FairQueue: Rethinking Prompt Learning for Fair Text-to-Image Generation

Christopher T. H. Teo*
christopher_teo@mymail.sutd.edu.sg

Milad Abdollahzadeh*
milad_abdollahzadeh@sutd.sg

Xinda Ma
xinda_ma@sutd.edu.sg

Ngai-Man Cheung†
ngaiman_cheung@sutd.edu.sg

Singapore University of Technology and Design (SUTD)

Abstract

Recently, prompt learning has emerged as the state-of-the-art (SOTA) for fair text-to-image (T2I) generation. Specifically, this approach leverages readily available reference images to learn inclusive prompts for each target Sensitive Attribute (tSA), allowing for fair image generation. In this work, we first reveal that this prompt learning-based approach results in degraded sample quality. Our analysis shows that the approach's training objective–which aims to align the embedding differences of learned prompts and reference images–could be sub-optimal, resulting in distortion of the learned prompts and degraded generated images.

To further substantiate this claim, **as our major contribution,** we deep dive into the denoising subnetwork of the T2I model to track down the effect of these learned prompts by analyzing the cross-attention maps. In our analysis, we propose novel prompt switching analysis: I2H and H2I. Furthermore, we propose new quantitative characterization of cross-attention maps. Our analysis reveals abnormalities in the early denoising steps, perpetuating improper global structure that results in degradation in the generated samples. Building on insights from our analysis, we propose two ideas: (i) *Prompt Queuing* and (ii) *Attention Amplification* to address the quality issue. Extensive experimental results on a wide range of tSAs show that our proposed method outperforms SOTA approach's image generation quality, while achieving competitive fairness. More resources at Project Page.

1 Introduction

There has been significant progress in the quality of text-to-image (T2I) generation [1–3] resulting in increasing adoption in different applications [4–10]. With this comes concerns regarding the fairness of these T2I models and their societal impacts [11–15].

Fair T2I Generation. T2I models may inherit biases present in their training data. Several approaches have been proposed to mitigate these biases [16–19] (See related work in Supp). Particularly, **Inclusive T2I Generation (ITI-GEN)** [16]–the existing SOTA–suggests that fair T2I approaches based on hard prompts (HP) (*e.g.,* "`A headshot of a person with fair skin tone`") are limited by linguistic ambiguity. For example, `Skin Tone` is often challenging to define and interpret based on HP, resulting in sub-optimal performance. To overcome this linguistic ambiguity, ITI-GEN adopts the notion that "a picture is worth a thousand words" and leverages readily available reference

*Equal Contribution
†Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0720

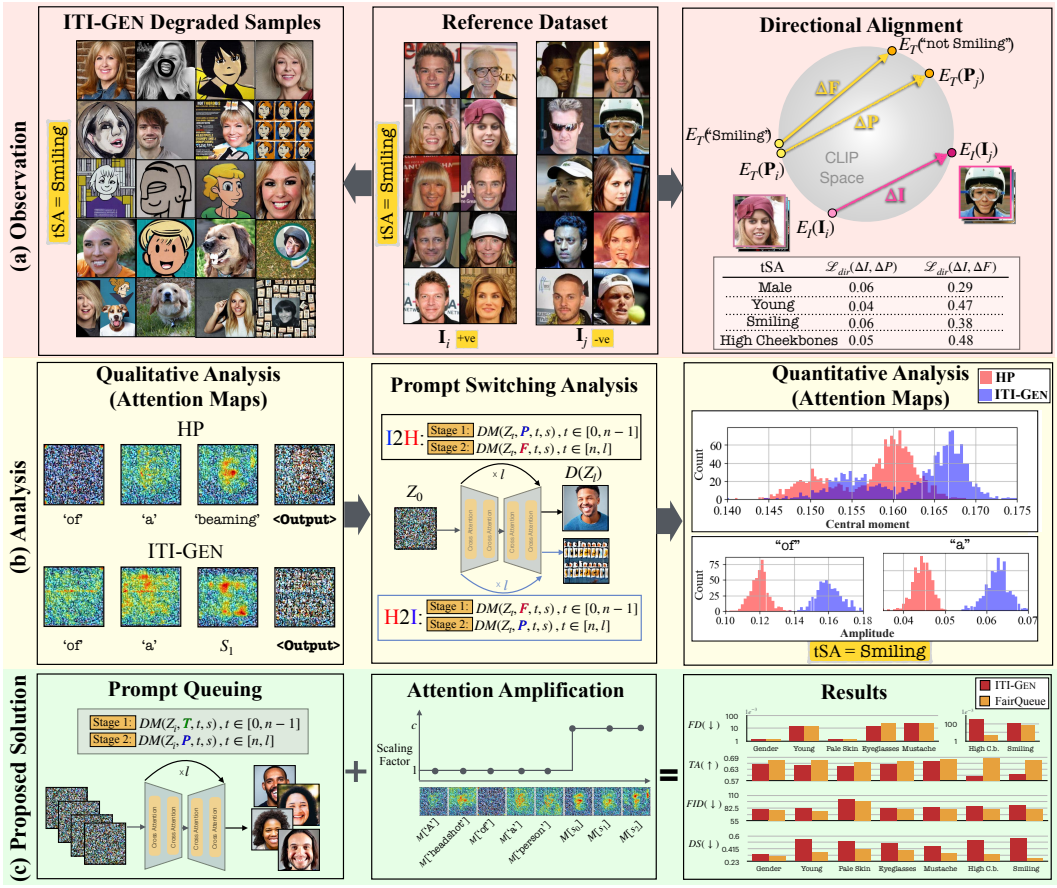


Figure 1: Our work re-visits SOTA fair T2I generation, ITI-GEN. We question ITI-GEN's central idea of prompt learning via alignment between the directions of prompt embeddings and reference image embeddings. (a) We observe degradation in images generated through ITI-GEN's learned prompts. We note that the direction of reference image embeddings could include unrelated concepts beyond tSA differences (*e.g.,* variations in accessories) resulting in learning of distorted prompts using ITI-GEN. Furthermore, we observe misalignment between the direction of credible hard prompts and that of reference images/learned prompts. (b) As our main contribution and to further understand how these distorted prompts affect the image generation process, we deep dive into the denoising network and analyze the cross-attention maps, revealing their abnormalities *e.g.,* higher activity for maps associated with non-tSA tokens (''of'', ''a''). We examine the degraded global structures resulting from these distorted prompts in the early denoising steps. Moreover, we propose I2H and H2I (Eq.2) analysis to understand impact of these degraded global structures and abnormalities in later denoising steps. In addition, we propose metrics (Eq.3) on cross-attention maps to quantify these abnormalities. (c) Building on insights from our analysis, we propose a solution to address distorted prompts while maintaining competitive fairness. Our solution FairQueue includes two ideas: prompt queuing and attention amplification. E_T and E_I are CLIP text and image encoder resp. [20]. $\boldsymbol{T}$, $\boldsymbol{F}$, $\boldsymbol{P}$ are the base prompt, hard prompt with minimal linguistic ambiguity, and ITI-GEN prompt, resp.

images to learn an inclusive prompt for each tSA category. This approach translates visual attribute differences present in the reference images into prompt differences, enabling the learned prompts to be used to generate images of all tSA categories, regardless of their linguistic ambiguity. Fairness is achieved by uniformly sampling the learned prompts to condition the T2I generation. *Central to this approach is the enforcement of directional alignment between learned prompt embeddings and reference image embeddings corresponding to a pair of tSA categories.*

In this work, we question the central idea of prompt learning via alignment between the direction of prompt embeddings and the direction of reference image embeddings in the context of fair T2I generative models. Our work starts with examining the generated images and observes that a moderate amount of degraded images are generated based on ITI-GEN. We argue that using the direction of reference image embeddings as guidance could be sub-optimal, as the difference between reference images could include additional unrelated concepts other than the tSA difference

https://doi.org/10.52202/079017-0720

(Fig 1). For example, reference images of ''`A headshot of a person smiling`'' and ''`A headshot of a person not smiling`'' could contain differences in poses, accessories, hairstyling, in addition to the difference in `smiling`. Therefore, the direction of reference image embeddings could be noisy and include additional unrelated concepts other than the tSA difference. We perform an analysis on the direction of embeddings to further understand the issue. **We hypothesize that using the direction of reference image embeddings as guidance could lead to distortion in the learned prompts, resulting in artifacts and quality degradation in the images generated by T2I models.**

To further substantiate this claim, **as our major contribution**, we deep dive into the denoising subnetwork of the T2I model to analyze ITI-GEN prompts in the generation pipeline. Our analysis include examination of the cross-attention maps of the learned prompts at individual time steps of the denoising process. We propose novel prompt switching analysis: **ITI-GEN to HP (I2H)**, and **HP to ITI-GEN (H2I)**. We further propose new quantitative metrics for cross-attention map characterization. Our analysis reveals cross-attention maps of the learned prompts have abnormalities in the initial time steps of the denoising process. This results in synthesizing improper global structures. Interestingly, we find that the learned prompts have a minimum abnormality in the later steps–the learned prompts perform adequately in generating the desired tSA category *provided that proper global image structures could be synthesized in the initial denoising steps.* To justify our analysis of cross-attention, we remark that cross-attention contextualizes prompt embeddings with the latent representation of images and has been shown to play a key role in T2I models [21, 22].

Building on the insights of our analysis, we propose a solution to address degraded generated images without compromising fairness and diversity. Particularly, we propose Prompt Queuing to apply base prompts (without tSA tokens) in the initial time steps and ITI-GEN learned prompts in the later time steps of the denoising process. We further propose Attention Amplification to balance the quality and fairness of the T2I generation. Overall, our solution can effectively address the degraded quality issue in ITI-GEN while maintaining competitive fairness. Our contributions are:

- We examine the generated images from the prompt learning-based fair T2I generation approach and reveal a moderate amount of generated images with degradation (Sec 3.1).
- We argue that the direction of reference image embeddings could be noisy and include unrelated concepts in addition to tSA difference, and prompt learning based on alignment with the direction of reference image embeddings could be sub-optimal (Sec 3.1).
- We deep dive into the denoising subnetwork of the T2I model and analyze cross-attention maps with our proposed prompt switching analysis I2H and H2I, and our proposed quantitative metrics for cross-attention maps. Our analysis reveals and characterizes abnormalities in cross-attentions of ITI-GEN prompts in the denoising process (Sec 3.2).
- We propose FairQueue, a solution based on prompt queuing and attention amplification to improve generation quality while maintaining competitive fairness (Sec 4).

2 Preliminaries

T2I Generation. SOTA T2I generation is based on diffusion model (DM) [1–3]. In the forward diffusion process, Gaussian noise is incrementally added to the training data to train the DM. Then, during reverse diffusion, the DM generates samples by randomly sampling latent noise $Z_0 \sim N(0, \boldsymbol{I})$ as an input. For more control, text-conditioning [1, 23–25] was introduced, where we denote the reverse diffusion (denoising) of a single step t by $Z_{t+1} \leftarrow DM(Z_t, \boldsymbol{R}, t, s)$. Here, Z_t is the latent of the noisy image, $\boldsymbol{R}$ the input prompt, $t \in [0, l]$ the denoising step, and s a random seed. Central to text conditioning is the cross-attention mechanism which contextualizes prompt embeddings with the image latent [21, 26]. Specifically the **cross-attention map** $\boldsymbol{M} \in \mathbb{R}^{r \times m \times n}$–where r is the number of tokens in the prompt, and $m \times n$ shows map size for each token–is computed by:

$$\boldsymbol{M} = SoftMax(\frac{QK^T}{\sqrt{d}}) \tag{1}$$

where, $Q = \ell_q(\phi(Z_t))$ is the linear projection of the latent spatial features $\phi(Z_t)$, and $K = \ell_q(E_T(\boldsymbol{R}))$ is the linear projection of the textual embedding $E_T(\boldsymbol{R})$ (usually CLIP text encoder [20]). For ease of notation, we refer to the token-specific attention maps as $\boldsymbol{M}[.]$ *e.g.,* $\boldsymbol{M}[$''`of`''$] \in \mathbb{R}^{m \times n}$ refers to the cross-attention map for the token ''`of`'' in $\boldsymbol{R}$. As our work focuses on the reverse diffusion process, we utilize Z_0 as the noisy latent input and Z_l as the final latent output. This Z_l is then finally passed into the DM decoder to output generated image, $D(Z_l)$.

https://doi.org/10.52202/079017-0720

Fairness in Generative Models. In generative models, fairness is defined as *equal representation* [27, 28], where for a tSA with K categories, a fair generator will generate an equal number of samples for each category. As an example, for a T2I model G with text prompt ''`A headshot of a person`'' as input, we consider G as fair model *w.r.t.* tSA = `Young`–with two categories `{Young, Old}` [29, 27, 28]– if it generates an equal number of samples for each categories of this tSA [30, 31].

Hard Prompts for Fair T2I Generation. A baseline for achieving fairness in T2I models is to append the tSA-related prompt to the *base prompt* [17, 19]. Considering the same tSA=`Young`, and the base prompt ''`A headshot of a person`'', adding a tSA-related prompt for each category results in the HPs: ''`A headshot of a person young/old`''. For a fair generation, we query T2I with each of these HPs uniformly. Note that HP although very effective with certain tSA, in most cases, is ineffective due to the tSAs having linguistic ambiguity [32]–having misleading or deceptive language.

Prompt Learning for Fair Text-to-Image Generation. To resolve the issue of ambiguous tSAs, inspired by the recent success of prompt learning [33, 34], ITI-GEN [16] aims to achieve fairness in a pre-trained T2I model by learning inclusive tokens for each category of the tSA. Assuming the tokenized *base prompt* as $\boldsymbol{T} \in \mathbb{R}^{p\times d}$, where p is the number of tokens and d is the dimension of the embedding space, for each category $k \in \{1, ...K\}$ of tSA, it learns q additional tokens. $\boldsymbol{S}^k = [\boldsymbol{S}^k_0, \boldsymbol{S}^k_1, \ldots, \boldsymbol{S}^k_{q-1}] \in \mathbb{R}^{q\times d}$. In [16], q is set to 3. Then, *ITI-GEN prompt* is constructed by appending these learned tokens to the original tokens: $\boldsymbol{P}_k = [\boldsymbol{T}; \boldsymbol{S}^k] \in \mathbb{R}^{(p+q)\times d}$. These tokens are learned using a set of labeled reference images (*w.r.t.* tSA) $\mathcal{D}_{ref} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N; y_i \in \{1, ..., K\}$ to provide stronger signals for describing tSA. More specifically, for a pair of categories (i, j) of tSA, a directional loss [35] is used to match the direction of learned prompts and images for this pair in CLIP embedding [20] space *i.e.,* $\min_{\boldsymbol{S}^i,\boldsymbol{S}^j} \mathcal{L}_{dir} = 1 - \frac{(\Delta \boldsymbol{I}_{(i,j)} \cdot \Delta \boldsymbol{P}_{(i,j)})}{(|\Delta \boldsymbol{I}_{(i,j)}||\Delta \boldsymbol{P}_{(i,j)}|)}$, where $\Delta \boldsymbol{I}_{(i,j)}$ ($\Delta \boldsymbol{P}_{(i,j)}$) denotes the direction between images (text prompts) of two categories i and j in CLIP's embedding space, and directional loss $\mathcal{L}_{dir}$ is minimized to learn tSA tokens $\boldsymbol{S}^i, \boldsymbol{S}^j$ for these categories. Finally, using $\boldsymbol{P}_k$ as input prompt, fairness is achieved by uniformly sampling the K categories of the tSA. We will omit the category index k when it is clear from context, and denote learned prompt and tokens by $\boldsymbol{P}$ and $\boldsymbol{S}_0, \ldots, \boldsymbol{S}_{q-1}$ resp.

3 A Closer Look at Prompt Learning for Fair Text-to-Image Generation

In this section, we take a closer look at ITI-GEN [16]. First, in Sec. 3.1, we analyze ITI-GEN performance where we find quality degradation in moderate number of generated samples. We attribute this to the sub-optimal learning objective in ITI-GEN, which captures unrelated concepts that distort the learned tokens in $\boldsymbol{P}$. Then, in Sec. 3.2, we analyze ITI-GEN prompts during sample generation by inspecting the cross-attention mechanism. Our analysis reveals that ITI-GEN prompts give rise to abnormality particularly damaging to the early steps of the denoising process.

Remark. To conduct the following analysis on ITI-GEN prompts' behavior we require a strong baseline as a pseudo-gold standard to compare against. To address this, we found that when considering certain tSA with minimal linguistic ambiguity (MLA) [32]–a few tSA that can be described without misleading or deceptive language–HPs can serve as this strong baseline. Therefore, in this section, we focus on tSAs with minimum linguistic ambiguity. Later, in experiment section, we will include all tSAs, with or without ambiguity.

3.1 Limitations of Prompt Learning for Fair T2I Generation

Although ITI-GEN [16] improves fairness in T2I generation, a closer examination of its outputs reveals a potential trade-off: compromised image quality. In this section, first, we perform a systematic experiment to showcase these quality issues and then explore the potential root causes behind them.

Experimental Setup. To evaluate our generated samples, we utilize the metrics: i) Fairness Discrepancy (FD) [27, 31, 11, 36] to measure fairness, ii) Text-Alignment (TA) [37, 22] and FID [38] to measure quality, and iii) DreamSim (DS) [39] to measure semantic preservation. Next, we determine a set of tSA with MLA to compare ITI-GEN with HP (as a pseudo-gold standard). Specifically, we follow [16] and use pre-trained *Stable Diffusion* (SD) [1] as T2I model. Then as mentioned in Sec. 2, for HP, we append the tSA-related prompts to the base prompt. We empirically found that tSAs `{Smiling, High Cheekbones}`, are unambiguous by classifying 500 generated sample per HP utilizing CLIP classifier [20], where on average they both achieve a 98% accuracy (Experiment details in Supp). Then, for ITI-GEN [16], we strictly follow [16] and use publicly available fair

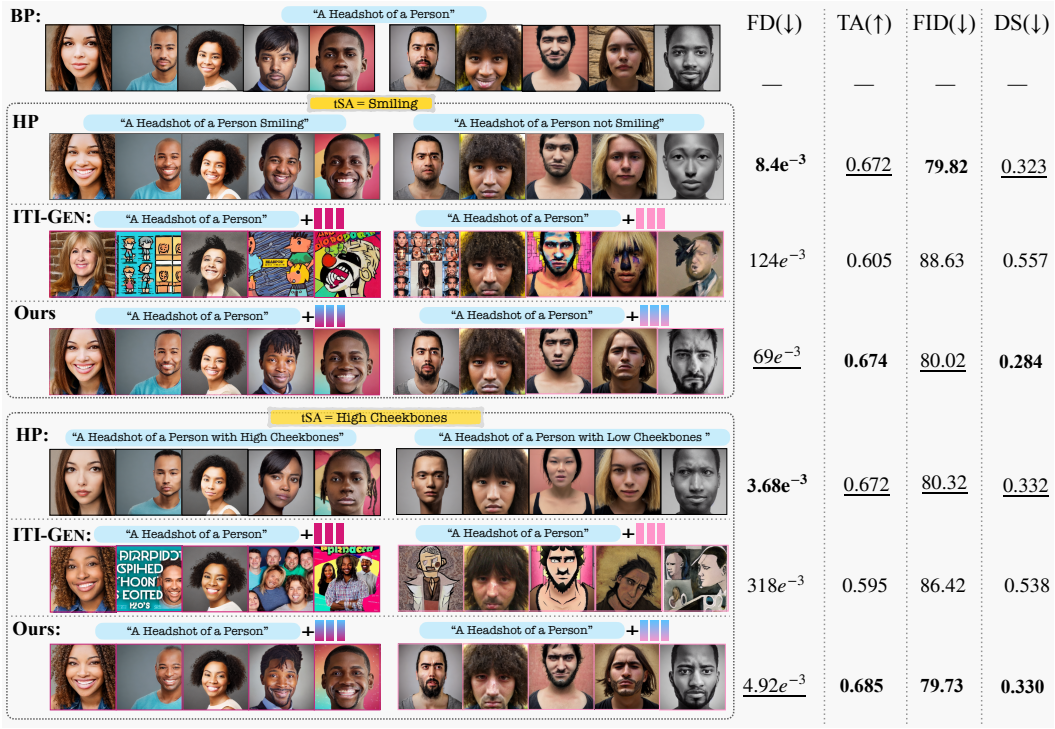


Figure 2: **T2I generation performance of HP, ITI-GEN [16], and our proposed FairQueue for target Sensitive Attributes (tSAs) with minimal linguistic ambiguities.** Samples generated by HP demonstrate outstanding performance with good fairness (FD), high quality (FID and FD), and good semantic preservation (DS). Meanwhile, ITI-GEN moderately degrades sample quality, impacting fairness and semantic preservation. FairQueue demonstrates comparable performance to HP, even surpassing HP in both quality and semantic preservation in many cases. Note that HP only performs well for unambiguous tSAs, and can not be used for general fair T2I generation purposes, as it can not be defined well for ambiguous tSAs (See Supp for detailed discussion).

image dataset–sampled from CelebA [29]–as reference images to learn inclusive tokens, $\boldsymbol{S}$. Finally, we generate and evaluate ITI-GEN samples based on the same latent noise input as HP. See Supp for experiment and metric details.

Fig. 2 shows some generated samples together with quantitative results. A moderate number of generated images with ITI-GEN have quality degradation often with unrelated content (*e.g.,* generating dog, multiple degraded faces, vague cartoons, etc.). Quantitative results show that for both tSAs, HP performs better in fairness (lower FD), quality (higher TA, lower FID), and semantic preservation (lower DS). We postulate degraded samples stem from ITI-GEN's sub-optimal training objective.

Issue of Directional Loss for Fair T2I Generation. We hypothesize that directional loss is sub-optimal in learning tSA-related tokens $\boldsymbol{S}$. Particularly, the differences in reference images $\Delta\boldsymbol{I}$ can include unrelated concepts in addition to variation in tSA categories. For example, considering tSA=`Smiling` in Fig. 1a (col 2), the reference images used for learning these two categories contain differences in pose, accessories, etc., in addition to the difference in `Smiling`. We further explore this potential of encoding unrelated concepts in $\Delta\boldsymbol{I}$ by taking a closer look into the CLIP embedding space, where ITI-GEN's learning process happens. Recall, that as we utilize tSA with MLA and the HPs only differ in the tSA categories, we can utilize them as references in our analysis. For example, considering tSA=`Smiling`, the related tokenized prompts of HP in CLIP space can be computed as follows: $\boldsymbol{F}_i = E_T$ (''`A headshot of a person smiling`''), and $\boldsymbol{F}_j = E_T$ (''`A headshot of a person not smiling`''), with E_T denoting CLIP's text encoder. Then $\Delta\boldsymbol{F} = \boldsymbol{F}_i - \boldsymbol{F}_j$ shows the direction of the tokenized prompts in the CLIP embedding space.

Our results in Fig. 1a (col3) shows the directional loss between $\Delta\boldsymbol{I}$ and $\Delta\boldsymbol{F}$ *i.e.,* $\mathcal{L}_{dir}(\boldsymbol{I}, \boldsymbol{F})$, for different tSAs using the reference images for each tSA. Note that $\mathcal{L}_{dir} = 0$ means perfect alignment. Our comparison reveal considerable misalignment between $\Delta\boldsymbol{I}$ and $\Delta\boldsymbol{F}$ implying that unrelated concepts are potentially encoded in $\Delta\boldsymbol{I}$. Meanwhile, $\Delta\boldsymbol{I}$ and $\Delta\boldsymbol{P}$ near perfect alignment implies that

https://doi.org/10.52202/079017-0720

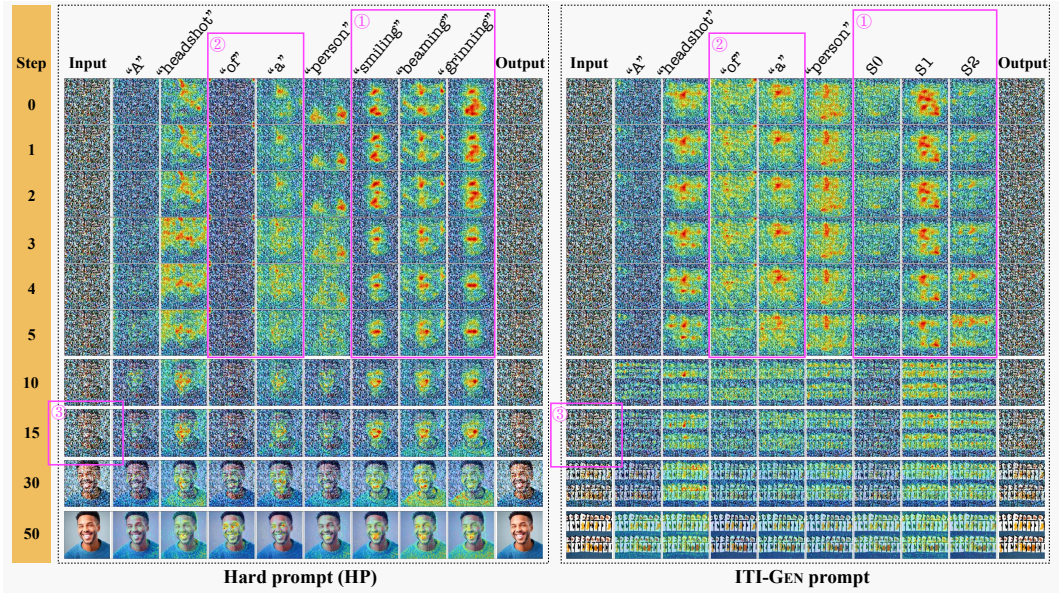


Figure 3: **Comparison of cross-attention maps during the denoising process with HP (left) and ITI-GEN (right).** Here, we use tSA=`Smiling` and plot the denoising process for one sample generation. Each denoising process consists of $l = 50$ steps initiated with the same noisy input. Each cell depicts the attention map for the respective token (column) at the respective step (row) overlaid on the input. We highlight 3 key observations: ① ITI-GEN tokens $\boldsymbol{S}_i$ have abnormal activities compared to the corresponding tSA-related tokens in HP by attending to unrelated regions (backgrounds) or scattered attention. ② non-tSA tokens like ''`of`'' and ''`a`'' are abnormally more active in the presence of ITI-GEN tokens. ③ As compared to the HP counterpart, issues created by ITI-GEN tokens (① & ②) degrade the global structure in the early denoising steps (*e.g.,* Step 15), for example, human face in HP vs some unrelated structure in ITI-GEN. The same behavior is observed for some other samples and tSAs (see Supp for more samples, and other tSAs with more denoising steps).

these unrelated concepts are potentially transferred to $\boldsymbol{P}$ via ITI-GEN's learning objective, resulting in distorted learned token $\boldsymbol{S}$.

3.2 Analyzing the Effect of ITI-GEN Prompts in T2I Generation

In the previous section, we observed degraded sample quality in ITI-GEN which we attribute to the sub-optimal training objective that results in learning distorted tokens. In this section, we take a step further to answer the question: *"Given a pre-trained T2I model and some distorted learned prompts as input, how do these distorted prompts affect the image generation process of the T2I model?"*

To answer this, we deep dive into the latent denoising network [1] and analyze the cross-attention mechanism [40]–the bridge for text and image modules in T2I models [1, 23–25]. In this analysis, *we visualize the cross-attention maps to investigate potential anomalies caused by distorted tokens in the denoising process.* Specifically, we compare cross-attention maps of ITI-GEN prompt against HP with minimal linguistic ambiguity (as reference). To allow fair token-to-token comparison, in this experiment, we lengthen HP by including additional tokens containing synonyms of the tSA. Note that this did not augment HP's behavior, and similar results are seen in the original HP. See Supp for more details.

Visualizing Cross-attention Maps. We follow DAAM [41] for visualizing cross-attention maps by tracing attention scores in the cross-attention module to demonstrate how an input token within a prompt influences parts of the generated image. Specifically, to visualize the cross-attention map of a token, DAAM interpolates and accumulates the attention scores over all scales (layer of the U-Net [42] as the denoising network [1]), and all denoising steps. However, we tailor DAAM to the requirements of our fine-grained analysis by introducing further controls. First, we isolate the attention maps for each denoising step to allow for both step-wise and multi-step analysis. Second, we introduce a prompt-switching mechanism, allowing for the interchangeable tracing of different prompts at any particular denoising step.

https://doi.org/10.52202/079017-0720

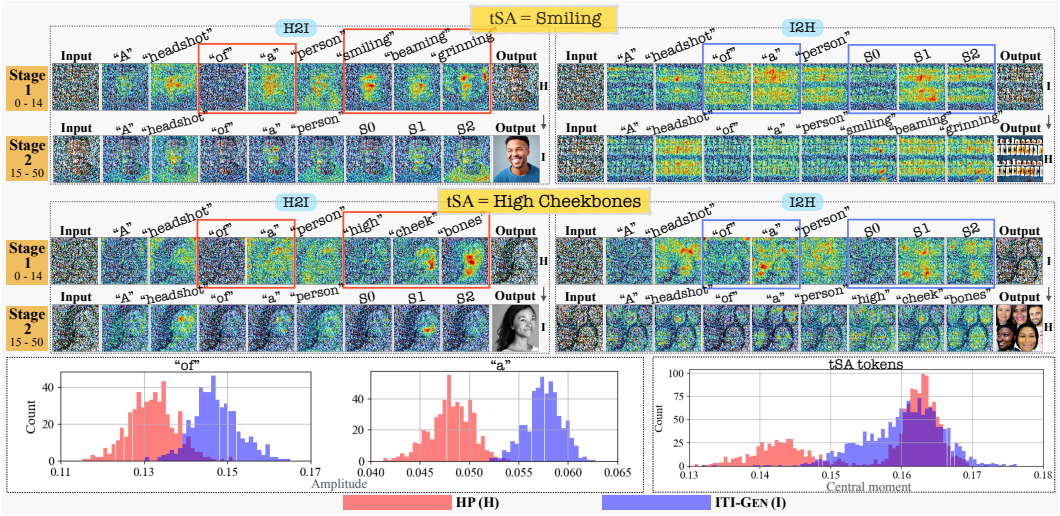


Figure 4: **Analyzing the accumulated cross-attention maps for the denoising process in our proposed prompt switching analysis I2H and H2I.** Here, we use two tSAs: `Smiling`, `High Cheekbones`. For each tSA, we show the accumulated cross-attention maps for H2I and I2H, with some quantitative results. In the H2I (I2H) experiment, the first row shows the accumulated cross-attention maps during the early denoising steps with HP (ITI-GEN) as the input prompt, and the second row shows the maps during later steps, after switching to ITI-GEN (HP). **Observation 1:** *Learned tokens in* ITI-GEN *affect early denoising steps, degrading global structure synthesis; such degraded global structure disrupts the final output.* This is observed in I2H. **Observation 2:** *Learned tokens in* ITI-GEN *works decently in the later stage of the denoising process if the global structure is synthesized properly.* This is observed in H2I. As we show in Supp, similar observations can be made for other samples and other tSAs. Bottom: Histograms of our proposed metrics on cross-attention maps demonstrate the abnormalities in many samples.

Abnormalities in the Presence of Distorted Tokens. To investigate potential anomalies arising from distorted tokens learned by ITI-GEN, we comprehensively analyze cross-attention maps of all denoising time steps on 500 generated samples per category of each tSA, for both ITI-GEN and HP. Fig. 3 shows one of these cross-attention maps comparing ITI-GEN and HP for tSA=`Smiling` (more examples in Supp). Our empirical investigation reveals four points: i) global structure is synthesized in the early steps of the denoising process aligning with previous works [21] that the denoising process progressively synthesizes the image. ii) Learned ITI-GEN tokens have abnormal attention compared to the tSA-related tokens in HP (‘‘`Smiling`’’ in col. 7 of HP), *e.g.*, $\boldsymbol{M}[\boldsymbol{S_i}]$ contain unrelated or scatter activation (**Issue 1**). iii) In the presence of the ITI-GEN tokens, other non-tSA tokens (like $\boldsymbol{M}[\text{‘‘a’’}]$ and $\boldsymbol{M}[\text{‘‘of’’}]$) are abnormally more active (**Issue 2**). We remark that tokens interact with each other in the denoising steps. iv) Considering **Issues 1 & 2**, we observe that degraded global structure is synthesized in the early steps of denoising, and eventually a degraded sample is generated at the end of the denoising. Note that similar issues occur with many other samples and tSAs (details in Supp).

To further understand issues, we isolate effect of distorted tokens by proposing two analyses focusing on different denoising steps. These two analyses dissect the influence of distorted tokens in key denoising steps (Recall denoising of a single step t is denoted by $Z_{t+1} \leftarrow DM(Z_t, \boldsymbol{R}, t, s)$, Sec. 2):

$$\text{I2H} = \begin{cases} DM(Z_t, \boldsymbol{P}, t, s) & t \in [0, n-1] \\ DM(Z_t, \boldsymbol{F}, t, s) & t \in [n, l] \end{cases}, \text{H2I} = \begin{cases} DM(Z_t, \boldsymbol{F}, t, s) & t \in [0, n-1] \\ DM(Z_t, \boldsymbol{P}, t, s) & t \in [n, l] \end{cases} \quad (2)$$

To do this, as seen in Fig.1b (col 2), we first propose **Analysis 1: switching prompt from ITI-GEN to HP (I2H)** during the denoising process. This allows for a better understanding of how the global structure's degradation in early denoising steps may affect the final generated output. Specifically, as in Eq. 2, I2H first utilizes *ITI-GEN prompt* $\boldsymbol{P}$ in early denoising steps which potentially leads to degraded global structure. This is then followed by utilizing *hard prompt* $\boldsymbol{F}$ for the remaining denoising steps. Next, to investigate if ITI-GEN tokens will create the same issues in the later steps of the denoising process, we propose **Analysis 2: HP to ITI-GEN (H2I)**. Converse to the previous experiment, we utilize $\boldsymbol{F}$ in early steps of denoising, and then switch to using $\boldsymbol{P}$ as input prompt. For each experiment, we plot and analyze the cumulative cross-attention maps for early steps (0 to $n-1$) and later steps (n to l) separately. Fig. 4 shows an example of the cross-attention maps for

https://doi.org/10.52202/079017-0720

these two experiments with tSA={`Smiling`, `High Cheekbones`}. See Supp for more samples and details. Considering results in Fig. 4 the following observations can be made:

Observation 1: *Learned tokens in* ITI-GEN *affect the early steps of the denoising process leading to degradation in synthesizing global structure.* More specifically comparing the first row of the cross-attention map between I2H and H2I in Fig. 4, we can have the following observations: i) ITI-GEN tokens have more scattered attention or attending to unrelated regions compared to tSA-related tokens in HP; ii) non-tSA tokens like ''`a`'', and ''`of`'' are more active in the presence of the ITI-GEN tokens. These two abnormalities result in degraded global structure in the early steps. In addition, considering the second row of the I2H in Fig. 4, the degraded global structure in the early steps leads to disrupted final output even though the (non-distorted) HP prompt is used in later steps.

Observation 2: *Learned tokens in* ITI-GEN *works decently in the later steps of the denoising process if the global structure is synthesized properly.* More specifically, considering H2I in Fig. 4, when HP prompts synthesize proper global structure in early steps, the ITI-GEN tokens attend to proper regions and contribute to adding the finer details related to tSA, as shown in the second row of H2I.

Quantitative Metrics for Cross-attention Maps. In addition to the visual demonstration, we propose two metrics to support further our observed abnormalities of ITI-GEN tokens in the early steps of denoising for a large number of generated samples. Specifically, for each generated sample: i) To quantify abnormally active attention associated with non-tSA tokens, we compute the expectation of **attention amplitude**: $\mathbb{E}_{(x,y)}\{\boldsymbol{M}[\boldsymbol{J}]\}$, where $\boldsymbol{J}$ is a non-tSA token such as ''`of`''. ii) We analyze the scatter in attention by measuring the second **central moment** [43] for each tSA token $\boldsymbol{K}$:

$$\mu(\boldsymbol{K}) = \sum_{x,y}\{[(x-\bar{x})^2 + (y-\bar{y})^2]\tilde{\boldsymbol{M}}[\boldsymbol{K}]_{(x,y)}\} \tag{3}$$

Here, $\tilde{\boldsymbol{M}}[\boldsymbol{K}] = (\boldsymbol{M}[\boldsymbol{K}]/\sum_{x,y}\boldsymbol{M}[\boldsymbol{K}])$, and $(\bar{x}, \bar{y})$ is the centroid. The two metrics are computed on the accumulated cross-attention maps from stage 1 of I2H (for ITI-GEN) and H2I (for HP). The histograms of these two metrics for 500 generated samples in Fig. 1b (col 3) and Fig. 4 demonstrate **Issues 1 & 2** in many generated samples.

Remark. Our thorough analysis in this section shows that distorted tokens learned by ITI-GEN only have destructive performance in the early steps of denoising, and they generally have decent performance in later steps when the global structure is formed properly (H2I). *We remark that even though H2I has decent performance in fair and high-quality T2I generation, it is only applicable to tSA with minimal linguistic ambiguity. In the next section, we will discuss our proposed method to address fair and high-quality T2I generation encompassing both ambiguous and unambiguous tSA.*

4 Proposed Method

In this section, we present our proposed method, FairQueue a new generation framework consisting of two additions: *Prompt Queuing* and *Attention Amplification* to improve the sample quality when implementing fair T2I generation. In addition to quality improvements, FairQueue also allows for better semantic preservation of the original sample generated from the base prompt $\boldsymbol{T}$.

Prompt Queuing. Recall that when utilizing ITI-GEN prompt $\boldsymbol{P}$–which is tuned to generate samples containing the tSA–degraded global structure occurs in early denoising steps for a moderate number of samples. Conversely, utilizing HP with minimal linguistic ambiguity enables high-quality and fair T2I generation. However, as such HPs are not available for all tSAs [16], we naturally consider the next best available option–the base prompt $\boldsymbol{T}$ (a natural language prompt without the distorted trainable tokens)–and propose prompt queuing. Specifically, as seen in Fig. 1(c) , prompt queuing first utilizes $\boldsymbol{T}$ in the early n denoising steps, thereby allowing for the global structures to form properly. Next, we transit to ITI-GEN prompt $\boldsymbol{P}$ for the remaining $(l-n)$ steps. This allows the more fine-grained tSA semantics to be developed on top of the already well-defined global structures.

Attention Amplification. By implementing prompt queuing, the output samples may experience a reduction in tSA expression due to the reduced exposure to the ITI-GEN prompt $\boldsymbol{P}$. To address this, we propose Attention Amplification, an intuitive solution that emphasizes the expression of the tSA by scaling the ITI-GEN token's cross-attention maps, *i.e.,* $c * \boldsymbol{M}[S_i]$ where $c > 1$.

5 Experiments

In this section, we evaluate our proposed (FairQueue) against the existing SOTA ITI-GEN [16] over various tSA. Then, we conduct an ablation study by first evaluating the contribution brought by each

Table 1: **Evaluating Proposed FairQueue against ITI-GEN .**We utilize *FD:* Fairness Discrepancy (↓), *TA:* Text-Alignment (↑), *FID* (↓), and *DS:* DreamSim (↓) to determine the fairness, quality, and semantic preservation, respectively. For FD a combination of CLIP [20], off-the-shelf classifier [44, 45] and human evaluator were utilized as tSA classifier. For TA we utilize CLIP [20] as the feature extractor. Overall, our proposed method demonstrates the best ability to balance between sample quality and fairness, while preserving the semantics of the original base-prompt $\boldsymbol{T}$.

tSA		FD (↓)	TA (↑)	FID (↓)	DS (↓)
Single tSA (CelebA)					
`Gender`	ITI-GEN	$\mathbf{6.41e^{-3} \pm 4.2e^{-3}}$	$0.655 \pm 1.2e^{-2}$	78.9 ± 1.3	$0.337 \pm 1.4e^{-2}$
	Ours	$\mathbf{6.41e^{-3} \pm 3.8e^{-3}}$	$\mathbf{0.676 \pm 5.2e^{-3}}$	$\mathbf{78.3 \pm 1.5}$	$\mathbf{0.308 \pm 1.2e^{-2}}$
`Young`	ITI-GEN	$\mathbf{13.1e^{-3} \pm 8.1e^{-3}}$	$0.653 \pm 9.4e^{-3}$	82.9 ± 1.4	$0.552 \pm 3.2e^{-2}$
	Ours	$15.5e^{-3} \pm 3.8e^{-3}$	$\mathbf{0.678 \pm 8.1e^{-3}}$	$\mathbf{75.3 \pm 2.1}$	$\mathbf{0.370 \pm 2.7e^{-2}}$
`Smiling`	ITI-GEN	$124e^{-3} \pm 9.2e^{-3}$	$0.605 \pm 1.2e^{-2}$	88.6 ± 0.9	$0.557 \pm 2.2e^{-2}$
	Ours	$\mathbf{69.0e^{-3} \pm 4.2e^{-3}}$	$\mathbf{0.674 \pm 1.7e^{-2}}$	$\mathbf{80.0 \pm 1.3}$	$\mathbf{0.284 \pm 1.0e^{-2}}$
`High Cheekbones`	ITI-GEN	$318e^{-3} \pm 12.0e^{-3}$	$0.595 \pm 1.2e^{-3}$	86.40 ± 2.1	$0.538 \pm 1.6e^{-2}$
	Ours	$\mathbf{4.92e^{-3} \pm 3.6e^{-3}}$	$\mathbf{0.685 \pm 7.2e^{-3}}$	$\mathbf{79.7 \pm 2.4}$	$\mathbf{0.330 \pm 2.2e^{-2}}$
`Pale Skin`	ITI-GEN	$\mathbf{1.41e^{-3} \pm 1.2e^{-3}}$	$0.646 \pm 1.8e^{-2}$	101.3 ± 4.6	$0.525 \pm 2.8e^{-2}$
	Ours	$\mathbf{1.41e^{-3} \pm 1.2e^{-3}}$	$\mathbf{0.666 \pm 1.9e^{-2}}$	$\mathbf{97.0 \pm 3.2}$	$\mathbf{0.408 \pm 3.0e^{-2}}$
`Eyeglasses`	ITI-GEN	$\mathbf{14.1e^{-3} \pm 2.6e^{-3}}$	$0.654 \pm 3.3e^{-3}$	83.5 ± 1.4	$0.486 \pm 1.4e^{-2}$
	Ours	$25.4e^{-3} \pm 1.9e^{-3}$	$\mathbf{0.670 \pm 6.1e^{-3}}$	$\mathbf{79.4 \pm 2.3}$	$\mathbf{0.391 \pm 1.6e^{-2}}$
`Mustache`	ITI-GEN	$26.2e^{-3} \pm 1.8e^{-3}$	$0.670 \pm 4.2e^{-3}$	85.0 ± 3.3	$0.452 \pm 1.9e^{-3}$
	Ours	$\mathbf{22.6e^{-3} \pm 1.2e^{-3}}$	$\mathbf{0.680 \pm 5.3e^{-3}}$	$\mathbf{80.2 \pm 3.0}$	$\mathbf{0.345 \pm 3.1e^{-3}}$
`Chubby`	ITI-GEN	$\mathbf{112e^{-3} \pm 8.8e^{-3}}$	$0.647 \pm 2.2e^{-3}$	79.2 ± 1.5	$0.551 \pm 3.6e^{-3}$
	Ours	$119e^{-3} \pm 7.2e^{-3}$	$\mathbf{0.675 \pm 2.3e^{-3}}$	$\mathbf{78.3 \pm 1.4}$	$\mathbf{0.387 \pm 3.0e^{-3}}$
`Gray Hair`	ITI-GEN	$286e^{-3} \pm 6.8e^{-3}$	$0.640 \pm 4.3e^{-3}$	87.3 ± 2.1	$0.533 \pm 2.9e^{-3}$
	Ours	$\mathbf{266e^{-3} \pm 7.1e^{-3}}$	$\mathbf{0.669 \pm 3.7e^{-3}}$	$\mathbf{82.2 \pm 2.3}$	$\mathbf{0.417 \pm 3.1e^{-3}}$
Multi tSA (CelebA)					
`Gender` × `Young`	ITI-GEN	$39.1e^{-3} \pm 1.2e^{-3}$	$0.668 \pm 7.1e^{-3}$	72.6 ± 3.1	$0.458 \pm 7.8e^{-3}$
	Ours	$\mathbf{12.4e^{-3} \pm 2.3e^{-3}}$	$0.686 \pm 5.7e^{-3}$	$\mathbf{71.7 \pm 2.5}$	$\mathbf{0.373 \pm 4.4e^{-3}}$
`Gender` × `Young` × `Eyeglasses`	ITI-GEN	$257e^{-3} \pm 8.7e^{-3}$	$0.654 \pm 3.3e^{-3}$	65.2 ± 1.6	$0.475 \pm 1.1e^{-3}$
	Ours	$\mathbf{208e^{-3} \pm 7.3e^{-3}}$	$\mathbf{0.671 \pm 4.1e^{-3}}$	$\mathbf{61.5 \pm 2.7}$	$\mathbf{0.360 \pm 6.3e^{-3}}$
`Gender` × `Young` × `Eyeglasses` × `Smiling`	ITI-GEN	$190e^{-3} \pm 1.7e^{-2}$	$0.643 \pm 7.7e^{-3}$	65.5 ± 2.7	$0.475 \pm 9.1e^{-3}$
	Ours	$\mathbf{168e^{-3} \pm 1.0e^{-2}}$	$\mathbf{0.661 \pm 2.4e^{-3}}$	$\mathbf{60.8 \pm 1.1}$	$\mathbf{0.379 \pm 9.7e^{-3}}$
Multi tSA (Fairface & Fair Benchmark)					
`Gender` × `Age`	ITI-GEN	$142e^{-3} \pm 4.2e^{-3}$	$0.659 \pm 7.2e^{-3}$	$\mathbf{58.24 \pm 3.4}$	$0.445 \pm 1.2e^{-3}$
	Ours	$\mathbf{108e^{-3} \pm 4.3e^{-3}}$	$\mathbf{0.672 \pm 1.1e^{-3}}$	58.81 ± 3.3	$\mathbf{0.359 \pm 3.5e^{-3}}$
`Gender` × `Skin Tone`	ITI-GEN	$166e^{-3} \pm 3.7e^{-3}$	$0.670 \pm 2.2e^{-3}$	59.56 ± 3.6	$0.463 \pm 7.7e^{-3}$
	Ours	$\mathbf{116e^{-3} \pm 4.4e^{-3}}$	$\mathbf{0.686 \pm 2.3e^{-3}}$	$\mathbf{54.66 \pm 2.7}$	$\mathbf{0.390 \pm 1.8e^{-3}}$

component for FairQueue *i.e.,* Prompt queuing, and Attention Scaling. Then we revisit the task initially proposed by ITI-GEN : Training Once-for-All Token. Overall, we show that FairQueue achieves new SOTA performance.

Experimental Setup. Following [16], we utilize the publicly available reference dataset from CelebA [29], FairFace [45] and FAIR benchmark [44]. For CelebA, we perform both single tSA and multi-tSA experiments. Note that in this dataset each tSA has two categories. In FairFace and FAIR datasets, tSAs have more categories, *e.g.,* `Age` and `Skin tones` contain 9 and 6 classes, respectively. Therefore, fair generation is more challenging in these datasets. For these experiments we use $\boldsymbol{T}$=E_t(`"A headshot of a person"`) as base prompt, and for a fair comparison, we utilize exactly the same learned $\boldsymbol{P}$ from ITI-GEN's original code [46] for both ITI-GEN and proposed FairQueue. In addition, we randomly sample a set of 500 latent codes and use the same latent codes for both approaches. As earlier discussed in Sec. 3, we utilize fairness discrepancy (FD) to evaluate fairness, Text-Alignment (TA) and FID for quality, and DreamSim (DS) to measure semantic preservation. See Supp for more details. We repeat this process 5 times and report the mean and standard deviation.

Our results in Tab. 1 demonstrate that FairQueue is able to match ITI-GEN fairness performance closely, and in some cases even improve upon it. For example, for tSA=`Smiling`, FairQueue indicates a significantly lower bias (FD=$6.9e^{-3}$) than ITI-GEN (FD=$124e^{-3}$). In addition, considering sample quality, FairQueue achieves an overall better performance than ITI-GEN for all datasets. For example, in CelebA, FairQueue's TA$\geq$ 0.666 while ITI-GEN's TA$\leq$ 0.655, with the worst performance with `High Cheekbones` (TA=0.595). These results are similarly reflected in FID. We remark that this quality degradation largely contributes to ITI-GEN fairness degradation. Finally, when considering semantic preservation (DS ↓) of the original sample generated with $\boldsymbol{T}$, FairQueue achieves unparalleled performance by ITI-GEN.

https://doi.org/10.52202/079017-0720

Ablation study: evaluating prompt queuing and attention scaling. To evaluate the contribution brought by FairQueue , we consider the same setup as Sec. 5 in the main manuscript focusing on the tSA `Smiling`. Here, we compare the performance when utilizing different attention amplification scaling factors, c, and different prompt queuing transition points *i.e.,* switching from $\boldsymbol{T}$ to $\boldsymbol{P}$. Specifically, we consider gradual increments in $c \in [0, 12]$ and shifting of the transition points, step $\in \{0, 0.1l, 0.2l, 0.3l\}$ when $l = 50$.

Our results in Fig.5 illustrate that generally when c increases, fairness improves. However, a saturation point ($c = 10$) exists where quality and semantic preservation beyond this point degrades. Then when considering different transition points, we find that at step $0.2l$ FairQueue achieves the best quality and semantic preservation performance while still achieving good fairness measurements. However, increasing beyond this point results in significant fairness degradation.

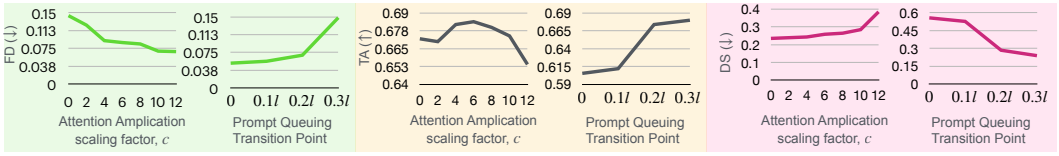


Figure 5: **Ablation Study:** Comparing FairQueue performance when varying i) attention amplification factor, c or ii) Prompt Queuing transition point from $\boldsymbol{T} \rightarrow \boldsymbol{P}$ for tSA `Smiling`.

Ablation study: revisiting training once-for-all token. Utilizing FairQueue we follow [16] and re-visit adapting pre-trained ITI-GEN tokens, S_i to a new Base Prompt $\boldsymbol{T'} = E_t$(`"A headshot of a doctor"`) by pre-pending. Then we generate samples utilizing both FairQueue and ITI-GEN with the same noise input. As seen in Fig. 6 FairQueue demonstrates better performance than ITI-GEN , achieving both better quality and semantic preservation of the sample generated by $\boldsymbol{T}'$ while still having good tSA representation—more illustration in Supp.

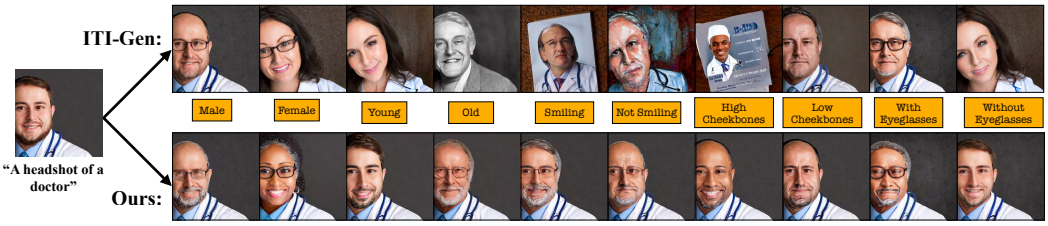


Figure 6: **Illustration of samples generated by ITI-GEN and FairQueue with a new Base Prompt $\boldsymbol{T'} = E_t$`{"A headshot of a doctor"}` via pre-pending.** FairQueue improves sample quality and ability to preserve the original sample's semantics while mainly adapting only the tSA.

6 Conclusion

In this paper, we reveal quality degradation in ITI-GEN –the existing SOTA fair T2I prompt learning approach. Our analysis reveals that this quality degradation is due to the distorted learned tokens in ITI-GEN prompt impacting cross-attention in the early steps of the denoising (reverse diffusion) process. To address this, we propose FairQueue a simple but effective solution consisting of: Prompt Queuing and Attention Amplification. Overall, our extensive experimentation demonstrates FairQueue achieves new SOTA performance in balancing quality, fairness, and semantic preservation. **Limitation, related work and additional experiments can be found in the Supp.**

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation, February 2021.

[3] 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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[4] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[5] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[6] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems*, 32, 2019.

[7] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[8] 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. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[9] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022.

[10] Alexander Quinn 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. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[11] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.

[12] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2023.

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

[14] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. *arXiv preprint arXiv:2312.01261*, 2023.

[15] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6957–6966, 2023.

[16] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023.

[17] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022.

 https://doi.org/10.52202/079017-0720

[18] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[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. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[20] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[23] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[26] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023.

[27] Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. On measuring fairness in generative models. *Advances in Neural Information Processing Systems*, 36, 2023.

[28] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[30] Christopher TH Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. Fair generative models via transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2429–2437, 2023.

[31] Christopher TH Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. Fairtl: A transfer learning approach for bias mitigation in deep generative models. *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[32] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. We're afraid language models aren't modeling ambiguity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=w3hL7wFgb3.

[33] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[35] 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 (TOG)*, 41(4): 1–13, 2022.

[36] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[37] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[38] 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. *Advances in neural information processing systems*, 30, 2017.

[39] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.310.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[43] Rafael C Gonzales, Richard E Woods, and Steven L Eddins. *Digital image processing using MATLAB*. Pearson Prentice Hall, 2004.

[44] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision*, 2022.

[45] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[46] Itigen. https://github.com/humansensinglab/ITI-GEN.

[47] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14367–14388, 2023.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[49] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[50] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[52] Midjourney. http://www.midjourney.com/. Accessed: 2024-01-11.

[53] 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 *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

 https://doi.org/10.52202/079017-0720

[54] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=AFDcYJKhND. Featured Certification.

[55] Eva Cetinic and James She. Understanding and creating art with ai: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–22, 2022.

[56] Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

[57] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2023.

[58] Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Magnet: Uniform sampling from deep generative network manifolds without retraining. In *International Conference on Learning Representations*, 2021.

[59] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[60] Milad Abdollahzadeh, Touba Malekzadeh, Christopher TH Teo, Keshigeyan Chandrasegaran, Guimeng Liu, and Ngai-Man Cheung. A survey on generative modeling with limited data, few shots, and zero shot. *arXiv preprint arXiv:2307.14397*, 2023.

[61] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[62] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

[63] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19840–19851, 2023.

Supplementary

This supplementary provides additional experiments as well as details that are required to reproduce our results. These were not included in the main paper due to space limitations. The supplementary is arranged as follows:

https://doi.org/10.52202/079017-0720

A More Experimental Results

A.1 Identifying FairQueue as the optimal combination

Table 2: Analysis of **all possible different combinations** for Prompt Queuing (PQ) and tSA Attention Amplification (AA). We summarize our findings from main paper for the tSA "Smiling". Note that $\alpha(S)$ notates AA for tSA tokens, ITI-GEN prompt P=[T;S], and results in bold and italics are the best and second best. Notice that C6:FairQueue (PQ+AA) provides the best combination: it achieves both outstanding sample quality (C6: TA=0.674 & FID=80.02 similar to C1: TA=0.681 & FID=76.9 with the best quality but poor fairness) and fairness (C6: FD=0.069 similar to C4: FD=0.05 with the best fairness but poor quality).

	Prompt Queuing (PQ)	Attention Amplification (AA)	Stage 1 (Prompt)	Stage 2 (Prompt)	FD(↓)	TA(↑)	FID(↓)	DS(↓)	remarks
C1: **no PQ, no AA** for Base Prompt	No	No	T	T	0.211	**0.681**	**76.9**	-	
C2: **no PQ, no AA** for ITI-Gen	No	No	[T;S]	[T;S]	0.124	0.605	88.63	0.557	
C3: **AA only** for Base Prompt	No	Yes	T	T	N.A	N.A	N.A	N.A	Unimplementable combination due to the absence of tSA tokens for AA
C4: **AA Only** for ITI-Gen	No	Yes	[T;S]	[T;$\alpha(S)$]	**0.05**	0.610	89.41	0.550	
C5: **PQ Only**	Yes	No	T	[T;S]	0.145	*0.674*	80.15	**0.240**	
C6: **PQ + AA** (Our proposed: FairQueue)	Yes	Yes	T	[T;$\alpha(S)$]	*0.069*	*0.674*	*80.02*	*0.284*	Both PQ and AA are present *i.e.,* FairQueue

In this section, we discuss in more detail how we identified FairQueue – with its two mechanisms: Attention Amplification and Prompt Queuing – as the best-performing solution. Specifically, we summarize our findings, discussed throughout the paper, when exhaustively considering all possible combinations. Our results are as follows:

- **C1: Base Prompt T Only** (no AA no PQ): It lacks tSA-specific knowledge and results in poor fairness. Additionally, without tSA tokens S, **AA is not applicable for C3.**
- **C2: ITI-Gen prompt P Only,** in Tab. 1 (no AA no PQ). Our analysis in Sec 3.2 shows it has poor quality due to distortion in global structure during sample generation. Without PQ, the issue of distorted global structure persists for some tSAs
- **C4: Attention Amplification (AA) Only**, in Fig. 5 when PQ transition point=0. It results in poor quality since only ITI-Gen is used. We remark that utilizing only AA for ITI-Gen may deceptively improve fairness, but the generated samples have poor quality e.g., Smiling cartoons. The reason is (similar to C2): without PQ, the issue of distorted global structure persists for some tSAs.
- **C5: Prompt Queuing (PQ) Only,** in Fig. 5 when $c = 0$. By replacing the distorted ITI-Gen prompt with the Base prompt in Stage 1, PQ leads to improved quality, but without AA, the fairness remains poor given reduced exposure to tSA tokens in the denoising process.
- **C6: FairQueue (PQ+AA)**, in Tab 1 Our proposed solution with optimal quality and fairness. Specifically, it combines the effects of Prompt Queuing– enabling the global structure to be properly formed resulting in good quality samples, and Attention Amplification–enhancing the tSA-specific expression for better fairness.

Overall, our results in Tab. 2 reveal that FairQueue (C6) is the superior combination balancing between fairness and quality. Specifically, Prompt Queuing is necessary whereby utilizing either only ITI-GEN (C2) or only Base Prompt (C1) results in quality and fairness degradation, respectively. Furthermore, our results show that both PQ and AA are necessary to obtain high-quality samples with good fairness performance, as without PQ (C4) sample quality is poor, and without AA (C5) fairness performance is degraded.

https://doi.org/10.52202/079017-0720

A.2 Cross-attention analysis

Sec. 3.2 analyzes the effect of inclusive tokens $\boldsymbol{S}^k$ by comparing the accumulated cross-attention maps of individual tokens between HP and ITI-GEN . It is observed that distorted tokens learned by ITI-GEN negatively affect the development of global structure in the early steps of denoising. The destructive effect arises with abnormally high activity of non-tSA tokens (e.g., ‘‘`of`’’), and the tSA-tokens attend to unrelated regions with scattered attention. A quantitative analysis is performed over 500 sample generations for different tSAs to affirm the observations. The below details how token-specific accumulated cross-attention maps are obtained from the SD pipeline, discusses interaction among tokens, and presents additional representative results for tSAs `Smiling`, `High Cheekbones`, `Gray Hair`, and `Chubby`.

Details of visualizing token-specific accumulated cross-attention map. Cross-attention is often used to contextualize prompt embeddings with latent representations per sample generation step. Following DAAM [41], coordinate-aware attention scores $\boldsymbol{M}[\boldsymbol{S}_i]$ are extracted from the latent diffusion network (i.e., U-Net) for the token $\boldsymbol{S}_i$ at the layers where cross-attentions take place. These token-specific attention scores, each with the same spatial dimensions as the latent representation, are upscaled bicubically to the image size (512×512 in this case) to reveal where attention is paid per token and accumulated within the assigned step(s). The resulting 2D matrix is visualized in Fig. 3 and Fig. 4 and referred to as an “accumulated cross-attention map”.

Interaction among tokens. We remark that the cross-attention map of a given token is dependent on the others in the prompt. There are two channels where the effect of tokens may interact: 1) via latent representation, as it is a function of input tokens and serves as the query in the cross-attention (see Sec.2); 2) softmax operation, as a component in the attention pipeline, softmax is taken across all tokens when processing attention scores. These two effects become increasingly apparent as we move through different cross-attention layers of the U-Net and perform more denoising steps.

HP vs. ITI-GEN : qualitative analysis. To investigate potential abnormality of ITI-GEN embeddings, images of different tSA are generated conditioning on HP ($\boldsymbol{F}$) and ITI-GEN ($\boldsymbol{P}$) respectively. The cross-attention map is employed as a tool to explore the cause of degraded generations. In pursuit of a fair token-to-token comparison, for some tSAs the original HPs (“HP1”, see Tab. 6) are extended to align with $\boldsymbol{P}$ in the number of tSA tokens (“HP2”). Nonetheless, as one can find in the samples in Fig.7 to 14, the extension does not change the behavior of HPs significantly.

Fig.7 to 17 give an overview of cross-attention maps during the denoising process. One may find that the tSA tokens in the HP(s) tend to concentrate on the region(s) semantically associated with the tSA, e.g, mouth for tSA `Smiling`, cheek for tSA `High Cheekbones`, and hair for tSA `Gray Hair`. On the other hand, ITI-GEN tSA tokens' activity tends to be less focused and attends broadly. With more steps than Fig. 3, it is clearer that the global structure of the images is synthesized in the early steps, which motivates the prompt switching experiments.

Prompt switching experiments and quantitative analysis. To further investigate HP and ITI-GEN prompts' behaviors in the early steps, the prompt switching experiments (i.e., I2H and H2I) are proposed in Sec. 3.2. Fig.18 to 21 present representative outcomes of the experiments. One can find that the destructive effect caused by ITI-GEN prompts only occurs at the early steps, i.e., Stage 1 in the figures.

In addition, the activation patterns are more clear in the accumulated cross-attention maps. The non-tSA tokens in ITI-GEN prompts are in general more active, and the tSA tokens tend to attend more broadly, which may explain the drastic semantic deviations from HPs in Fig.7 to 17. The latter observation is particularly evident for tSA `Smiling`, a highly localized facial expression, which is supported by the histogram of central moments in Fig.1. The other tSAs, though may not be directly associated with a specific facial feature, share the same trend, as manifested statistically by the histograms in Fig.22 and Fig.23.

Utilizing Base Prompt (T) in FairQueue . In Prompt Queuing the use of T, in place of the HP, is similarly grounded on the I2H/H2I analysis, as both T and HP are natural language prompts – free of learned tokens. This can be seen in the embedding analysis in Supp B.6 where the HP and T are seen to be close to one another. As a result, the sample generated by T is expected to be of similar quality as the HP.

https://doi.org/10.52202/079017-0720

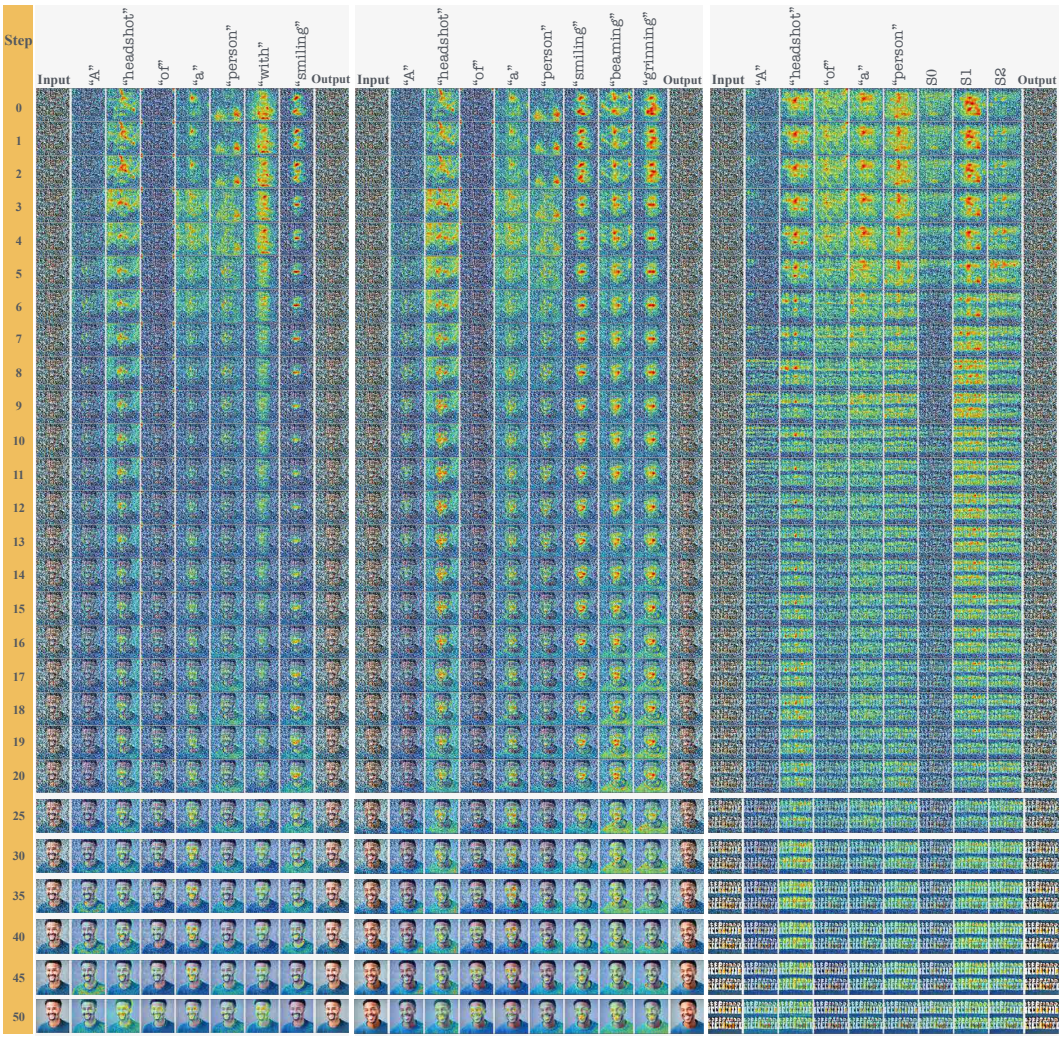


Figure 7: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`Smiling`.

In Fig. 24, we provide further visualizations of T's effectiveness in generating the global structure in early denoising steps. Specifically, we compare the cross-attention maps of FairQueue with ITI-Gen during sample generation, together with quantitative analysis. Results in col 2 vs 3 illustrate T's effectiveness in synthesizing the global structure in stage 1, and non-abnormal attention (in Fig. 25), resulting in effective global synthesis than ITI-Gen and better sample quality.

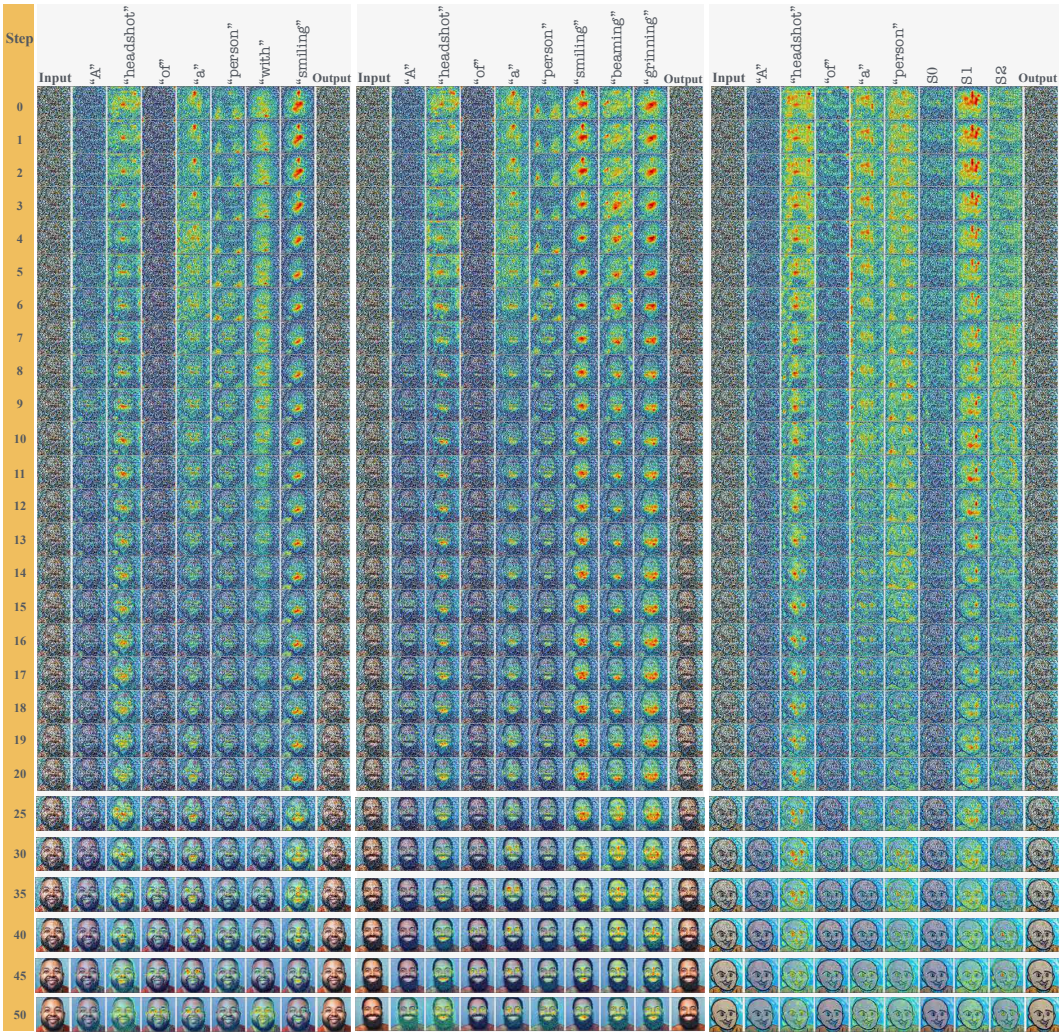


Figure 8: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`Smiling`.

https://doi.org/10.52202/079017-0720

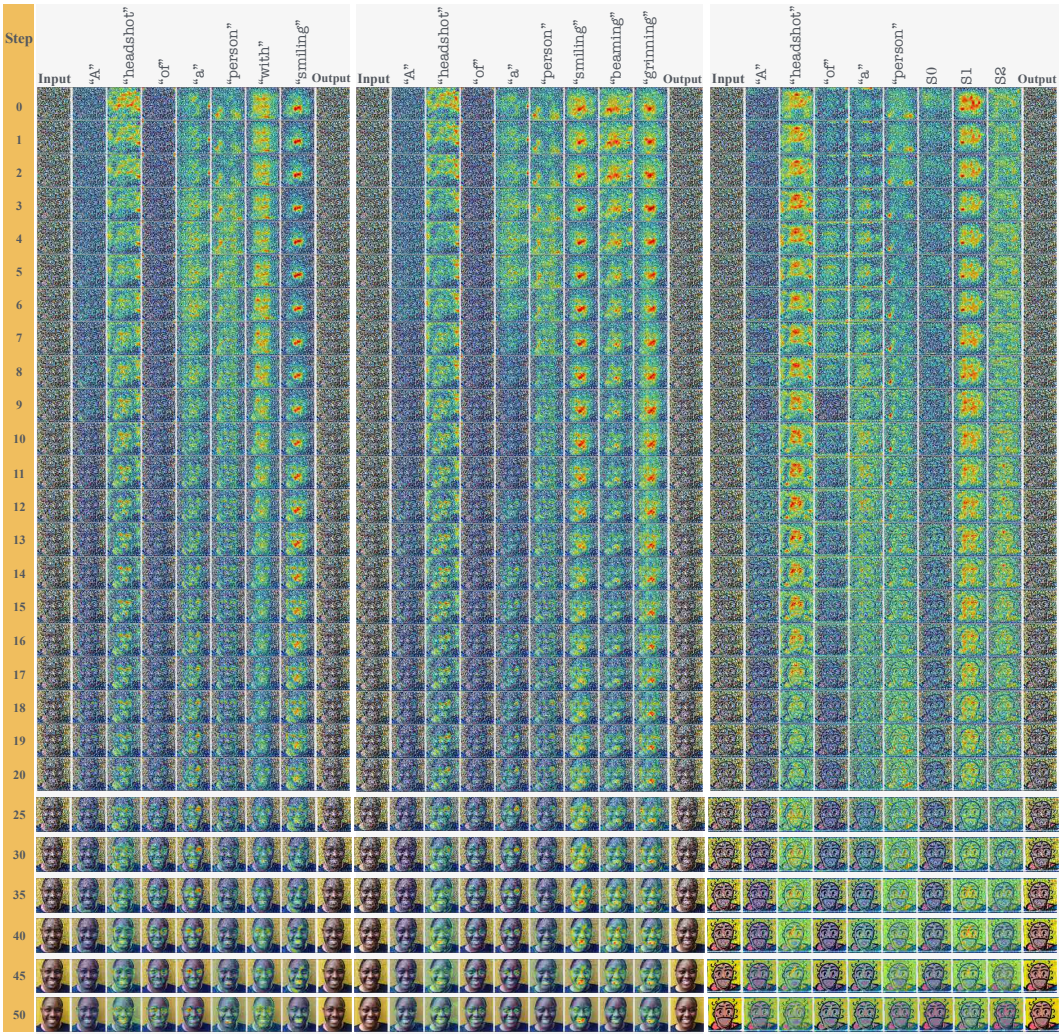


Figure 9: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`Smiling`.

https://doi.org/10.52202/079017-0720

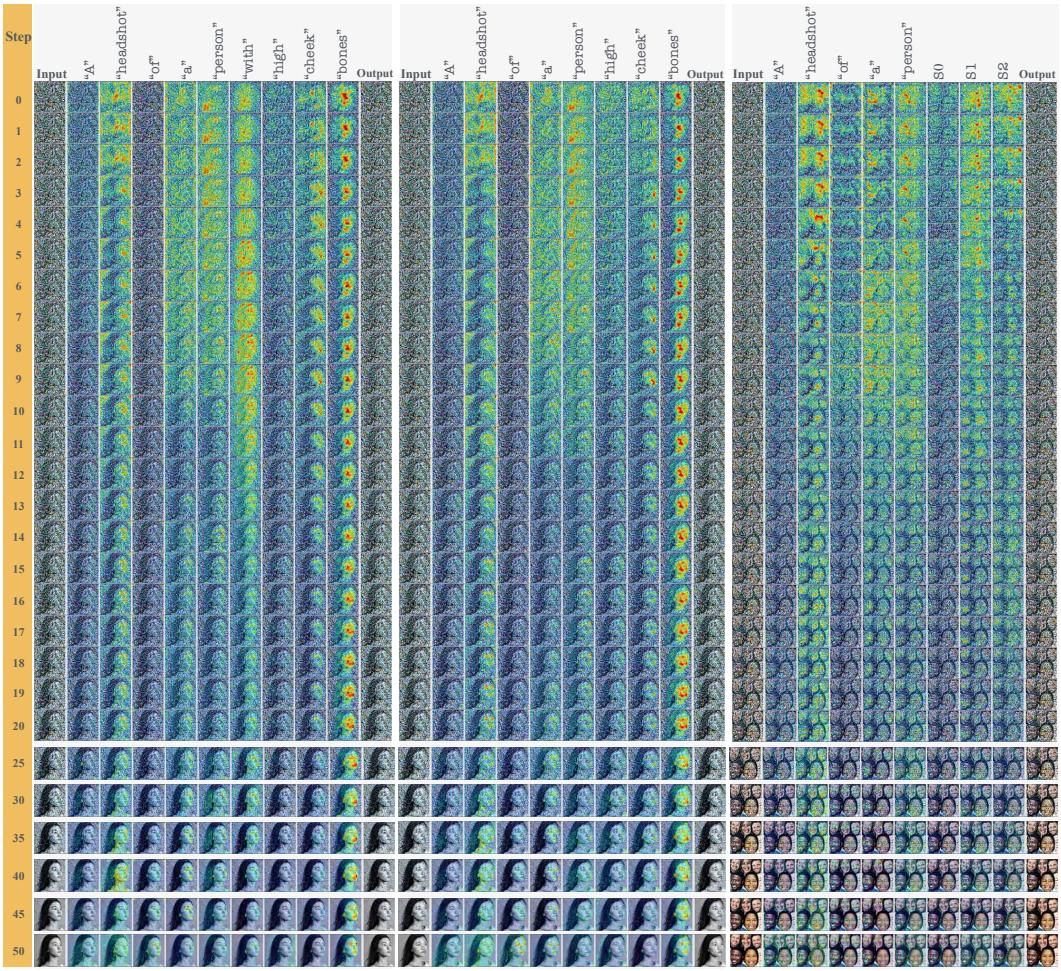


Figure 10: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`High Cheekbones`.

https://doi.org/10.52202/079017-0720

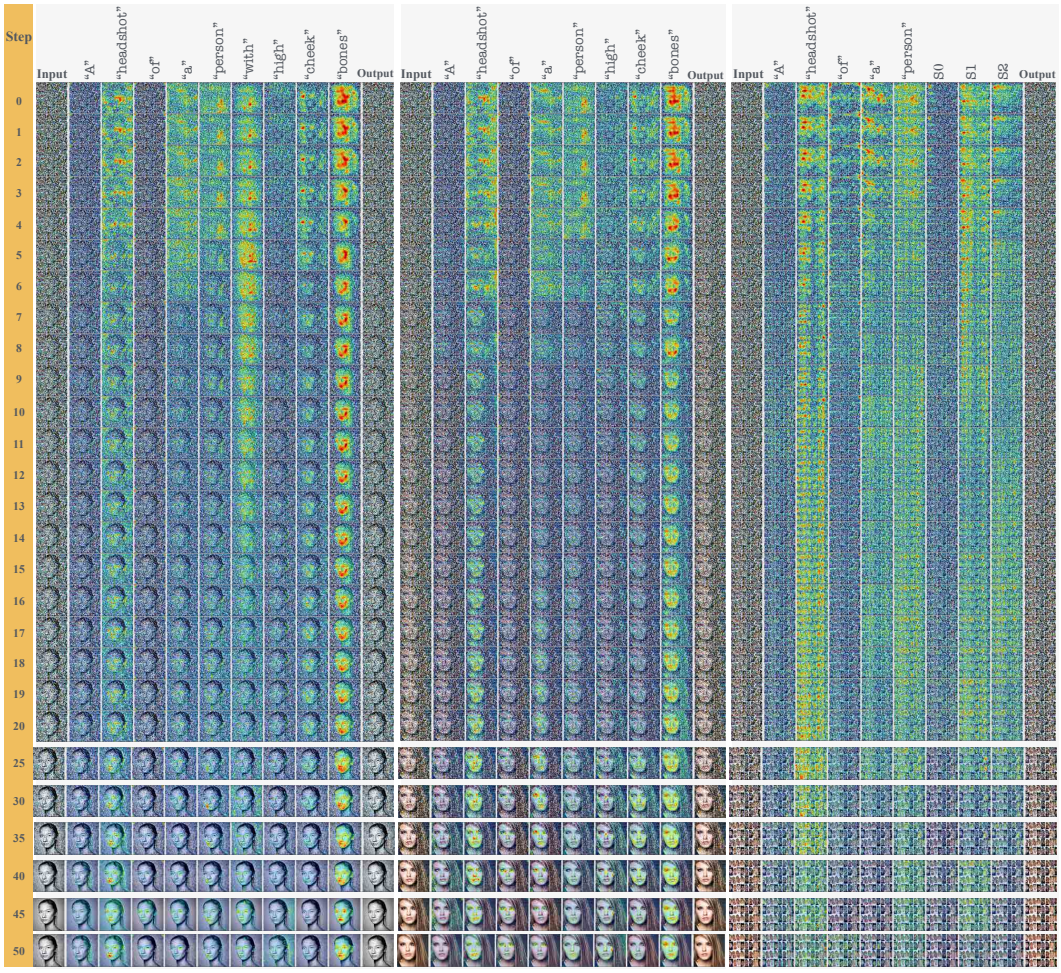


Figure 11: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`High Cheekbones`.

https://doi.org/10.52202/079017-0720

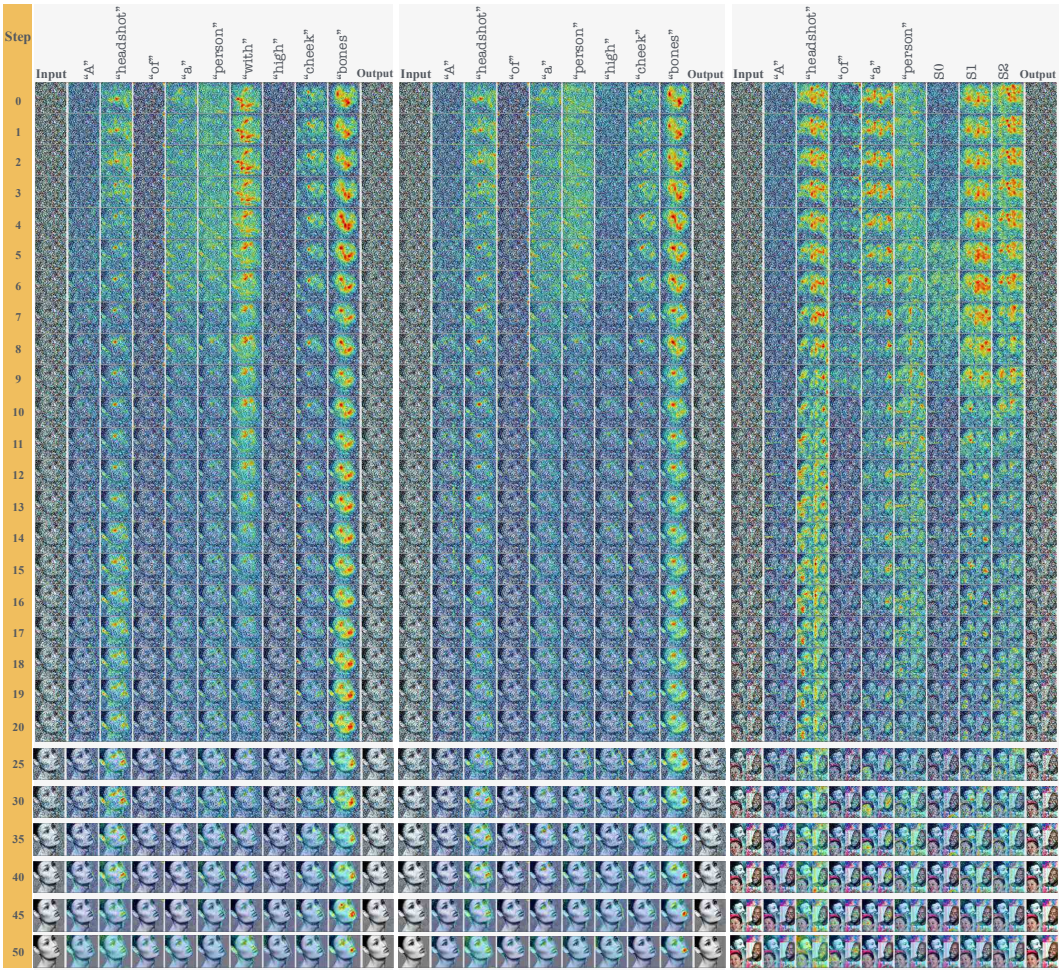


Figure 12: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=`High Cheekbones`.

https://doi.org/10.52202/079017-0720

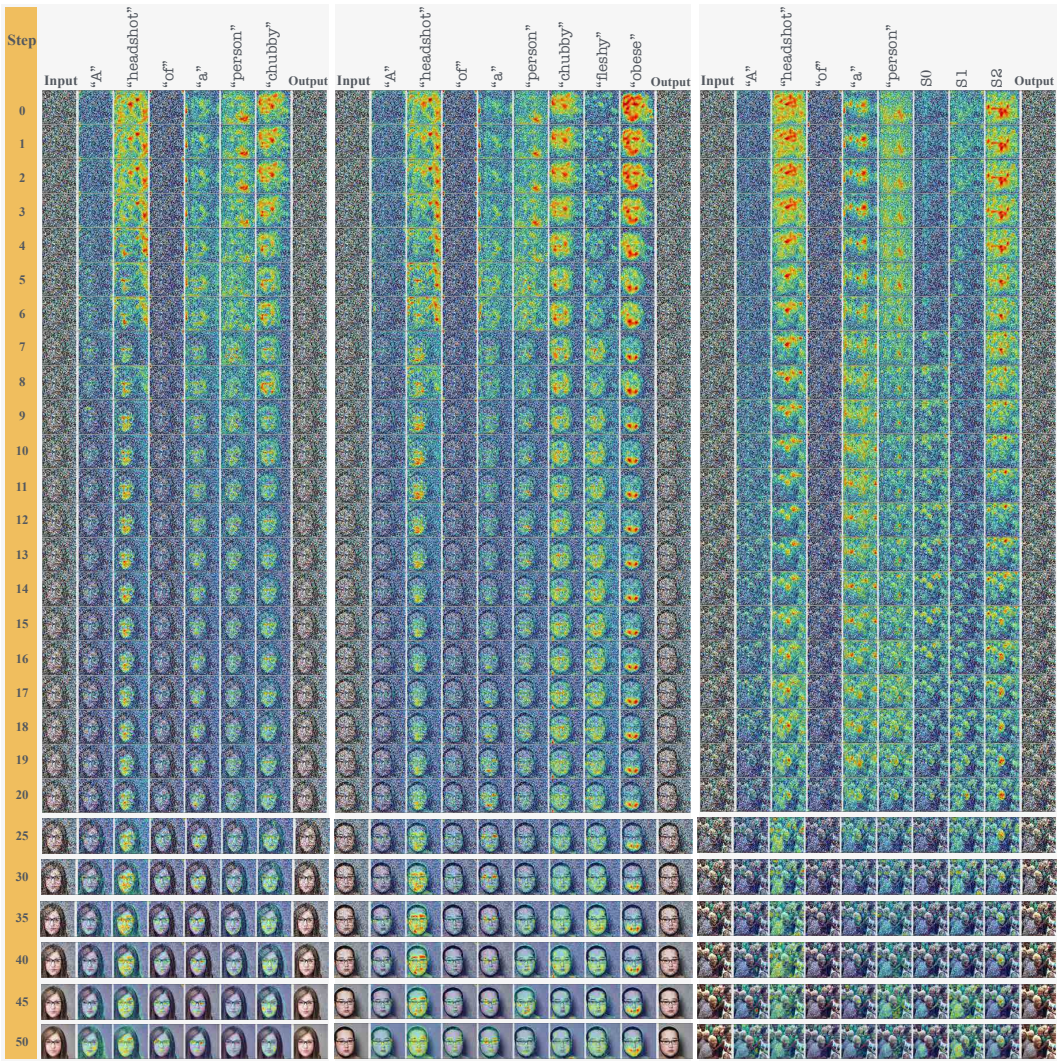


Figure 13: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=Chubby.

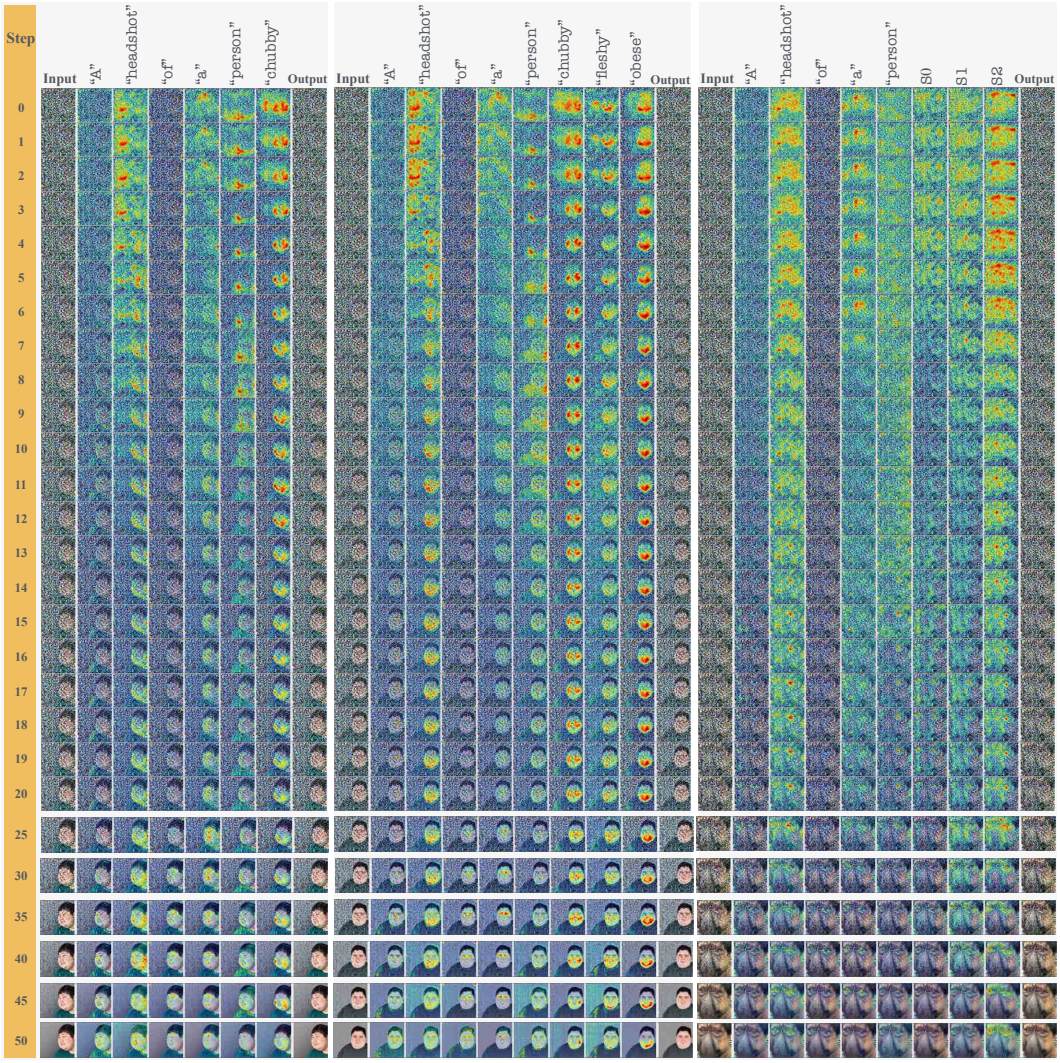


Figure 14: **Cross-attention maps during the denoising process with HP1 (left), HP2 (middle, equal #tSA tokens to ITI-GEN), and ITI-GEN (right) prompts.** tSA=Chubby.

https://doi.org/10.52202/079017-0720

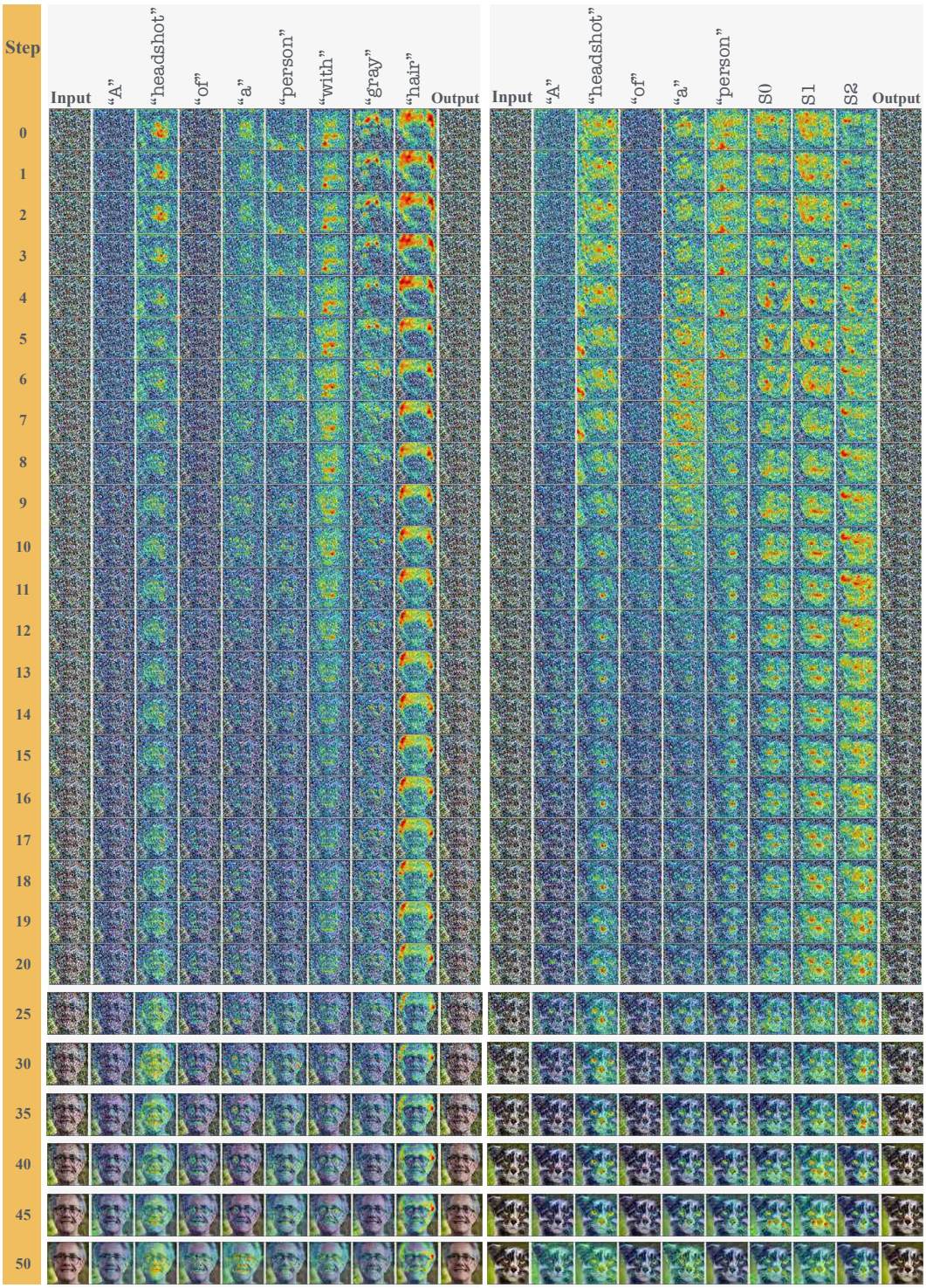


Figure 15: **Cross-attention maps during the denoising process with HP (left) and ITI-GEN (right) prompts.** tSA=`Gray Hair`.

https://doi.org/10.52202/079017-0720

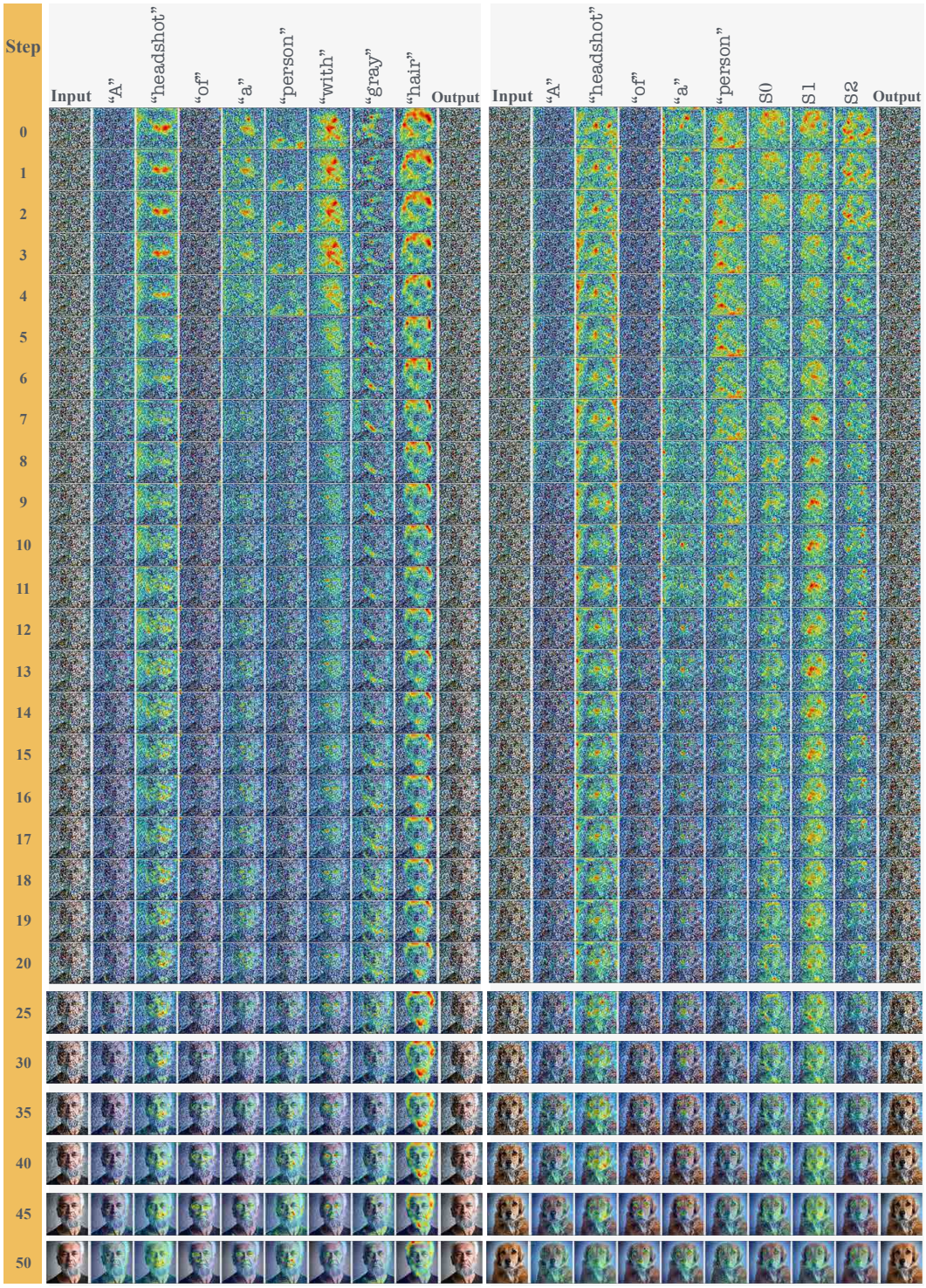


Figure 16: **Cross-attention maps during the denoising process with HP (left) and ITI-GEN (right) prompts.** tSA=`Gray Hair`.

https://doi.org/10.52202/079017-0720

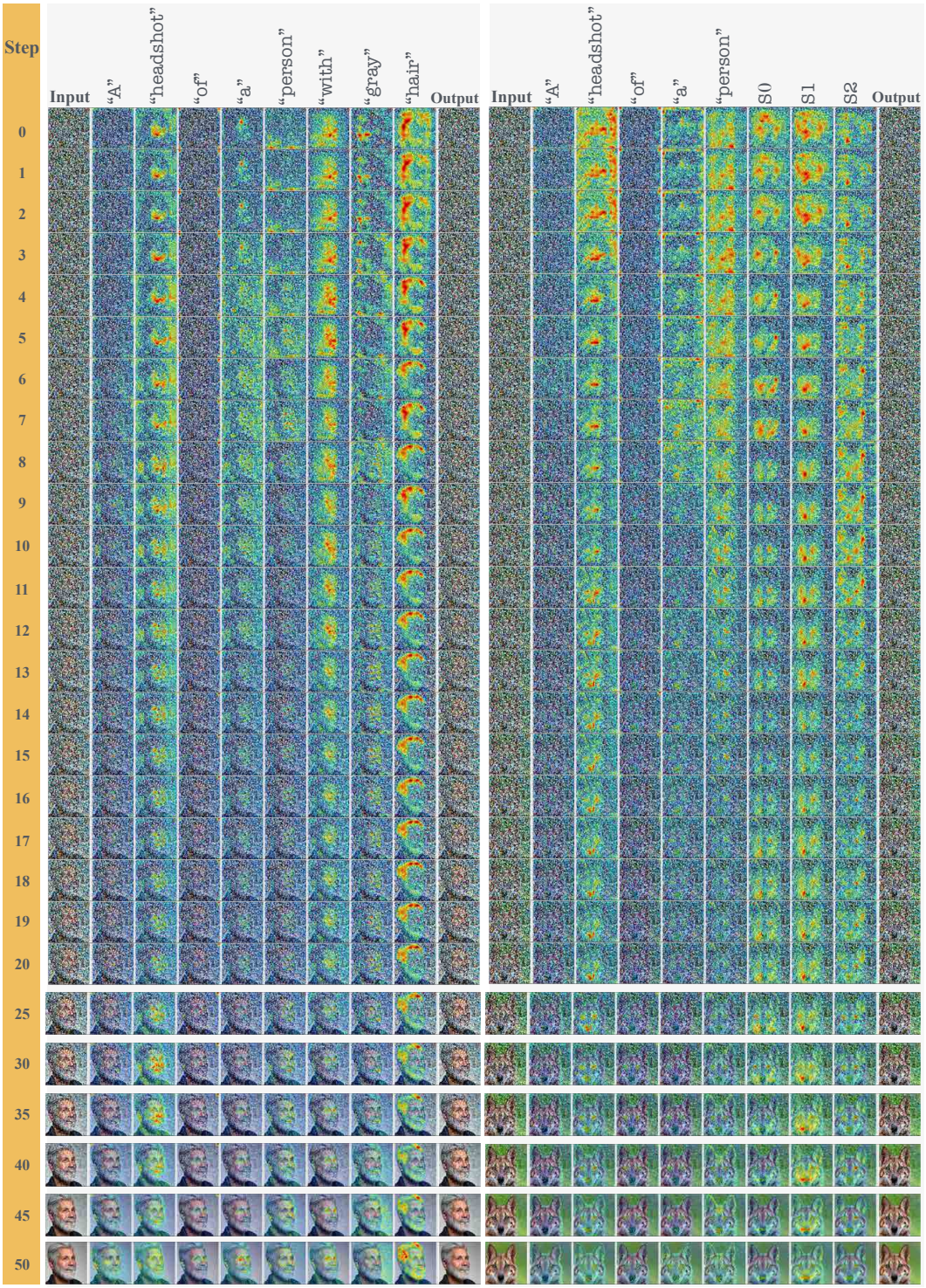


Figure 17: **Cross-attention maps during the denoising process with HP (left) and ITI-GEN (right) prompts.** tSA=`Gray Hair`.

Figure 18: **Accumulated cross-attention maps for the denoising process in our proposed prompt switching experiments I2H and H2I: tSA =** `Smiling`**.**

https://doi.org/10.52202/079017-0720

Figure 19: **Accumulated cross-attention maps for the denoising process in our proposed prompt switching experiments I2H and H2I: tSA =** `High Cheekbones`**.**

Figure 20: **Accumulated cross-attention maps for the denoising process in our proposed prompt switching experiments I2H and H2I: tSA =** `Chubby`**.**

https://doi.org/10.52202/079017-0720

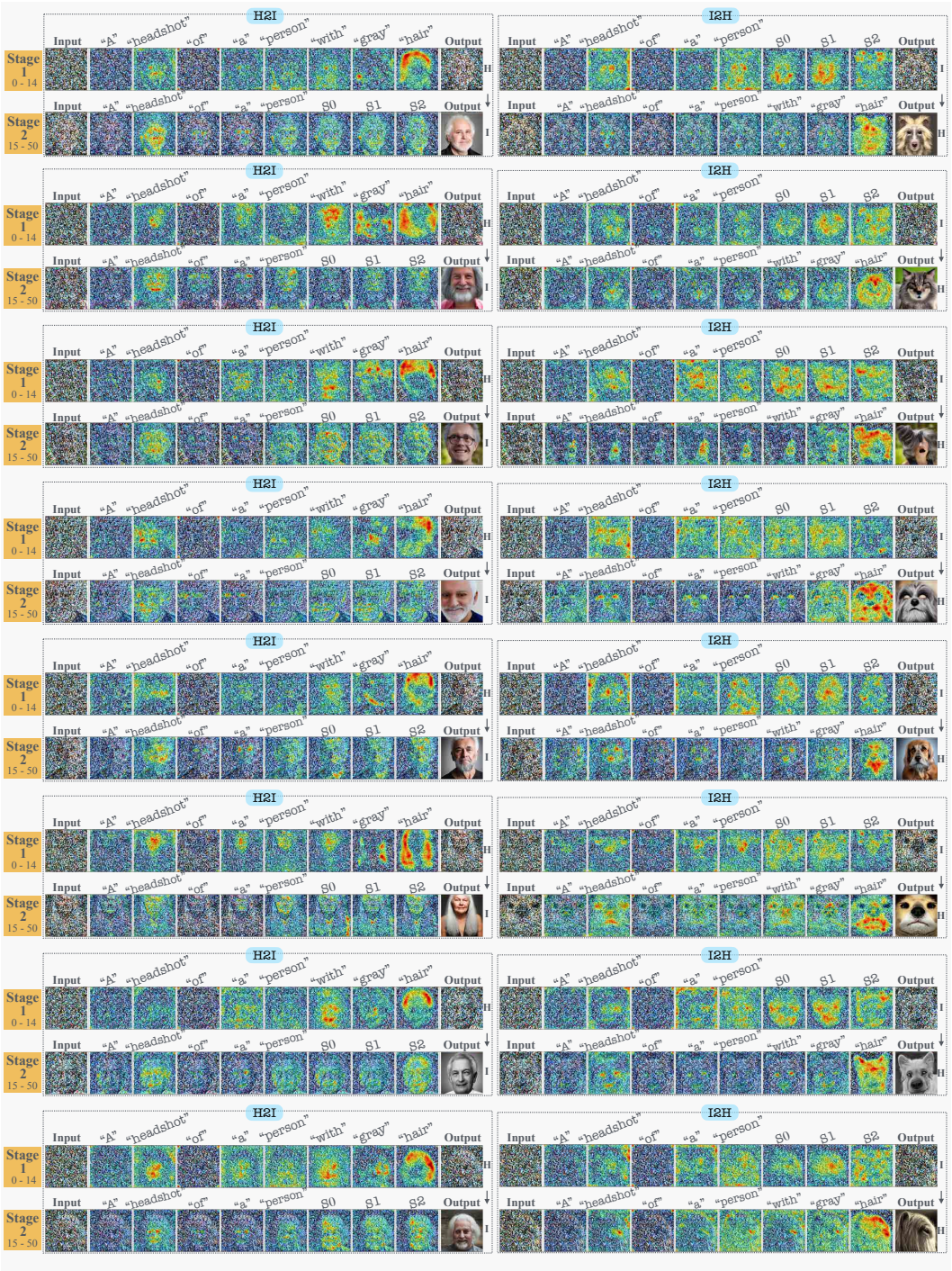


Figure 21: **Accumulated cross-attention maps for the denoising process in our proposed prompt switching experiments I2H and H2I: tSA =** `Gray Hair`**.**

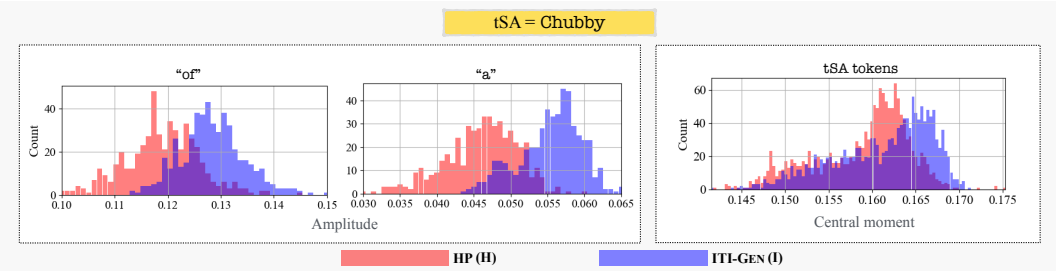


Figure 22: **Histograms for cross-attention analysis in prompt switching experiments I2H and H2I: tSA = `Chubby`.**

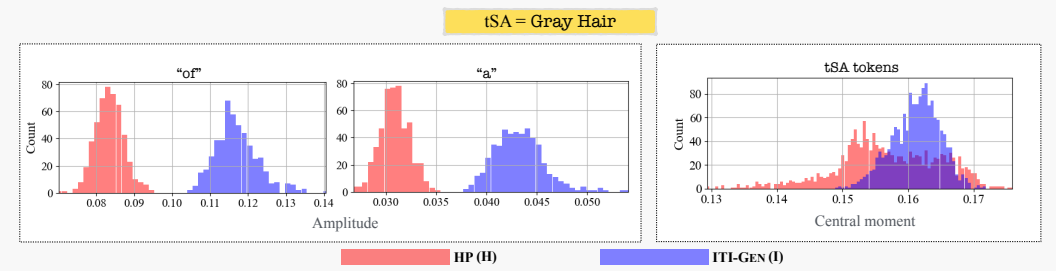


Figure 23: **Histograms for cross-attention analysis in prompt switching experiments I2H and H2I: tSA = `Gray Hair`.**

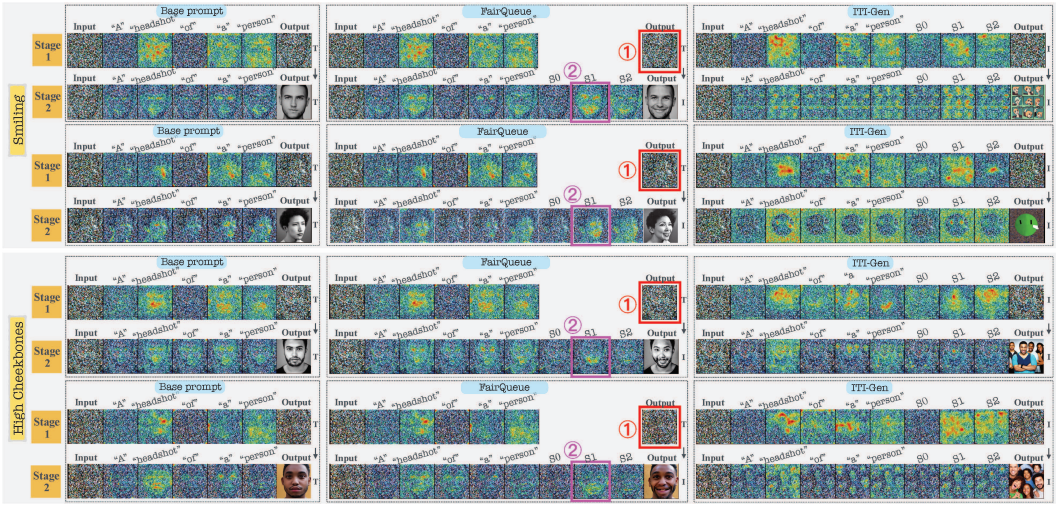


Figure 24: Accumulated cross-attention maps for Base prompt, FairQueuing and ITI-Gen, tSA=Smiling, High Cheekbones. Same setup as Sec 3.2 of the main manuscript (e.g., Fig.4). The mid column presents FairQueue, with Base prompt (T) in Stage 1 and ITI-Gen (I) in Stage 2; the left column is only based on Base prompt and the right is only based on ITI-Gen. Note Base prompt behaves similarly to the HP in forming good global structures in the first stage (annotated by red frames); in the second stage, the tSA token of ITI-Gen can attend to tSA-related regions (e.g., eyes and mouth for Smiling, or the lower half of the face and cheeks for High Cheekbones, see magenta frames) and enhance associated facial features.

https://doi.org/10.52202/079017-0720

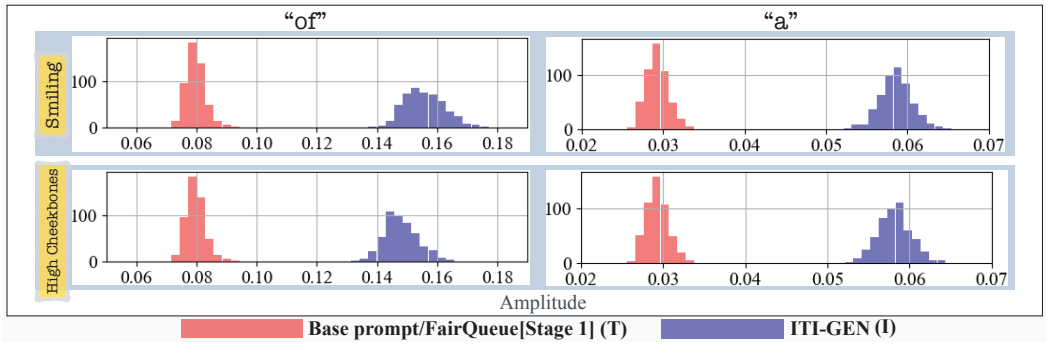


Figure 25: Histograms for non-tSA tokens in the first stages, tSA=Smiling and High Cheekbones for Fig. 24.

https://doi.org/10.52202/079017-0720

A.3 More on Ablation Studies

A.3.1 Analyzing the Effects of Attention Amplification

In this section, we provide more illustrations for the ablation study. Our results in Fig. 26, illustrate the effect of attention application with different scaling factors (c). Notice that at $c = 0$ the tSA expression may still be lacking but increasing c results in emphasized tSA expression.

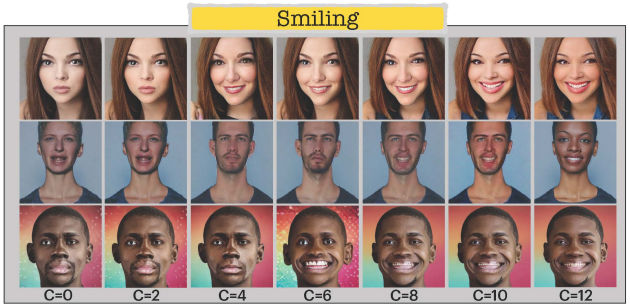
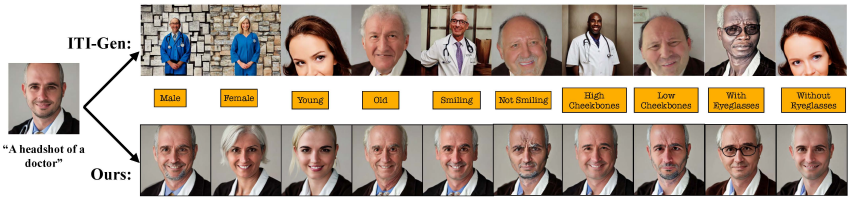


Figure 26: Illustration of FairQueue samples utilizing different attention amplification scaling factors (c), tSA=Smiling. In each row, we utilize the same seed and prompt queuing transition point.

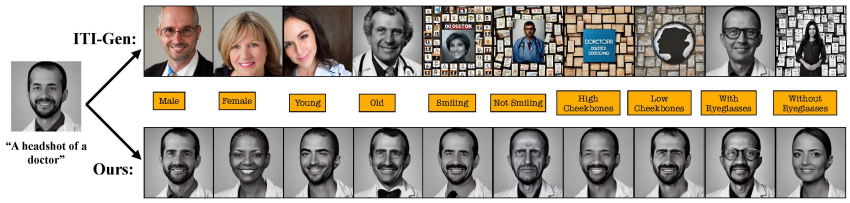
A.3.2 More Illustrations for Training Once-fo-All Tokens

In Fig. 27, we provide more illustrations for the analysis on Revisiting Training Once-for-All Token.

(a) Random Sample 2



(b) Random Sample 3



(c) Random Sample 4

Figure 27: **Illustration of samples generated by ITI-GEN and FairQueue with a new Base Prompt T' = E_t{"A headshot of a doctor"} via pre-pending.** Notice that FairQueue improved sample quality and ability to preserve the original sample's semantics while mainly adapting only the tSA.

https://doi.org/10.52202/079017-0720

A.3.3 Human-recognized Assessment Comparing ITI-Gen and FairQueue Quality

In this section, we carried out a human user study to evaluate the sample quality and fairness of the generated sample from FairQueue against ITI-GEN . Specifically, we utilize the same seed to generate 100 sample pairs with ITI-Gen and FairQueue for { `Smiling`, `High Cheekbones`, `Gender`, and `Young` }. Then, utilizing Amazon Mechanical Turk we conduct 2 tasks:

- **Quality comparison by A/B testing:** Human labelers select the better quality sample between ITI-Gen and FairQueue (from the same seed). Each sample was given to 3 labelers.
- **Fairness comparison by human-recognized tSA:** labelers identified the tSA class for each sample. The final label was based on the majority of 3 labelers. labelers were also given an "unidentifiable" option if the class could not be determined. Finally, the labels were used to measure FD.

Our results in Tab. 3 reveal that FairQueue generates better quality samples than ITI-Gen (>62.0% preference) and Tab. 4 shows that FairQueue achieves competitive fairness with ITI-Gen. Overall, this aligns with our quantitative results in Tab. 1.

Table 3: A/B testing: Human assessment comparing quality between ITI-Gen vs FairQueue for 200 samples per tSA. Col 2 and 3 indicate the percentage of labelers that prefer the method's sample quality. A larger value is better.

	ITI-GEN	FairQueue
`Smiling`	1.3%	98.7%
`High Cheekbones`	2.7%	97.3%
`Gender`	33.0%	67.0%
`Young`	38.0%	62.0%

Table 4: Fairness comparison by human-recognized tSA: Human assessment to compare FD for ITI-Gen vs FairQueue for 200 samples per tSA.

	ITI-GEN FD(↓)	FairQueue FD(↓)
`Smiling`	0.106	0.014
`High Cheekbones`	0.144	0.021
`Gender`	0.014	0.014
`Young`	0.014	0.028

https://doi.org/10.52202/079017-0720

A.4 More Illustration

In this section, we provide more samples generated by FairQueue based on the setup in Sec. 5. Recall that here we utilize the base prompt $\boldsymbol{T} = E_T$''`A headshot of a person`'' and consider the tSA$\in$ {`Male`, `Young`, `Smiling`, `Low Cheekbones`, `Pale Skin`, `Eyeglasses`, `Mustache`}. Each sample is then generated based on the same 10 fixed noise inputs.

Figure 28: Base-Prompt images $\boldsymbol{T}$ with fixed latent noise input

Figure 29: FairQueue with tSA=`Female` with fixed latent noise input

Figure 30: FairQueue with tSA=`Male` with fixed latent noise input

Figure 31: FairQueue with tSA=`Old` with fixed latent noise input

Figure 32: FairQueue with tSA=`Young` with fixed latent noise input

Figure 33: FairQueue with tSA=`not Smiling` with fixed latent noise input

Figure 34: FairQueue with tSA=`Smiling` with fixed latent noise input

https://doi.org/10.52202/079017-0720

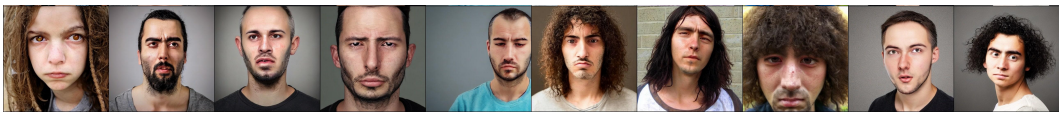


Figure 35: FairQueue with tSA=`Low Cheekbones` with fixed latent noise input

Figure 36: FairQueue with tSA=`High Cheekbones` with fixed latent noise input

Figure 37: FairQueue with tSA=`not Pale Skin` with fixed latent noise input

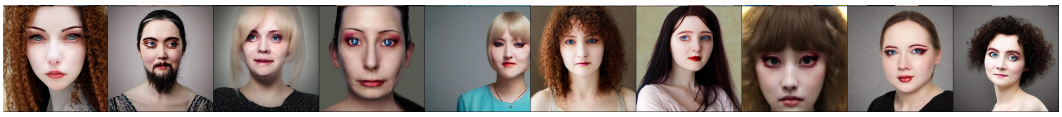


Figure 38: FairQueue with tSA=`Pale Skin` with fixed latent noise input

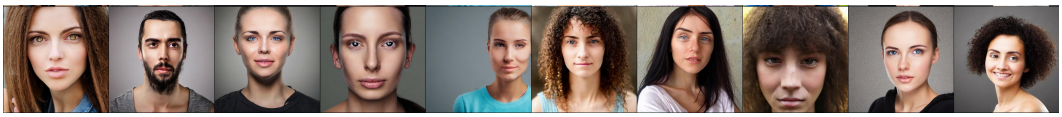


Figure 39: FairQueue with tSA=`no Eyeglasses` with fixed latent noise input

Figure 40: FairQueue with tSA=`with Eyeglasses` with fixed latent noise input

Figure 41: FairQueue with tSA=`no Mustache` with fixed latent noise input

Figure 42: FairQueue with tSA=`with Mustache` with fixed latent noise input

A.5 Evaluating Minimal Linguistic Ambiguity for tSA

In this section, we provide an experiment to search for tSA with minimal linguistic ambiguity. To do this, we utilize the tSA and their respective HPs in Tab. 6. Then, with these HPs, we generate 500 samples per category of the individual tSA. Finally, we classify the respective samples and evaluate the accuracy of the T2I model in generating samples with the respective categories of the tSA. Our results in Tab. 5 report the accuracy of the T2I generative model in accurately

interpreting the HPs to generate the respective tSAs. Based on our results, we can determine that the tSA$\in \{$`Male`, `Young`, `Smiling`, `High Cheekbones`$\}$ are with minimal linguistic ambiguity, as their HPs can be easily interpreted by the T2I generator as seen by the high accuracy.

Table 5: Accuracy of the T2I model in generating the tSA with the HP. See Tab. 6 for list on all the HPs.

tSA	Positive Prompts	Negative Prompts
`Male`	99.8%	99.2%
`Young`	97.8%	98.4%
`Smiling`	98.2%	97.8%
`High Cheekbones`	96.6%	98.8%
`Pale Skin`	91.3%	7.8%
`Eyeglasses`	97.4%	2.4%
`Mustache`	90.4%	13.6%
`Gray Hair`	97.2%	22.4%
`Chubby`	98.2%	30.4 %

B Experimental Details

B.1 Details of Calculating Directional Loss for Prompt Tuning

In this section, we provide more details on the directional loss, $\mathcal{L}_{dir}$ utilized by ITI-GEN . Recall that in ITI-GEN the direction loss takes in two components: 1) direction of Image embeddings, $\Delta \boldsymbol{I}$, and 2) direction of ITI-GEN token embedding, $\Delta \boldsymbol{P}$. For ease of discussion, we utilize a binary tSA, but the same concept can be simply scaled to multi-class tSA.

Specifically, we first measure $\Delta \boldsymbol{I}$ where ITI-GEN first evaluates the mean image embedding for each category of the tSA *i.e.,* $\alpha = \mathbb{E}_k \left[E_I(x^k)\right]$, where E_I is a CLIP Image encoder [20] and x^k are the reference image for k samples in a given mini-batch. Then, we calculate the directional Image embeddings: $\Delta \boldsymbol{I}_{i,j} = \alpha_i - \alpha_j$ where i and j are the different categories of the selected tSA. Then when considering direction of ITI-GEN token embedding we simply calculate $\Delta \boldsymbol{P} = E_T(\boldsymbol{P}_i) - E_T(\boldsymbol{P}_j)$, where E_T is the CLIP text encoder.

B.2 Details of the Ambiguities in Text Prompts

Ambiguities arise because of the potentially multiple interpretations of the same utterance. Ambiguity is a well-known concept in Large Language Models (LLMs), and recently there have been some attempts to understand the impact of these ambiguities in the intersection of the LLM and image generation models, a.k.a T2I models. Among different types of ambiguity, three major types can affect the quality of the T2I generation [47]:

- *Syntax:* where there could be different interpretations of the same text. As an example, in input prompt ‘‘`the mouse looks at the cat standing on rug`’’, it is not clear whether the mouse is standing on the rug or the cat.
- *Semantics:* where the words within a text have multiple meanings. As an example, ‘‘`a photo of a bat`’’, it is not clear whether it refers to a nocturnal flying mammal or the flat wooden club used in sports like baseball or cricket (with ‘cricket’ itself having two different meanings ;)).
- *underspecification* where the used text prompt can not completely describe the required attributes in the image. For example, ‘‘`a woman with light hair`’’ can not specifically determine the category of the hair color.

Within the context of this paper, ambiguity is generally used with the Hard Prompts (HP) where we append the description of a specific tSA to the base prompt. For example, considering the base prompt ‘‘`a headshot of a person`’’, and considering the tSA=`Smiling`, one can create the HP as ‘‘`a headshot of a person smiling`’’, and ‘‘`a headshot of a person not smiling`’’. In our paper, we empirically realize that for some tSA, creating HP like this will result in a strong baseline for fair T2I generation, observed by our quality and fairness metrics. We call these, tSAs with minimum linguistic ambiguity, and as an example, we empirically found that tSAs like `Young`, `Smiling`, `Gender` fall within this category. Due to their superior performance, HP with these tSAs is used as our pseudo-gold standard in the analysis of the quality degradation issue in ITI-GEN is

https://doi.org/10.52202/079017-0720

Sec. 3. However, other tSAs suffer ambiguity in either finding proper terms to define them or inferior performance in terms of the quality or fairness of the generation. This enforces using more advanced techniques like prompt learning to overcome these ambiguities.

B.3 Details of Model Hyper Parameters

Models and training hyper-parameters. In our experiments, we follow the same setup as ITI-GEN [16] and utilize Stable Diffusion v1.4 [1] as our T2I generator. Then when implementing ITI-GEN we utilize an $\boldsymbol{S}_i$ with a token length of 3 per tSA which is optimized based on a learning rate of $lr = 0.01$. For reference datasets, we utilize [16] readily available datasets with contain 200 reference images per category of each tSA. An Adam [48] optimizer is utilized during prompt learning. For sample generation, we follow the recommended diffusion steps of $l = 50$ and utilize an Attention scale of $c = 10$ and an Attention Queuing transitioning step =10.

Hard Prompt. Tab. 6 illustrates the list of HPs utilized in our experiments. We remark that not all tSAs have a clear HP. For example, with the tSA=`Skin Colour` it is difficult to find a reasonable HP to describe an individual with a specific skin tone.

Table 6: Hard Prompts utilized in T2I generation.

tSA	Positive Prompts	Negative Prompts
Male	"A headshot of a person Male"	"A headshot of a person Female"
Young	"A headshot of a person Young"	"A headshot of a person Old"
Smiling	"A headshot of a person with Smiling"	"A headshot of a person with no/without Smiling"
High Cheekbones	"A headshot of a person with high cheekbones"	"A headshot of a person with low cheekbones"
Pale Skin	"A headshot of a person with pale skin"	"A headshot of a person with no/without pale skin "
Eyeglasses	"A headshot of a person with eyeglasses"	"A headshot of a person with no/without eyeglasses"
Mustache	"A headshot of a person with mustache"	"A headshot of a person without mustache"
Gray Hair	"A headshot of a person with gray hair"	"A headshot of a person without gray hair"
Chubby	"A headshot of a person chubby"	"A headshot of a person no chubby"

Stable Diffusion Version. We verified that the problems of poor performance with Hard Prompts as indicated by Zhang et al. [16] would still persist even in more recent versions of Stable Diffusion e.g., SD 3.0. This issue can be observed by simply inputting the prompt "A headshot of a person without glasses" to SD 3.0 where the generated samples still frequently have the wrong tSA class ("with glasses") indicating this ambiguity still exists. For example, when utilizing the same setup as Tab 1 with HP (from Tab. 6) and SD3.0, our results show poor fairness performance for `Eyeglasses` (FD=$670e^{-3}$) and `Pale Skin` (FD=$580e^{-3}$).

Increasing the size of the reference dataset (2k reference images per category per tSA). We verified that increasing the size of the reference dataset does not resolve ITI-GEN's quality degradation. Specifically, we repeated the experiment in Tab. 1 with tSA Smiling with 2k reference samples per class. Our results measured an FD=$127e^{-3}$, TA=0.591, FID=89.2, and DS=0.532 which is similar to the results in Tab. 1 (based on 200 reference images). This indicates that the core problem may not be the data size, and may need specific data curation (including sample pairs with only semantic differences in tSA, and similar semantics elsewhere), which poses scalability and applicability challenges.

B.4 Computation Resources

Tab. 7 illustrates the amount of the compute including the GPU Hours for different steps of our research and the estimated carbon emission for our experiments.

Table 7: **Estimated Computation time**. The carbon emission values are computed using `https://mlco2.github.io/impact`.

Experiment	Hardware	GPU Hours	Carbon emitted (kg)
T2I Sample Generation	RTX3090	25.0	4.87
Directional Alignment analysis	RTX3090	0.25	0.048
Cross Attention Analysis	RTX3090	1.0	0.195
Prompt Learning	RTX3090	1.0	0.195
Total:		27.25	5.308

B.5 Details of the Evaluation Metrics

In this section, we provide more detail on the evaluation metrics used in our work.

Fairness. We use the *fairness discrepancy* (FD) metric which compares tSA distribution in generated images with an ideal uniform distribution [27]. In this metric FD=0 would indicate perfect fairness. Following [16, 11, 36] we use a combination of CLIP [20] models, human evaluators, and off-the-shelf models [3] for predicting the distribution of the tSA in target images. We remark that to evaluate FD we utilize classifiers that achieve reasonably high accuracy when evaluated on CelebA [29] reference dataset, as seen in Tab. 8.

Table 8: Accuracy of the tSA classifier when evaluated with CLIP/off-shelf classifier.

tSA	Accuracy
`Male`	99.6%
`Young`	98.8%
`Smiling`	98.6%
`High Cheekbones`	97.2%
`Pale Skin`	98.4%
`Eyeglasses`	96.7%
`Mustache`	87.4%
`Gray Hair`	88.3%
`Chubby`	90.4%

Quality. Text alignment (TA) [37, 22] and Fréchet Inception Distance (FID) [38] are utilized as quality metrics. Specifically, *Text-alignment* is based on the notation that a sample generated based on ITI-GEN prompt ($\boldsymbol{P}$) or HP ($\boldsymbol{F}$) for a given tSA, should retain the semantics of the original base prompt $\boldsymbol{T}$, unless the sample has degraded in quality. For example, samples generated based on $\boldsymbol{F} = E_T(\text{``}\texttt{A headshot of a person young}\text{''})$ should still retain the semantic of $\boldsymbol{T} = E_T(\text{``}\texttt{A headshot of a person}\text{''})$. Therefore, to evaluate quality, we compare the generated images (based on the $\boldsymbol{P}$ or $\boldsymbol{F}$) with base prompt $\boldsymbol{T}$ using text-image cosine similarity in CLIP's feature space [37, 22]. FID [38] then compares the feature statistics of the generated samples against a reference fair FFHQ [49] dataset from the CleanFID [50] library.

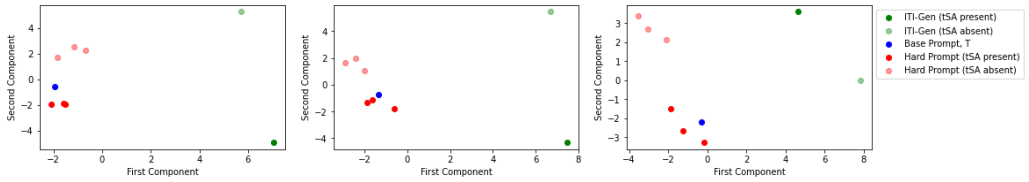
Semantic Preservation: When a fairness approach enforces fairness *w.r.t.* a tSA, changes related to that tSA are more favorable. Considering our setup, for the same latent code Z, the generated images with base prompt and fairness scheme are more favorable to be different only in features related to tSA, and similar in other features. To measure this behavior, for each input latent code Z, the generated image by querying the T2I model with base prompt $\boldsymbol{T}$ is considered as the reference image. Then, for the same latent code Z, we measure the similarity of the generated image by querying the T2I with HP and ITI-GEN prompts with that reference image using *DreamSim* [39] features. DreamSim improves upon existing semantic measurement by considering both low-level features, mid-level, and high-level features for a more holistic semantic measurement.

[3] https://trust.is.tue.mpg.de/; https://github.com/dchen236/FairFace; https://github.com/SonyResearch/apparent_skincolor

https://doi.org/10.52202/079017-0720

B.6 Visualizing the Learned Embedding vs Base Prompt

In addition to the direction loss values provided in the main paper to show that $\Delta \boldsymbol{P}$ is not aligned well with $\Delta \boldsymbol{T}$ as pseudo-gold standard direction–which increases the possibility of encoding unrelated knowledge and leading to distorted learned tokens. Here, we provide an additional visualization of the embedding in Fig. 43 to show the difference between the learned embedding in ITI-GEN and the embeddings of the base prompt and hard prompt. As one can see, even though the embeddings of the base prompt and three different variants of the hard prompt are clustered well, the embeddings of the learned tokens in ITI-GEN are quite far from this cluster increasing the possibility of the encoded unrelated concepts and degrading the generation quality. One may argue that using a regularizer can push these embeddings towards this cluster. However, our experimental results suggested that this will result in a compensated performance in terms of the fair T2I generation.



(a) Gender (Male) (b) Young (c) Smiling

Figure 43: **PCA Analysis on CLIP-text embedding for ITI-GEN , well-defined HP and Base Prompt.** Utilizing a pre-trained CLIP text encoder we attain the text-embedding for the i) Base Prompt T="A headshot of a person", ii) ITI-GEN tokens, and iii) selected well-defined Hard Prompts for SA$\in\{$`Gender (Male)`, `Young`$\}$. Then we apply Principle Component Analysis (PCA) for dimensional reduction. We remark that these same text embeddings are later used in the SDM for sample generation.

C Limitations and Broader Impact

In this section, we discuss some limitations regarding our work as well as some potential societal impacts that it may have.

Limitations. Firstly, FairQueue work follows ITI-GEN which utilizes a reference dataset to first optimize a ITI-GEN prompt $\boldsymbol{P}$. This setup requires dozens of reference images for each category of the tSA which may not always be readily available. Second, although FairQueue provides better quality and semantic preservation of the original base prompt sample, its sample generation is still prone to experiencing entanglement in certain tSA. Specifically, entanglement in tSA could result in some unwanted augmentation of non-tSA during sample generation.

Broader impact. Our work FairQueue takes a significant step towards enhancing fairness in text-to-image generation. By improving the quality of samples generated through a fair text-to-image algorithm, we facilitate greater adoption of these techniques by the general public. This increased adoption can help prevent the perpetuation of unwanted biases in everyday applications, promoting a more equitable and inclusive use of technology in society.

D Related Work

Text-to-Image Generation. There was a surge in text-to-image (T2I) models in the last years, exemplified by models like DALL-E [7, 51], Stable Diffusion [1], Midjourney [52] (with over 16 million active users in July 2023), and many others [53, 24, 54]. These models have demonstrated their capability to generate high-quality samples across different domains and new applications are defined around these models in different areas such as art [55], design, and even medical imaging [56]. In contrast to the earlier generative models which primarily served research purposes within research and scientific settings, current T2I models offer much broader accessibility [57]. However, this increased accessibility also amplifies the potential consequences of bias within these generative models [27]. Our work aims to address this issue by mitigating bias and prompting fair T2I generation.

Fairness in Generative Models. In generative modeling, fairness is usually defined as *equal representation* where all categories of a tSA are supposed to be represented with a similar probability. Different approaches are proposed to improve the fairness of the conventional generative models including weak supervision to achieve fairness with the importance weighting of a fair dataset [28], transfer learning from a large biased dataset to a small fair dataset [30, 31], or enforce uniform sampling from the latent feature space [58]. In the context of fair T2I generation, using the pretrained T2I models and adapting prompts for a fair generation has recently attracted a lot of attention. Bansal et al. [17] proposes the use of "ethical interventio" prompts for text-to-image generators to encourage the concept of independence w.r.t. the the sensitive attributes. Specifically, these ethical intervention prompts can be appended to the original prompt *e.g.,* "a photo of a bride *from diverse cultures*" and "a photo of a person wearing a hat *irrespective of their [SA]*" Furthermore, the work substantiates this approach by mentioning that these neutral language prompts exist in the training data. Overall, these additional prompts have empirically been shown to produce diverse samples (a different definition of diversity *w.r.t.* SA) and high-quality samples (human-voted). Chuang et al. [36] proposes to project out of the biased direction of the text embeddings as a form for bias mitigation. Specifically, given a known sensitive attribute, we are able to minimize the equalization loss and find a text embedding that is of equal distance from the biased prompt's embedding. Then utilizing this optimized (debiased) text-embedding, we are able to generate samples with a fairer SA distribution. In addition, recently prompt learning has been proposed to learn the tSA knowledge from a set of reference images in ITI-GEN [16]. In this work, we address the issues that arise due to unrelated concepts being encoded in these prompts, leading to defects in the cross-attention maps and degrading the image generation quality.

Prompt Learning. Prompt learning has recently been shown to be a very effective and efficient way of adapting pretrained large vision-language models to different downstream tasks like zero-shot classification [34, 33, 59], image editing [21], personalization of diffusion models [22, 60–62], and few-shot image generation [63]. In this work, we analyze the potential issue of prompt learning in the context of fair T2I generation by analyzing the cross-attention module in the presence of the learned prompts and propose a simple yet efficient approach to integrate prompt learning more efficiently for fair T2I generation.

https://doi.org/10.52202/079017-0720

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] *The main contribution of this paper is the analysis of the shortcomings in prompt learning for fair T2I generation, and then proposing a new approach to address these shortcomings.*

   Justification: [Yes] *The claims are justified with both provided extensive analysis in cross-attention map (in the main paper and Supp), together with experimental results.*

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] *There are some limitations inherited from the setup used in the literature regarding the required dataset to learn the prompts and also the possible entanglement between different target Sensitive Attributes (tSAs)*

   Justification: [Yes] *Please see Sec. C for the detailed discussion.*

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] *We don't have theoretical assumptions and proofs.*

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] *In addition to the basic experimental details provided in the main paper, the detailed setup for enabling reproducibility is discussed in Sec. B. In addition, to facilitate reproducibility we have also provided the anonymous link for the code used in this paper.*

   Justification: [Yes] *Please check Supp for the code, and Sec. B for the details of the experimental setup.*

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-0720

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] *The anonymous link to the code is provided in the Supp. In addition, all datasets used in this paper are publicly available. The link for datasets is also added in the readme file of the uploaded code.*

Justification: [Yes] *Please refer to Supp.*

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes] *All the details necessary to understand the results are included in Sec. B.*

Justification: [Yes] *Please see Supp Sec. B.*

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes] *For the experimental results provided for the comparison of our proposed FairQueue with SOTA approach* ITI-GEN*, as mentioned in Sec. 5 the experiments are run 5 times and the error bars are included in the Tab. 1.*

Justification: [Yes] Please refer to Tab. 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] *The detailed amount of compute resources together with carbon emission estimations are provided in Supp Sec. B.4.*

   Justification: [Yes] *Please check Supp Sec. B.4.*

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] *We carefully read the NeurIPS's code of ethic to make sure we follow it to the best of our ability.*

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] *The broader impact of our work is discussed in Sec. C which enables a safer adoption of T2I models by addressing the fairness concerns.*

    Justification: [Yes] *Please see Supp Sec. C*

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations

https://doi.org/10.52202/079017-0720

(e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] *We use the publicly available stable diffusion as our T2I model in our research, and considering that this model already has the safe-checker to prevent generating NSFW content, we do not add any additional safeguards in our research.*

    Justification: [NA] *Please read our answer above.*

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] *All dataset, codes and other assets used in this paper are cited properly throughout the whole paper.*

    Justification: [Yes] *Please check the citations.*

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] *We are not introducing new assets in our reserach.*

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] *We do not have human subjects in our studies.*

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] *We do not have human subjects in our studies.*

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0720