FairRAG: Fair Human Generation via Fair Retrieval Augmentation

Robik Shrestha[1*] Yang Zou[2] Qiuyu Chen[2] Zhiheng Li[2] Yusheng Xie[2] Siqi Deng[2]
[1]AWS AI Labs [2]Amazon AGI
robikshrestha@gmail.com {yanzo,qychen,lzhiheng,yushx,siqideng}@amazon.com

Abstract

Existing text-to-image generative models reflect or even amplify societal biases ingrained in their training data. This is especially concerning for human image generation where models are biased against certain demographic groups. Existing attempts to rectify this issue are hindered by the inherent limitations of the pre-trained models and fail to substantially improve demographic diversity. In this work, we introduce Fair Retrieval Augmented Generation (FairRAG), a novel framework that conditions pre-trained generative models on reference images retrieved from an external image database to improve fairness in human generation. FairRAG enables conditioning through a lightweight linear module that projects reference images into the textual space. To enhance fairness, FairRAG applies simple-yet-effective debiasing strategies, providing images from diverse demographic groups during the generative process. Extensive experiments demonstrate that FairRAG outperforms existing methods in terms of demographic diversity, image-text alignment, and image fidelity while incurring minimal computational overhead during inference.

1. Introduction

Generative artificial intelligence has witnessed rapid growth and adoption in a short span of time. In particular, diffusion-based text-to-image models are able to produce high-quality, photo-realistic images from textual prompts [11, 32, 33, 36] and are thus increasingly being integrated into practical applications [16, 28]. However, this growing adoption also underscores the need to investigate and address fairness concerns. Specifically, text-to-image generation systems tend to mirror or even amplify societal biases in their training data, which is especially evident in human image generation [2, 26, 30]. They exhibit biases against specific demographic groups in terms of age, gender and skin tone. For example, Stable Diffusion [33] produces individuals with darker skin tones when prompted for workers from lower-paying occupations [29]. These tendencies result in adverse outcomes and diverge from the goals of equitable representation. There are some attempts to address this issue [2, 10], however, they do not adequately mitigate the dataset biases ingrained within the pre-trained models.

Figure 1. The proposed (FairRAG) framework improves demographic diversity (fairness in image generation) by conditioning generative models on external human reference images. As defined in Eq. 3, the diversity metric measures representation from different age, gender and skin tone groups.

To address this, we introduce *Fair Retrieval Augmented Generation* (FairRAG), which harnesses an external data

*Work done during an internship at AWS AI Labs.

source consisting of real human images from diverse age, gender, and skin tone groups to improve fairness. The FairRAG framework allows us to fix bias without the costly processes of fixing pre-training data or re-training backbone. Additionally, by expanding the external dataset, it can easily generalize to newer concepts, making it an extensible framework. FairRAG utilizes lightweight, yet effective mechanisms to improve fairness. First, FairRAG requires a way to condition the generative model on reference images. For this, we train a single linear layer that projects reference images into textual space to condition a frozen backbone. This circumvents the computational overhead in existing conditioning approaches, which either re-train the model [4, 7] or require test-time parameter tuning [12]. At inference, directly retrieving and conditioning on a set of images with the highest similarity score for a text prompt does not improve fairness because biases exist in the external database too. To address this, FairRAG consists of a fair retrieval system that utilizes efficient, post-hoc debiasing strategies to sample from diverse demographic groups. Compared to previous approaches [2, 10] that fully rely on internal knowledge in the models, which can be biased, FairRAG is more steerable, explainable, and transparent in controlling demographic distributions for image generation.

We compare FairRAG against multiple methods in terms of the demographic diversity metric (*cf*. Eq. (3)), which assigns higher scores for fairer demographic representations. Compared to the best non-RAG method, FairRAG improves the diversity metric from 0.341 to 0.438. We also observe improvements in image-text alignment (CLIP score [31]): 0.144 to 0.146 and image fidelity (FID [14]): 74.1 to 51.8.

The contributions of this paper are as follows.

- We propose FairRAG, a novel framework to improve demographic diversity in human generation by leveraging reference images drawn from external sources.
- FairRAG employs lightweight conditioning and fairness-enhanced retrieval mechanisms that require minimal computational overhead.
- Experimental results show improvements over existing methods in terms of diversity, alignment and fidelity.

2. Related Works

Societal Biases in Diffusion Models. Diffusion-based text-to-image generative models produce high fidelity, realistic images and have seen increasing adoption [11, 16, 28, 32, 33, 36]. However, they are trained on large-scale image-text datasets that contain harmful biases [3, 38]. Several works study how this causes the text-to-image generative systems to also be biased against specific demographic groups [3, 10, 29, 30]. Some recent works attempt to mitigate these issues, for instance, by editing the text prompt to encourage diversity [2] or by guiding the generative process to balance out the representations from different groups [10, 13]. However, such methods do not substantially mitigate the effects of the biased associations embedded within the models. To tackle this, FairRAG leverages external references that lessen such biases, *i.e*., contain samples from diverse groups to improve fairness in generation.

Conditioning Text-to-Image Diffusion Models. There are existing approaches to condition on visual references [12, 35, 43, 45]. Some employ test-time tuning which is computationally expensive since it requires changing model parameters at inference [12, 22, 35]. Tuning-free methods avoid this by employing a separate adaptor module that is already trained for conditioning [27, 40, 43, 45]. FairRAG is also a tuning-free method. However, compared to the heavier adaptor modules used in prior works, FairRAG uses a lightweight linear conditioning layer. Another concurrent work: ITI-Gen [44] learns prompt embeddings from visual references for conditioning. However, this entails learning a separate embedding per concept, which is not scalable. FairRAG, on the other hand, trains the conditioning module once and re-uses it to transfer demographic attributes and contextual information from new images at inference, making it more general.

Retrieval Augmented Generation (RAG). RAG-based methods retrieve relevant items from external sources to condition the generative process [6, 23, 42]. RAG has only recently been explored for diffusion-based models. For example, a recent RAG-based method [4] shows improvements in image quality and style transfer. Another work, Re-Imagen [7], shows the efficacy of RAG in generating rare and unseen entities. Compared to these works, FairRAG is more suitable for fair human generation. Unlike previous approaches, FairRAG has a retrieval mechanism designed to improve demographic diversity and does not require costly retraining of backbone to support conditioning.

3. Fair Retrieval Augmented Generation

We propose the FairRAG framework to improve fairness in human image generation by using demographically diverse reference images. To achieve this, FairRAG requires a mechanism to condition the pre-trained backbone, which is enabled by training a lightweight linear encoder while keeping the backbone frozen (*cf*. Sec. 3.1 and Fig. 2). At inference, FairRAG uses simple, post-hoc debiasing strategies to improve fairness, including balanced sampling and query modification to for fair retrieval (Sec. 3.2) and a transfer instruction to enhance the generative process (Sec. 3.3).

3.1. Linear Conditioning Mechanism

In this section, we first give the background of the backbone generative model, which is kept frozen for both training and inference. Then we describe our mechanism that conditions the frozen backbone on the references (Fig. 2).

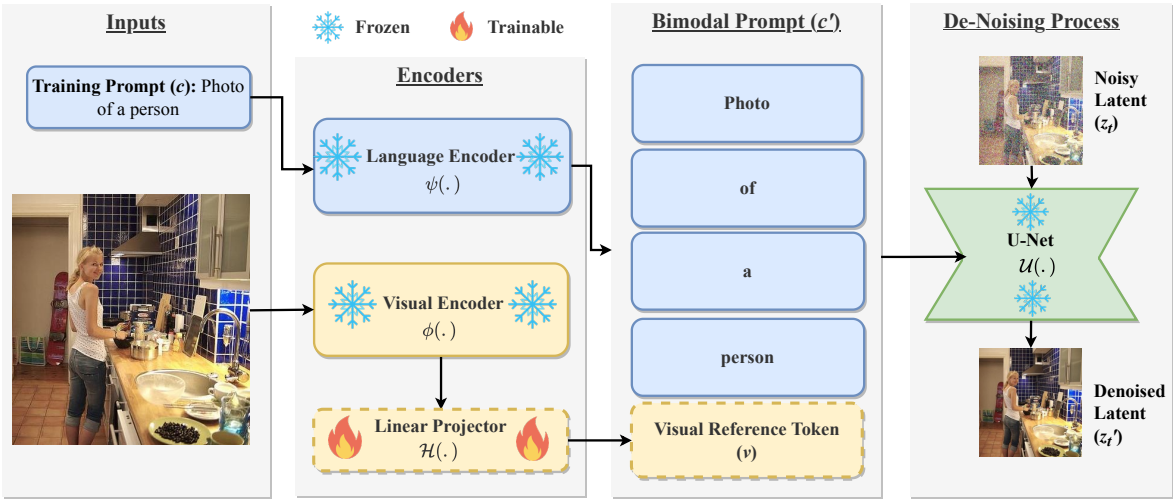

Figure 2. We train the linear projector $\mathcal{H}(.)$ using a denoising loss on the latent space while keeping the backbone model frozen. To train $\mathcal{H}(.)$, we sample images uniformly from each demographic group, pairing each image with the prompt: *Photo of a person*.

Frozen Backbone. Our frozen backbone is a pre-trained text-to-image latent diffusion model—Stable Diffusion (SD) [33]. It reverses noises applied to the latent embeddings of images. SD contains a variational autoencoder (VAE) [19]: $\mathcal{E}(.)$, a text encoder: $\Psi(.)$ and a U-Net [34]: $\mathcal{U}(.)$. Specifically, VAE encodes images x to produce latent representations z. During the forward diffusion process, SD uses a noise scheduler to sample a timestep t, and applies Gaussian noise: $\epsilon_t \sim \mathcal{N}(0,1)$ to z. During the backward diffusion process, $\mathcal{U}(.)$ estimates the noise (ϵ_t') added to the latent, enabling image generation via iterative denoising. The denoising process can also be conditioned on text prompt: c encoded by the text encoder. Specifically, c is fed alongside the noisy latent: z_t into $\mathcal{U}(.)$ to control the denoising process. During inference, one can feed in random Gaussian noise and text prompt through the model to generate images.

Conditioning Module. As shown in Fig. 2, we use a linear projector: $\mathcal{H}(.)$ to condition the backbone on retrieved human references. $\mathcal{H}$ projects the reference image into a text-compatible token, augmenting the text prompt with additional information for conditioning. Let x be the reference image, $v = \phi(x)$ be the visual embedding obtained from a CLIP image encoder [31] and $c = (w_1, w_2, ..., w_n)$ be the text prompt encoded via a CLIP text encoder $\Psi(.)$. The linear projector: $\mathcal{H}(.)$ projects v into a conditioning vector (token): $\mathcal{H}(v)$, which is concatenated with c to obtain a bimodal prompt: $c' = (w_1, w_2, ..., w_n, \mathcal{H}(v))$. This retrieval-augmented text prompt is then fed into the U-Net to condition the denoising process.

Training Procedure. We train $\mathcal{H}(.)$ with the denoising loss in the latent space (Fig. 2). At timestep t, we train $\mathcal{H}(.)$ with the following denoising loss [33]:

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon'(z_t, t, c')||_2^2\right]. \quad (1)$$

We pair the images with a simple text prompt: *Photo of a person* during training, avoiding the usage of detailed captions that may not always be accessible.

3.2. Fair Retrieval System

During inference, FairRAG ensures that the reference images are demographically diverse by using simple post-hoc debiasing techniques that do not require model re-training. The conventional approach of retrieving the Top-K most similar images for a given query, as employed in prior RAG frameworks [4, 7], does not ensure diversity. To address this limitation, FairRAG adopts a two-step process. First, it retrieves a larger set of N candidate images ($N > K$), then, it performs balanced group sampling to obtain a balanced set of K references to condition the model.

Top-N Retrieval with Debiased Query. To obtain N demographically diverse candidate images, FairRAG constructs a *debiased query* by appending the original text prompt with the following phrase: *with any age, gender, skin tone*. This simple query modification improves fairness in retrieval while maintaining consistency with the prompt.

Top-K Selection via Balanced Sampling. While the debiased query improves diversity, the candidate images may not be ordered in a balanced manner. For this, FairRAG applies a *balanced sampling strategy* and selects a set with K images for conditioning. We store the prediction for each demographic attribute: $a \in \{$*age, gender, skin tone*$\}$ for each image and use them for balanced sampling. For this, let us denote *an intersectional group* as g, which is a tuple of the attribute values *e.g.*, g = *(25 year, dark-skinned, male)*

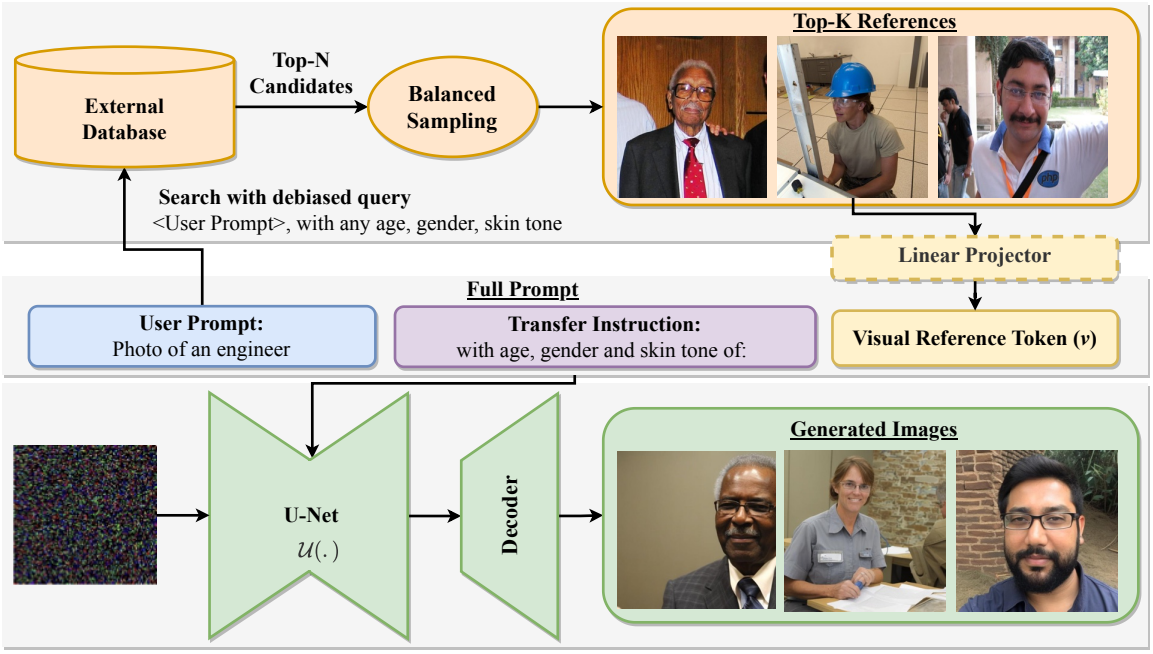

Figure 3. During inference, FairRAG constructs a debiased query to retrieve Top-N candidates for a given prompt. Using their demographic group annotations, FairRAG then selects a balanced set of K images with high demographic diversity for conditioning. The full bimodal prompt consists of: a) the original user prompt, b) a transfer instruction and c) the projected visual reference token. This bimodal prompt is used within the cross-attention layers of the U-Net to condition the generative process.

and *an individual group* corresponding to the demographic attribute (a), as: $g[a]$. For instance, $g[age]$ = *25 year* is an individual age group. Next, let G be the set of unique intersectional groups in the Top-N candidates and $m_{g[a]}$ be the number of times an individual group $g[a]$ appears in G. Then, the sampling weight for g is:

$$w_g = \left[\sum_a \frac{m_{g[a]}}{n_a} \right]^{-1}, \tag{2}$$

which is high if g has individual groups that are rare and low if they are frequent. Here, n_a is the total number of possible values for a, used for normalization (*e.g.*, $n_{gender} = 2$). This strategy thus provides higher priority to the demographic groups that are underrepresented.

3.3. Image Generation

As shown in Fig. 3, FairRAG projects the reference images through $\mathcal{H}$ and adds a text instruction to enhance attribute transfer while generating images. Given an example prompt: *Photo of an engineer*, FairRAG constructs an extended bimodal prompt: *Photo of an engineer, with age, gender and skin tone of:* v, which contains the instruction: *with age, gender and skin tone of:* to improve the conditioning process. This method does not explicitly specify the exact age, gender or skin tone values, yet helps the model transfer those attributes from the reference images. This extended prompt is used within the cross-attention layers of $\mathcal{U}(.)$ to condition the model at each denoising step.

4. Experiments

4.1. Experiment Setup

We compare FairRAG against other baselines for the task of diverse human generation using neutral text prompts that do not specify any demographic groups, but still exhibit bias. We evaluate diversity among three *demographic attributes*: age, gender and skin tone. We use a modified version of the *demographic groups* presented in FairFace [18] for: age (< *20, 20-29, 30-39, 40-49, 50-59, 60+*) and gender (*male* and *female*). For skin tone analysis, we use the 10-point Monk Skin Tone (MST) scale [1].

Evaluation Prompts. The evaluation set of FairRAG consists of test prompts with professions that exhibit bias with respect to different demographic groups. These prompts are classified into 8 categories, including: 6 artists (*e.g.*, a dancer), 6 food and beverage (F&B) workers (*e.g.*, a cook), 9 musicians (*e.g.*, a guitarist), 6 security personnel (*e.g.*, a guard), 9 sports players (*e.g.*, a tennis player), 12 STEM professionals (*e.g.*, an engineer), 7 workers (*e.g.*, a laborer) and 25 from other professions (*e.g.*, a politician). Please refer to Appendix B for the full list of prompts. For the main experiments, we use the template *Photo of <profession>*, *e.g.*, *Photo of a doctor* to create the prompt.

Demographic Diversity Metric. We use normalized entropy as our diversity metric, reporting the intersectional and individual values, where higher values indicate more equitable representation across demographic groups. Intersectional diversity is computed over the unique combinations of age, gender and skin tone groups, and individual diversity score is computed separately per attribute, *e.g.*, the gender diversity score is the normalized entropy for *male* and *female* categories. Specifically, let p_i be the proportion of images generated for i^{th} group, then, the diversity score is the entropy of the group memberships normalized by a uniform distribution, with a maximum value of 1:

$$\mathcal{D} = \frac{\sum_{i=1}^{n} p_i \log [p_i]}{\frac{1}{n} \log(\frac{1}{n})}, \tag{3}$$

where, n is the total number of possible groups. We compute the metrics over the images generated for each prompt, reporting the average across all prompts.

We employ existing prediction systems for age, gender and skin tone classification. For age classification, DeepFace [39] predicts an integer age value, which we map to our age-range categories. We follow Dall-Eval [8] for gender and skin tone classification. Specifically, we use the CLIP model (ViT/L-14) with two classifier prompts: *photo of a male or a man or a boy* and *photo of a female or a woman or a girl*, using the highest scoring prompt to determine the gender. For skin tone classification, we detect skin pixels within the facial region, and determine the Monk Skin Tone value that is closest to the average color of the skin region [20]. When computing the diversity metrics, we account for the cases where the methods fail to generate any human face. In such instances, we treat the images as belonging to the most prevalent demographic group, thereby imposing a penalty on the diversity score.

While we follow previous research [8, 10] for demographic group classification, we acknowledge their limitations as well. First, we perform skin tone analysis, but refrain from making inferences about the race. This choice is driven by the recognition that racial identity can be influenced by social and political factors [8] and whether one can predict race from visual information alone is debatable. We leverage skin tone, a lower-level image feature, in an attempt to conduct more objective assessment. More precisely speaking, we consider the *apparent* skin tone from RGB images, in the absence of access to true skin tone where more rigorous process needs to be applied (*e.g.*, lab controlled data collection via spectrophotometers). Second, we employ a simplified binary gender classification, even though gender is known to encompass a broader spectrum [17]. This is because estimating gender from a wider range of possibilities based solely on appearance can potentially reinforce appearance-related stereotypes. While our studies and discussions of gender diversity in this work are limited to apparent binary genders, the framework we devise may be generalized to alternative definitions.

Alignment and Fidelity Metrics. We use CLIP score [31], *i.e.*, the cosine similarity between the text embeddings and the image embeddings to compute image-text alignment. We report the Fréchet Inception Distance (FID) [14] between the generated images and real distribution, approximated by sampling a fixed number of real images per prompt.

Training and Retrieval Sets. We use images containing humans from two datasets: MSCOCO [24] and OpenImages-v6 [21] to: a) train the linear projector and b) retrieve images during inference. Since the work focuses on human generation, we run a face detector [9] and only keep the images with human faces. Since the datasets contain low quality images *e.g.*, blurry scenes, we run an aesthetic scorer [37] to filter out images with low scores. We combine the MSCOCO and OpenImages datasets, splitting into non-overlapping training and retrieval sets, consisting of 173,289 and 330,777 images respectively. For retrieval, we index the image embeddings using CLIP VIT-L/14 [15]. Since these embeddings are pre-computed and stored, FairRAG avoids using CLIP image encoder during inference.

4.2. Comparison Methods

We compare FairRAG against these methods:

- ***SDv2.1 [33]*** is the baseline method used without applying any debiasing technique.
- **Ethical Interventions (*Interven [2]*)** attempts to improve diversity by augmenting the original prompts with ethical phrases, *e.g.*, *Photo of a doctor if all individuals can be a doctor irrespective of their age, gender and skin tone*.
- **Fair Diffusion (*FairDiff [10]*)** applies semantic-guidance [5] to steer the model towards a specific intersectional group. The groups are sampled with a uniform prior.
- **Text Augmentation (*TextAug*)** creates multiple variants of the prompt by explicitly mentioning the demographic groups, *e.g.*, *Photo of a doctor. This person is 55-year old, dark-skinned, female*. Surprisingly, despite its simplicity, past studies do not study this method [2, 10]. We find TextAug to be a strong baseline in our experiments.
- **Baseline RAG (*Base RAG*)** is an ablated version of FairRAG that removes the fairness interventions, *i.e.*, does not use debiased query, balanced sampling or text instruction. It still relies on the linear module for conditioning.

FairDiff [10] and TextAug require explicit specifications of the demographic groups. For this, we use the template: *This person is <age>-year old, <skin tone>, <gender>*. Age is the mid-point of the corresponding group *e.g.*, *25-year-old* for the group: *20–29 years old*; skin tone is specified as: *light-skinned*, *medium skin colored* or *dark-skinned*

Table 1. Breakdown of diversity scores for individual and intersectional (intersec.) groups, showing how leveraging external images can help improve the metrics.

Method	Age	Gender	Skin Tone	Intersec.
SDv2.1 [33]	0.220	0.273	0.224	0.188
Interven [2]	0.439	0.451	0.362	0.333
FairDiff [10]	0.225	0.371	0.223	0.196
TextAug	0.426	0.766	0.334	0.341
Base RAG	0.417	0.582	**0.439**	0.374
FairRAG	**0.559**	**0.800**	0.416	**0.438**

and gender is either *male* or *female*. Note that FairRAG avoids such explicit attribute specification, relying instead on the text instruction for implicit attribute transfer.

4.3. Results

In this section, we discuss the overall quantitative and qualitative results. Table 2 summarizes the intersectional diversity, alignment and fidelity metrics alongside the absolute gains over SDv2.1 [33]. FairRAG outperforms other methods in terms of the diversity and alignment scores and is close to Base RAG in terms of image fidelity, showing the benefits of the proposed setup. As shown in Fig. 4, FairRAG effectively leverages human-specific attributes from the reference images to condition the generated images, resulting in enhanced demographic diversity. As presented in Table 1, it is able to improve diversity metrics for all three attributes: age, gender and skin tone. In terms of the baselines, we find that Base RAG is also able to improve diversity, alignment and fidelity scores, however, the additional fairness interventions used in FairRAG, *i.e.*, *query debiasing*, *balanced sampling* and the *transfer instruction* help boost the diversity scores further. In terms of non-RAG baselines, we find TextAug to be the most effective, obtaining improvements in all three metrics over other non-RAG methods. However, qualitatively, it produces synthetic, unrealistic contexts, an issue observed for other baselines as well. FairRAG on the other hand generates more realistic images due to the conditioning from real images. Next, Interven [2] and FairDiff [10] also improve the diversity scores to some extent, but are well below FairRAG. As shown in Fig. 4, extra text intervention used in Interven [2] results in grids of smaller sub-images, which is an undesired side-effect. FairRAG also uses a text instruction, but this does not lead to such inadvertent consequences. Therefore compared to the baseline methods, FairRAG stands out as more effective.

Diversity for different prompt categories. We present example outputs for different categories in Fig. 5 and present the intersectional diversity values for each of the 8 categories in Table 3. The improved diversity metrics shows that FairRAG generalizes to different professions.

Table 2. Quantitative results from all the comparison methods, highlighting the **best** and the second-best scores. We also show the improvement and deterioration in terms of absolute difference from SDv2.1. Compared to other baselines, FairRAG shows improvements in diversity and alignment, and rivals Base RAG in terms of the fidelity score.

Method	Diversity (↑)	CLIP (↑)	FID (↓)
SDv2.1 [33]	0.188	0.142	85.3
Interven [2]	0.333 (+.145)	0.132 (-.011)	93.9 (+08.6)
FairDiff [10]	0.196 (+.008)	0.142 (-.000)	87.8 (+02.5)
TextAug	0.341 (+.153)	0.144 (+.002)	74.1 (-11.2)
Base RAG	0.374 (+.186)	**0.146** (+.003)	**49.4** (-35.9)
FairRAG	**0.438** (+.250)	**0.146** (+.003)	51.8 (-33.5)

Minimal Increase in Latency. FairRAG involves: a) text-to-image retrieval with debiased query, b) balanced sampling, c) visual reference projection to obtain the bimodal prompt, and d) conditional image generation. The first three steps are specific to FairRAG, but add minimal computational overhead compared to SDv2.1. On a single NVIDIA A10G Tensor Core GPU, SDv2.1 and FairRAG require 2.8 secs and 3 secs respectively, to generate a single image with 20 denoising steps. We re-iterate that FairRAG is also more efficient than prior methods that use test-time tuning [12] or heavier conditioning modules [45].

Face and Body Size. Most past studies focus on close-up views of faces neglecting analysis of images with the human body taking a larger portion of the image [10, 44]. To test if FairRAG works well for both the cases, we employ two different prompt prefixes: *Headshot of* and *Full body of*, controlling the face/body size. As shown in Fig. 6, we find that FairRAG is able to transfer the demographic attributes in both the cases. It also generates contexts that are more realistic than the SDv2.1 baseline, which is especially evident for full body images.

Ablation Study. In Table 4, we present ablated variants of FairRAG to investigate the effects of different components. Retrieval-time interventions: *debiased query* and *balanced sampling* and generation-time intervention: *text instruction*, all contribute positively to the intersectional diversity score, thereby validating our decisions to incorporate these mechanisms. All three mechanisms enhance the age and diversity scores. For skin tone diversity, we do not find additional benefit from text instruction, but debiased query and balanced sampling contribute positively to skin tone diversity as well. We also present the diversity scores for the real distribution, *i.e.*, the retrieved images. FairRAG still has a room for improvement, which can potentially be achieved by improving the transfer of attributes.

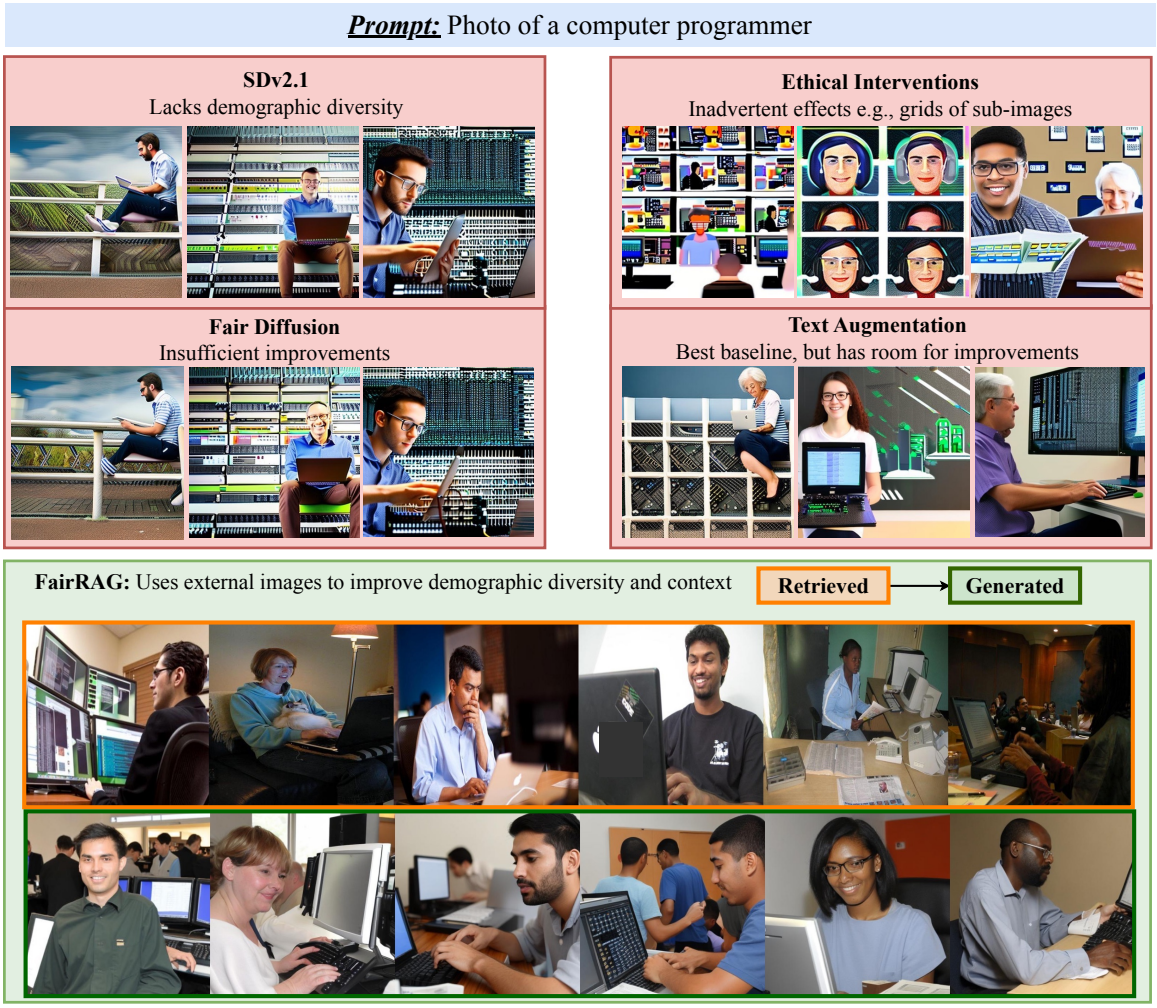

Figure 4. Example outputs from different methods for the text prompt *Photo of a computer programmer*. Baseline methods, barring Text Augmentation, fail to produce images with high demographic diversity. FairRAG improves demographic diversity with the help of external visual references. Apart from that, it also improves alignment and fidelity.

Figure 5. FairRAG improves demographic diversity for different categories of professions.

Table 3. Intersectional diversity metrics on the eight concept types used in our evaluation set. FairRAG shows improvements in each concept type, showcasing the generality of the approach.

	Artists	F&B Workers	Musicians	Security Personnel	Sports Players	STEM Profes.	Workers	Others
SDv2.1 [33]	0.261	0.237	0.168	0.137	0.175	0.199	0.197	0.175
Interven [2]	0.385	0.284	0.282	0.314	0.299	0.359	0.282	0.370
FairDiff [10]	0.259	0.273	0.164	0.133	0.161	0.240	0.210	0.177
TextAug	0.391	0.269	0.322	0.348	0.314	0.342	0.349	0.357
Base RAG	0.401	**0.428**	0.404	0.394	0.336	0.357	0.362	0.402
FairRAG	**0.436**	0.413	**0.440**	**0.458**	**0.416**	**0.432**	**0.419**	**0.454**

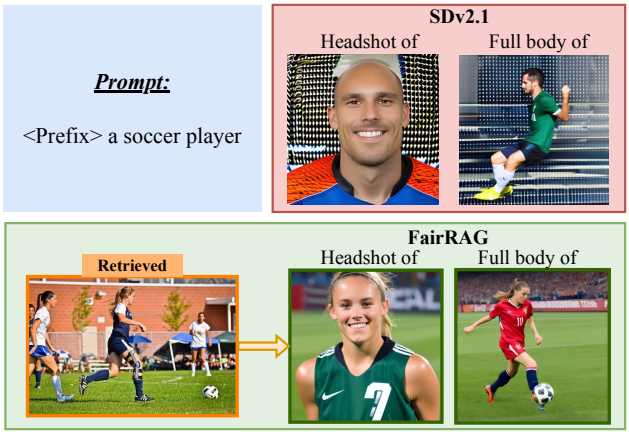

Figure 6. While most past works focus on close-up views of faces, we find FairRAG can transfer attributes when asked to generate full body images as well.

Table 4. Ablation studies showing how debiased query (debiased q), balanced sampling (bal. sampl.) and text instruction (text instr.) help boost the diversity scores. We also present the metrics for retrieved images, *i.e.*, the real distribution, showcasing room for further improvement.

Method	Age	Gender	Skin Tone	Intersec.
SDv2.1 [33]	0.220	0.273	0.224	0.188
TextAug	0.426	0.766	0.334	0.341
Base RAG	0.440	0.562	0.437	0.386
Ablated variants of FairRAG				
No Debiased Q	0.525	0.764	0.411	0.414
No Bal. Sampl.	0.538	0.734	0.392	0.420
No Text Instr.	0.481	0.771	0.416	0.407
FairRAG	**0.559**	0.800	0.416	0.438
Retrievals	0.546	**0.902**	**0.526**	**0.478**

5. Limitations and Future Directions

In this section, we discuss some limitations and layout potential future directions for further improvement. To begin with, FairRAG uses one-to-one image mapping *i.e.* uses single reference image for each generated image. An alternative would be to use multiple images to summarize the concepts to be transferred to enhance the conditioning process. Multiple references could also help in cases where single retrieval does not encompass all of the concepts mentioned in the prompt, by aggregating different concepts from different images. Second, despite conditioning on real images, the samples generated by FairRAG can contain disfigurements especially in small faces, limbs and fingers. We hypothesize that fixing this issue requires a better way to incorporate knowledge on human anatomy within the models, which likely entails re-training or tuning the backbone. We discuss this issue in greater detail in A.3.

There are other considerations before a framework such as FairRAG can be deployed in practice. For one, FairRAG is limited to human image generation and thus cannot tackle non-human prompts. A more general RAG framework could utilize references from a broader range of categories. However, even with a larger data source, a practical framework should also be capable of tackling cases in which the references lack consistency with the user's prompt. A worthwhile future direction to tackle this issue is to devise conditioning mechanisms that avoid transferring concepts in references that conflict with the prompt. Furthermore, while the scope of this study is limited to generating humans, we note that the proposed conditioning mechanism and the fairness interventions can be extended to and employed in other domains, making these mechanisms more general with broader applicability.

6. Conclusion

In this work, we developed the FairRAG framework to condition pre-trained generative models on external images to improve demographic diversity. We showed that a lightweight, linear layer can be trained to project visual references for conditioning the backbone and post-hoc debiasing methods can enhance fairness in generation. These mechanisms add minimal overhead during inference, yet, help FairRAG surpass prior methods in terms of diversity, alignment and fidelity.

References

[1] Improving skin tone evaluation in machine learning. https://skintone.google/. 4

[2] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1, 2, 5, 6, 8, 13

[3] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 2

[4] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 2, 3

[5] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[6] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3417–3419, 2022. 2

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2, 3

[8] 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. 5

[9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 5

[10] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 1, 2, 5, 6, 8, 13

[11] 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. 1, 2

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 6

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2

[14] 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. 2, 5

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 5

[16] Moreno Johan. With its latest ai innovations, adobe doesn't want to cut out humans out of the picture just yet. https://www.forbes.com/sites/johanmoreno/2022/10/21/with-its-latest-ai-innovations-adobe-doesnt-want-to-cut-out-humans-out-of-the-picture-just-yet/?sh=f8e73d33491f. 1, 2

[17] Kiku Johnson. Sexual orientation, gender identity, and expression; affirming approach and expansive practices. https://www.health.ny.gov. 5

[18] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 4

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] Seema Kolkur, Dhananjay Kalbande, P Shimpi, Chaitanya Bapat, and Janvi Jatakia. Human skin detection using rgb, hsv and ycbcr color models. *arXiv preprint arXiv:1708.02694*, 2017. 5

[21] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 5

[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. 2

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 11

[26] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[28] Lomas Natasha. Shutterstock to integrate openai's dall-e 2 and launch fund for contributor artists. https://tinyurl.com/shutterstockai. 1, 2

[29] Leonardo Nicoletti and Dina Bass. Humans are biased. generative ai is even worse. https://www.bloomberg.com/graphics/2023-generative-ai-bias/. 1, 2

[30] Malsha V Perera and Vishal M Patel. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*, 2023. 1, 2

[31] 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. 2, 3, 5

[32] 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. 1, 2

[33] 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. 1, 2, 3, 5, 6, 8, 13

[34] 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. 3

[35] 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. 2

[36] 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. 1, 2

[37] Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 5

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[39] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 5

[40] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 11

[42] David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*, 2021. 2

[43] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[44] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 2, 6

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 6

Appendix

A. Additional Results

A.1. Qualitative Analysis

We present more qualitative examples including the reference (highlighted in orange) and generated (highlighted in green) images from FairRAG in Fig. 8. We observe that FairRAG is able to utilize the reference images to improve demographic diversity.

A.2. Quantitative Analysis

We present comprehensive results in Table 5 with all the metrics for all the non-RAG baselines alongside the different variants of FairRAG. We present results for both retrieved and generated images. First, we observe that the intersectional diversity scores improve for both real and generated distributions with *debiased query*, *balanced sampling* and *text instruction*. We also observe some trade-offs between the diversity and alignment/fidelity metrics. CLIP score increases when debiased query is not used and FID value improves when text instruction is not used, while both showcase improvements in the diversity score. This leaves a room for improvement in both alignment and fidelity with the additional mechanisms.

A.3. Disfigurements

As shown in Fig. 7, the generated images can contain disfigurements for small faces, limbs and fingers. We address this issue to a limited extent by using a negative prompt: *bad, disfigured, cropped, bad anatomy, poorly drawn hands, poorly drawn fingers* for all the methods. Simply conditioning frozen backbone on real images does not solve this issue. We hypothesize further improvements require incorporation of the knowledge on human anatomy within the models, which likely entails re-training or tuning the backbone. We leave this for future research efforts.

A.4. Varying number of candidates (N)

In Table 6, we analyze the effects of the initial number of candidates N used to gather the subset of K references. Diversity score increases from N = 100 to N = 750, and saturates after that. We do not observe clear trends for CLIP and FID scores. For all the experiments, we set $N = 250$, using results from preliminary experiments without tuning N on the test set. We set $K = 20$ to compute all the metrics.

B. Evaluation Set

The evaluation set consists of 80 prompts that exhibit bias with respect to different demographic groups. They are classified into 8 categories, including:

- **6 Artists**: craftsperson, dancer, makeup artist, painter, puppeteer, sculptor

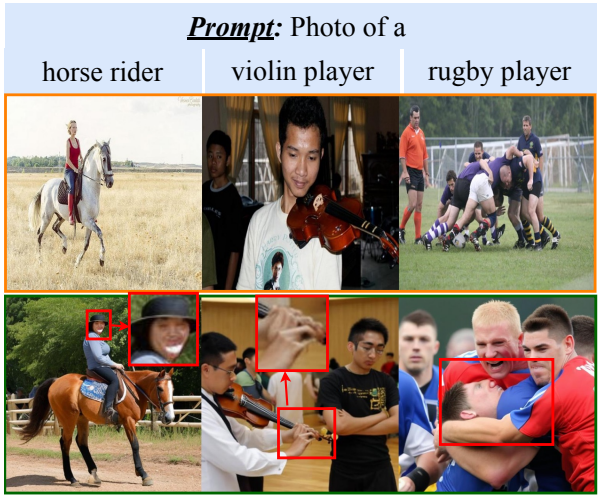

Figure 7. Despite conditioning on real images, the outputs from FairRAG can still contain disfigurements as depicted within the red boxes. Fixing this issue likely requires improved mechanisms to incorporate the knowledge on human anatomy in the models.

- **6 Food and Beverage Workers**: bartender, butcher, chef, cook, fast-food worker, waiter
- **9 Musicians**: disk jockey, drummer, flutist, guitarist, harp player, keyboard player, singer, trumpeter, violin player
- **6 Security Personnels**: firefighter, guard, lifeguard, police officer, prison officer, soldier
- **9 Sports Players**: baseball player, basketball player, gymnast, horse rider, rugby player, runner, skateboarder, soccer player, tennis player
- **12 STEM Professionals**: architect, astronaut, computer programmer, dentist, doctor, electrician, engineer, mechanic, nurse, pilot, scientist, surgeon
- **7 Workers**: carpenter, farmer, gardener, housekeeper, janitor, laborer, person washing dishes
- **25 Others**: backpacker, cashier, CEO, cheerleader, climber, flight attendant, hairdresser, judge, lawyer, lecturer, motorcyclist, patient, politician, public speaker, referee, reporter, retailer, salesperson, sailor, seller, social worker, solicitor, student, tailor, teacher

C. Implementation Details

We train the linear encoder: $\mathcal{H}$ for 50K iterations using the AdamW optimizer [25] ($\beta_1 = 0.9, \beta_2 = 0.999$), with a learning rate of 1e-3 and a weight decay of 0.01. We use balanced sampling during training with a uniform prior over each intersectional group (age, gender and skin tone). During training, we clip the gradients if the norm is greater than 1.0. To generate images during inference, we use the DDIM noise scheduler [41], with 20 de-noising steps conditioned on the text prompt, textual instruction and the projected visual reference.

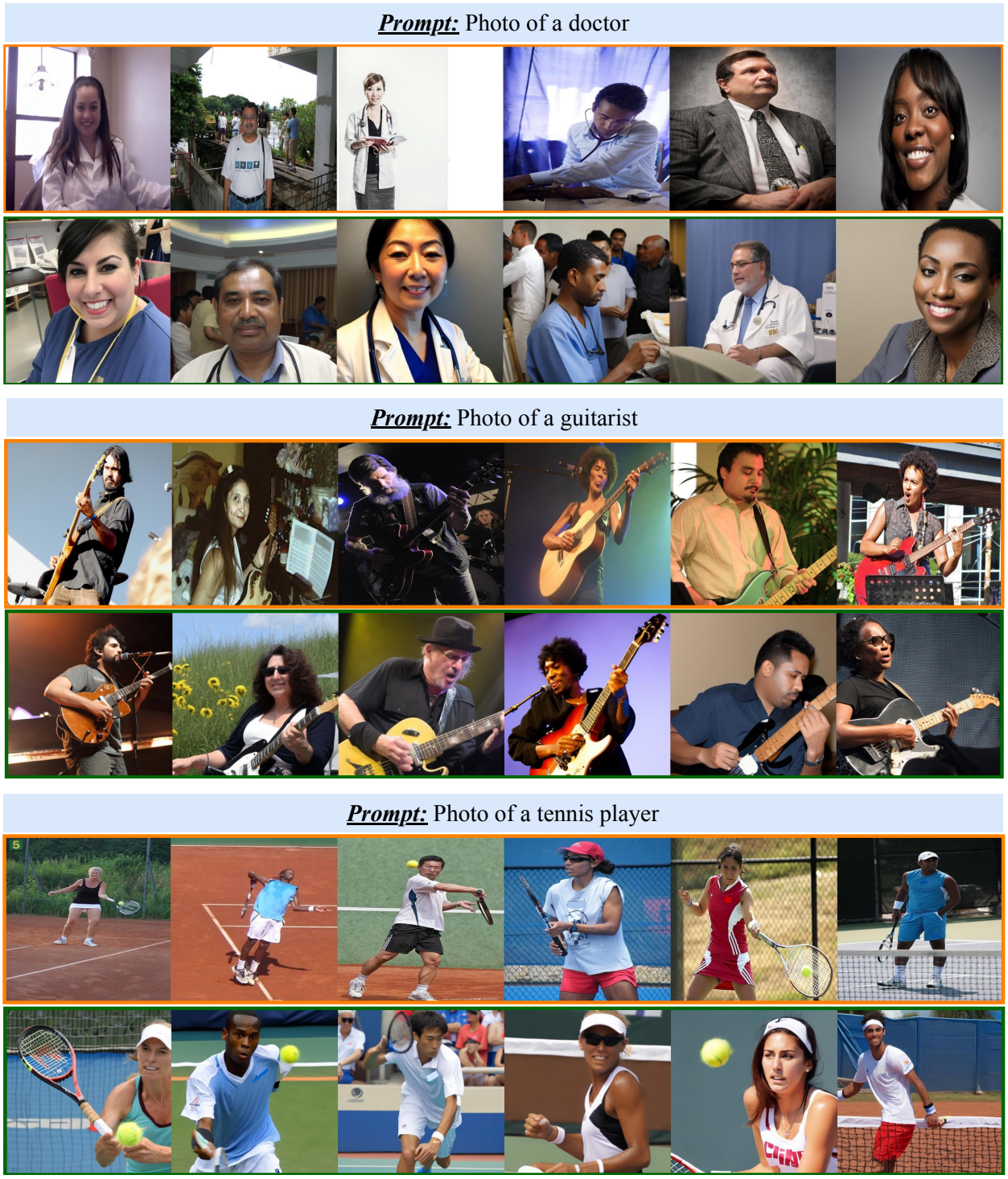

Figure 8. Example outputs illustrating how FairRAG uses the reference images (highlighted in orange) to improve diversity of the generated images (highlighted in green).

Table 5. Presenting diversity, alignment and fidelity metrics for all the baselines and ablated versions of FairRAG. We present results for both retrieved and generated images for FairRAG.

	Diversity				CLIP	FID
	Age	**Gender**	**Skin Tone**	**Intersec.**		
SDv2.1 [33]	0.220	0.273	0.224	0.188	0.142	85.3
Interven [2]	0.439	0.451	0.362	0.333	0.132	93.9
FairDiff [10]	0.225	0.371	0.223	0.196	0.142	87.8
TextAug	0.426	0.766	0.334	0.341	0.144	74.1
Ablated variants of FairRAG						
BaseRAG						
Retrieved	0.475	0.622	0.558	0.447	0.167	33.1
Generated	0.440	0.562	0.437	0.386	0.146	49.4
Without Debiased Query						
Retrieved	0.477	0.867	0.530	0.460	0.166	31.9
Generated	0.525	0.764	0.411	0.414	0.150	50.5
Without Balanced Sampling						
Retrieved	0.528	0.741	0.522	0.458	0.159	30.0
Generated	0.538	0.734	0.392	0.420	0.146	53.0
Without Text Instruction						
Retrieved	0.544	0.902	0.526	0.478	0.158	26.5
Generated	0.481	0.771	0.416	0.407	0.145	48.9
FairRAG						
Retrieved	0.544	0.902	0.526	0.478	0.158	26.5
Generated	0.559	0.800	0.416	0.438	0.146	51.8

Table 6. Diversity, image-text alignment and image fidelity metrics for different values of N used for retrieval.

	Diversity				CLIP	FID
	Age	**Gender**	**Skin Tone**	**Intersec.**		
SDv2.1 [33]	0.220	0.273	0.224	0.188	0.142	85.3
Interven [2]	0.439	0.451	0.362	0.333	0.132	93.9
FairDiff [10]	0.225	0.371	0.223	0.196	0.142	87.8
TextAug	0.426	0.766	0.334	0.341	0.144	74.1
Top-N						
N=100	0.547	0.785	0.409	0.433	0.145	52.7
N=250	0.559	0.800	0.416	0.438	0.146	51.8
N=500	0.580	0.816	0.407	0.443	0.145	51.5
N=750	0.586	0.824	0.415	0.447	0.144	52.2
N=1000	0.572	0.850	0.418	0.445	0.146	52.8