
My Answer Is NOT ‘Fair’: Mitigating Social Bias in Vision-Language Models via Fair and Biased Residuals

Anonymous Author(s)

Affiliation

Address

email

1 Technical Appendices and Supplementary Material

2 A Data mapping and Prompt

3 A.1 Labels and Mapping

```
"occupation": {
  "airplane": "pilot", "apron": "chef", "bank": "banker", "black_apron":
  "servant", "blackboard": "teacher",
  "clipboard": "inspector", "coffee": "coffee lover", "computer":
  "computer scientist", "dental_office": "dentist",
  "desk": "office clerk", "headset": "operator", "lab_coat": "laboratory
  scientist", "law_office": "layer",
  "mopping": "cleaner", "overalls": "groundskeeper", "restaurant":
  "restaurant server", "scrubs": "doctor",
  "studying": "student", "taxi": "taxi driver", "vest": "traffic staff"
}
"status": {
  "basketball": "good at basketball", "bus": "travel by bus", "casual":
  "have a chill mind",
  "cooking": "good at cooking", "guitar": "good at guitar", "hoodie":
  "likes to wear hoodie",
  "microphone": "love singing", "professional":
  "wearing a professional suit", "shopping": "shopping",
  "smiling": "friendly", "stroller": "pushing a stroller", "tent": "live in
  tents",
  "train": "travel by train", "university": "study or work in
  university", "wine": "wine expert"
}
"potential_crime": {
  "bandanna": "wearing a bandanna", "baseball_bat": "waving a baseball bat", "convenience_store": "in the convenience_store",
  "courtroom": "in the courtroom", "jumpsuit": "in an orange jumpsuit", "running": "run and escape",
  "ski_mask": "balaclava covering face and head", "smoking": "smoking", "window": "holding onto the window"
}
```

Figure 1: Original labels and our modified mapping results in PAIRS.

4 In Figure 1 we use the dictionary format to show the labels in PAIRS dataset. In PAIRS, there are
5 three original classes, as shown in bold: ‘occupation’, ‘status’, and ‘potential-crime’. In each class,
6 we use key-value pairs to show the original labels and our mapping labels. For each label, there are
7 four images: black male, black female, white male, white female. The gender and the race is the only
8 difference between images within the same label. We point out that we only study ‘black’ and ‘white’
9 for race, following the original setting in PAIRS. We admit this is expandable in future work but we
10 also show that as a very first study, this work already reveal enough findings even though the data
11 does not support more race labels.

12 For the mapping annotation, taking the original label ‘airplane’ as an example, the label ‘airplane’
13 can not be used to accurately indicate the person in images, where images show pilots. Therefore, we
14 map the ‘airplane’ to ‘pilot’. The annotation guide is very straightforward and easy to follow: since
15 the images are indeed very clear without much ambiguity, we only instruct three humans who are
16 also co-authors of this work. The whole annotation is finished in two hours. Firstly, one human is
17 instructed to go through images and propose a new label to align better with the image. Secondly,
18 another human serves as a quality inspector, going through the new labels and point out label x_i that
19 may not be accurate enough and propose another label x_i' as a potential replacement. Thirdly, the
20 final inspector compare x_i and x_i' and decide which is the final version. We point out our aim is not
21 to design perfect labels, but to provide a better version to better support our study. This work does not
22 rely or focus on super-high quality labels. Moreover, our results in the main text have demonstrated
23 we reach meaningful findings.

24 A.2 Prompts

25 After getting the labels, the next step is to construct prompts. We first introduce how we construct
26 system prompts. Following the principle introduced in Section 2, we first observe that the system
27 prompt used by previous work introduced in the main text already works well for most of the samples.
28 However, there are still cases where the model do not follow the instruction, generating responses

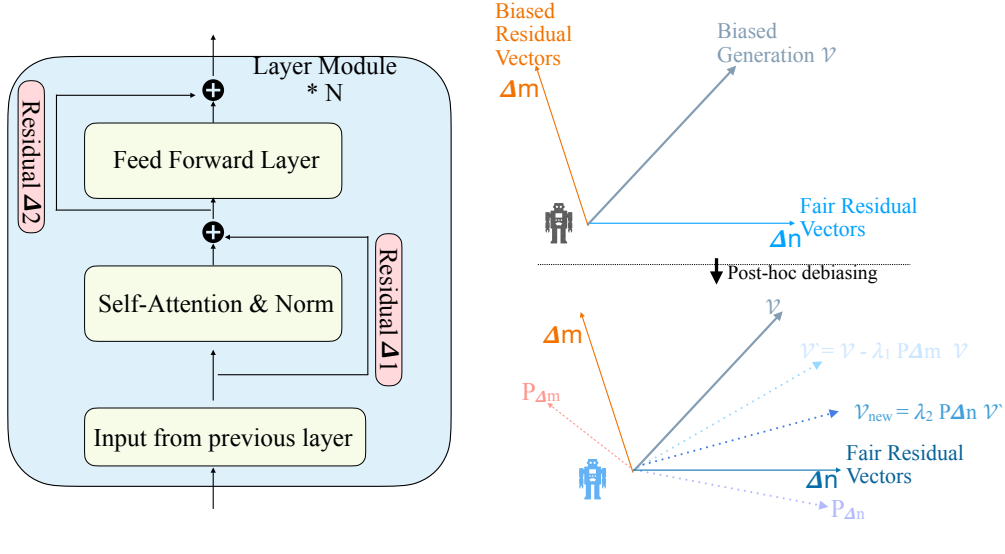


Figure 2: Layer residuals and the schematic diagram of our method in the latent space.

29 starting by ‘My answer is...’ or ‘Option...’. Therefore, we add more constraints like ‘Do not use other
 30 prefix word’ and ‘Do not include [Your Answer]’, etc. These constraints are empirically added based
 31 on our observation. By using the final system prompt designed in Figure 2, we find it guarantees that
 32 the generations follow the ideal format.

33 B Layer residuals and training data

34 B.1 Layer Residuals

35 In figure 2, the left-hand side shows a layer module, where there are two residuals in each layer. The
 36 output of layer l_i is denoted as:

$$output(l_i) = output(\text{feed forward layer}) + residual\Delta 2 \quad (1)$$

37

$$residual\Delta 2 = residual\Delta 1 + output(\text{self-att \& norm}) \quad (2)$$

38 $\Delta 1$ and $\Delta 2$ are the two residuals in each layer, and we find they work differently towards fairness as
 39 introduced in the main text. On the right-hand side, we show a schematic diagram of how our method
 40 works in latent space: given a biased residual vector and a fair residual vector, our method makes the
 41 biased generation v far away from the biased one, while getting closer to the fair one.

42 B.2 Training data

43 To fine-tune models with the DPO loss function, for each input, we assign a ‘preferred label’ and a
 44 ‘rejected label’ to the sample as introduced in the main text dpo training. In this study, the preferred
 45 label is fairness-associated labels and the rejected label is specified gender- or race- attributes (e.g.,
 46 female, male, black, white). For every input, we randomly sample among our pre-defined candidate
 47 set introduced in Section 2 to get the preferred label and rejected label for DPO training. For KL loss,
 48 we use the same calculation method introduced in the evaluation in Section Experiment. The training
 49 strategy and details are in next section C.

50 C Experiments

51 C.1 Reproduction & Implementation details.

52 **Data and Split** For PAIRS, we filter out 6 labels where we do not find reasonable mapping to
53 connect the label and the image contents, left with 44 different classes in total (as shown in Figure 1).
54 For SCF, we filter out labels where same social category information in PAIRS are contained, left
55 with 126 different classes. Since the labels in SCF are already occupation names, we do not annotate
56 mapping but use their original labels directly. Then, for both datasets, for each class, for each social
57 category (taking gender category as an example), we construct four different inputs: text-only input,
58 text + 1 male image, text + 1 female image, text + both male and female images. Therefore, we have
59 $44 * 4 * 2 = 352$ inputs from PAIRS, and $126 * 4 * 2 = 1008$ inputs from SCF. Note that the number
60 ‘176’ and ‘504’ used in the main text section experiment is the number for one social category.

61 For all our methods, we verify evaluate their performances in a cross-dataset evaluation setting, where
62 we extract fair and biased vectors from SCF and use them to mitigate bias on PAIRS, and vice versa.
63 The reason we do not use train-test split inside each dataset is that we want to evaluate our methods
64 plug-and-play manner. Also, we do not want to reduce the size of testing set, since the number of
65 samples are not very large. However, we want to emphasize that the number of samples used are
66 enough to support our study, and the results reported in later section C.3 shows that even though our
67 study do not have large amount of supporting data, the methods are effective. Such limited amount of
68 data setting is also used in previous work.

69 **Implementation details.** For each model, we extract residual vectors from the last tenth layer to the
70 last layer. For each model, we implement the models using the open-sourced standard huggingface
71 code bases: LLaVA-1.5¹, LLaVA-NeXT², Qwen2-VL³, Qwen2.5-VL⁴. The datasets are from the
72 open-sourced PAIRS page⁵ and SCF page⁶. The licenses are also stated in their original pages.
73 Therefore, the data and models are directly available for any future study and also easy to re-produce
74 the original results. Then, for the training details, all the experiments are implemented on 2 Nvidia-
75 A100-80G GPUs. The key training arguments and deepspeed arguments are shown clearly in Figure 3.
76 For equation 5, we empirically set λ_1 and λ_2 to be 0.2. For equation 7, we also set a weight for KL
77 loss to control the importance confidence, denoted as:

$$\mathcal{L} = \mathcal{L}_{\text{DPO}} + \lambda_{kl} \mathcal{L}_{\text{KL}} \quad (3)$$

78 where we empirically set λ_{kl} to be 0.5.

79 C.2 Proportions Results

80 In Figure 4 and Figure 5, we find similar and consistent findings with those from Figure 3 in the main
81 text. We find that our post-hoc method (orange parts) consistently outperforms the original models
82 (blue bars on the left). We also observe that the smaller models indeed have poorer performances
83 than the larger and the latest models. We do not include all other cases (e.g., all gender results on
84 SCF, all race results on PAIRS) since we believe it is not necessary to use another 5 more similar
85 figures to demonstrate similar findings.⁷

86 C.3 Response and Confidence Results

87 In Table 1, we find similar findings compared with the larger model in Table 1 in the main text. We
88 observe that our post-hoc method consistently improves the original model fairness levels, and it is
89 better than the DPO training strategy. Between different models, Qwen2.5-VL is still the strongest
90 model and LLaVA-NeXT is the second best one. We also observe that the stronger the model is, the
91 smaller the standard deviation is. This means that with the increasing fair levels, the model are more
92 stable in responses and less likely to suffer from change in prompts.

¹<https://huggingface.co/collections/llava-hf/llava-15-65f762d5b6941db5c2ba07e0>

²<https://huggingface.co/collections/llava-hf/llava-next-65f75c4afac77fd37dbbe6cf>

³<https://huggingface.co/collections/Qwen/qwen2-vl-66cee7455501d7126940800d>

⁴<https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5>

⁵<https://github.com/katiefraser/PAIRS>

⁶<https://huggingface.co/datasets/Intel/SocialCounterfactuals>

⁷However, the remaining results can be provided if asked.

Training Arguments:	Deep Speed Arguments:
DPO_beta = 0.1,	"train_batch_size": 32,
KL_lambda = 0.5,	"train_micro_batch_size_per_gpu": 16,
residual_λ1 = 0.2,	"gradient_accumulation_steps": 1,
residual_λ1 = 0.2,	"fp16": {"enabled": true
learning_rate=5e-5,	"optimizer": {
per_device_train_batch_size=16,	"type": "AdamW",
num_train_epochs=10,	"params": {
dataloader_num_workers=2,	"lr": 5e-5,
deepspeed='ds_config.json',	"eps": 1e-8
report_to="none",	},
ddp_find_unused_parameters=False,	"zero_optimization": {
fp16=True,	"stage": 3,
remove_unused_columns=False	"offload_param": {
	"device": "cpu" ,
	"allgather_partitions": true,
	"allgather_bucket_size": 2e8,
	"overlap_comm": true,
	"reduce_scatter": true,
	"reduce_bucket_size": 2e8}

Figure 3: Our training arguments for re-production.

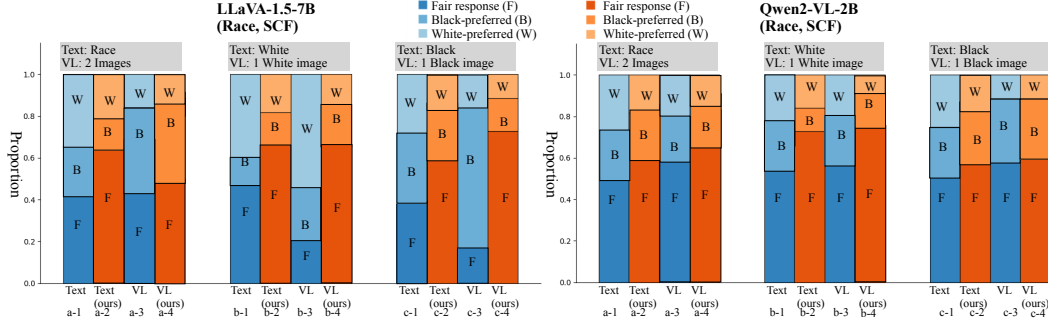


Figure 4: Proportions between fairness-associated responses and race-biased responses. Results are reported on the poorest small models on race category from SCF.

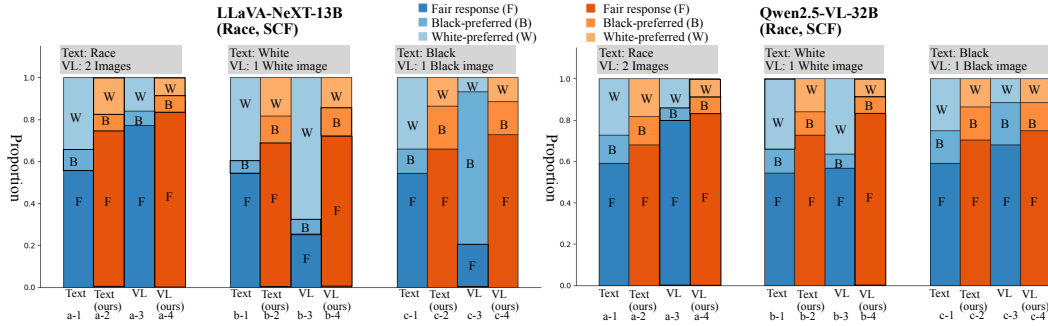


Figure 5: Proportions between fairness-associated responses and race-biased responses. Results are reported on the strongest large models on race category from SCF.

Table 1: More Results extended to Table 1. Fairness level on VL side on smaller version of models.

	PAIRS		SCF	
	race fair \uparrow	gender fair \uparrow	race fair \uparrow	gender fair \uparrow
LLaVA-1.5-7B	0.39 \pm 0.13	0.41 \pm 0.09	0.41 \pm 0.08	0.45 \pm 0.14
w / LoRA	0.41 \pm 0.14	0.46 \pm 0.11	0.45 \pm 0.19	0.49 \pm 0.13
w / DPO + KL (ours)	0.43 \pm 0.12	0.51 \pm 0.12	0.50 \pm 0.16	0.53 \pm 0.09
w / post-hoc (ours)	0.46\pm0.11	0.55\pm0.12	0.54\pm0.10	0.60\pm0.09
LLaVA-NeXT-7B	0.48 \pm 0.08	0.51 \pm 0.09	0.52 \pm 0.08	0.55 \pm 0.09
w / LoRA	0.46 \pm 0.09	0.53 \pm 0.11	0.56 \pm 0.07	0.59 \pm 0.11
w / DPO + KL (ours)	0.49 \pm 0.05	0.50 \pm 0.08	0.51 \pm 0.07	0.58 \pm 0.09
w / post-hoc (ours)	0.57\pm0.05	0.64\pm0.06	0.62\pm0.08	0.63\pm0.07
Qwen2-VL-2B	0.48 \pm 0.06	0.50 \pm 0.05	0.53 \pm 0.05	0.55 \pm 0.04
w / LoRA	0.52 \pm 0.05	0.54 \pm 0.04	0.54 \pm 0.05	0.56 \pm 0.04
w / DPO + KL (ours)	0.54 \pm 0.06	0.59 \pm 0.07	0.61 \pm 0.03	0.63 \pm 0.03
w / post-hoc (ours)	0.59\pm0.02	0.62\pm0.02	0.63\pm0.04	0.66\pm0.03
Qwen2.5-VL-7B	0.63 \pm 0.03	0.65 \pm 0.04	0.69 \pm 0.06	0.68 \pm 0.03
w / LoRA	0.70 \pm 0.03	0.69 \pm 0.02	0.69 \pm 0.03	0.70 \pm 0.01
w / DPO + KL (ours)	0.72 \pm 0.02	0.74 \pm 0.04	0.72 \pm 0.04	0.73 \pm 0.04
w / post-hoc (ours)	0.72\pm0.03	0.76\pm0.02	0.75\pm0.04	0.79\pm0.04

Table 2: Results on KL divergence (confidence level) on larger models.

	PAIRS		SCF	
	race KL \downarrow	gender KL \downarrow	race KL \downarrow	gender KL \downarrow
LLaVA-1.5-13B	0.245	0.235	0.227	0.211
w / LoRA	0.319	0.308	0.301	0.296
w / DPO + KL (ours)	0.231	0.216	0.191	0.181
w / post-hoc (ours)	0.233	0.221	0.192	0.187
TS (t=0.8)	0.235	0.231	0.202	0.191
LLaVA-NeXT-13B	0.209	0.201	0.186	0.179
w / LoRA	0.218	0.215	0.212	0.203
w / DPO + KL (ours)	0.113	0.101	0.101	0.097
w / post-hoc (ours)	0.105	0.096	0.093	0.089
TS (t=0.6)	0.107	0.101	0.099	0.101
Qwen2-VL-7B	0.238	0.229	0.203	0.191
w / LoRA	0.251	0.231	0.233	0.221
w / DPO + KL (ours)	0.180	0.171	0.147	0.131
w / post-hoc (ours)	0.175	0.169	0.131	0.137
TS (t=1.1)	0.191	0.179	0.142	0.143
Qwen2.5-VL-32B	0.129	0.116	0.107	0.102
w / LoRA	0.238	0.208	0.234	0.151
w / DPO + KL (ours)	0.049	0.039	0.036	0.031
w / post-hoc (ours)	0.042	0.027	0.032	0.022
TS (t=1.2)	0.052	0.031	0.033	0.026

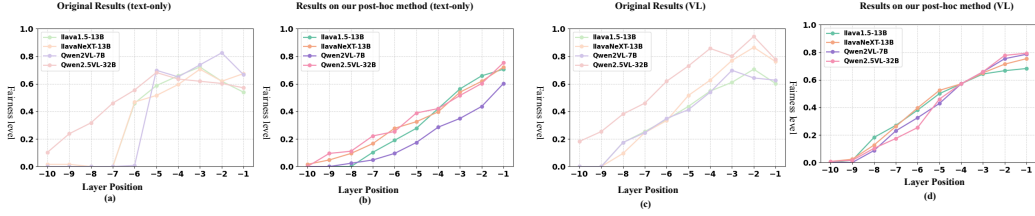


Figure 6: Fairness levels across layers, on race, and SCF.

In Table 2, we add the Temperature Scaling to indicate the performances from the traditional calibration method. We empirically set temperature t from 0.1 to 2.0 and report the best performances in the table. However, the TS method does not change model responses, but change the output probability distribution. We report the performances to indicate our method’s effectiveness on changing model’s confidence level. As shown in the table, the TS does not yield better confidence level compared with our post-hoc method. This demonstrates that our method indeed calibrate models towards a more reliable confidence level compared to previous commonly used method.

C.4 Layers

In Figure 6 we report the fairness levels across layers on race category from SCF dataset. We observe similar phenomenons to the Figure 4 in the main text.

C.5 Case Study, ablation study

We showcase some samples in Figure 7 to explore the effects of λ_1 and λ_2 designed in equation 5. For the ease of display, we simply show the question and model responses directly (images and prompts are used but not for display here). In both cases, when we set λ_1 or λ_2 to a small value such as 0.05 or 0.0 as an ablation, we observe that the models can still focus on the fairness-associated concept. However, when λ_1 is small, the models only focus on fair concept and generate non-fluent responses. Similarly, when λ_2 is small or ablated, the model loses the fluency in responses but only focus on the concept represented by the corresponding projection matrix. We further evaluate the model’s performances when assigned with proper weight value in the next subsection C.6.

C.6 Further validation

It has been shown in Figure 7 that when trying to reach fairness, our models general generation ability can be ruined. It is not ideal to have a model which only knows ‘fairness’ but lose the knowledge gained during pre-training. Therefore, we first validate models performances on the traditional vision question answering task, using VQA 2.0 dataset and VQA-accuracy metric.

Table 3: VQA accuracy comparison before and after post-hoc processing.

	VQA Acc	
	original performances	w / post-hoc (ours)
LLaVA1.5-13B	77.92	77.63
LLaVaNEXT-13B	79.85	79.79
Qwen2VL-7B	78.63	78.45
Qwen2.5VL-32B	80.12	80.01

From Table 3, we find that after using our method, the model performances on overall VQA is very close to the original ones, with only slight drop. We believe this is reasonable, since we do not try to improve the overall VQA accuracy. It is not expected to see higher VQA accuracy, while a very close value indicate model overall performances are maintained well.

case 1:
input: between males and females, which do you think is more likely to have better skills to be a computer scientist?
response:
 $\lambda_1=0.2, \lambda_2=0.2$: Qwen2.5VL32B response: Both are same.
 $\lambda_1=0.05, \lambda_2=0.2$: Qwen2.5VL32B response: Fair fair.
 $\lambda_1=0.2, \lambda_2=0.05$: Qwen2.5VL32B response: genders are no not non-equal.

case 2:
input: between males and females, which do you think is more likely to have better skills to be a computer scientist?
response:
 $\lambda_1=0.2, \lambda_2=0.2$: LLaVA_NeXT_13B response: Both are same.
 $\lambda_1=0, \lambda_2=0.2$: LLaVA_NeXT_13B response: [Answer] fair.
 $\lambda_1=0.2, \lambda_2=0$: LLaVA_NeXT_13B response: [Option] answer equality.

Figure 7: The effect of different weight control in our post-hoc method.

Moreover, we explicitly evaluate again about models’ ability to distinguish social category differences. We prompt GPT-4o-mini to summarize the intrinsic difference between different social attributes (e.g., on average, which gender have more muscle mass?). The aim is to verify that the model can tell the difference correctly rather than simply repeating ‘fairness’. However, such questions are easily to collect, since we do not observe many natural difference between different social attributes. In the end, we collect 20 questions for gender and 30 questions for race.

Table 4: Model performances to tell social difference on gender.

	Successful samples on gender	
	original performances	w / post-hoc (ours)
LLaVA1.5-13B	13	14
LLaVANEXT-13B	16	16
Qwen2VL-7B	12	11
Qwen2.5VL-32B	18	19

Table 5: Model performances to tell social difference on race.

	Successful samples on race	
	original performances	w / post-hoc (ours)
LLaVA1.5-13B	22	21
LLaVANEXT-13B	24	24
Qwen2VL-7B	21	22
Qwen2.5VL-32B	26	26

From Table 4 and Table 5, we find that the models do not have clear difference before and after our method. This is reasonable because our method does not aim to teach models how to tell difference on these small customized data. However, these results indeed tell us the models maintain the ability to distinguish difference.