A EMPIRICAL VALIDATION ASSUMPTION 1

We use ERM to train a ViT on the CelebA dataset. We save the model that achieves the highest validation accuracy. From the last encoder, we extract $\mathbf{q}_{\text{CLS}}$ and $\mathbf{K}_{10}$, where "CLS" refers to a special token and $\mathbf{K}_{10}$ represents the key vector of the 10-th token. We randomly select one dimension from $\mathbf{q}_{\text{CLS}}$ and $\mathbf{K}_{10}$ to generate two conditional distribution plots based on a.

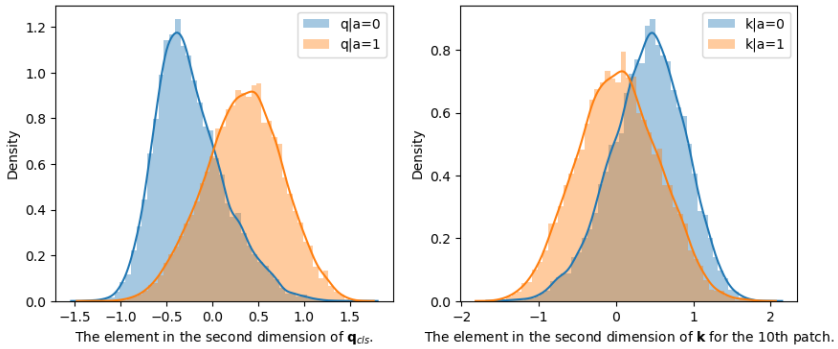

Figure 3: Histogram of a random selected dimension of $\mathbf{q}_{\text{cls}}$ and $\mathbf{k}$

B PROOF FOR THEROREM 1

We first consider a 1-D case and next generalize the derivation to d_k dimensions.

For the 1-D derivation, now we focus only on q, then

$$q \sim \lambda_0 \mathcal{N}(\mu_0, \sigma_0^2) + \lambda_1 \mathcal{N}(\mu_1, \sigma_1^2) \tag{9}$$

where $\lambda_0 + \lambda_1 = 1$. Then the expectation of q is

$$\mu := \mathbb{E}[q] = \lambda_0 \mu_0 + \lambda_1 \mu_1 \tag{10}$$

and the variance can be written as

$$\begin{aligned}
\sigma^2 := \text{Var}[q] =& \mathbb{E}[q^2] - (\mathbb{E}[q])^2, && (11)\\
=& \lambda_0 \mathbb{E}_{\mathcal{N}(\mu_0, \sigma_0^2)}[q^2] + \lambda_1 \mathbb{E}_{\mathcal{N}(\mu_1, \sigma_1^2)}[q^2] - (\lambda_0 \mu_0 + \lambda_1 \mu_1)^2 && (12)\\
=& \lambda_0(\mu_0^2 + \sigma_0^2) + \lambda_1(\mu_1^2 + \sigma_1^2) - (\lambda_0 \mu_0 + \lambda_1 \mu_1)^2 && (13)\\
=& (\lambda_0 \sigma_0^2 + \lambda_1 \sigma_1^2) + \lambda_0 \lambda_1 (\mu_0 - \mu_1)^2 && (14)\\
=& (\lambda_0 \sigma_0^2 + \lambda_1 \sigma_1^2) + \lambda_0 \lambda_1 \Delta^2 \quad \text{WLOG, let } \Delta = \mu_0 - \mu_1 > 0 && (15)
\end{aligned}$$

After normalization with $q^{norm} = \frac{q - \mathbb{E}[q]}{\sqrt{\text{Var}[q]}}$, the conditioned variable $q^{norm}|a = 0$ and $q^{norm}|a = 1$ are still Gaussian, and the conditioned expectations can be written as

$$\mu_{0,norm} := \mathbb{E}[q^{norm}|a = 0] = \frac{\mu_0 - \mu}{\sigma} = \frac{\lambda_1 \Delta}{\sqrt{(\lambda_0 \sigma_0^2 + \lambda_1 \sigma_1^2) + \lambda_0 \lambda_1 \Delta^2}} \tag{16}$$

$$\mu_{1,norm} := \mathbb{E}[q^{norm}|a = 1] = \frac{\mu_1 - \mu}{\sigma} = -\frac{\lambda_0 \Delta}{\sqrt{(\lambda_0 \sigma_0^2 + \lambda_1 \sigma_1^2) + \lambda_0 \lambda_1 \Delta^2}} \tag{17}$$

and the corresponding variances are

$$\sigma_{i,norm} = \frac{\sigma_i}{\sigma}, \quad i = 0, 1 \tag{18}$$

After taking the absolute values, that is $q^{de} = |q^{norm}|$, the conditioned expectations can be written as

$$\mathbb{E}[q^{de}|a=i] = -\int_{-\infty}^{0} tp(t)\mathrm{d}t + \int_{0}^{\infty} tp(t)\mathrm{d}t, \quad p(t) = \mathcal{N}(\mu_{i,norm}, \sigma_{i,norm}^2), i = 0, 1 \tag{19}$$

$$= \sigma_{i,norm}\sqrt{\frac{2}{\pi}}\exp(-\frac{\mu_{i,norm}^2}{2\sigma_{i,norm}^2}) + \mu_{i,norm}\mathrm{erf}(\frac{\mu_{i,norm}}{\sqrt{2}\sigma_{i,norm}}) \tag{20}$$

$$\tag{21}$$

Recall that $\mu_{i,norm} = \frac{\mu_i - \mu}{\sigma}$ and $\sigma_{i,norm} = \sigma_i/\sigma$, we have

$$\mathbb{E}[q^{de}|a=i] = \frac{\sigma_i}{\sigma}\sqrt{\frac{2}{\pi}}\exp(-\frac{(\frac{\mu_i-\mu}{\sigma})^2}{2(\frac{\sigma_i}{\sigma})^2}) + \frac{\mu_i-\mu}{\sigma}\mathrm{erf}(\frac{\frac{\mu_i-\mu}{\sigma}}{\sqrt{2}\frac{\sigma_i}{\sigma}}) \tag{22}$$

$$= \frac{1}{\sigma}\Big[\sigma_i\sqrt{\frac{2}{\pi}}\exp(-(\frac{\mu_i-\mu}{\sqrt{2}\sigma_i})^2) + (\mu_i-\mu)\mathrm{erf}(\frac{\mu_i-\mu}{\sqrt{2}\sigma_i})\Big] \tag{23}$$

$$= \frac{1}{\sigma}\Big[\sigma_i\sqrt{\frac{2}{\pi}}\exp(-(\frac{\lambda_{1-i}\Delta}{\sqrt{2}\sigma_i})^2) + (\lambda_{1-i}\Delta)\mathrm{erf}(\frac{\lambda_{1-i}\Delta}{\sqrt{2}\sigma_i})\Big] \tag{24}$$

which is because $\mu_1 - \mu = \mu_1 - \lambda_1\mu_1 - \lambda_0\mu_0 = \lambda_0(\mu_1 - \mu_0) = -\lambda_0\Delta$ and $\mu_0 - \mu = \mu_0 - \lambda_1\mu_1 - \lambda_0\mu_0 = \lambda_1(\mu_0 - \mu_1) = \lambda_1\Delta$. And $(-x)\mathrm{erf}(-x) = x\mathrm{erf}(x)$.

Note that $\frac{\lambda_{1-i}\Delta}{\sqrt{2}\sigma_i} > 0$, we have $\exp(-(\frac{\lambda_{1-i}\Delta}{\sqrt{2}\sigma_i})^2) < 1$ and $\mathrm{erf}(\frac{\lambda_{1-i}\Delta}{\sqrt{2}\sigma_i}) < 1$. As a result, the expectation can be bounded by

$$\mathbb{E}[q^{de}|a=i] \leq \frac{1}{\sigma}\Big[\sigma_i\sqrt{\frac{2}{\pi}} + \lambda_{1-i}\Delta\Big] \tag{25}$$

$$= \frac{\sigma_i\sqrt{\frac{2}{\pi}}}{\sqrt{(\lambda_0\sigma_0^2 + \lambda_1\sigma_1^2) + \lambda_0\lambda_1\Delta^2}} + \frac{\lambda_{1-i}\Delta}{\sqrt{(\lambda_0\sigma_0^2 + \lambda_1\sigma_1^2) + \lambda_0\lambda_1\Delta^2}} \tag{26}$$

$$\leq \lim_{\sigma_i\to\infty}\frac{\sigma_i\sqrt{\frac{2}{\pi}}}{\sqrt{(\lambda_0\sigma_0^2 + \lambda_1\sigma_1^2) + \lambda_0\lambda_1\Delta^2}} + \lim_{\Delta\to\infty}\frac{\lambda_{1-i}\Delta}{\sqrt{(\lambda_0\sigma_0^2 + \lambda_1\sigma_1^2) + \lambda_0\lambda_1\Delta^2}} \tag{27}$$

$$= \sqrt{\frac{2}{\pi\lambda_i}} + \sqrt{\frac{\lambda_{1-i}}{\lambda_i}} \tag{28}$$

Similar results hold for $\mathbf{k}$, too. Now the original statement can be re-written as

$$|\mu_{q_0,de}\mu_{k_0,de} - \mu_{q_1,de}\mu_{k_1,de}| \tag{29}$$

$$= \big|\mathbb{E}[q^{de}|a=0]\mathbb{E}[k^{de}|a=0] - \mathbb{E}[q^{de}|a=1]\mathbb{E}[k^{de}|a=1]\big| \tag{30}$$

$$\leq \mathbb{E}[q^{de}|a=0]\mathbb{E}[k^{de}|a=0] + \mathbb{E}[q^{de}|a=1]\mathbb{E}[k^{de}|a=1] \tag{31}$$

$$\leq \Big(\sqrt{\frac{2}{\pi\lambda_0}} + \sqrt{\frac{\lambda_1}{\lambda_0}}\Big)^2 + \Big(\sqrt{\frac{2}{\pi\lambda_1}} + \sqrt{\frac{\lambda_0}{\lambda_1}}\Big)^2 \tag{32}$$

Now consider the high-dimensional scenario, we have

$$|\delta^{\text{de}}| = \left| \mathbb{E}[\frac{\mathbf{q}^{\text{de}\top}\mathbf{k}^{\text{de}}}{\sqrt{d_k}}|a=0] - \mathbb{E}[\frac{\mathbf{q}^{\text{de}\top}\mathbf{k}^{\text{de}}}{\sqrt{d_k}}|a=1] \right| \tag{33}$$

$$= \frac{1}{\sqrt{d_k}} \left| \mathbb{E}[\sum_{j=1}^{d_k}(q_j^{de}k_j^{de})|a=0] - \mathbb{E}[\sum_{j=1}^{d_k}(q_j^{de}k_j^{de})|a=1] \right| \tag{34}$$

$$= \frac{1}{\sqrt{d_k}} \left| \sum_{j=1}^{d_k} \mathbb{E}[q_j^{de}k_j^{de}|a=0] - \sum_{j=1}^{d_k} \mathbb{E}[q_j^{de}k_j^{de}|a=1] \right| \tag{35}$$

$$= \frac{1}{\sqrt{d_k}} \left| \sum_{j=1}^{d_k} (\mathbb{E}[q_j^{de}k_j^{de}|a=0] - \mathbb{E}[q_j^{de}k_j^{de}|a=1]) \right| \tag{36}$$

$$\leq \frac{1}{\sqrt{d_k}} \sum_{j=1}^{d_k} \left| \mathbb{E}[q_j^{de}k_j^{de}|a=0] - \mathbb{E}[q_j^{de}k_j^{de}|a=1] \right| \tag{37}$$

$$= \frac{1}{\sqrt{d_k}} \sum_{j=1}^{d_k} \left| \mathbb{E}[q_j^{de}|a=0]\mathbb{E}[k_j^{de}|a=0] - \mathbb{E}[q_j^{de}|a=1]\mathbb{E}[k_j^{de}|a=1] \right| \tag{38}$$

$$\leq \sqrt{d_k} \Big[\big(\sqrt{\frac{2}{\pi\lambda_0}} + \sqrt{\frac{\lambda_1}{\lambda_0}}\big)^2 + \big(\sqrt{\frac{2}{\pi\lambda_1}} + \sqrt{\frac{\lambda_0}{\lambda_1}}\big)^2 \Big]. \square \tag{39}$$

C IMPLEMENTATION DETAILS

We summarize the implementation details as follows:

- **ERM**: We take the transformer with 8 stack encoders as the baseline model. Each encoder has 8 attention heads. We implement the ERM model using the Huggingface library without pre-trained weights. We resize the image from CelebA and UTK datasets to 64×64, and divide it into 16 patches. For CelebA and UTK, we take the AdamW as the optimizer with a learning rate of 10^{-4}, and no scheduler is applied for the fair comparison. For NLP tasks, we take the AdamW as the optimizer with a learning rate of 10^{-5}. We share all methods with the same configuration as the ERM model.
- Distributionally robust optimization (**DRO**): We adapt the backbone to the same as the ERM model. We tune the hyper-parameter η at the validation set to achieve the highest accuracy. For CelebA and UTK experiments, we set $\eta = 0.15$ and $\eta = 0.10$ respectively. For HateXplain and MultiNLI, we set $\eta = 0.25$.
- Adversarially reweighted learning (**ARL**): We adapt the learner network to the same as the ERM model. For the adversary network, we apply a 6-layer stack encoder in the transformer. This is a smaller configuration compared to the learner network, as recommended by the authors.
- Fairness without demographics through knowledge distillation (**KD**): We adapt the student model that is the same as the ERM model. For the teacher network, we follow the suggestion of the authors that use a larger network, hence we adopt the vision transformer with 12 stacked encoder layers. The student network is trained by the output of the teacher model with softmax activation.
- Just train twice (**JTT**): We adapt the backbone network is the same as the ERM model. For CelebA, we follow the suggestion in the paper choose the number of epochs of training the identification model $T = 1$. During the second training, we choose the upsampling factor $\lambda_{\text{up}} = 50$. For UTK, we set $T = 10$, $\lambda_{up} = 50$. For HateXplain, we set $T = 20$, $\lambda_{up} = 50$. For MultiNLI, we set $T = 2$, $\lambda_{up} = 20$.
- Learning from failure (**LfF**): We take the networks for a biased model and a debiased model are the same as the ERM model. We set the amplification coefficient in generalized cross entropy loss $q = 0.7$ as suggested in the original paper.

- **Ours**: For vision tasks, our method employs a backbone identical to the ERM model, which comprises an 8-stacked encoder vision transformer. In the final encoder, we normalize and apply the absolute value to $\mathbf{q}$ and $\mathbf{k}$ for each head. From this encoder layer, we select the $\mathbf{v}$ vectors associated with the two highest attention weights and execute a local alignment. For NLP tasks, we extend a pre-trained model with an additional encoder layer, applying our method specifically to this added layer. Similarly, in the NLP tasks, we choose the $\mathbf{v}$ vectors with the two highest attention weights for alignment.

For all methods, we maintain a consistent batch size. Specifically, for vision tasks, the batch size is set to 256. For BERT Large, it's 32, and for BERT Base, we use a batch size of 64.

D EVALUATION METRICS

The group fairness metrics are measurements of the performance of different sensitive groups. We focus on three specific metrics: Demographic Parity, Equal Opportunity, and Equalized Odds. Demographic Parity (DP): DP focuses on the equality of the outcomes across different demographic groups, regardless of their abilities. It ensures that each group receives positive outcomes at the same rate. Equal Opportunity (EOp): EOp ensures that samples who should receive a positive outcome have an equal chance of being correctly identified, regardless of their group. Equalized Odds (EOd): EOd requires both that samples that should receive a positive outcome have an equal chance of being correctly identified (like EOp), and also that samples that should receive a negative outcome have an equal chance of being correctly identified, across all sensitive groups. The computations for the fairness metrics are as follows:

$$\begin{aligned} \text{DP} &= |PP_i - PP_j|, \quad \text{EOp} = |TPR_i - TPR_j|, \\ \text{EOd} &= \frac{1}{2}(|TPR_i - TPR_j| + |FPR_i - FPR_j|), \quad i, j \in \mathcal{A} \end{aligned} \tag{40}$$

where PP, TPR, and FPR are the positive prediction rate, the true positive rate, and the false positive rate.

E ATTENTION WEIGHT VISUALIZATION

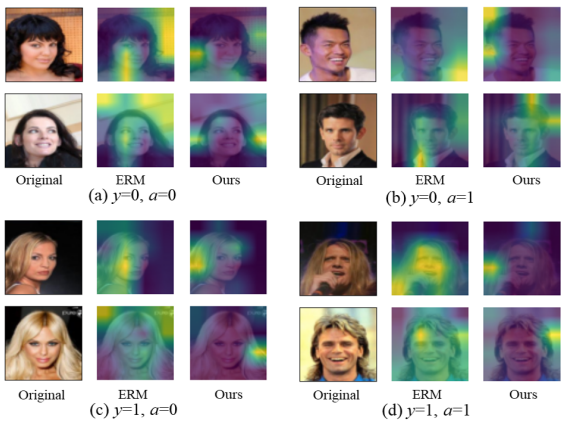

Figure 4: Visualization of attention weight. y =blond hair, a =male in CelebA.

We present a visualization of the average attention weight in the last encoder layer for both the ERM and our proposed models in Fig 4. We opted for models that demonstrated the highest validation accuracy. The ERM model achieved an accuracy of 94.27%, our model achieved an accuracy of 94.08%. We observe inconsistencies in attention allocation using the ERM training objective function. Despite its high accuracy, the ERM model predominantly focuses on facial features, failing to distribute attention adequately. In contrast, our model provides a more uniform attention allocation, effectively reducing the focus on facial features or other irrelevant features.

F ABLATION STUDY

We perform an ablation experiment from three perspectives to understand their importance. **w/o local alignment**: We train our model without incorporating the local value alignment technique. The training process is solely guided by the cross-entropy loss. **w/o debias attention**: While training our model, we exclude the normalization and exclude the absolute values in vectors $\mathbf{q}$ and $\mathbf{k}$ when calculating the attention weight. **w/o absolute value**: We train our model using normalized vectors $\mathbf{q}$ and $\mathbf{k}$ to compute the attention weight, but we exclude applying the absolute value on these vectors. We use CelebA dataset with a =male and y =Blond hair. All the methods share the same seed with the Baseline model.

Method	DP↓	EOp ↓	EOd ↓	ACC ↑
Baseline	16.99	43.04	23.05	94.20
w/o local alignment	18.70	39.59	21.72	93.94
w/o debias attention	16.59	43.36	22.92	**94.63**
w/o absolute value	18.24	37.44	20.59	94.39
Ours	**15.81**	**36.49**	**19.47**	93.96

Table 7: Ablation study on CelebA dataset.

Table 7 demonstrates the efficacy of our design modules. Notably, without the debias attention, the outcomes are close to those of the Baseline model. This highlights the significance of debias attention as an essential component for debiasing attention. Concurrently, the local value alignment technique further improves fairness, with a large impact on EOd. When the network integrates both techniques, it achieves optimal fairness with a slight drop in accuracy.