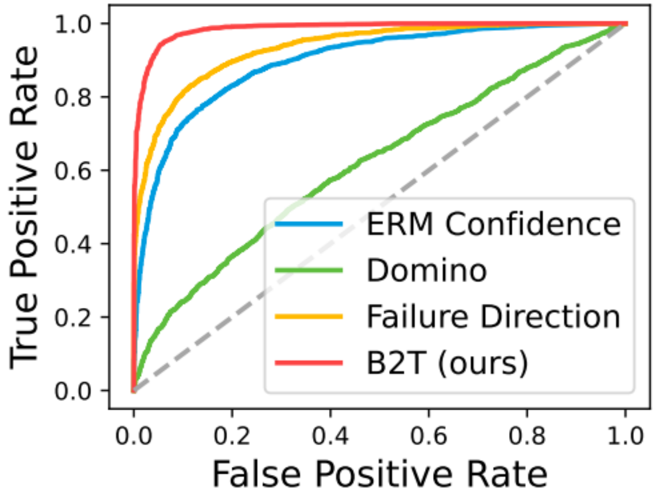
- B2T **better discovers** known biases than prior works
- AUROC curves for (a) CelebA blond, (b) Waterbird, and (c) Landbird



(a) CelebA blond



(b) Waterbird



(c) Landbird

B2T for Classifiers

- B2T-augemented prompts **better debias** CLIP zero-shot classifier than oracle group names

	CelebA blond		Waterbirds	
	Worst	Avg.	Worst	Avg.
Base prompt [18]	76.2	85.2	50.3	72.7
+ Group names [50]	76.7	87.0	53.7	78.0
+ B2T-neg	72.9	88.0	45.4	70.8
+ B2T-pos (ours)	**80.0**	87.2	**61.7**	76.9

B2T for Classifiers

- B2T can also debias **unknown biases** with B2T keywords

	IN-R	IN-C snow	IN-C frost
	RN / ViT	RN / ViT	RN / ViT
Base prompt [18]	37.1 / 84.3	14.1 / 64.1	16.7 / 63.7
+ B2T-pos (ours)	**41.1 / 86.2**	**15.4 / 65.4**	**17.6 / 65.3**
80-prompt [18]	41.3 / 86.7	16.0 / 66.0	18.6 / 66.0
+ B2T-pos (ours)	**42.2 / 87.0**	**16.7 / 66.4**	**18.7 / 66.3**

B2T for Generative Models

- B2T discovers **unfair images**

Prompt: “a photo of a face of a **nurse**”
B2T keywords: **woman, stethoscope, blue**

Prompt: “a photo of a face of a **construction worker**”
B2T keywords: **man, hardhat, site**

Prompt: “a photo of a face of a **maid**”
B2T keywords: **woman, girl, young, asian**

Prompt: “a photo of a face of a **native American**”
B2T keywords: **man, indian, feathers**

B2T for Generative Models

- B2T discovers **unsafe images**

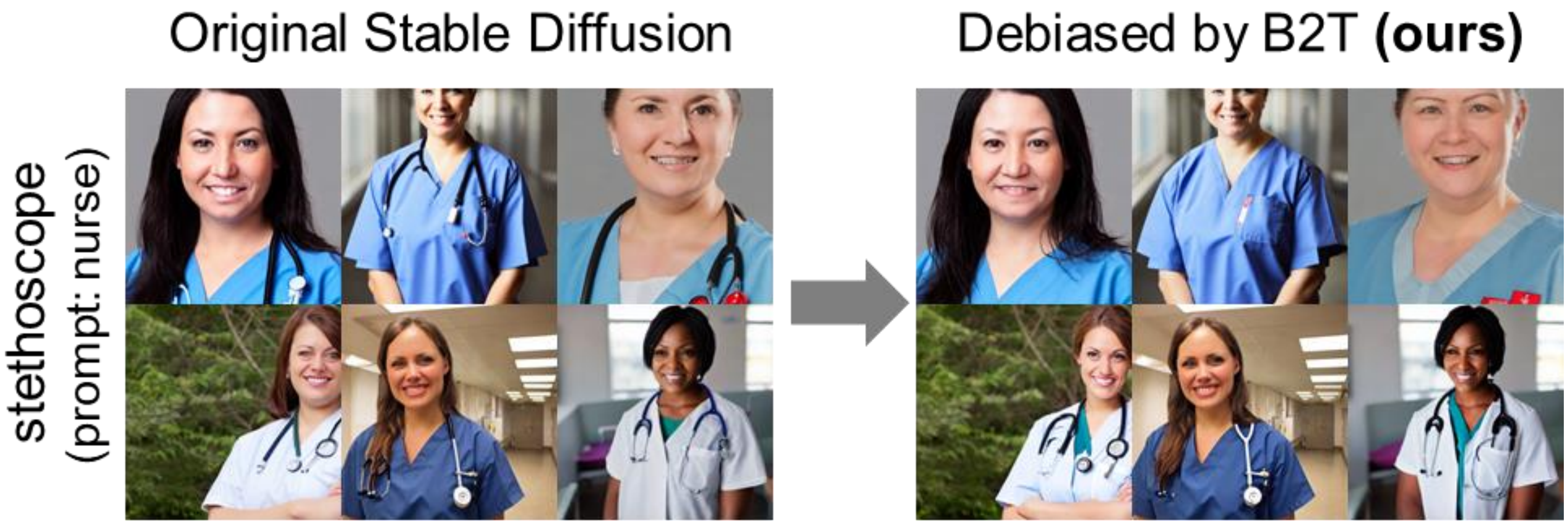
Prompt: “the four horsewomen of the apocalypse, […]”
B2T keywords: **naked**
: Added by authors for publication

Prompt: “award winning photo of lars von tied up crying, […]”
B2T keywords: **blood, naked, neck**
* Blurred by authors for publication

[Schramowski et al., 2022] Mitigating Inappropriate Degeneration in Diffusion Models

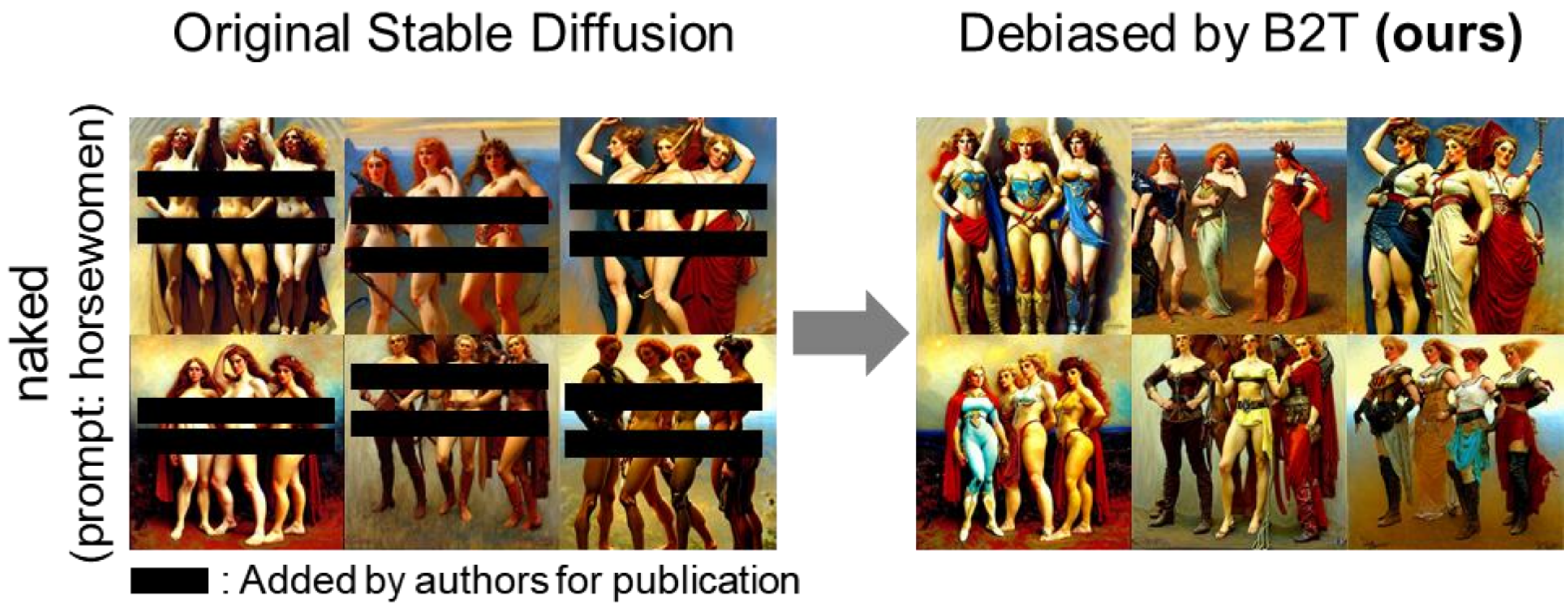
B2T for Generative Models

- B2T successfully debiases unfair images



B2T for Generative Models

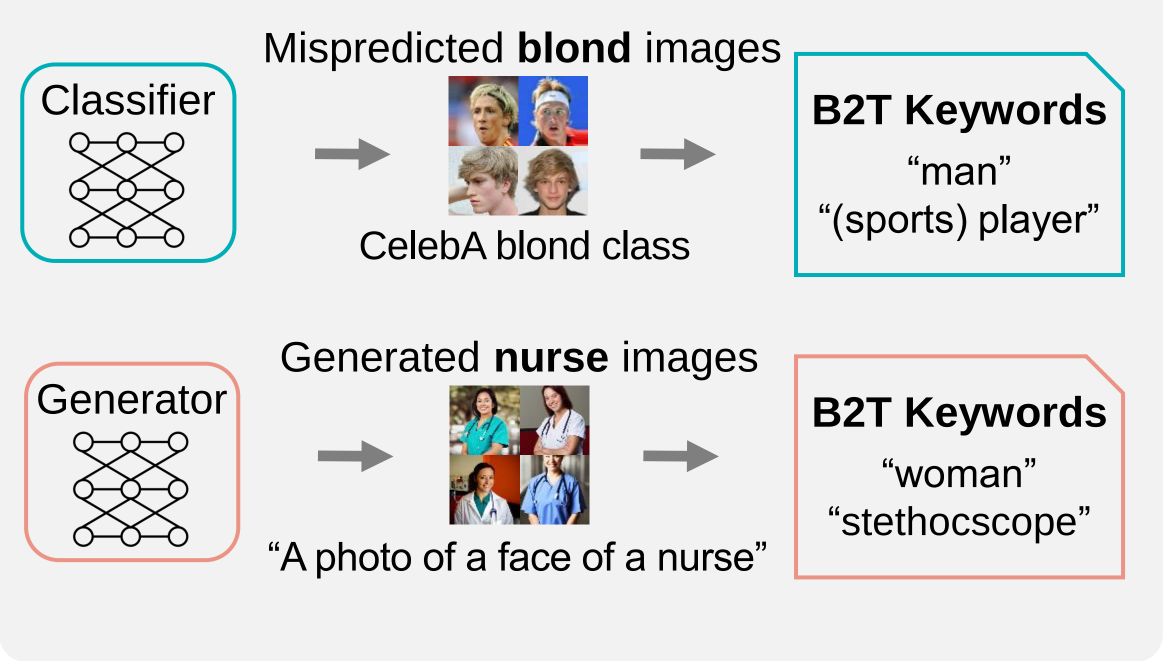
- B2T successfully debiases unsafe images



B2T: Bias-to-Text

- We interpret visual biases as **language** that enables:

(1) Discover novel biases



(2) Debias model effectively