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] Fernando De la Torre [1]

[1] Carnegie Mellon University [2] Google

Abstract

Text-to-image generative models often reflect the biases of the training data, leading to unequal representations of underrepresented groups. This study investigates ***inclusive*** *text-to-image generative models that generate images based on human-written prompts and ensure the resulting images are* ***uniformly distributed*** *across attributes of interest. Unfortunately, directly expressing the desired attributes in the prompt often leads to sub-optimal results due to linguistic ambiguity or model misrepresentation. Hence, this paper proposes a drastically different approach that adheres to the maxim that* "a picture is worth a thousand words". *We show that, for some attributes, images can represent concepts more expressively than text. For instance, categories of skin tones are typically hard to specify by text but can be easily represented by example images. Building upon these insights, we propose a novel approach,* **ITI-GEN**[1]*, that leverages readily available reference images for* ***I****nclusive* ***T****ext-to-****I****mage* ***GEN****eration. The key idea is learning a set of prompt embeddings to generate images that can effectively represent all desired attribute categories. More importantly,* ITI-GEN *requires no model fine-tuning, making it computationally efficient to augment existing text-to-image models. Extensive experiments demonstrate that* ITI-GEN *largely improves over state-of-the-art models to generate inclusive images from a prompt.*

1. Introduction

In recent years we have witnessed a remarkable leap in text-based visual content creation, driven by breakthroughs in generative modeling [70, 28, 60, 59, 64] and the access to large-scale multimodal datasets [68, 36]. Particularly, publicly released models, such as Stable Diffusion [64], have matured to the point where they can produce highly realistic images based on human-written prompts.

However, one major drawback of existing text-to-image models is that they inherit biases from the training data [6, 59, 64, 12, 5] and thus have yet to exhibit *inclusiveness* — the generated images based on the input text may reflect stereotypes, leading to the exclusion of certain attributes or minority groups. For instance, given the prompt "a headshot of a person", Figure 1(a) shows how a state-of-the-art system generates about 92% images of subjects without eyeglasses, and only 8% with eyeglasses, showing a clear bias towards people without eyeglasses. Alternatively, as shown in Figure 1(b), one could specify the attribute in the prompt, resulting in better outcomes; however, this will still result in a sub-optimal solution due to linguistic ambiguity. While inclusiveness has been critical to responsible AI, existing text-to-image models are still lagging [12, 5, 56, 54, 47]. In this work, we propose a new method that achieves inclusive-

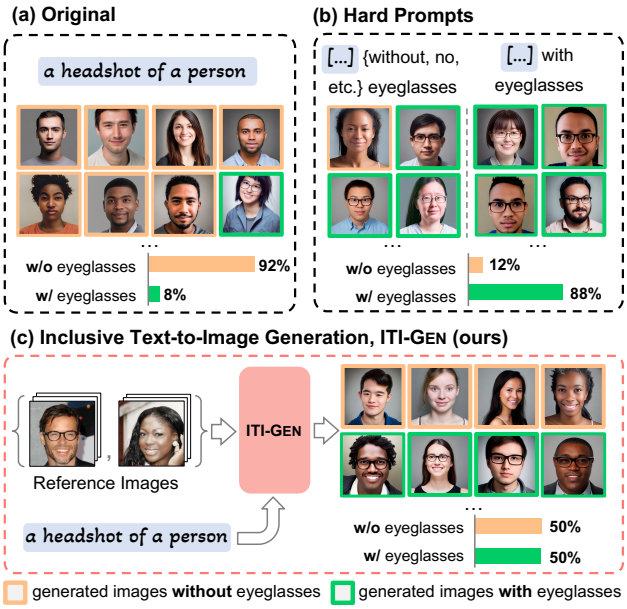


Figure 1. **(a)** Given a human-written prompt ("*a headshot of a person*"), existing text-to-image models [64] can hardly synthesize pictures representing minority groups (*i.e.*, people with eyeglasses in this example). **(b)** Conventional hard prompt searching [19] is sub-optimal due to linguistic ambiguity. **(c)** We address these problems by leveraging a small set of reference images for inclusive text-to-image generation (ITI-GEN).

[1]Project page: https://czhang0528.github.io/iti-gen

ness[2] in text-to-image generation using only a few example images, as illustrated in Figure 1(c).

To advance inclusive generation, a straightforward way is to retrain or fine-tune the model upon request, using *truly* inclusive training data [18, 84]. Doing so, however, is insurmountably challenging as collecting large-scale training data that is balanced/inclusive across all attributes of interest is impractical, and training generative models is highly compute-intensive [68, 66, 18]. Another principled approach towards inclusiveness is to specify or enumerate each category in natural language (*i.e.*, hard prompt searching) [19, 56]. However, many categories are difficult to specify with natural language (*e.g.*, skin tone) or cannot be well synthesized by the existing models due to linguistic ambiguity or model misrepresentation [30].

At first glance, these seem to paint a grim picture for inclusive text-to-image generation. However, we argue that instead of specifying attributes explicitly using descriptive natural language, images can represent specific concepts or attributes more efficiently. Observing the availability of a shared vision-language embedding in many multimodal generative models [57], we raise the question: *can we learn inclusive prompt embeddings using images as guidance?*

To achieve this goal, we introduce **ITI-GEN**, a novel and practical framework that creates discriminative prompts based on readily available reference images for **I**nclusive **T**ext-to-**I**mage **GEN**eration. Concretely, we leverage the vision-language pre-trained CLIP model [57] to obtain the embeddings of the reference images and learnable prompts. In the joint embedding space, we design a new training objective to align the directions of the image and prompt features. The core idea is to translate the visual attribute differences into natural language differences such that the generated images based on the learned prompts can effectively represent all desired categories. By equalizing the sampling process over the learned prompts, our method guarantees inclusiveness for text-to-image generation.

We validate our framework with Stable Diffusion [64]. ITI-GEN can leverage reference images from different domains, including human faces [44, 35, 21] and scenes [69], to achieve inclusive generation in single or multiple attributes of interest. ITI-GEN needs neither prompt specification nor model fine-tuning, bypassing the problems of linguistic ambiguity as well as computational complexity. Moreover, ITI-GEN is compatible with the existing text-based image generation models (*e.g.*, ControlNet [83] and instruction-based image editing models [7]) in a plug-and-play manner. To the best of our knowledge, this is the first method that allows inclusive text-to-image generation over a frozen model and obtains competitive results throughout.

[2]Few works [12, 5] have studied fairness issues in text-to-image generation but mainly focused on social biases (*e.g.*, perceived gender, ethnicity). This paper incorporates a broader spectrum of attributes.

2. Related Work

Text-to-Image Generative Models. Text-based image generation has been widely studied with numerous model architectures and learning paradigms [49, 63, 72, 60, 24, 81, 19, 20, 9, 70, 80, 16, 18, 39]. Recently, the overwhelming success of diffusion-based text-to-image models [59, 67, 59, 52] has attracted significant attention. A key factor to this success is their ability to deal with large-scale multimodal datasets [68, 36, 11]. Thus, questions concerning inclusiveness while learning with biased datasets remain a crucial open problem [12, 5, 3].

Bias Mitigation in Text-to-Image Generation. While fairness has been studied extensively in discriminative models [75, 76, 77, 43], research on developing fair generative models is limited [85, 31, 23, 14, 47]. Most efforts focus on GAN-based models [13, 58, 32, 61, 82, 37, 79, 71, 34, 48], restricting their applicability to the emerging diffusion-based text-to-image models. Recently, there have been some efforts to address this limitation. For instance, Bansal *et al.* [5] proposed to diversify model outputs by ethical intervention[3]. Ding *et al.* [19] proposed to directly add attribute words to the prompt. However, these hard prompt searching methods have limitations such as being opaque and laborious [5], and not always generating diverse images reliably [30, 5]. In this work, we incorporate a broad spectrum of attributes beyond social groups. Moreover, we learn inclusive prompts in the continuous embedding space, requiring no hard prompt specification.

To learn a fair generative model, Wu *et al.* [78] employed off-the-shelf models, such as CLIP [57] and pre-trained classifiers, as guidance. Choi *et al.* [13] used a reference dataset to train the model via sample re-weighting. In contrast, we use reference data in a drastically different way — treating the images as proxy signals to guide prompt learning but without retraining the text-to-image model.

Image-Guided Prompt Tuning. Our method is inspired by Prompt Tuning (PT) [42, 33]. Typically, PT methods insert small learnable modules (*e.g.*, tokens) into the pre-trained models and fine-tune these modules with downstream tasks while freezing the model parameters. Recently, PT has been leveraged in personalized text-to-image generation [25, 65, 40]. By providing several reference images with the customized subject, they use a special token to represent the object by optimizing the token embedding [25, 40] or the diffusion models [65, 40]. This motivates us to learn the specific token embedding for each attribute category for inclusiveness. However, we note that the previously mentioned methods for personalization do not effectively capture the attributes in the images. Thus, we propose to optimize the directions of the attribute-specific

[3]*e.g.*, appending "irrespective of their gender" to the end of a neutral prompt "a photo of a lawyer" for generating diverse pictures w.r.t. genders.

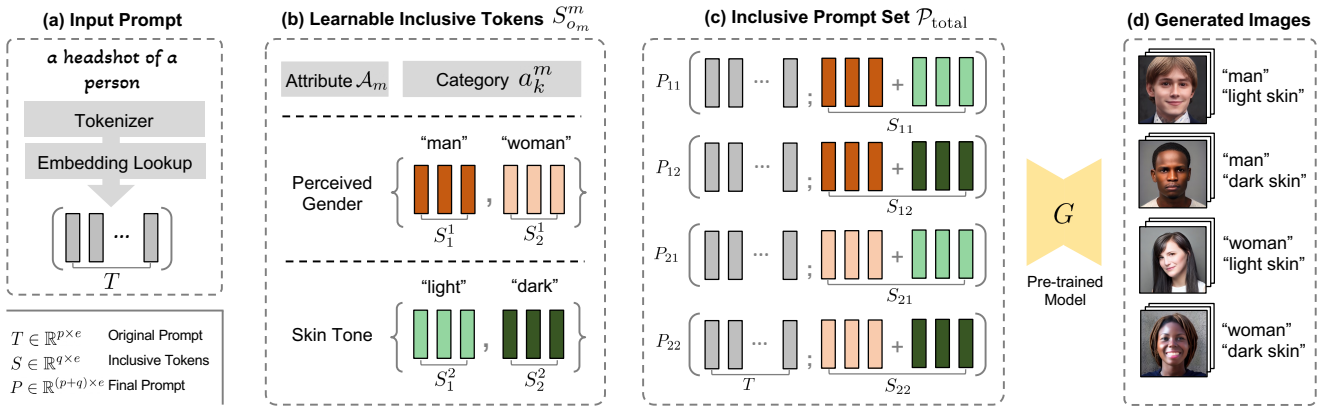


Figure 2. **Illustration of Inclusive Text-to-Image GENeration (ITI-GEN)** with the example of two binary attributes: *perceived gender* and *skin tone*. **(a)** Given an input prompt, **(b)** ITI-GEN learns discriminative token embeddings to represent each category of every target attribute. **(c)** By injecting the learned tokens after the original input prompt, ITI-GEN synthesizes an inclusive prompt set that can be used to **(d)** sample equal (or controllable) numbers of images for any category combination. Further, our framework can be easily extended to multi-category multi-attribute scenarios of inclusive text-to-image generation. Note that, in practice, multi-category skin tones beyond {"light", "dark"} as in this example may be challenging to specify with language (see Figure 3). Please see Section 3.1 for details.

prompts in the joint vision-language embedding space, bypassing training text-to-image generative models.

3. Inclusive Text-to-Image Generation

To drive the progress of Inclusive Text-to-Image Generation, we propose ITI-GEN, which creates inclusive prompts that represent various attributes and their combinations. This is particularly challenging for attributes that are difficult to describe in language or underrepresented. To address this, ITI-GEN uses readily available reference images as guidance, enabling unambiguous specification of different attributes. Figure 2 illustrates the overall framework. In this section, we first introduce the framework of ITI-GEN in Section 3.1, then describe the details of the learning strategy in Section 3.2, and finally discuss the key properties of ITI-GEN in Section 3.3.

3.1. Overview

Problem Statement. Given a pre-trained text-to-image generative model G and a human-written prompt (*e.g.*, *"a headshot of a person"*) tokenized as $\boldsymbol{T} \in \mathbb{R}^{p\times e}$, where p is the number of tokens and e is the dimension of the embedding space, we aim to sample equal (or controllable) numbers of images that can represent any *category* combination given the *attribute* set $\boldsymbol{A}$. Formally,

$$\boldsymbol{A} = \{\mathcal{A}_m | 1 \le m \le M\}; \mathcal{A}_m = \{a^m_k | 1 \le k \le K_m\} \quad (1)$$

contains M different attributes (*e.g.*, perceived gender, skin tone, *etc.*), where a^m_k records a mutually exclusive category (*e.g.*, a specific type of skin tone) in attribute $\mathcal{A}_m$ and K_m denotes the number of categories in $\mathcal{A}_m$. Note that K_m may vary among different attributes.

Inclusive Prompt Set. Inspired by [42, 33], we propose prompt tuning for inclusive generation. Specifically, for a given category a^m_k within attribute $\mathcal{A}_m$, we inject q *learnable* tokens $\boldsymbol{S}^m_k \in \mathbb{R}^{q\times e}$ after the original $\boldsymbol{T}$ to construct a new prompt $\boldsymbol{P}^m_k = [\boldsymbol{T}; \boldsymbol{S}^m_k] \in \mathbb{R}^{(p+q)\times e}$. By querying the model G with $\boldsymbol{P}^m_k$, we can generate images exhibiting the characteristics of the corresponding category a^m_k. To differentiate the new tokens $\boldsymbol{S}^m_k$ from the original prompt $\boldsymbol{T}$, we refer to them as *inclusive tokens*.

When jointly considering M attributes, we aggregate M separate inclusive tokens $\boldsymbol{S}^1_{o_1}, \boldsymbol{S}^2_{o_2}, \ldots, \boldsymbol{S}^M_{o_M}$ to represent a specific category combination $(a^1_{o_1}, a^2_{o_2}, \ldots, a^M_{o_M})$, *e.g.*, the concept of ("woman", "dark skin", ..., "young"). We thus expect to create a unique $\boldsymbol{S}_{o_1 o_2 \ldots o_M}$,

$$\boldsymbol{S}_{o_1 o_2 \ldots o_M} = f(\boldsymbol{S}^1_{o_1}, \boldsymbol{S}^2_{o_2}, \ldots, \boldsymbol{S}^M_{o_M}) \quad (2)$$

that can be injected after $\boldsymbol{T}$ to generate images for this particular category combination. The aggregation function f in Equation 2 should be able to take various numbers of attributes while maintaining the permutation invariant property[4] with respect to attributes. Common options include element-wise average, sum, and max operations. Following [50], we adopt element-wise sum to preserve the text semantics without losing information[5]. Finally, we define the *inclusive prompt set* as follows:

$$\mathcal{P}_{\text{total}} = \{\boldsymbol{P}_{o_1 o_2 \ldots o_M} = [\boldsymbol{T}; \sum_{m=1}^{M} \boldsymbol{S}^m_{o_m}] \in \mathbb{R}^{(p+q)\times e} \mid 1 \le o_1 \le K_1, \ldots, 1 \le o_M \le K_M\}. \quad (3)$$

[4] That is, the output of f should be the same even if we permute the indices m of the attributes in $\boldsymbol{A}$ (cf. Equation 1).

[5] Please see Appendix E.2 for more analysis and other options for aggregating multiple tokens, *e.g.*, concatenation.

By uniformly sampling the prompts from $\mathcal{P}_{\text{total}}$ as the conditions to generate images using the generative model G, we achieve inclusiveness across all attributes (see Figure 2). *More generally speaking, the distribution of the generated data is directly correlated to the distribution of the prompts, which can be easily controlled.*

In contrast to specifying the category name in discrete language space [5, 19], we optimize prompts entirely in the *continuous* embedding space. Additionally, we only update the attribute-specific embeddings — the colors • and • in Equation 3 indicate frozen and learnable parameters, respectively. This decoupled optimization mechanism thus provides the advantage of using the learned inclusive tokens in a plug-and-play manner across various applications, as will be demonstrated in Section 3.3 and Section 4.3. We elaborate on the learning process in the following section.

3.2. Learning Inclusive Prompts

Reference Image Set. We propose using reference images to guide prompt learning, as they can provide more expressive signals to describe attributes that may be challenging to articulate through language. Specifically, we assume the availability of a reference image set $\mathcal{D}_{\text{ref}}^m = \{(\boldsymbol{x}_n^m, y_n^m)\}_{n=1}^{N_m}$ for a target attribute $\mathcal{A}_m$, where N_m is the dataset size and $y_n^m \in \mathcal{A}_m$ (defined in Equation 1) indicates the category to which $\boldsymbol{x}_n$ belongs. When considering multiple attributes, we only need a reference dataset for each attribute, rather than one large balanced dataset with all attribute labels. *This property is extremely beneficial, as it is much easier to obtain a dataset that captures only the distribution of one attribute (i.e., the marginal distribution) rather than one that captures the joint distribution of all attributes.*

Aligning Prompts to Images with CLIP. Given reference image sets for the target attributes, can we learn prompts that align the attributes in the images? Recently, pre-trained large-scale multimodal models have demonstrated strong capabilities in connecting vision and language. One such model is CLIP [57], which aligns visual concepts with text embeddings by jointly training a text encoder E_{text} and an image encoder E_{img}. The output of the pre-trained CLIP text encoder has also been used as the condition for text-guided image generation [64, 59], opening up an opportunity to align prompts to reference images without the need to modify the text-to-image models.

One straightforward solution is to maximize the similarity between the prompt and the reference image embeddings in the CLIP space, as suggested by [57]. However, we found it deficient for two reasons. First, this objective forces the prompt to focus on the overall visual information in the images, rather than the specific attribute of interest. Second, the generated images from the learned prompt often exhibit adversarial effects or significant quality degradation, potentially due to image features distorting the prompt embedding. To address these, we propose direction alignment and semantic consistency losses, as described below.

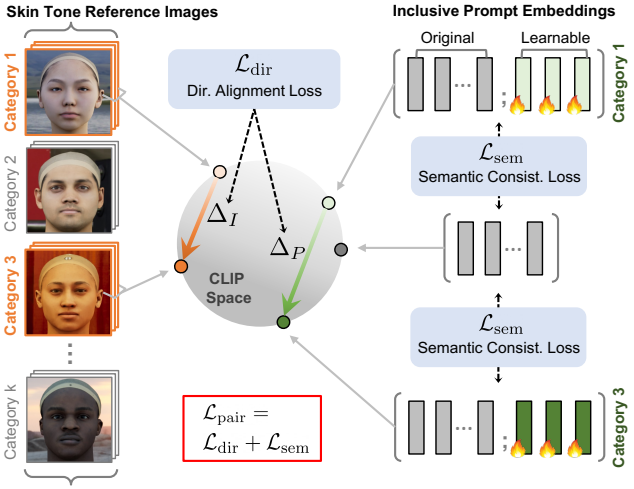


Figure 3. **Translating visual differences into text embedding differences.** Given reference images of a multi-category attribute (*e.g.*, skin tone), we learn the inclusive tokens by direction alignment between **images** and **prompts**, ensuring that the visual difference matches the learned language description. In addition, we propose semantic consistency loss to address language drift. Images are from FAIR benchmark [21]. Details are in Section 3.2.

Direction Alignment Loss. Instead of directly maximizing the similarity between the prompts and the images, we draw inspiration from [55, 26] to induce the direction between the prompt $\boldsymbol{P}_i^m$ and $\boldsymbol{P}_j^m$ to be aligned with the direction between the averaged embeddings of the reference images corresponding to *a pair of categories* a_i^m and a_j^m in $\mathcal{A}_m$. This alignment of pairwise categories direction serves as a proxy task for guiding the prompts to learn the visual difference among images from category a_i^m and a_j^m (Figure 3).

Specifically, we define the direction alignment loss $\mathcal{L}_{\text{dir}}$ to maximize the cosine similarity between the image direction and the prompt direction as follows:

$$\mathcal{L}_{\text{dir}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m) = 1 - \left\langle \Delta_{\boldsymbol{I}}^m(i,j), \Delta_{\boldsymbol{P}}^m(i,j) \right\rangle. \tag{4}$$

Here, the image direction $\Delta_{\boldsymbol{I}}$ is defined as the difference of the averaged image embeddings between two categories of the attribute $\mathcal{A}_m$. Let $\mathfrak{X}_k^m = \frac{1}{|\mathcal{B}_k|} \sum_{y_n^m = a_k^m} E_{\text{img}}(\boldsymbol{x}_n^m)$ be the averaged image embedding for category a_k^m; $|\mathcal{B}_k|$ is the number of images from category a_k^m in each mini-batch. We denote the image direction as follows:

$$\Delta_{\boldsymbol{I}}^m(i,j) = \mathfrak{X}_i^m - \mathfrak{X}_j^m. \tag{5}$$

Similarly, the prompt direction $\Delta_{\boldsymbol{P}}$ is defined as the difference of the averaged prompt embeddings between two categories. Let $\mathfrak{P}_k^m = \frac{1}{|\mathcal{P}_k^m|} \sum_{\boldsymbol{P} \in \mathcal{P}_k^m} E_{\text{text}}(\boldsymbol{P})$ be the averaged prompt embedding for attribute a_k^m. Specifically,

$\mathcal{P}_k^m = \{\boldsymbol{P} \in \mathcal{P}_{\text{total}} \mid o_m = k\}$ is a collection of prompts containing all the category combinations for other attributes given the category a_k^m for attribute $\mathcal{A}_m$ (cf. Equation 3). Finally, we denote the prompt direction as follows:

$$\Delta_{\boldsymbol{P}}^m(i,j) = \mathfrak{P}_i^m - \mathfrak{P}_j^m. \tag{6}$$

By inducing the direction alignment, we aim to facilitate the prompt learning of more meaningful and nuanced differences between images from different categories.

Semantic Consistency Loss. We observe that direction alignment loss alone may result in language drift [46, 41, 65] — the prompts slowly lose syntactic and semantic properties of language as they only focus on solving the alignment task. To resolve this issue, we design a semantic consistency objective to regularize the training by maximizing the cosine similarity between the learning prompts and the original input prompt (see Figure 3):

$$\mathcal{L}_{\text{sem}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m) = \max\Big(0, \lambda - \big\langle E_{\text{text}}(\boldsymbol{P}), E_{\text{text}}(\boldsymbol{T})\big\rangle\Big) \tag{7}$$

where $\boldsymbol{P} \in \mathcal{P}_i^m \cup \mathcal{P}_j^m$ and λ is a hyperparameter (see an analysis in Section 4.3). This loss is crucial for generating high-quality images that remain faithful to the input prompt.

Optimization. Building upon $\mathcal{L}_{\text{dir}}^m$ and $\mathcal{L}_{\text{sem}}^m$, our total training loss for learning the inclusive tokens of a pair of categories in attribute $\mathcal{A}_m$ is written as follows:

$$\mathcal{L}_{\text{pair}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m) = \mathcal{L}_{\text{dir}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m) + \mathcal{L}_{\text{sem}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m). \tag{8}$$

At each iteration, we update the embeddings of inclusive tokens of all the categories from *only one attribute* but freeze the parameters of inclusive tokens for all other attributes. The final objective during the whole learning process is:

$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M} \sum_{1 \le i < j \le K_m} \mathcal{L}_{\text{pair}}^m(\boldsymbol{S}_i^m, \boldsymbol{S}_j^m), \tag{9}$$

where the inner summation enumerates all pairwise categories for one attribute $\mathcal{A}_m$ at each iteration, while the outer summation alters the attribute across the iteration.

3.3. Key Properties of ITI-GEN

Generalizability. Unlike personalization methods that train the embeddings for a specific model (because they use diffusion losses [25, 40, 65]), *the tokens learned by* ITI-GEN *are transferable between different models.* We highlight two use cases for these tokens. (1) *In-domain generation.* We use the user-specified prompt $\boldsymbol{T}$ to learn the inclusive tokens and then apply them back to $\boldsymbol{T}$ to generate inclusive images. (2) *Train-once-for-all.* As shown in Equation 3, the newly introduced inclusive tokens do not change the original prompt $\boldsymbol{T}$, which implies that the learned tokens can be compatible with a different human-written prompt. For human face images, an example $\boldsymbol{T}$ for training can be any neutral prompt, *e.g.*, *"a headshot of a person"*. After training, inclusive tokens can be used to handle out-of-domain prompts (*e.g.*, *"a photo of a doctor"*) or facilitate different models [83, 7] in a plug-and-play manner, justifying the generalizability of our approach.

Data, Memory, and Computational Efficiency. ITI-GEN uses averaged image features to guide prompt learning, indicating that (1) only a few dozen images per category are sufficient, and (2) a balanced distribution across categories within an attribute is *not* required. ITI-GEN keeps the text-to-image model intact and only updates the inclusive tokens, allowing it to circumvent the costly back-propagation step in the diffusion model. Training with a single attribute takes approximately 5 minutes (1 A4500 GPU). In practice, we set the length[6] (q in Equation 3) of inclusive tokens to 3 (which is less than 10KB) for all attribute categories of interest in our study. Hence, when scaling up to scenarios with multiple attributes, ITI-GEN always has low memory requirements for both training and storing inclusive tokens.

Comparison to Image Editing Methods. Our direction alignment loss may be reminiscent of the directional CLIP loss employed in image editing methods [26, 38]. However, they are fundamentally different. First, our ITI-GEN is designed to promote the inclusiveness, while image editing methods focus on single image manipulation. Second, image editing methods modify the source image according to the change in texts (from source to target), whereas ITI-GEN learns prompts by leveraging changes in images from one category to another. This key difference suggests a significant distinction: the two methods are learning the task from completely different directions.

4. Experiments

We validate ITI-GEN for inclusive text-to-image generation on various attributes and scenarios. We begin by introducing the experimental setup in Section 4.1, then present the main results in Section 4.2, and finally, show detailed ablation studies and applications in Section 4.3. Please see Appendix for additional details, results, and analyses.

4.1. Setup

Datasets. We construct reference image sets and investigate a variety of attributes based on the following datasets. **(1) CelebA** [44] is a face attributes dataset and each image with 40 binary attribute annotations. We experiment with these binary attributes and their combinations. **(2) FAIR benchmark (FAIR)** [21] is a recently proposed synthetic face dataset used for skin tone estimation. Following [21],

[6]The token length used here is generalizable across the attributes we studied in this paper. See Appendix E.1 for a detailed ablation study.

Table 1. **Comparison with baseline methods with (a) single attribute and (b) multiple attributes.** Reference images are from CelebA. We use CLIP [57] as the attribute classifier [12, 14]. ITI-GEN achieves competitive results for both settings. **SD**: vanilla stable diffusion. **EI**: ethical intervention. **HPS**: hard prompt searching. **PD**: prompt debiasing. **CD**: custom diffusion. See Appendix F for full results.

Method	(a) Single Attribute						(b) Multiple Attributes		
	$\mathbb{D}_{\text{KL}}^{\text{male}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{young}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{pale skin}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{eyeglass}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{mustache}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{smile}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{male}\times\text{young}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{male}\times\text{young}\times\text{eyeglass}} \downarrow$	$\mathbb{D}_{\text{KL}}^{\text{male}\times\text{young}\times\text{eyeglass}\times\text{smile}} \downarrow$
SD [64]	0.343	0.578	0.308	0.375	0.111	0.134	0.882	1.187	1.406
EI [5]	0.143	0.423	0.644	0.531	0.693	0.189	0.361	1.054	1.311
HPS [19]	1×10^{-5}	0.027	2.8×10^{-3}	0.371	0.241	4.4×10^{-3}	3.5×10^{-3}	0.399	0.476
PD [14]	0.322	0.131	0.165	0.272	0.063	0.146	–	–	–
CD [40]	0.309	0.284	0.074	0.301	0.246	0.469	–	–	–
ITI-GEN	$\mathbf{2 \times 10^{-6}}$	$\mathbf{2 \times 10^{-4}}$	**0**	$\mathbf{2 \times 10^{-4}}$	$\mathbf{4.5 \times 10^{-4}}$	$\mathbf{2.5 \times 10^{-3}}$	$\mathbf{1.3 \times 10^{-4}}$	**0.061**	**0.094**

Figure 4. **Qualitative results of the combination of four binary attributes** (the last column in Table 1). The input prompt ($\boldsymbol{T}$) is *"a headshot of a person"*. By using the learned inclusive tokens (cf. Equation 3), ITI-GEN can inclusively generate images with all attribute combinations. Images across each tuple are sampled using the same random seed. More examples are included in Appendix F.

Figure 5. **Examples of reference images.** CelebA [44] and FairFace [35] are real-face datasets with different resolutions and focuses. FAIR benchmark [21] is a synthetic dataset used for skin tone estimation. Landscape (LHQ) [69] contains images from natural scenes. ITI-GEN can leverage various image sources to benefit inclusive text-to-image generation for various attributes.

we use the ground-truth albedos to classify each facial crop into one of six skin tone levels [22] and use FAIR for inclusiveness on skin tone type. **(3) FairFace** [35][7] contains face images with annotations for 2 perceived gender and 9 perceived age categories. **(4) Landscapes HQ (LHQ)** [69] provides unlabeled natural scene images. With the annotation tool from [74], each image can be labeled with 6 quality (*e.g.*, colorfulness, brightness) and 6 abstraction (*e.g.*, scary, aesthetic) attributes. Figure 5 shows example images.

Experimental Protocols. We only require that a reference image set captures a marginal distribution for each attribute (cf. Section 3.2). Note that, while images from CelebA and FairFace are annotated with multiple attributes, we use only the attribute label for each target category but not others. We randomly select 25 reference images per category as our default setting (and ablate it in Section 4.3). For attribute settings, we consider *single binary attribute*, *multi-category attributes*, and *multiple attributes* in the domains of human faces and scenes. We study both in-domain and train-once-for-all generations (cf. Section 3.3) and further provide qualitative and quantitative analyses for each setup.

Quantitative Metrics. We use two metrics to quantify distribution diversity and image quality. (1) *Distribution Discrepancy* ($\mathbb{D}_{\text{KL}}$). Following [12, 14], we use the CLIP model to predict the attributes in the images. For attributes that CLIP might be erroneous, we leverage pre-trained classifiers [35] combined with human evaluations. Specifically, for skin tone, which is extreme difficult to obtain an accurate scale [1, 2, 29], we adopt the most commonly used Fitzpatrick skin type [10] combined with off-the-shelf

[7] We note that, while the FairFace dataset contains race categories, we focus instead on skin tone in this study. This is because skin tone is more readily inferable from pixels, whereas racial identities are better understood as social concepts that are neither immutable nor biological in nature [8, 15, 62, 4]; furthermore, phenotypic variation of skin tone within racial identification groups is well documented [51].

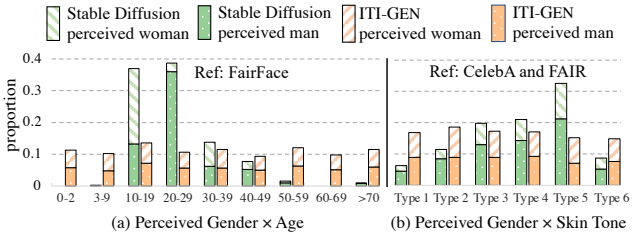


Figure 6. **Multi-category distribution** with "*a headshot of a person*". For a reliable evaluation, the results of (a) are evaluated using classifiers in [35], and (b) are evaluated using existing models [10, 21]. The generated images from ITI-GEN are more uniformly distributed across different sub-groups than the baseline Stable Diffusion. See Figure 7 for qualitative results.

models [21] for evaluation. (2) *FID*. We report the FID score [27, 53] (FFHQ [36]) to measure the image quality. Please see Appendix E for more details.

Baselines. We compare ITI-GEN to the following methods. (1) *Stable Diffusion* (SD) [64] without any modification. (2) *Ethical Intervention* (EI) [5] that edits the prompt by adding attribute-related interventions. (3) *Hard Prompt Searching* (HPS) [19] that directly expresses the desired attribute category in the prompt. (4) *Prompts Debiasing* (PD) [14] that calibrates the bias in the text embedding by using the attribute category names. (5) *Custom Diffusion* (CD) [40] that fine-tunes the text-to-image model with reference images based on Textual Inversion [25, 65].

Implementation Details. We use Stable Diffusion [64] (sd-v1-4) as the base model for all methods and show compatibility with ControlNet [83] and InstructPix2Pix [7]. ITI-GEN is model agnostic as long as they take token embeddings as the inputs. We set $\lambda = 0.8$ in $\mathcal{L}_{\text{sem}}$ across all experiments and show that λ can be robustly selected according to the prior knowledge (see Section 4.3). All the inclusive tokens are initiated as zero vectors[8]. We set the length of the inclusive tokens to 3 in all experiments. *There is no additional hyper-parameter in our framework.* The total number of the parameters for the inclusive tokens that need to be optimized is $\sum_{m=1}^{M} K_m \times 3 \times 768$, where M is the number of attributes, K_m is the category number for attribute m, and 768 is the dimension of the embedding (e in Equation 3). We train the models with 30 epochs on a batch size of 16 and a learning rate of 0.01. During training, we leverage image augmentations used in the CLIP image encoder.

4.2. Main Results

Single Binary Attribute. To demonstrate the capability of ITI-GEN to sample images with a variety of face attributes, we construct 40 distinct reference image sets based on attributes from CelebA [44]. Each represents a specific

Figure 7. **Results of ITI-GEN on multi-category attributes** for Gender×Age (Figure 6(a)) and Gender×Skin Tone (Figure 6(b)). Examples are randomly picked with "*a headshot of a person*".

binary attribute and contains an equal number of images (50%) for the positive and negative categories[9]. Table 1(a) shows a comparison to state-of-the-art methods. We evaluate 5 text prompts — "*a headshot of a* {*person, professor, doctor, worker, firefighter*}" — and sample 200 images per prompt for each attribute, resulting in 40K generated images. We highlight the averaged results across 5 prompts of 6 attributes. We provide complete results in Appendix F.2. ITI-GEN achieves near-perfect performance on balancing each binary attribute, justifying our motivation: using separate inclusive tokens is beneficial in generating images that are uniformly distributed across attribute categories.

Multiple Attributes. Given multiple reference image sets (each captures the marginal distribution for an attribute), can ITI-GEN generate diverse images across any category combination of the attributes? We provide an affirmative answer and present results in Table 1(b) and Figure 4. As we observe, ITI-GEN produces diverse and high-quality images with significantly lower distribution discrepancies compared to baseline methods. We attribute this to the aggregation operation of inclusive tokens (Equation 3), allowing ITI-GEN to disentangle the learning of different inclusive tokens with images in marginal distributions.

Multi-Category Attributes. We further investigate multi-category attributes including perceived age and skin tone. Specifically, we consider two challenging settings: (1) Perceived Gender × Age (Figure 6(a)), and (2) Perceived Gender × Skin Tone (Figure 6(b)). ITI-GEN achieves inclusiveness across all setups, especially on extremely underrepresented categories for age (< 10 and > 50 years old in Figure 6(a)). More surprisingly (Figure 6(b)), ITI-GEN can leverage synthetic images (from FAIR) and jointly learn

[8] We investigated other options such as random initialization but did not see notable differences in both generation quality and training speed.

[9] We found that different ratios do not lead to notable differences. We provide an analysis of learning with imbalanced data in Appendix E.3.

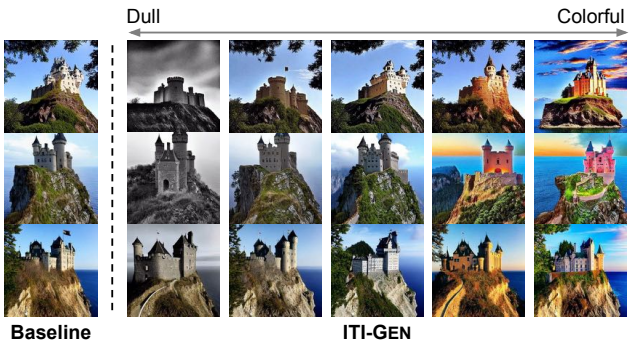


Figure 8. **ITI-GEN with perception attributes on scene images.** The tokens of "colorfulness" are trained with "*a photo of a natural scene*" and applied to "*a castle on the cliff*" in this example (*train-once-for-all* in Section 3.3). ITI-GEN (right) enables the baseline Stable Diffusion (left) to generate images with different levels of colorfulness. Same seed for each row. Better viewed in color. See Appendix F.5 for results of other attributes, *e.g.*, scary, brightness.

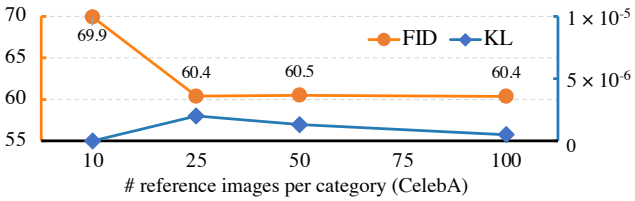


Figure 9. **Ablation on the quantity of reference images.** More reference images (> 10) help possibly due to more diversity and less noise. ITI-GEN is robust in the low data regime (Section 3.3).

Table 2. **Ablation on reference image sources and $\mathcal{L}_{\text{sem}}$.** ITI-GEN produces lower FID than the baseline Stable Diffusion. Semantic consistency loss $\mathcal{L}_{\text{sem}}$ plays a key role in quality control.

Method	Source	$\mathcal{L}_{\text{sem}}$	FID↓
Baseline [64]	–	–	67.40
ITI-GEN	CelebA [44]	✓	**60.38**
		✗	(+17.40) 77.78
	FairFace [35]	✓	**55.10**
		✗	(+9.01) 64.11
	FAIR [21]	✓	**51.83**
		✗	(+10.86) 62.69

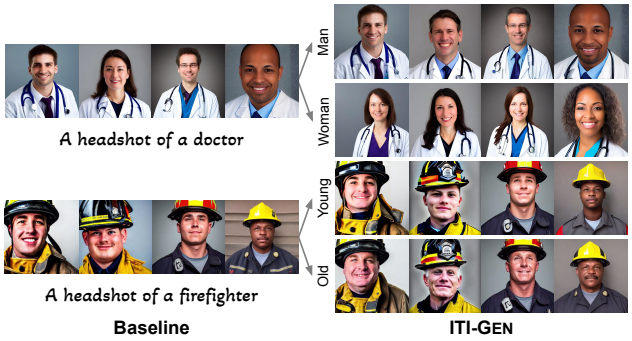


Figure 10. **Train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these two examples to alleviate stereotypes. See Appendix F.6 for more results.

from different data sources (CelebA for gender and FAIR for skin tone), demonstrating great potential for bootstrapping inclusive data generation with graphics engines.

Other Domains. Besides human faces, we apply ITI-GEN to another domain: scene images. We claim that the inclusive text-to-image generation accounts for attributes from not only humans but also scenes, objects, or even environmental factors. Specifically, we use images from LHQ [69] as guidance to learn inclusive tokens and generate images with diverse subjective perception attributes. As illustrated in Figure 8, ITI-GEN can enrich the generated images to multiple levels of colorfulness[10], justifying the generalizability of our method to the attributes in different domains.

4.3. Ablations and Applications

Reference Images. Figure 9 illustrates the impact of the *quantity* of reference images per attribute category, telling that ITI-GEN can produce high-quality images using very few reference data without sacrificing inclusiveness (KL). In addition, as indicated in Table 2, ITI-GEN consistently generates realistic images regardless of reference sources (see examples in Figure 4 and Figure 7). More interestingly, we found that using synthetic images (*i.e.*, FAIR [21]) is slightly better than real data [44, 35]. We hypothesize that the background noise in real images degrades the quality.

Semantic Consistency Loss $\mathcal{L}_{\text{sem}}$. Again in Table 2, we compare ITI-GEN with and without $\mathcal{L}_{\text{sem}}$. With the help of the semantic constraint (Figure 3), we regularize the learned embeddings not too far from the original prompt. We show evidence to verify this insight: the averaged CLIP similarity scores of text features between the hard prompts of 40 attributes in CelebA and the original prompt is 0.8 (the λ we used), suggesting that the hyper-parameter can be robustly chosen based on prior linguistic knowledge.

Train-once-for-all Generalization. As shown in Figure 8, inclusive tokens can be applied to user-specified prompts in a plug-and-play manner (Section 3.2). In Figure 10, we provide more examples of professional prompts to demonstrate the ability of train-once-for-all generation.

Compatibility with ControlNet [83]. ITI-GEN achieves inclusiveness by learning attribute-specific prompts without modifying the original text-to-image model, potentially benefiting various downstream vision-language tasks. In Figure 11, we demonstrate its compatibility with ControlNet [83], a state-of-the-art model capable of conditioning

[10]Note that the subjective attributes we explore here are different from artistic styles (*e.g.*, painting, cartoon) in image-to-image translation (*e.g.*, [26]). Understanding the attributes related to *quality* and *look* of images may be intuitive for humans but remain non-trivial for generative models.

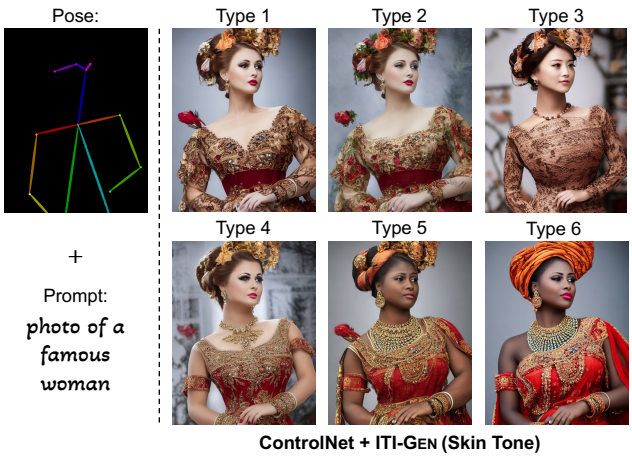


Figure 11. **Compatibility with models using additional conditions**, *e.g.*, human pose (left). ITI-GEN promotes inclusiveness of ControlNet [83] by using the inclusive tokens of six skin tone types (right). The tokens are trained with "*a headshot of a person*" guided by images from FAIR dataset [21], and applied here in a *train-once-for-all* manner (Section 3.3). See Appendix F.7 for additional results on versatile conditions, *e.g.*, depth, segmentation.

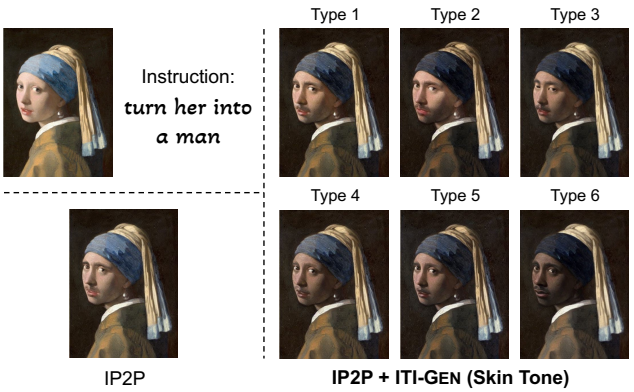


Figure 12. **Compatibility with instruction-based image editing methods.** Given an image and a written instruction (top-left), InstructPix2Pix (IP2P) [7] follows the instruction to edit the image (bottom-left). ITI-GEN (right) enables inclusive instruction-based image editing. Similar to Figure 11, the inclusive tokens used in this example are trained in a train-once-for-all manner.

on a variety of inputs beyond text. Interestingly, we observe an intriguing feature where the newly introduced tokens may implicitly entangle other biases or contrasts inherent in the reference image sets, such as clothing style. Nevertheless, we emphasize that disentanglement of attributes is not the primary concern of this study. ITI-GEN achieves competitive results in distributional control for the *intended* attributes (*e.g.*, skin tone in Figure 11) — aggregating tokens learned from marginal distributions implicitly disentangles the *known* attributes of interest.

Compatibility with InstructPix2Pix (IP2P) [7]. Note that, achieving fully unsupervised disentanglement is a challenging task [45]. Previous attempts in image generation often resort to additional supervision, either through the use of reference data [13], classifiers learned from a joint distribution [71], or even more robust controls such as instruction-based image editing [7]. Here, we show that ITI-GEN can potentially disentangle the target attribute by incorporating InstructPix2Pix [7] — to improve the inclusiveness of IP2P on the target attribute, while ensuring minimal changes to other features such as clothing and background. Results are shown in Figure 12, telling that ITI-GEN can be an effective method to condition diffusion on contrastive image sets, *e.g.*, images taken by different cameras, art by unknown artists, and maybe even different identities of people.

5. Conclusion and Discussion

We present a new method for inclusive text-to-image generation. Our main contribution lies in a new direction: *leveraging readily available reference images to improve the inclusiveness of text-to-image generation.* This problem is timely and challenging [6, 5, 14, 23, 12]. Our key insight is learning separate token embeddings to represent different attributes of interest via image guidance. The proposed ITI-GEN method is simple, compact, generalizable, and effective on various applications. Specifically, ITI-GEN has several advantages: (1) scalable to multiple attributes and different domains using relatively small numbers of images; (2) can be used in a plug-and-play manner to out-of-distribution, relatively complex prompts; (3) efficient in both training and inference; (4) compatible with the text-to-image generative models that support additional conditions or instructions. We conduct extensive experiments to verify the effectiveness of the proposed method on multiple domains, offering insights into various modeling choices and mechanisms of ITI-GEN. We incorporate a broad spectrum of attributes in both human faces and scenes. We hope that our results and insights can encourage more future works on exploring inclusive data generation.

Limitations. ITI-GEN can handle a wide range of general attributes, such as perceived gender and skin tone, and excels in cases where "Hard Prompt" struggles. However, there remain several limitations. First, ITI-GEN does not always provide optimal results for very subtle facial attributes (Appendix F.2) or for the combinations of highly entangled attributes (Appendix F.3). Second, ITI-GEN still requires dozens of reference images for each category as guidance. It is possible that the reference images may introduce biases or inaccuracies. One mitigation strategy is to integrate ITI-GEN with models that offer robust controls [7], such as the one highlighted in Figure 12.

Acknowledgments. We thank Oliver Wang, Jianjin Xu, and Or Patashnik for their feedback on the drafts of this paper.

References

[1] Gender shades. http://gendershades.org/. 6, 14

[2] Google skin tone research. https://skintone.google/. 6, 14

[3] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *preprint arXiv:2108.02818*, 2021. 2

[4] Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. *arXiv preprint arXiv:2302.03629*, 2023. 6, 14

[5] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *EMNLP*, 2022. 1, 2, 4, 6, 7, 9, 16

[6] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT*, 2023. 1, 9

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 5, 7, 9

[8] Simone Browne. *Dark matters: On the surveillance of blackness*. Duke University Press, 2015. 6

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *preprint arXiv:2301.00704*, 2023. 2

[10] Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13(4):191–208, 1991. 6, 7, 14

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *preprint arXiv:1504.00325*, 2015. 2

[12] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *preprint arXiv:2202.04053*, 2022. 1, 2, 6, 9, 14, 17

[13] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*, 2020. 2, 9

[14] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *preprint arXiv:2302.00070*, 2023. 2, 6, 7, 9, 13, 14, 16, 17

[15] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021. 6

[16] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *preprint arXiv:2209.04747*, 2022. 2

[17] Sandra Del Bino and FJBJoD Bernerd. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology*, 169(s3):33–40, 2013. 14

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[19] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 1, 2, 4, 6, 7, 16, 17

[20] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *preprint arXiv:2204.14217*, 2022. 2

[21] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022. 2, 4, 5, 6, 7, 8, 9, 14, 16

[22] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871, 1988. 6

[23] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2, 9, 13

[24] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 2

[25] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *preprint arXiv:2208.01618*, 2022. 2, 5, 7, 13

[26] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022. 4, 5, 8

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7, 14

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[29] John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):550–560, 2021. 6, 14

[30] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks. In *AACL*, 2022. 2

[31] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *preprint arXiv:2001.09528*, 2020. 2

[32] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *ICML*, 2021. 2

[33] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2, 3, 13

[34] Cemre Efe Karakas, Alara Dirik, Eylül Yalçınkaya, and Pinar Yanardag. Fairstyle: Debiasing stylegan2 with style channel manipulations. In *ECCV*, 2022. 2

[35] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. In *WACV*, 2021. 2, 6, 7, 8, 14, 34

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 7, 13, 14

[37] Patrik Joslin Kenfack, Kamil Sabbagh, Adín Ramírez Rivera, and Adil Khan. Repfair-gan: Mitigating representation bias in gans using gradient clipping. *preprint arXiv:2207.10653*, 2022. 2

[38] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 5

[39] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 2

[40] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *preprint arXiv:2212.04488*, 2022. 2, 5, 6, 7, 13, 16

[41] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *preprint arXiv:1909.04499*, 2019. 5, 15

[42] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 2, 3, 13

[43] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022. 2

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 5, 6, 7, 8, 14, 16, 17

[45] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019. 9

[46] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *ICML*, 2020. 5

[47] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 1, 2

[48] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *ECCV*, 2022. 2

[49] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2016. 2

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 2013. 3

[51] Ellis P Monk Jr. The unceasing significance of colorism: Skin tone stratification in the united states. *Daedalus*, 150(2):76–90, 2021. 6

[52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *preprint arXiv:2112.10741*, 2021. 2

[53] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 7, 14

[54] Otávio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Debiasing methods for fairer neural models in vision and language research: A survey. *preprint arXiv:2211.05617*, 2022. 1

[55] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 4

[56] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *preprint arXiv:2211.12112*, 2022. 1, 2

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4, 6, 17

[58] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *CVPR*, 2021. 2

[59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *preprint arXiv:2204.06125*, 2022. 1, 2, 4

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 2

[61] Harsh Rangwani, Naman Jaswani, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Improving gans for long-tailed data through group spectral regularization. In *ECCV*, 2022. 2

[62] Victor Ray. *On critical race theory: why it matters & why you should care*. Random House, 2022. 6

[63] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 6, 7, 8, 16, 17

[65] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2022. 2, 5, 7, 13, 15

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2

[67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *preprint arXiv:2205.11487*, 2022. 2

[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1, 2

[69] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *ICCV*, 2021. 2, 6, 8, 14, 23, 24, 25, 26, 27, 28, 34

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2

[71] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *preprint arXiv:2012.04842*, 2020. 2, 9

[72] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, 2022. 2

[73] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes. In *ACCV*, 2017. 17

[74] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6, 14

[75] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *CVPR*, 2020. 2

[76] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019. 2

[77] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020. 2

[78] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando De la Torre. Generative visual prompt: Unifying distributional control of pre-trained generative models. In *NeurIPS*, 2022. 2

[79] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *ICBD*, 2018. 2

[80] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *preprint arXiv:2209.00796*, 2022. 2

[81] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *preprint arXiv:2206.10789*, 2022. 2

[82] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *ECCV*, 2020. 2

[83] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *preprint arXiv:2302.05543*, 2023. 2, 5, 7, 8, 9, 34

[84] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *ICML*, 2020. 2

[85] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In *NeurIPS*, 2018. 2

Appendix

A. Ethical and Social Impacts

One important consideration is the potential impact on privacy and data protection. In order to generate inclusive images, ITI-GEN relies on reference images that are often sourced from publicly available datasets. However, the utilization of these images raises concerns about privacy and the potential for unintended consequences, such as the misuse of personal data. It is crucial to consider ways to mitigate these risks, such as data anonymization or obtaining explicit consent from individuals whose images are used.

While ITI-GEN's directional loss avoids directly measuring the distance between the prompts and the reference images, it is possible that the reference images used to represent certain attributes may themselves contain biases or inaccuracies. To address this concern, it will be important to carefully evaluate the quality and representativeness of the reference images used in the model and to develop strategies for identifying and correcting biases when they arise.

Inclusive image generation has the potential to promote greater representation and diversity in various industries, which could in turn promote greater social equality and reduce discrimination. However, it is also possible that the technology could be misused or weaponized to promote negative or harmful stereotypes. Therefore, it will be important to consider the potential risks and benefits of ITI-GEN carefully for mitigating negative outcomes.

B. Additional Related Work and Comparisons

In this section, we provide a more comprehensive comparison between ITI-GEN and related methods.

Bias Mitigation Methods in Text-to-Image Generation. As mentioned in Section 2 of the main paper, ITI-GEN uses images as guidance while existing approaches focus on debiasing the prompts. Two concurrent works, Prompt Debiasing [14] and Fair Diffusion [23] require the category names of the target attributes for learning fair prompts. However, we argue that, for some attributes, attribute names might be hard to specify using language (*e.g.*, skin tone, different levels of brightness). ITI-GEN learns tokens without gradient propagation through the original text-to-image models, making it more efficient in both training and deployment.

Personalization. Both ITI-GEN and personalized text-to-image generation methods [40, 25] are inspired by prompt tuning [33, 42]. However, they are fundamentally different, as introduced in Section 2 of the main paper. We compare with custom diffusion [40] in Table 1 of the main paper mainly to provide a justification for whether the personalization methods [40, 65, 25] can be used in inclusive text-to-image generation. Specifically, we attempt to provide different numbers of reference images for Custom Diffusion [40] and select the best results to report. Moreover, unlike personalization methods that use diffusion losses to train the special tokens, the tokens learned by ITI-GEN are generalizable between different models.

Disentanglement. It is worth mentioning that the aggregation of multiple inclusive tokens learned with separate reference datasets in marginal distributions can implicitly disentangle attribute learning. However, we emphasize that the primary goal of ITI-GEN is *not* to achieve feature (or attribute) disentanglement [36]. Please see Section 4.3 and Figure 11 of the main paper for a detailed discussion.

Image-to-image Translation and Editing. As mentioned in Section 3.3 of the main paper, the goal of our work is to promote inclusiveness or diversity but not for image editing. In image-to-image translation or editing tasks, it is required to edit the desired attribute while keeping other features of the image intact. However, we *do not* have such a requirement for ITI-GEN. For example, in Figure 4, Figure 7, Figure 10, and Figure 11 of the main paper, while there are subtle changes to the clothing or background in the images, ITI-GEN *already achieves inclusiveness for the intended attribute. We show examples with the same random seeds in these figures mainly for a better comparison.*

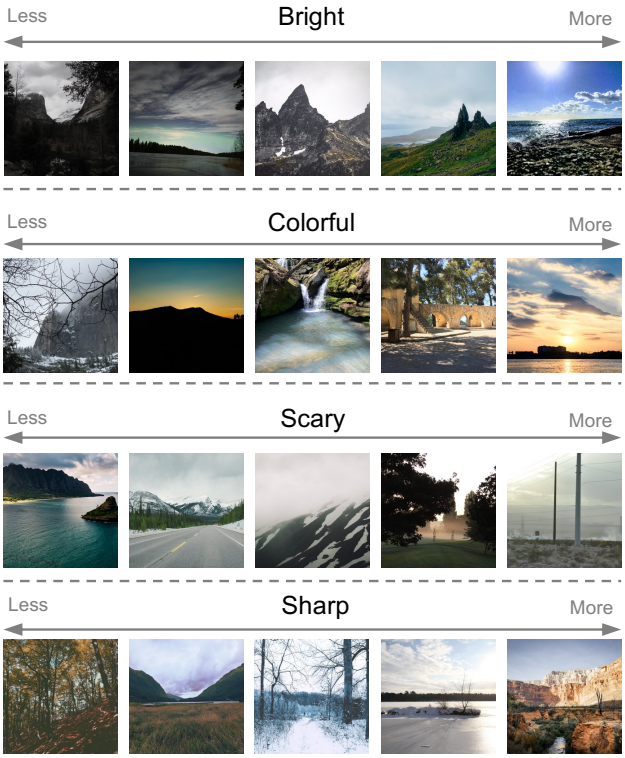


Figure 13. **Examples of reference images from LHQ [69].** We show randomly picked images for four attributes. Images within each category are classified into one of five groups.

C. Reference Images Preparation

In this section, we provide more details on the construction of reference image sets to complement Section 4.1 of the main paper. We use the following datasets as resources.

CelebA [44] is a benchmarked face attributes dataset and each image with 40 binary attribute annotations. We experiment with these binary attributes and their combinations.

FAIR Benchmark (FAIR) [21] is a recently proposed synthetic face dataset used for skin tone estimation. Specifically, we use images from the validation set containing 234 images and 702 facial crops. The validation set is released with ground-truth UV albedo maps. In order to obtain ground-truth skin tone types, we follow [21] to compute the Individual Typology Angle (ITA) score [10] of an albedo map to be the average of all pixel-wise ITA values with a pre-computed skin region area. For each image, ITA can be used to classify the skin tone type according to 6 categories, ranging from very light (*i.e.*, type 1) to dark (*i.e.*, type 6) [10, 17]. We randomly select 25 images per skin tone type as the reference images.

FairFace [35] contains face images with annotations for 2 perceived gender, 9 perceived age, and 7 race categories. As discussed in Section 4.1 of the main paper, although we value the contribution of the FairFace database to the community, we prefer using race labels and instead advocate for skin tone descriptions that recognize phenotypic diversity within broad racial categories [4]. Therefore, we only use their age annotations in our experiments.

Landscape (LHQ) [69] provides unlabeled natural scene images, allowing us to extend ITI-GEN to a different domain beyond human faces. With the annotation tool from [74], each image can be labeled with a score s ranging from 0 to 1, with a higher value indicating a closer match to the corresponding attribute. Using this score, we classify each image into one of the five degrees of the target attribute, resulting in a multi-category attribute. Figure 13 shows example reference images in the LHQ dataset. *Note that, the purpose of this experiment is not to justify LHQ as a perfect resource for learning tokens for perception attributes, but to investigate the capability of our* ITI-GEN *framework that can leverage the data from another domain as guidance.*

D. Evaluation Metrics.

Distribution Discrepancy ($\mathbb{D}_{\text{KL}}$). Following [12, 14], we use the CLIP model to predict the attributes in the images. For the attributes in which every category can be accurately specified by natural language, we input the original prompt combined with different names of categories into CLIP for obtaining the attribute label. For instance, if we want to evaluate the attribute "male" for the images generated from "a headshot of a person", we construct the input text of CLIP as ["a headshot of a man", "a headshot of a woman"]. For the attributes in which some of the categories can not be specified by natural languages, such as "eyeglasses" and "without eyeglasses" (due to the issue of negative prompt), we input the text ["a headshot of a person with eyeglasses", "a headshot of a person"]. For attributes that CLIP might be erroneous, we leverage pre-trained classifiers [35] combined with human evaluations. Specifically, for the skin tone, which is extremely difficult to obtain an accurate scale [1, 2, 29], we adopt the most commonly used Fitzpatrick skin type [10] combined with off-the-shelf models [21] for evaluation.

Fréchet Inception Distance (FID) [27]. We report the FID score to measure image quality. Specifically, we use the CleanFID library [53] to calculate the FID relates to statistics in FFHQ [36].

E. Additional Ablations and Analyses

E.1. Tokens Length

In our experiments, we set the length of inclusive tokens as 3 (q in Equation 3 of the main paper). Here, we provide further analyses on the size of q and show results in Figure 14. We see that fewer than 3 tokens may hurt the

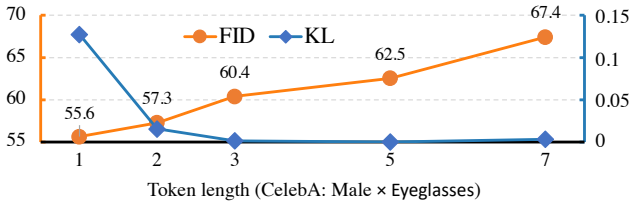


Figure 14. **Ablation study on tokens length.** Using fewer tokens is not sufficient enough to capture the concepts of attributes, leading to a relatively high distribution discrepancy (*i.e.*, KL). On the other hand, using more tokens may degrade the image quality due to language drifts (*i.e.*, relatively high FID scores).

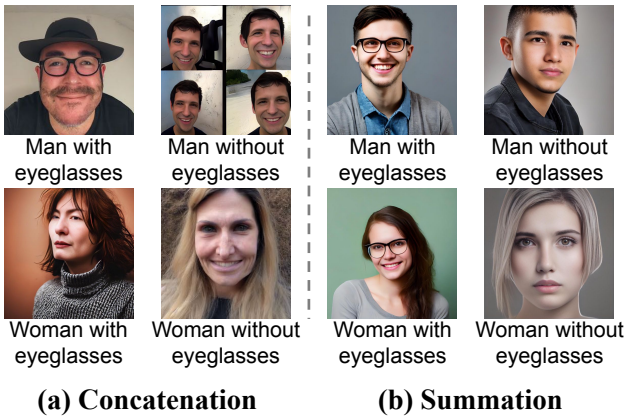


Figure 15. **Concatenation *vs.* summation on inclusive tokens aggregation.** We show an example of the combination of "Male" and "Eyeglasses" attributes. **(a)** Simply concatenating may reduce the image quality or fail to generate the images with corresponding attributes (*e.g.*, "Woman with eyeglasses") potentially because of the language drifts [41, 65]. **(b)** ITI-GEN provides better results with a conceptually simpler summation.

performance — cannot generate images with the desired attributes — potentially due to less representation capacity in capturing the concepts in the reference images. On the other hand, more tokens may result in adversarial effects or collapse. We hypothesize that prepending too many tokens after the original prompts leads to language drifts [41, 65]. This cannot be alleviated even with the semantic consistency loss (Equation 7 of the main paper) because simply forcing the two prompts with very different lengths to be close in the embedding space is ineffective.

E.2. Tokens Aggregation

As mentioned in Section 3.1 of the main paper, we use summation operation to aggregate the inclusive tokens of multiple attributes to achieve permutation invariance. Here, we provide another option — concatenation. Specifically, we ignore the positional encodings before feeding the inclusive tokens in the CLIP text encoder. Thus, the attention mechanism applied to prompt tokens is permutation invariant. Figure 15 shows comparison results. We notice that

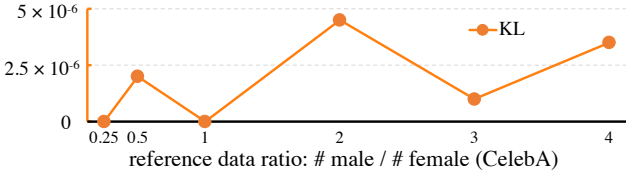


Figure 16. **Ablation study on the ratio of different categories in the reference set.** We study on the perceived gender attribute in CelebA by changing the ratio of images from the "male" and "female" categories. ITI-GEN is robust (*i.e.*, with very small distribution discrepancy, KL) to the ratio of different categories in the reference image set.

Figure 17. Results of ITI-GEN with (a) mutually exclusive and (b) overlapped reference images for attributes: *gender*×*eyeglasses*.

ITI-GEN (with token summation) not only achieves better results than concatenation but also offers a simpler and cleaner solution for token aggregation.

E.3. Imbalanced Reference Images

As mentioned in Section 4.1 of the main paper, we select 25 reference images per category in our experiments. We also mentioned that ITI-GEN is robust to imbalanced data distributions in Section 3.3. Here, we provide additional results as evidence. We change the ratio of "male" images *vs.* "female" images for the Perceived Gender attribute in CelebA and show the results in Figure 16. ITI-GEN can always generate images with nearly a balanced distribution.

E.4. Overlapped Reference Images

As mentioned in Section 3.2 of the main paper, we need a reference dataset for each attribute. However, this does not pose a practical issue (which seems like an all-too-exhaustive list to cover), because each reference dataset does not have to be mutually exclusive. An existing dataset (*e.g.*, CelebA or smaller) can be divided into overlapped sub-datasets, either manually or using a classifier. To demonstrate this, we compare two settings: (a) *Exclusive* — two datasets, each containing 50 images with equal gender and eyeglasses distribution, respectively; (b) *Overlapped* — a single dataset of 50 images with equal numbers between man and woman labels, as well as with and without eyeglasses. The results in Figure 17 show that using a smaller, *overlapped* dataset does not affect the performance.

E.5. Corrupted Reference Images

In this subsection, we further study whether the quality of the provided reference image strongly affects the general-

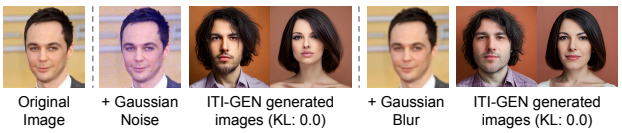


Figure 18. ITI-GEN with corrupted reference data. The attribute of interests is *gender*.

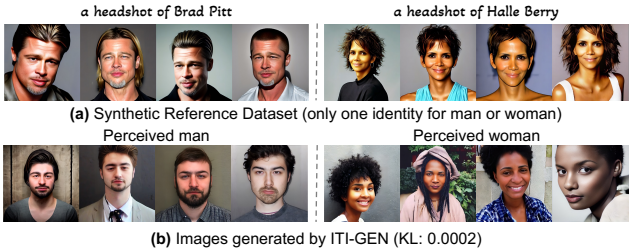


Figure 19. ITI-GEN can leverage less diverse reference images in (a) for inclusive generation for the *gender* attribute in (b).

Figure 20. **Qualitative results of ITI-GEN when** $K_m = 1$. When only "female" images are provided as the reference images (left in (a)), ITI-GEN can leverage the synthetic data generated by the original prompt ("*a headshot of a person*", right in (a)), together with the real data, to construct the reference image set. By jointly using these two sources, ITI-GEN learns inclusive tokens representing the concept of "female", which can be used to synthesize images for the desired category, as shown in (b). Section E.6 illustrates details.

ization and the application of the ITI-GEN. We provide the results with noisy or blurred reference images in Figure 18. We also experiment with less diverse reference images (only using the images with one identity) and show results in Figure 19. Both demonstrate the robustness of ITI-GEN to the quality of reference data.

E.6. Single Category Attribute ($K_m = 1$)

In the main paper, we mainly studied the attributes that have more than one category (K_m is larger than 1 in Equation 3 of the main text). What if we only have the reference images from one category of the target attribute ($K_m = 1$)? In light of our pairwise direction loss (Equation 4), there are at least two different categories needed in the reference images. Here, we show that ITI-GEN can leverage the synthetic data generated by the original prompt (*e.g.*, "*a headshot of a person*") as an additional category to compute the directional loss for the case of $K_m = 1$.

We verify this idea by using only images of the "female" category from the perceived gender attribute. From Figure 20, we can observe that by leveraging the female real images and another set of synthetic images generated from "*a headshot of a person*", ITI-GEN is able to synthesize female images. Further quantitative evaluation for the images generated by ITI-GEN indicates that 100% perceived woman is obtained.

Table 3. **FID (↓) comparison**. Reference images for ITI-GEN are from FAIR benchmark [21]. ITI-GEN produces lower FID than all the other baselines. **SD**: vanilla stable diffusion. **EI**: ethical intervention. **HPS**: hard prompt searching. **PD**: prompt debiasing. **CD**: custom diffusion.

SD [64]	EI [5]	HPS [19]	CD [40]	PD [14]	**ITI-GEN**
67.4	81.4	69.9	62.4	63.3	**51.8**

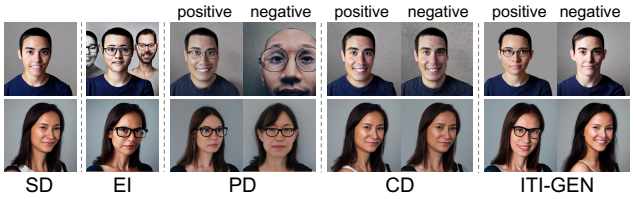


Figure 21. **Visualization of different methods.** The prompt is *"a headshot of a person"*. Attributes are *gender* × *eyeglasses*. Images across each line are sampled using the same random seed.

F. Additional Results

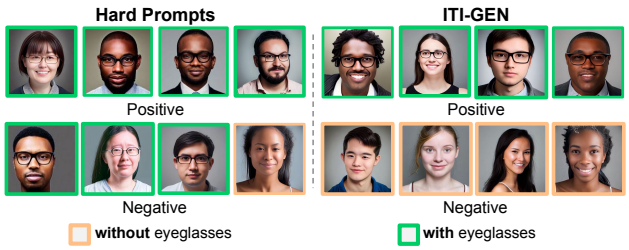
Due to space limitations, we only reported the results of several attributes mainly to cover the attributes relating to social factors and facial expressions in the main paper. In this section, we provide additional results and detailed comparisons to strong baseline methods.

F.1. Qualitative and FID Results for Baselines

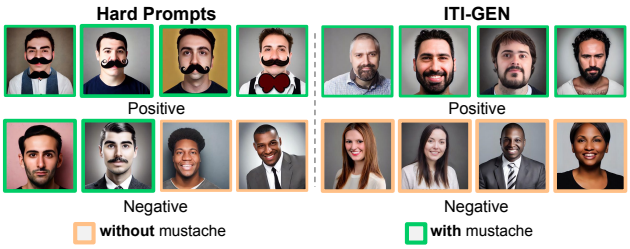
We only provide KL Divergence metric ($\mathbb{D}_{\text{KL}}$) in the main paper for different baselines. Here, we incorporate the comparisons of FID in Table 3 and visualizations in Figure 21 with other baselines.

F.2. Single Binary Results

We summarize the full results of single binary attributes with CelebA [44] in Table 4. We compare with the base-



(a) Target Attribute: eyeglasses



(b) Target Attribute: mustache

Figure 22. **Challenges of (a) linguistic ambiguity and (b) model misrepresentation.** While Hard Prompts demonstrated strong capabilities in generating images with desired attributes, they cannot handle some situations. **(a)** Vanilla text-to-image models can hardly understand *negative* prompts (*e.g.*, "not", "without") possibly due to *linguistic ambiguity*. **(b)** For some attributes (*e.g.*, mustache), directly using hand prompts results in misrepresented results caused by the model bias.

line Stable Diffusion model [64] and Hard Prompt Searching [19], which demonstrated strong performance in many attributes (cf. Table 1 of the main paper). From Table 4, we observe ITI-GEN achieves the best performance in nearly all 40 attributes except some subtle facial attributes (*e.g.*, "Wearing Necklace"). We use the prompt "*a headshot of a person*" in Table 4 and show qualitative results of other prompts (*e.g.*, other occupations such as politician and musician) in Figure 33. Furthermore, we list all the hard prompts used in our experiments in Table 5.

As mentioned in Section 1 of the main paper, we reiterate that ITI-GEN is designed to handle several cases (attributes) that Hard Prompts may struggle with. First, attributes with fine-grained categories may be difficult to express in language. Second, linguistic ambiguity, as shown in Figure 22 (a). Third, model misrepresentation, as illustrated in Figure 22 (b). More importantly, we argue that ITI-GEN is *not* to replace Hard Prompts (especially for attributes that are already can be handled by language) but to support complex prompts with multiple attributes, as illustrated in Figure 23.

F.3. Multiple Attributes

We now consider multi-attribute cases and show additional results in Figure 24. To fully characterize the performance of ITI-GEN, we study three additional settings

Table 4. **A full comparison with baseline methods with the 40 single attribute setting** ($\mathbb{D}_{\text{KL}}$ $\downarrow$). Reference images are from CelebA [44]. Following [12, 14], we use CLIP [57] as the attribute classifier. **SD**: vanilla stable diffusion [64]. **HPS**: hard prompt searching [19]. Given the strong capability of the existing text-to-image generative models, one can express the (most but not all) desired attributes directly using *Hard Prompts*. However, it faces challenges in certain attributes and ITI-GEN addresses most of these drawbacks. Please see Figure 22 for a side-by-side qualitative comparison between HPS and ITI-GEN. Please see Figure 23 for how ITI-GEN can be compatibly used with Hard Prompts.

Attribute	SD [64]	HPS [19]	ITI-GEN
5'o Clock Shadow	0.02957	**0.00847**	0.06882
Arched Eyebrows	0.32972	0.04570	**0.00892**
Attractive	0.11264	0.07405	**0.00000**
Bags Under Eyes	0.33325	0.10498	**0.01395**
Bald	0.51578	0.22175	**0.00892**
Bangs	0.33886	0.19975	**0.00000**
Big Lips	0.20984	0.02908	**0.00892**
Big Nose	0.32423	0.01629	**0.00056**
Black Hair	0.35189	0.12539	**0.00000**
Blond Hair	0.60804	0.00501	**0.00222**
Blurry	**0.01077**	0.25348	0.09707
Brown Hair	0.41683	0.14207	**0.05663**
Bushy Eyebrows	0.07108	0.29737	**0.02747**
Chubby	0.14293	0.40233	**0.00000**
Double Chin	0.28637	0.48016	**0.19274**
Eyeglasses	0.38773	0.32622	**0.00056**
Goatee	0.25933	0.04266	**0.00000**
Gray Hair	0.65905	0.27921	**0.17049**
Heavy Makeup	0.39293	0.10989	**0.04570**
High Cheekbones	0.47875	**0.00020**	0.03599
Male	0.01033	0.00005	**0.00000**
Mouth Slightly Open	0.07030	0.14207	**0.04570**
Mustache	0.02013	0.12009	**0.00000**
Narrow Eyes	0.14968	**0.00847**	0.08228
No Beard	0.22442	0.49463	**0.00222**
Oval Face	0.39526	0.03158	**0.02014**
Pale Skin	0.17394	0.00045	**0.00000**
Pointy Nose	0.48951	0.02221	**0.00000**
Receding Hairline	0.31784	0.61526	**0.02014**
Rosy Cheeks	0.46275	**0.03691**	0.14987
Sideburns	0.55409	0.04570	**0.02013**
Smiling	0.25059	0.02075	**0.00000**
Straight Hair	**0.08506**	0.61526	0.19274
Wavy Hair	0.47663	0.36806	**0.03599**
Wearing Earrings	0.32029	0.15998	**0.09707**
Wearing Hat	0.65144	0.12539	**0.01395**
Wearing Lipstick	0.50658	0.12539	**0.11323**
Wearing Necklace	0.63600	**0.05897**	0.49463
Wearing Necktie	0.46687	0.69315	**0.06882**
Young	0.65647	0.00056	**0.00000**

based on the attribute correlation matrix from the CelebA dataset [44] (see Figure 2 in [73] for the correlation matrix). Specifically, we select three attribute combinations with dif-

Table 5. **Hard Prompts used in our experiments.** Different attributes may not follow the same template and we carefully specify or express the attribute in the input prompt. The human-written hard prompts are used to generate images. Results are shown in Table 4.

Attribute	Positive Prompt	Negative Prompt
5'o Clock Shadow	a headshot of a person {with 5'o clock shadow}.	a headshot of a person {with no/without 5'o clock shadow}.
Arched Eyebrows	a headshot of a person {with arched eyebrows}.	a headshot of a person {with no/without arched eyebrows}.
Attractive	a headshot of {an attractive} person.	a headshot of {an unattractive} person.
Bags Under Eyes	a headshot of a person {with bags under eyes}.	a headshot of a person {with no/without bags under eyes}.
Bald	a headshot of a person {in bald}.	a headshot of a person {in no bald}.
Bangs	a headshot of a person {with bangs}.	a headshot of a person {with no/without bangs}.
Big Lips	a headshot of a person {with big lips}.	a headshot of a person {with small lips}.
Big Nose	a headshot of a person {with big nose}.	a headshot of a person {with small nose}.
Black Hair	a headshot of a person {with black hair}.	a headshot of a person {with no/without black hair}.
Blond Hair	a headshot of a person {with blond hair}.	a headshot of a person {with no/without blond hair}.
Blurry	a headshot of a person {in blurry}.	a headshot of a person {in no/without blurry}.
Brown Hair	a headshot of a person {with brown hair}.	a headshot of a person {with no/without brown hair}.
Bushy Eyebrows	a headshot of a person {with bushy eyebrows}.	a headshot of a person {with no/without bushy eyebrows}.
Chubby	a headshot of a {chubby} person.	a headshot of a {no chubby} person.
Double Chin	a headshot of a person {with double chin}.	a headshot of a person {with no/without double chin}.
Eyeglasses	a headshot of a person {with eyeglasses}.	a headshot of a person {with no/without eyeglasses}.
Goatee	a headshot of a person {with goatee}.	a headshot of a person {with no/without goatee}.
Gray Hair	a headshot of a person {with gray hair}.	a headshot of a person {with no/without gray hair}.
Heavy Makeup	a headshot of a person {with heavy makeup}.	a headshot of a person {with no/without heavy makeup}.
High Cheekbones	a headshot of a person {with high cheekbones}.	a headshot of a person {with low cheekbones}.
Male	a headshot of a {man}.	a headshot of a {woman}.
Mouth Slightly Open	a headshot of a person {with mouth slightly open}.	a headshot of a person {with mouth closed}.
Mustache	a headshot of a person {with mustache}.	a headshot of a person {with no/without mustache}.
Narrow Eyes	a headshot of a person {with narrow eyes}.	a headshot of a person {with no/without narrow eyes}.
No Beard	a headshot of a person {with no/without beard}.	a headshot of a person {with beard}.
Oval Face	a headshot of a person {with oval face}.	a headshot of a person {with no/without oval face}.
Pale Skin	a headshot of a person {with pale skin}.	a headshot of a person {with dark skin}.
Pointy Nose	a headshot of a person {with pointy nose}.	a headshot of a person {with no/without pointy nose}.
Receding Hairline	a headshot of a person {with receding hairline}.	a headshot of a person {with no/without receding hairline}.
Rosy Cheeks	a headshot of a person {with rosy cheeks}.	a headshot of a person {with no/without rosy cheeks}.
Sideburns	a headshot of a person {with sideburns}.	a headshot of a person {with no/without sideburns}.
Smiling	a headshot of a person {with smiling}.	a headshot of a person {with no/without smiling}.
Straight Hair	a headshot of a person {with straight hair}.	a headshot of a person {with no/without straight hair}.
Wavy Hair	a headshot of a person {with wavy hair}.	a headshot of a person {with no/without wavy hair}.
Wearing Earrings	a headshot of a person {wearing earrings}.	a headshot of a person {without wearing earrings}.
Wearing Hat	a headshot of a person {wearing hat}.	a headshot of a person {without wearing hat}.
Wearing Lipstick	a headshot of a person {wearing lipstick}.	a headshot of a person {without wearing lipstick}.
Wearing Necklace	a headshot of a person {wearing necklace}.	a headshot of a person {without wearing necklace}.
Wearing Necktie	a headshot of a person {wearing necktie}.	a headshot of a person {without wearing necktie}.
Young	a headshot of a {young} person.	a headshot of {an old} person.

ferent levels of attribute entanglement (*i.e.*, co-occurrence frequency) — a higher co-occurrence value means the attribute combination is more common in daily life while a lower co-occurrence value indicates a rare case in the original CelebA dataset. Admittedly, there are several cases ITI-GEN does not always generate images with a balanced distribution or faithfully generates images with specific attributes. Please see Figure 24 for details.

F.4. Multi-Category Attributes

In Figure 6 and Figure 7 of the main paper, we investigated the combinations of multi-category attributes. Here, we further study another challenging setup: Perceived Gender (CelebA) $\times$ Skin Tone (FAIR) $\times$ Age (FairFace) (108 different combinations of categories in total). Qualitative results are shown in Figure 25 and in Figure 26. As expected, ITI-GEN is capable of handling multiple fine-grained attribute categories to achieve inclusiveness.

F.5. Other Domains

As shown in Figure 8 of the main paper, ITI-GEN can generalize to a different domain for perception attributes on scene images. In this subsection, we demonstrate more results of other attributes in Figure 27 for "colorfulness", Figure 28 for "sharpness", Figure 29 for "scary", Figure 30 for "contrast", Figure 31 for "brightness", and Figure 32 for "'brightness". As we observe, ITI-GEN generates more diverse results than the baseline model even with very complex input prompts.

F.6. Train-once-for-all Generalization

We provide additional qualitative results with different occupation prompts in Figure 33, Figure 34, Figure 35, Figure 36, and Figure 37.

F.7. Compatibility with ControlNet

We provide additional examples of compatibility with ControlNet in Figure 38.

Figure 23. **Compatibility of ITI-GEN to hard prompts.** As mentioned in Section F.2 and Figure 22, Hard Prompts show accurate results with some attributes (*e.g.*, "young" and "perceived man" in the top row) but may result in misrepresented results for other attributes (*e.g.*, "mustache" in the middle row). ITI-GEN demonstrates strong compatibility with Hard Prompts to benefit a broad spectrum of attributes (bottom row).

G. Future Work

To establish the new direction and demonstrate its feasibility so that future works can easily build upon, we intentionally avoid sophisticated techniques to improve ITI-GEN in favor of simplicity and believe that additional modifications can further enhance the inclusive generative models.

Lifelong ITI-GEN. In this study, we assume all the attributes are accessible at the same time. In practice, we hope to show that ITI-GEN is capable of the continue learning setup. That is, adding new attributes while without forgetting or re-training the previous inclusive tokens.

Other Attributes. There are other attributes ITI-GEN might be able to control via appropriately prepared reference images. For example, the 3D geometry attributes such as head poses and materials such as normal and lighting.

(a) Male × Eyeglasses

(b) Male × Young

(c) Male × Heavy Makeup

Figure 24. **Additional results on multiple attributes.** We consider three settings based on the attribute co-occurrence matrix in the CelebA dataset (see Section F.3). The attribute combinations in (a) and (b) are relatively less entangled between the sub-categories whereas in (c) — a *failure* case of ITI-GEN— the category "with heavy makeup" is heavily entangled with the category "female" in CelebA, which indicates that other category combinations (*e.g.*, "man with heavy makeup") can rarely happen in our daily life. Therefore, the text-to-image model can hardly synthesize images with this underrepresented attribute combination.

Figure 25. **Results of ITI-GEN on multi-category attributes** for Perceived Gender (2) × Skin Tone (6) × Age (9). Examples are randomly picked with "*a headshot of a person*" for **Perceived Man** × Skin Tone (6) × Age (9). Please see Figure 26 for more results on Perceived Woman × Skin Tone (6) × Age (9).

Figure 26. **Results of ITI-GEN on multi-category attributes** for Perceived Gender (2) × Skin Tone (6) × Age (9). Examples are randomly picked with "*a headshot of a person*" for **Perceived Woman** × Skin Tone (6) × Age (9). Please see Figure 25 for more results on Perceived Man × Skin Tone (6) × Age (9).

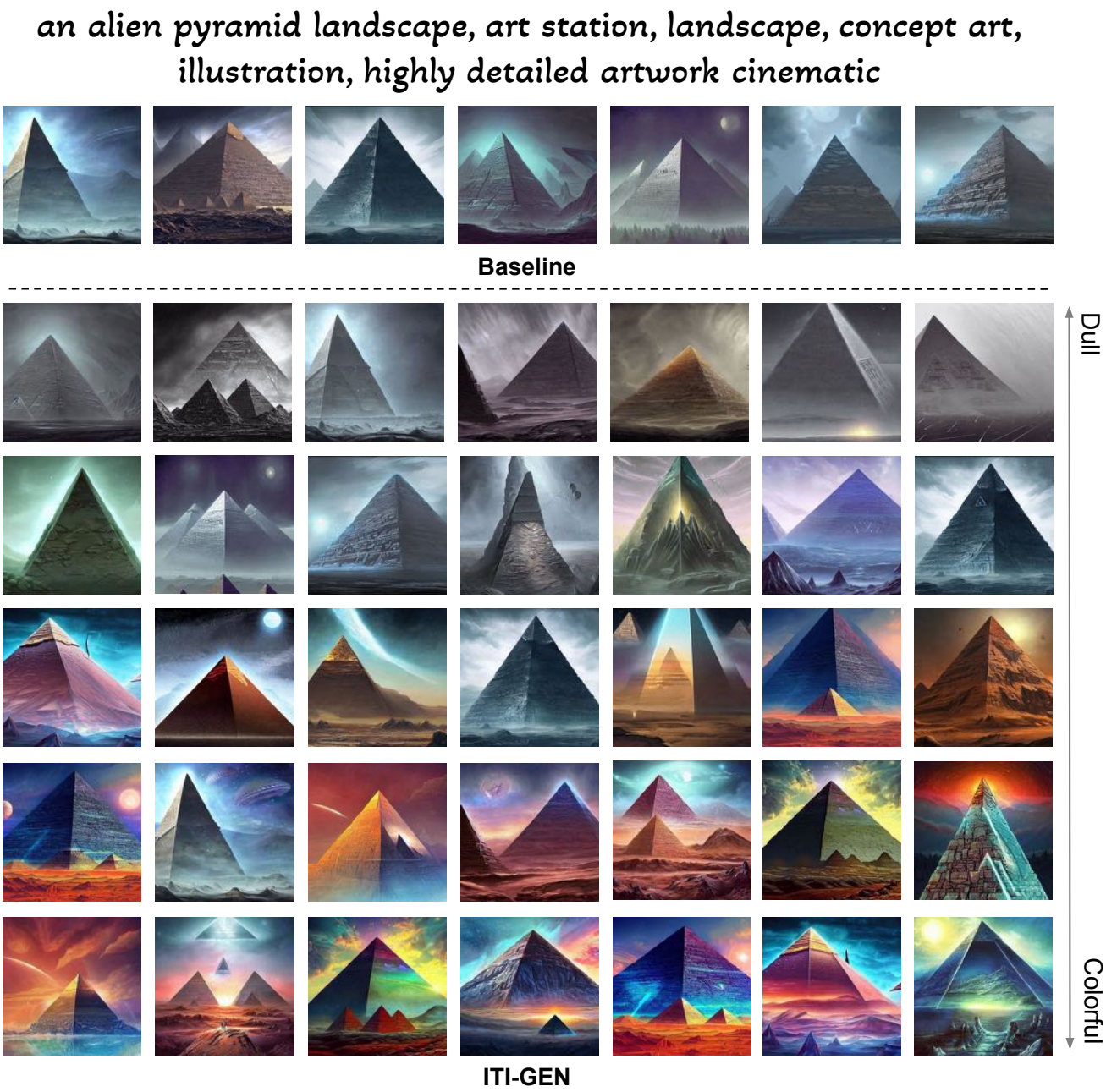


Figure 27. **ITI-GEN with perception attributes ("Colorfulness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of colorfulness. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

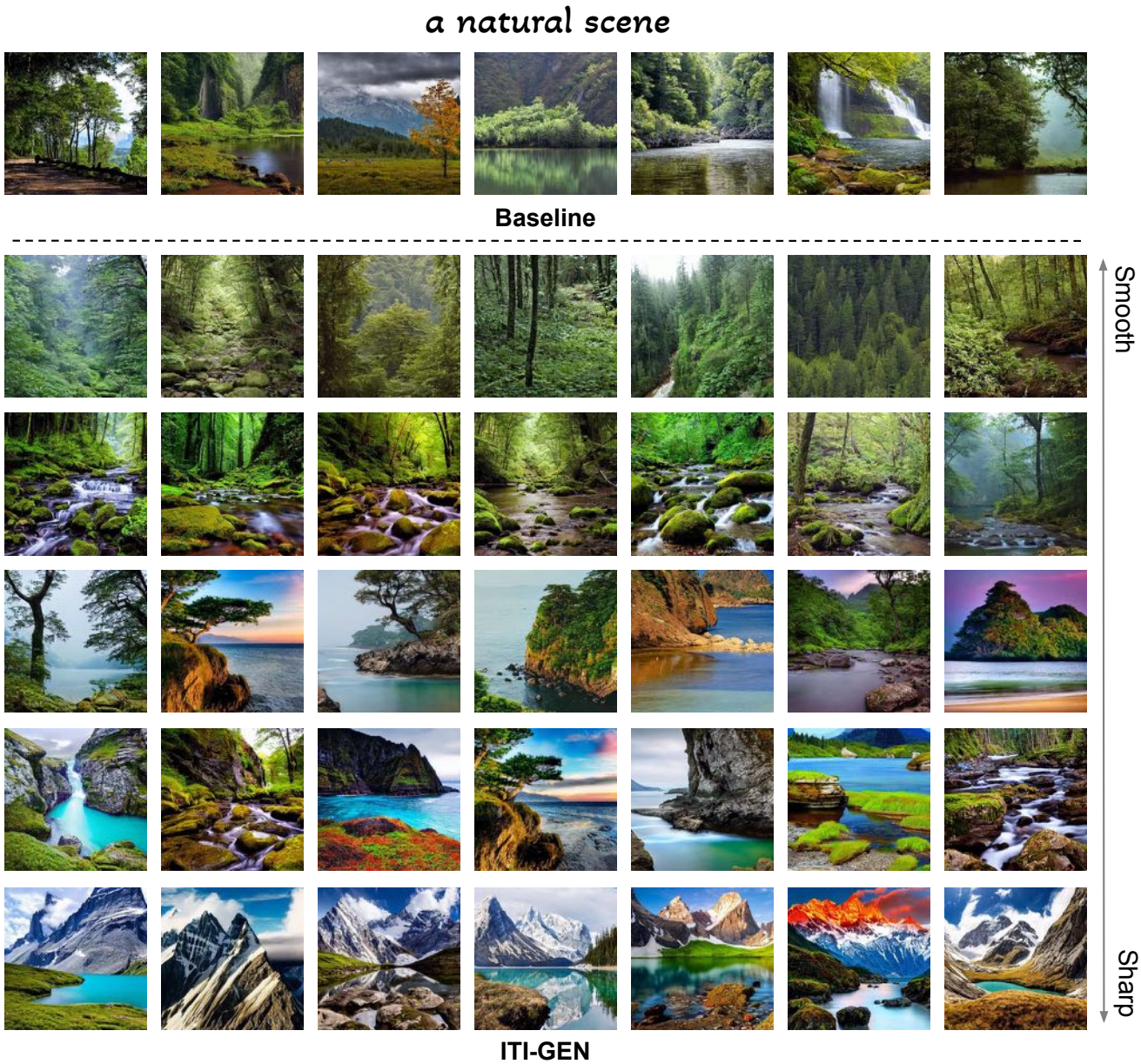


Figure 28. **ITI-GEN with perception attributes ("Sharpness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of sharpness. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

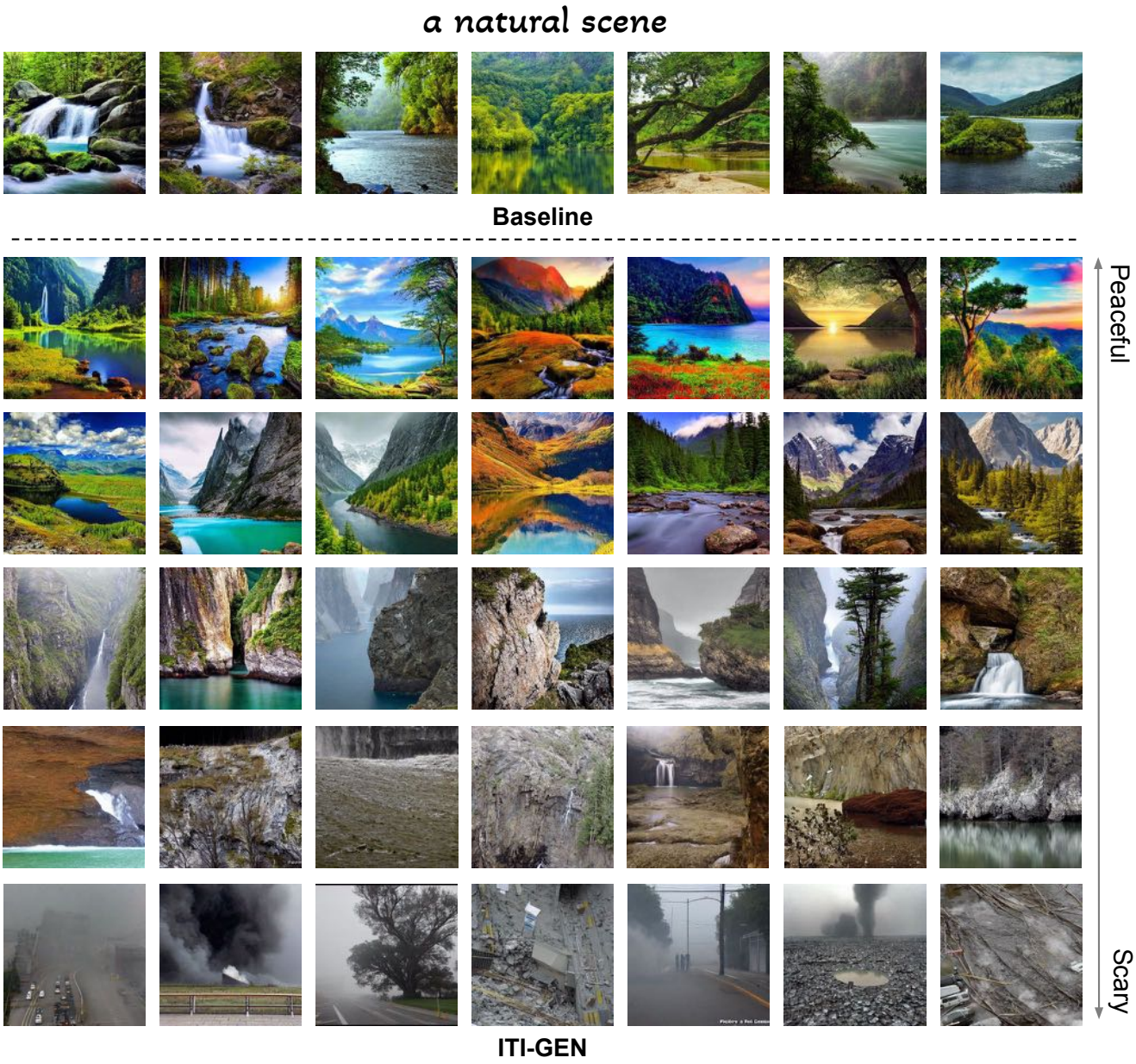


Figure 29. **ITI-GEN with perception attributes ("Scary") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of scary. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

Figure 30. **ITI-GEN with perception attributes ("Contrast") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of contrast. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

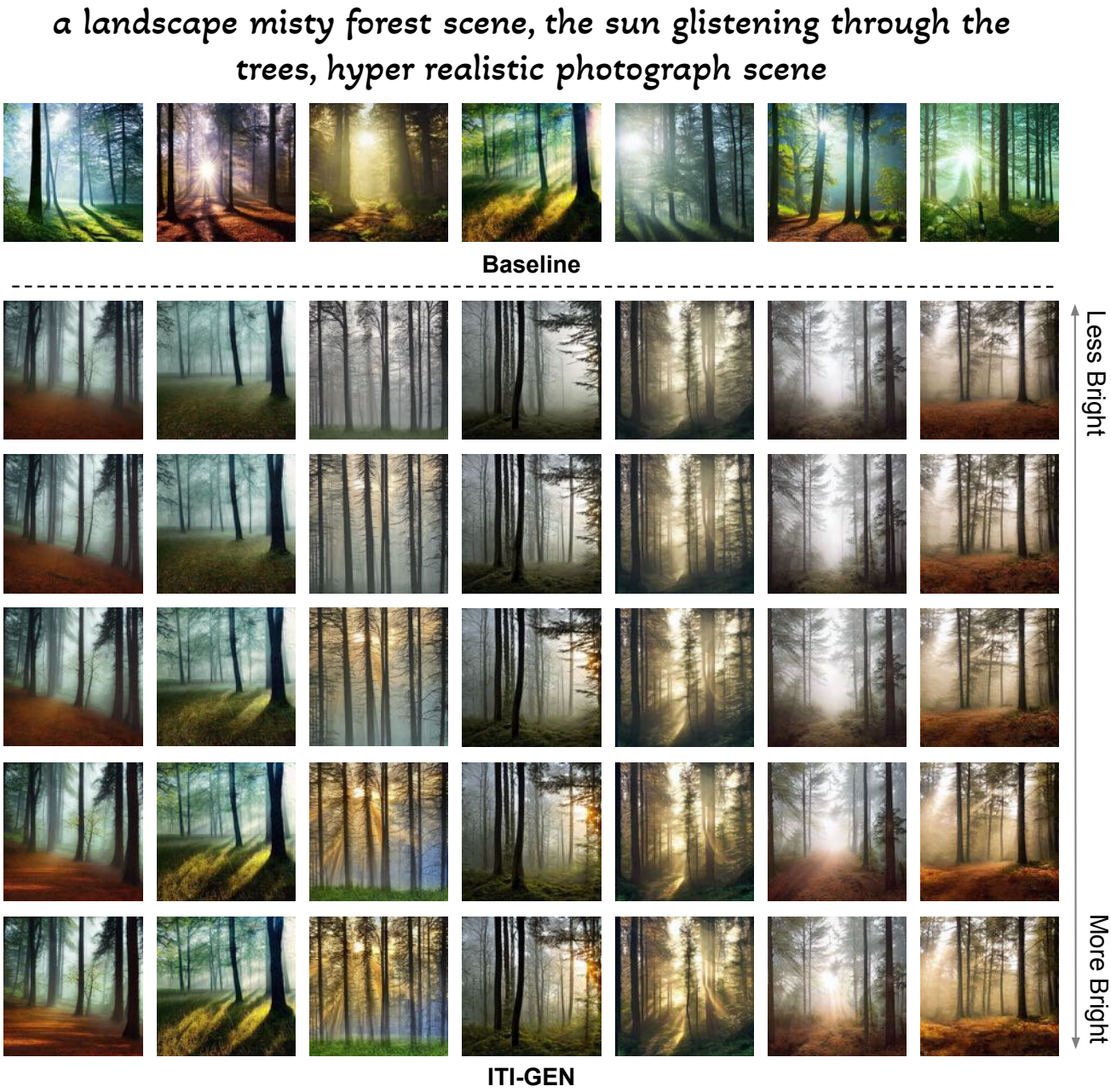


Figure 31. **ITI-GEN with perception attributes ("Brightness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of brightness. In this example, we intentionally pick images using the same random seed in each column for ITI-GEN. Please compare the first and last examples in each column for a clear change in brightness. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

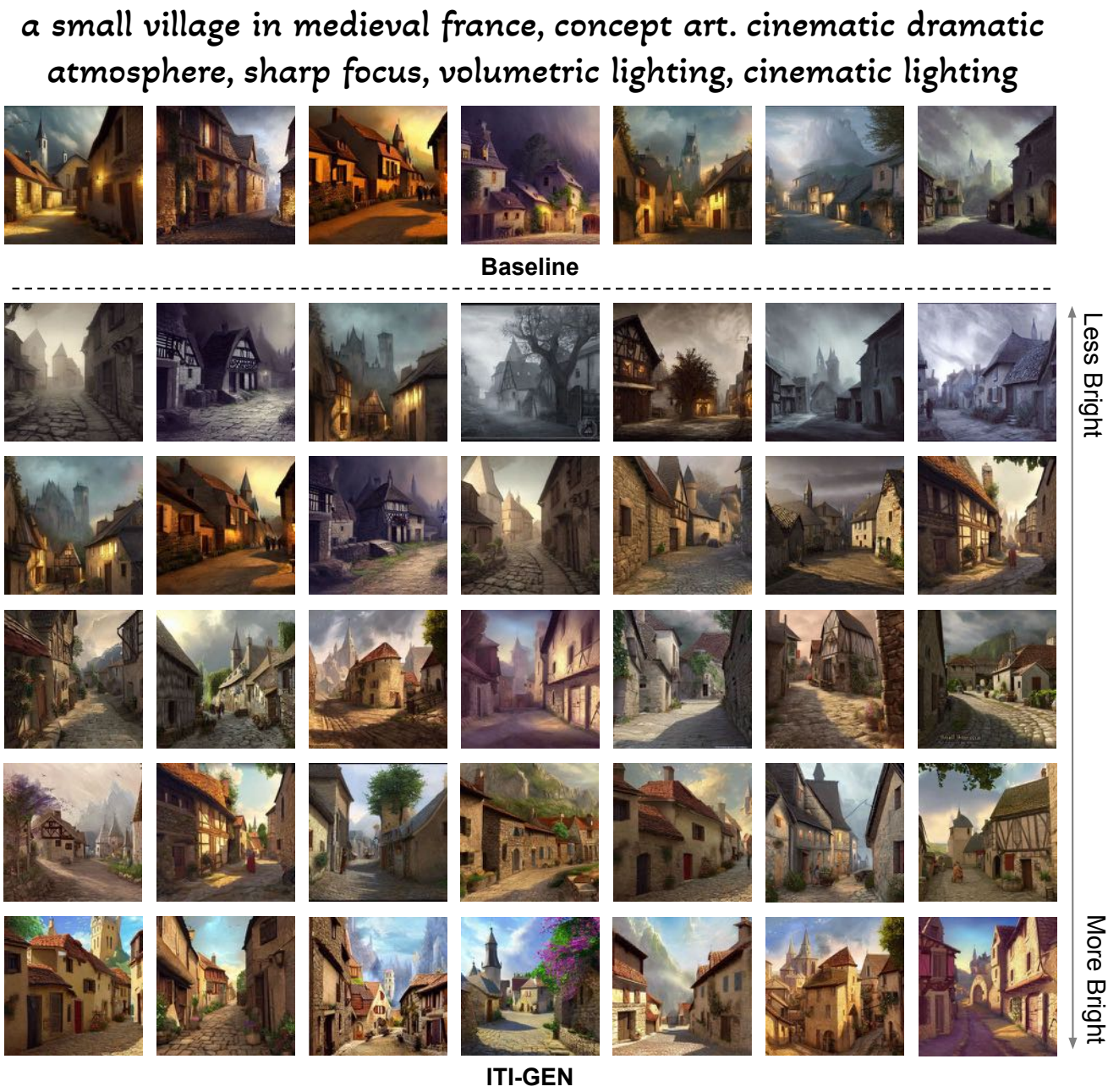


Figure 32. **ITI-GEN with perception attributes ("Brightness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of brightness. See Section C for details and Figure 13 for reference image examples from LHQ [69]. Better viewed in color.

Figure 33. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

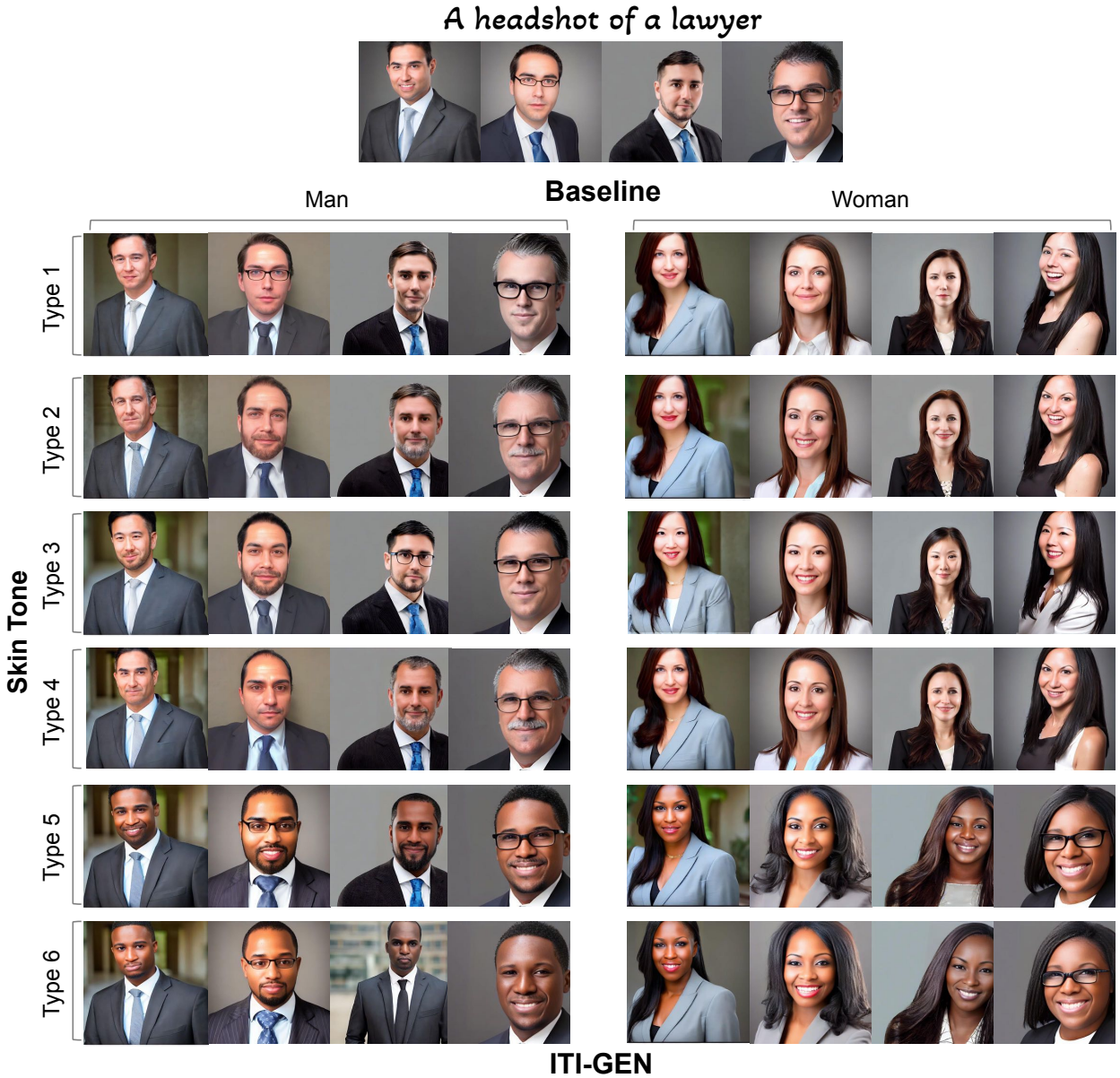


Figure 34. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

Figure 35. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

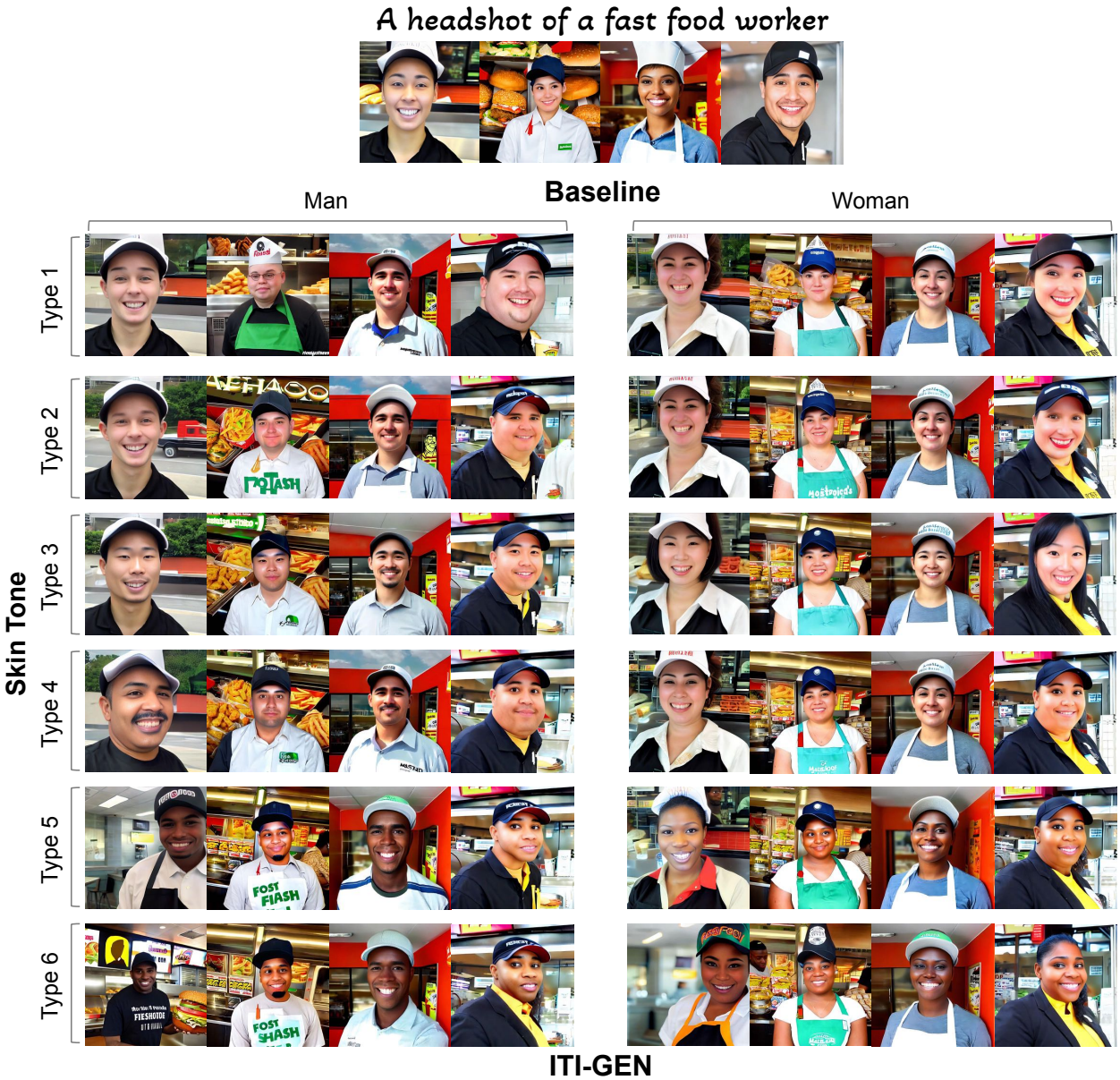
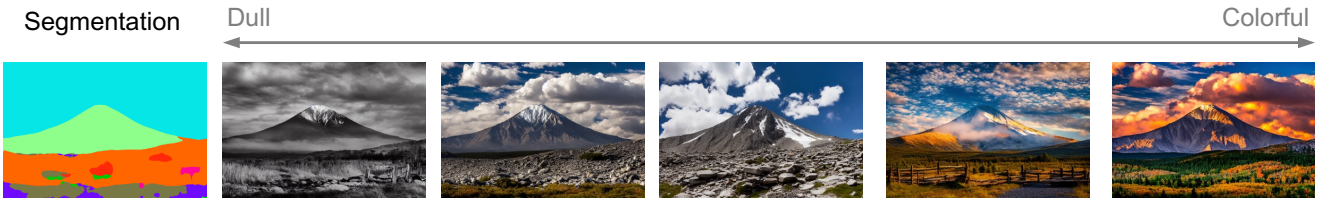
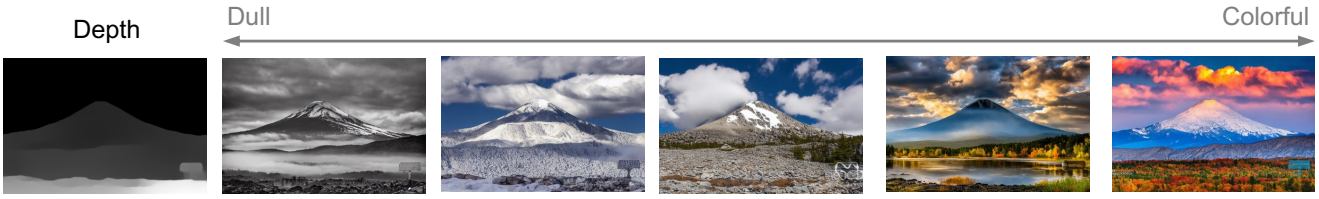


Figure 36. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

Figure 37. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

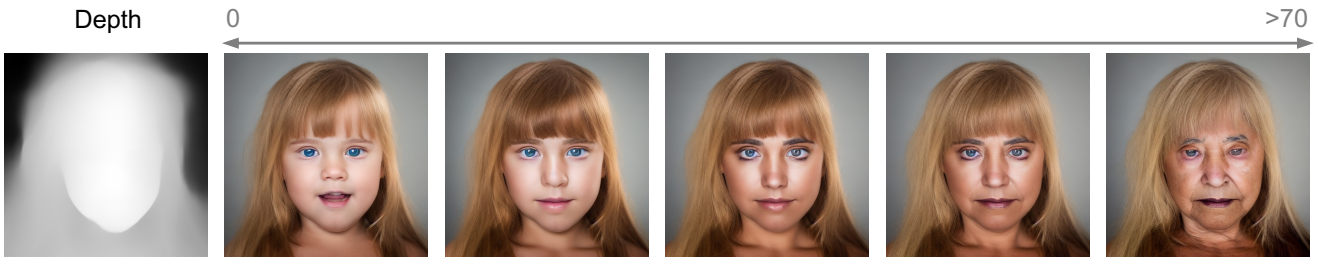


(a) Condition: segmentation map. Attribute: colorfulness.



(b) Condition: depth map. Attribute: colorfulness.

(c) Condition: canny edge map. Attribute: brightness.



(d) Condition: depth map. Attribute: age.

Figure 38. **Additional results on the compatibility with ControlNet [83].** All examples are based on *train-once-for-all* generation (Section 3.3 of the main paper). For scene images in (a), (b), and (c), the inclusive tokens are trained with "*a natural scene*" using LHQ images [69]. For human faces in (d), the tokens for age attribute are trained with "*a headshot of a person*" using FairFace images [35]. As discussed in Section B, our method is designed for improving inclusiveness but not for image editing.