ITI-GEN: Inclusive Text-to-Image Generation

Cheng Zhang[1], Xuanbai Chen[1], Siqi Chai[1], Chen Henry Wu[1], Dmitry Lagun[2], Thabo Beeler[2], and Fernando De la Torre[1]

Robotics Institute, Carnegie Mellon University [1] Google [2]

Webpage

Highlights

- ***ITI-GEN:*** *ensure generated images are uniformly distributed across single or multiple target attributes*
- ***How:*** *prompt optimization using a few reference images*
- ***Scope:*** *diverse attributes spanning humans & scenes*
- ***Train-once-for-all:*** *transferrable tokens; no model-specific fine-tuning needed*

1. Motivation

- Text-to-Image models demonstrate stereotypes

Our goal: *Inclusive* Text-to-Image Generation (ITI-GEN)

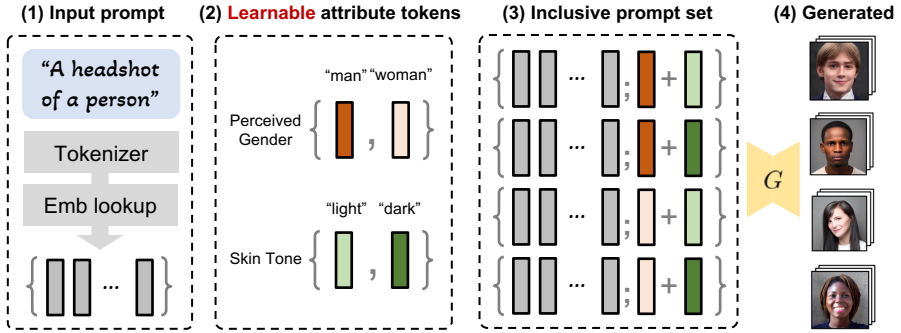
- Given a human-written prompt, the generated images should be *uniformly distributed* across attributes of interest

2. Challenges of Inclusive Generation

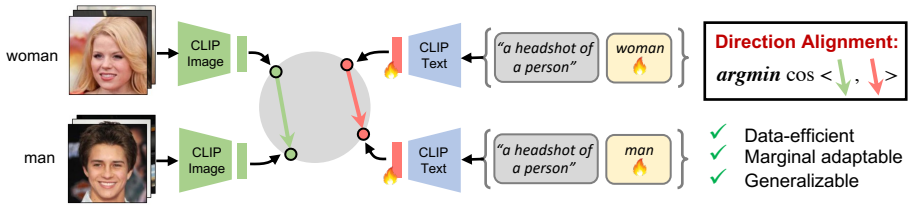
- **Model re-training**: impractical due to data imbalance, high compute cost
- **Text-based debiasing methods** [2,3,4]
  - ➢ **Ambiguity**: leads to clarity issues and model misunderstanding
  - ➢ **Specification gap**: fails to capture nuances, e.g., distinct skin tones

Our key insight: *visual attributes (e.g., blond hair, skin tone type, brightness)* are more expressively described by *images* than by *text*

3. Proposed ITI-GEN Framework [gender & skin tone]

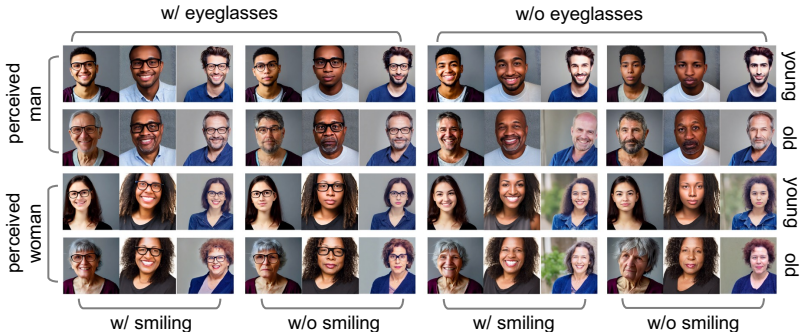


How to learn: translating *visual* differences into *embedding* differences



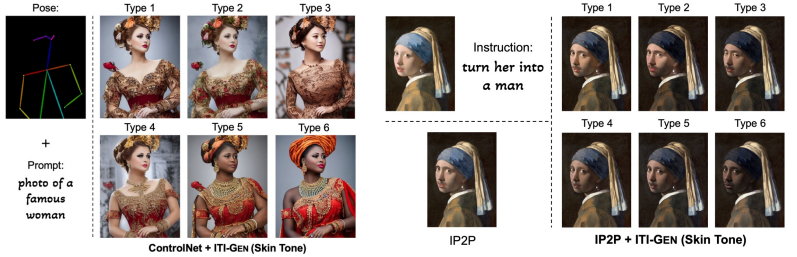
4. Experiments

- **Single** and **multiple** attributes



Method	(a) Single Attribute				(b) Multiple Attributes
	$\mathbb{D}_{KL}^{male}\downarrow$	$\mathbb{D}_{KL}^{young}\downarrow$	$\mathbb{D}_{KL}^{eyeglass}\downarrow$	$\mathbb{D}_{KL}^{smile}\downarrow$	$\mathbb{D}_{KL}^{male\times young\times eyeglass\times smile}\downarrow$
Stable Diffusion [1]	0.343	0.578	0.375	0.134	1.406
Ethical Intervention [2]	0.143	0.423	0.531	0.189	1.311
Hard Prompts [3]	1×10^{-5}	0.027	0.371	4.4×10^{-3}	0.476
Prompt Debiasing [4]	0.322	0.131	0.272	0.146	–
Custom Diffusion [5]	0.309	0.284	0.301	0.469	–
ITI-GEN	$\mathbf{2\times10^{-6}}$	$\mathbf{2\times10^{-4}}$	$\mathbf{2\times10^{-4}}$	$\mathbf{2.5\times10^{-3}}$	**0.094**

- **Compatibility** to ControlNet [6] and InstructPix2Pix [7]



[1] Rombach et al. "High-resolution …". CVPR 2022
[2] Bansal et al. "How well can text-to-image …". EMNLP 2022
[3] Ding et al. "Mastering text-to-image …". NeurIPS 2021
[4] Chuang et al. "Debiasing vision-language …". arXiv 2023
[5] Kumari et al. "Multi-concept customization …". CVPR 2023
[6] Zhang et al. "Adding conditional control …". arXiv 2023
[7] Brooks et al. "InstructPix2Pix: learning to …". CVPR 2023

Please see our paper for results on other attributes, scene domains, and more experimental analysis

ICCV23