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

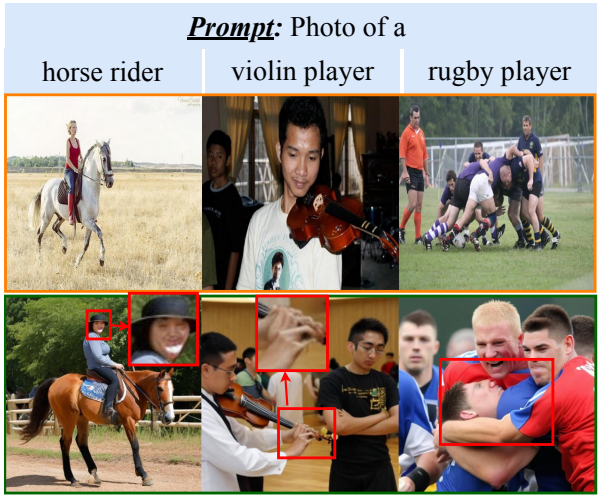

Figure 7. Despite conditioning on real images, the outputs from FairRAG can still contain disfigurements as depicted within the red boxes. Fixing this issue likely requires improved mechanisms to incorporate the knowledge on human anatomy in the models.

- **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.

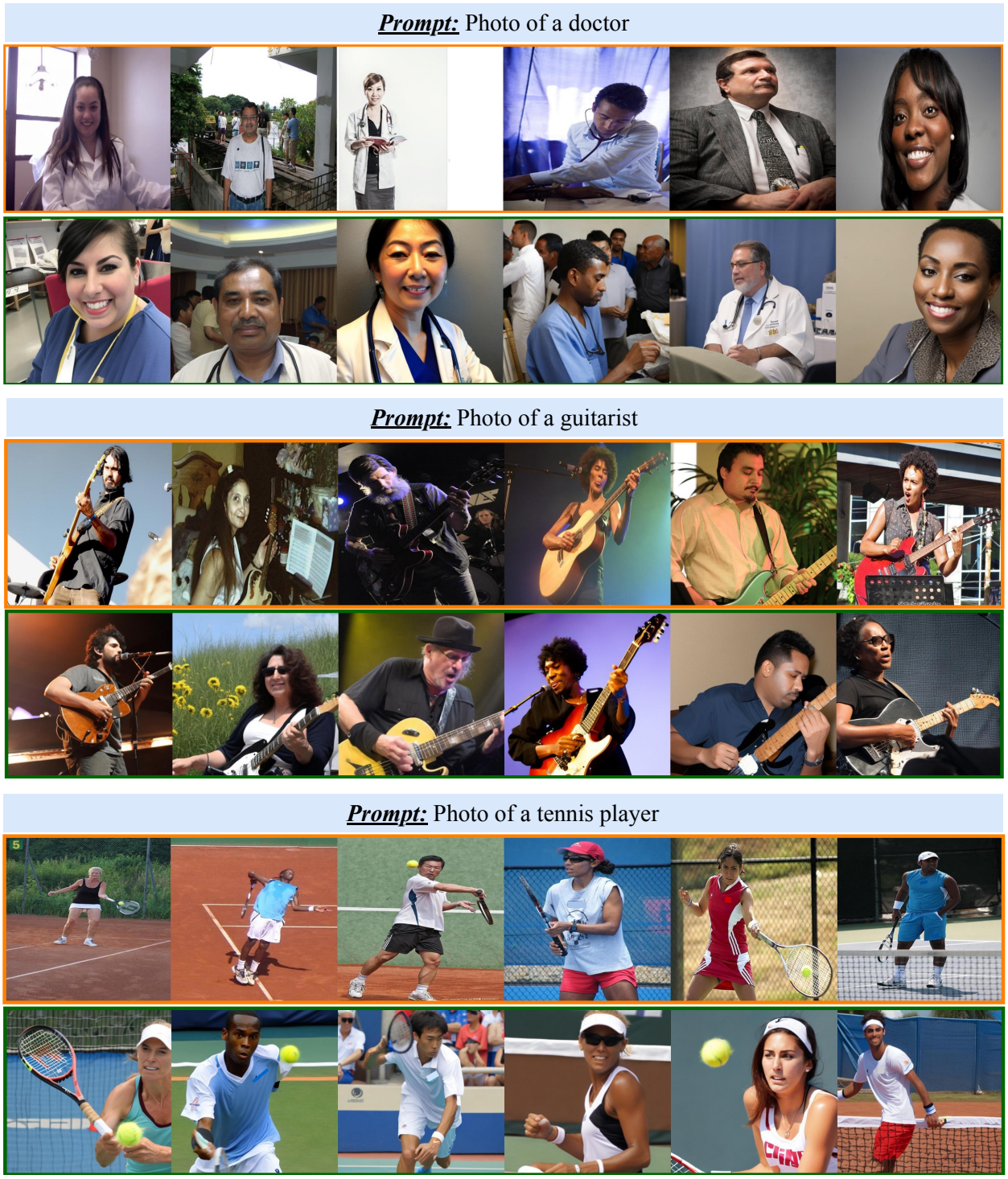

Figure 8. Example outputs illustrating how FairRAG uses the reference images (highlighted in orange) to improve diversity of the generated images (highlighted in green).

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