ITI-GEN: Inclusive Text-to-Image Generation

Cheng Zhang ¹ Xuanbai Chen ¹ Siqi Chai ¹ Chen Henry Wu ¹ 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-GEN¹, that leverages readily available reference images for Inclusive Text-to-Image GENERation. 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 [69, 27, 59, 58, 63] and the access to large-scale multimodal datasets [67, 35]. Particularly, publicly released models, such as Stable Diffusion [63], 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, 58, 63, 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”, existing text-to-image models [63] can hardly synthesize pictures representing minority groups (i.e., people with eyeglasses in this example). (b) Conventional hard prompt searching [18] 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).

Figure 1. (a) Given a human-written prompt (“a headshot of a person”), existing text-to-image models [63] can hardly synthesize pictures representing minority groups (i.e., people with eyeglasses in this example). (b) Conventional hard prompt searching [18] 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).

¹Project 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(e).

To advance inclusive generation, a straightforward way is to retrain or fine-tune the model upon request, using truly inclusive training data \([17, 82]\). 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 \([67, 65, 17]\). Another principled approach towards inclusiveness is to specify or enumerate each category in natural language (i.e., hard prompt searching) \([18, 55]\). 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 \([29]\).

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 \([56]\), we raise the question: can we learn inclusive prompt embeddings using images as guidance?

To achieve this goal, we introduce ITI-G, a novel and practical framework that creates discriminative prompts based on readily available reference images for Inclusive Text-to-Image GENERation. Concretely, we leverage the vision-language pre-trained CLIP model \([56]\) 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 \([63]\). ITI-G can leverage reference images from different domains, including human faces \([43, 34, 20]\) and scenes \([68]\), to achieve inclusive generation in single or multiple attributes of interest. ITI-G needs neither prompt specification nor model fine-tuning, bypassing the problems of linguistic ambiguity as well as computational complexity. Moreover, ITI-G is compatible with the existing text-based image generation models (e.g., ControlNet \([81]\) 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. Related Work

Text-to-Image Generative Models. Text-based image generation has been widely studied with numerous model architectures and learning paradigms \([48, 62, 71, 59, 23, 79, 18, 19, 9, 69, 78, 16, 17, 38]\). Recently, the overwhelming success of diffusion-based text-to-image models \([58, 66, 58, 51]\) has attracted significant attention. A key factor to this success is their ability to deal with large-scale multimodal datasets \([67, 35, 11]\). 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 \([73, 74, 75, 42]\), research on developing fair generative models is limited \([83, 30, 22, 14, 46]\). Most efforts focus on GAN-based models \([13, 57, 31, 60, 80, 36, 77, 70, 33, 47]\), 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. \([18]\) 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 \([29, 5]\). 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. \([76]\) employed off-the-shelf models, such as CLIP \([56]\) 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) \([41, 32]\). Typically, PT methods insert small learnable modules (e.g., tokens) into the pretrained 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 \([24, 64, 39]\). By providing several reference images with the customized subject, they use a special token to represent the object by optimizing the token embedding \([24, 39]\) or the diffusion models \([64, 39]\). 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

\(^2\)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.

\(^3\)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-G) with the example of two binary attributes: perceived gender and skin tone. (a) Given an input prompt, (b) ITI-G learns discriminative token embeddings to represent each category of every target attribute. (c) By injecting the learned tokens after the original input prompt, ITI-G 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.

3. Inclusive Text-to-Image Generation

To drive the progress of Inclusive Text-to-Image Generation, we propose ITI-G, 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-G 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-G in Section 3.1, then describe the details of the learning strategy in Section 3.2, and finally discuss the key properties of ITI-G 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 $T \in \mathbb{R}^{p \times e}$, 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 $A$. Formally,

$$A = \{A_m | 1 \leq m \leq M\}; A_m = \{a^m_k | 1 \leq k \leq K_m\}$$  \hspace{1cm} (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 $A_m$ and $K_m$ denotes the number of categories in $A_m$. Note that $K_m$ may vary among different attributes.

Inclusive Prompt Set. Inspired by [41, 32], we propose prompt tuning for inclusive generation. Specifically, for a given category $a^m_k$ within attribute $A_m$, we inject $q$ learnable tokens $S^m_k \in \mathbb{R}^{p \times e}$ after the original $T$ to construct a new prompt $P^m_k = [T; S^m_k] \in \mathbb{R}^{(p+q) \times e}$. By querying the model $G$ with $P^m_k$, we can generate images exhibiting the characteristics of the corresponding category $a^m_k$. To differentiate the new tokens $S^m_k$ from the original prompt $T$, we refer to them as inclusive tokens.

When jointly considering $M$ attributes, we aggregate $M$ separate inclusive tokens $S^1_{o_1}, S^2_{o_2}, \ldots, S^M_{o_M}$ 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 $S_{o_1o_2\ldots o_M}$ that can be injected after $T$ 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\footnote{That is, the output of $f$ should be the same even if we permute the indices $m$ of the attributes in $A$ (cf. Equation 1).} with respect to attributes. Common options include element-wise average, sum, and max operations. Following [49], we adopt element-wise sum to preserve the text semantics without losing information\footnote{Please see Appendix E.2 for more analysis and other options for aggregating multiple tokens, e.g., concatenation.}. Finally, we define the inclusive prompt set as follows:

$$S_{o_1o_2\ldots o_M} = f(S^1_{o_1}, S^2_{o_2}, \ldots, S^M_{o_M})$$  \hspace{1cm} (2)

$$\mathcal{P}_{\text{total}} = \{P_{o_1o_2\ldots o_M} = [T; \sum_{m=1}^{M} S^m_{o_m}] \in \mathbb{R}^{(p+q) \times e} | 1 \leq o_1 \leq K_1, \ldots, 1 \leq o_M \leq K_M\}$$ \hspace{1cm} (3)
By uniformly sampling the prompts from $P_{\text{total}}$ 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, 18], we optimize prompts entirely in the continuous embedding space. Additionally, we only update the attribute-specific embeddings — the colors $\bullet$ and $\ast$ 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 $D_{\text{ref}} = \{(x_n^m, y_n^m)\}_{n=1}^{N_m}$ for a target attribute $A_m$, where $N_m$ is the dataset size and $y_n^m \in A_m$ (defined in Equation 1) indicates the category to which $x_n$ 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 [56], which aligns visual concepts with text embeddings by jointly training a text encoder $E_{\text{text}}$ and an image encoder $E_{\text{img}}$. The output of the pre-trained CLIP text encoder has also been used as the condition for text-guided image generation [63, 58], 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 [56]. 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.

![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 [20]. Details are in Section 3.2.](image-url)
\( \mathcal{P}_k^m = \{ P \in \mathcal{P}_{\text{total}} \mid \alpha_m = k \} \) is a collection of prompts containing all the category combinations for other attributes given the category \( \alpha_m^0 \) for attribute \( \mathcal{A}_m \) (cf. Equation 3). Finally, we denote the prompt direction as follows:

\[
\Delta^m_j(i,j) = \mathbf{y}^m_i - \mathbf{y}^m_j. \tag{6}
\]

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 [45, 40, 64] — 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):

\[
\mathcal{L}_{\text{sem}}(\mathbf{S}^m_i, \mathbf{S}^m_j) = \max \left( 0, \lambda - \langle E_{\text{text}}(P), E_{\text{text}}(T) \rangle \right) \tag{7}
\]

where \( P \in \mathcal{P}_i^m \cup \mathcal{P}_j^m \) and \( \lambda \) 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:

\[
\mathcal{L}_{\text{pair}}^m(\mathbf{S}^m_i, \mathbf{S}^m_j) = \mathcal{L}_{\text{dir}}^m(\mathbf{S}^m_i, \mathbf{S}^m_j) + \mathcal{L}_{\text{sem}}^m(\mathbf{S}^m_i, \mathbf{S}^m_j). \tag{8}
\]

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:

\[
\mathcal{L}_{\text{total}} = \sum_{m=1}^{M} \sum_{1 \leq i < j \leq K_m} \mathcal{L}_{\text{pair}}^m(\mathbf{S}^m_i, \mathbf{S}^m_j), \tag{9}
\]

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 [24, 39, 64]), 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 \( T \) to learn the inclusive tokens and then apply them back to \( T \) 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 \( T \), which implies that the learned tokens can be compatible with a different human-written prompt. For human face images, an example \( T \) 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 [81, 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\(^6\) (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 [25, 37]. 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* [43] 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)* [20] is a recently proposed synthetic face dataset used for skin tone estimation. Following [20],

\(^6\)The token length used here is generalizable across the attributes we studied in this paper. See Appendix E.1 for a detailed ablation study.
CelebA is a real-face dataset with different resolutions and focuses. It is used in the CelebA-HQ dataset [43] and FairFace [34] is a real-face dataset with different resolutions. The HPS benchmark contains images from natural scenes. ITI-G can leverage various image sources to benefit inclusive text-to-image generation for various attributes.

Table 1. Comparison with baseline methods with (a) single attribute and (b) multiple attributes. Reference images are from CelebA. We use CLIP [56] as the attribute classifier [12, 14]. ITI-G 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 more results.

We use two metrics to quantify the distribution diversity and image quality. (1) Distribution Discrepancy ($D_{KL}$). 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 [34] combined with human evaluations. Specifically, for skin tone, which is extreme difficult to obtain an accurate scale [1, 2, 28], we adopt the most commonly used Fitzpatrick skin type [10] combined with off-the-shelf
models [20] for evaluation. (2) FID. We report the FID score [25, 52] (FFHQ [35]) to measure the image quality. Please see Appendix D for more details.

**Baselines.** We compare ITI-GEN to the following methods. (1) **Stable Diffusion** (SD) [63] without any modification. (2) **Ethical Intervention** (EI) [5] that edits the prompt by adding attribute-related interventions. (3) **Hard Prompt Searching** (HPS) [18] 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) [39] that fine-tunes the text-to-image model with reference images based on Textual Inversion [24, 64].

**Implementation Details.** We use Stable Diffusion [63] (sd-v1-4) as the base model for all methods and show compatibility with ControlNet [81] and InstructPix2Pix [7]. ITI-GEN is model agnostic as long as they take token embeddings as the inputs. We set $\lambda = 0.8$ in $L_{\text{sem}}$ across all experiments and show that $\lambda$ can be robustly selected according to the prior knowledge (see Section 4.3). All the inclusive tokens are initiated as zero vectors. 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 [43]. Each represents a specific

---

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

9We 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).

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 [68] 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 [20]) is slightly better than real data [43, 34]. 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 \(\lambda\) 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 [81]. 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 [81], a state-of-the-art model capable of conditioning

\[\begin{array}{|c|c|c|}
\hline
\text{Method} & \text{Source} & \mathcal{L}_{\text{sem}} & \text{FID} \\
\hline
\text{Baseline [63]} & – & – & 67.40 \\
\text{CelebA [43]} & ✓ & ✓ & 60.38 \quad (+17.40) \\
\hline
\vdots & \vdots & \vdots & \vdots \\
\text{FAIR [20]} & ✓ & ✓ & 51.83 \quad (+10.86) \\
\hline
\end{array}\]

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.

\(^{10}\)Note that the subjective attributes we explore here are different from artistic styles (e.g., painting, cartoon) in image-to-image translation (e.g., [25]). Understanding the attributes related to quality and look of images may be intuitive for humans but remain non-trivial for generative models.
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 [44]. Previous attempts in image generation often resort to additional supervision, either through the use of reference data [13], classifiers learned from a joint distribution [70], 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, 22, 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.

Figure 11. Compatibility with models using additional conditions, e.g., human pose (left). ITI-GEN promotes inclusiveness of ControlNet [81] 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 [20], 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.

3977
References


[16] Huiwen 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


