FADES: Fair Disentanglement with Sensitive Relevance

Supplementary Material

7. Contradiction on other methods

Here, we extend our claim of the conflict between the two objectives to invariant learning and correlation-aware learning as we discussed in Section 3 in main paper. Under the data bias where Y and A are not independent, it can be inferred that $\mathbf{z}_X$ and A will also be dependent when perfect predictiveness of Y from $\mathbf{z}_X$ is achieved. Thus, the predictiveness goal would conflict with the fairness objective, which is to learn $\mathbf{z}$ that is invariant with respect to the sensitive information A.

In correlation-aware learning, the goal is to disentangle the latent code into $\mathbf{z}_d$ and $\mathbf{z}_r$, which corresponds to the two sets of features: X_d which are descendants of A; and X_r which are irrelevant to A. For instance, if A is defined as *gender*, the descendant attributes would be defined as $X_d = \{\text{Makeup, Mustache}, \cdots\}$, and irrelevant attributes as $X_r = \{\text{Age}, \cdots\}$. Then the predictiveness objective is to recover X_d and X_r from corresponding latent codes, and the fairness objective is to learn two codes that are independent. In this scenario, it is possible that some features in X_d and X_r have causal relationships. For instance, *Age* is also the cause of *Mustache* along with *gender*, which leads to a correlation between A and X_r. As a result, it may not be possible to achieve both the independence of latent codes $\mathbf{z}_d$ and $\mathbf{z}_r$ and perfect recovery of X_d and X_r at the same time, as these objectives contradict each other. The conflict between performance and fairness objectives of invariant learning and correlation-aware learning is depicted in Figure 7.

8. Theoretical Analysis

Continuing from the setting we are interested in, we assume that target labels and sensitive information are not perfectly separable, making it more applicable to real-world situations. Although trade-offs and inherent correlations exist, prior research primarily focused on learning fair representations by requiring the latent representation to be independent of the sensitive attribute. Consequently, the dual objectives of fairness and performance are in conflict with each other, as illustrated in Figure 2 of main paper. Since two objectives (Goal1 and Goal2) are in conflict with each other, achieving the global optimum for each objective simultaneously becomes infeasible.

Here, we provide informal proof as the theoretical analysis of the impossibility of invariant learning (Fig 1-(a)). Given the fact that $Y \not\perp A$, let's first assume we trained the optimal classifier that predicts Y from Z by minimizing cross-entropy loss (Goal 1). Since the cross-entropy loss is lower bounded by conditional entropy as

$$CE(Y, \hat{Y}) = H_Y(\hat{Y}) \geq H(Y|\hat{Y}), \tag{9}$$

when optimal classifier f_θ^* minimizes cross-entropy loss, we can say $Y \approx \hat{Y}$. Also, this implies maximizing mutual information between Z and $\hat{Y}$ (or Y), which yields

$$I_\theta(Z; \hat{Y}) = H_\theta(\hat{Y}) - H_\theta(\hat{Y}|Z) \approx H_\theta(\hat{Y}), \tag{10}$$

which suggests $H_\theta(\hat{Y}|Z) \approx 0$, meaning that $\hat{Y}$ is deterministic with the optimal classifier f_θ^* given Z.

Since $\hat{Y}$ is solely determined by Z,

$$P(\hat{Y}|Z, A) = P(\hat{Y}|Z). \tag{11}$$

If we assume that we can learn Z that is independent of A (Goal 2), we have

$$P(Z|A) = P(Z). \tag{12}$$

When we multiply two equations on each side, we have

$$\begin{aligned} & P(\hat{Y}|Z, A)P(Z|A) && = P(\hat{Y}|Z)P(Z) \\ \Leftrightarrow \quad & P(\hat{Y}, Z|A) && = P(\hat{Y}, Z) \\ \Leftrightarrow \quad & \textstyle\int_Z P(\hat{Y}, Z|A) && = \displaystyle\int_Z P(\hat{Y}, Z) \\ \Leftrightarrow \quad & P(\hat{Y}|A) && = P(\hat{Y}), \end{aligned}$$

which implies that $\hat{Y}$ and A are independent, contradicting the fact that $Y \not\perp A$, since $\hat{Y} \approx Y$.

Also for disentanglement learning (Fig 1-(b)), we prove the case without the connection between Y and $\mathbf{z}_A$ [7,8] for simplicity. Similar to the proof of invariant learning scenario, let's assume we have optimal classifiers $f_{\theta_y}^*$ and $f_{\theta_a}^*$ that perfectly predicts Y and A from $\mathbf{z}_X$ and $\mathbf{z}_A$ (Goal 1), respectively, such that $\hat{Y}$ (resp. $\hat{A}$) is solely determined by $\mathbf{z}_X$ (resp. $\mathbf{z}_A$).

Then we can write

$$P(\hat{Y}|\mathbf{z}_X, \hat{A}) = P(\hat{Y}|\mathbf{z}_X), \tag{13}$$

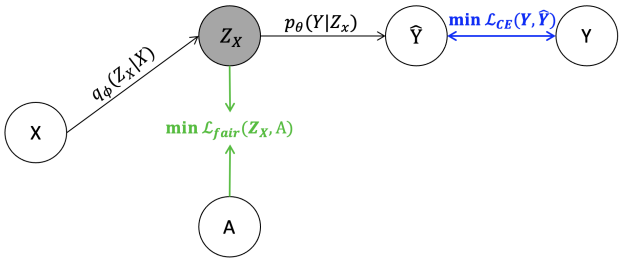
since $\hat{Y}$ is determined only by $\mathbf{z}_X$ regardless of $\hat{A}$.

If we assume fairness objective is satisfied (Goal 2), we have

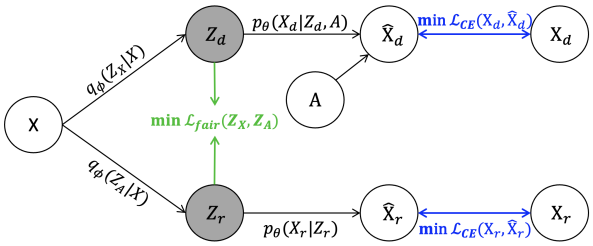
$$P(\mathbf{z}_X|\mathbf{z}_A) = P(\mathbf{z}_A). \tag{14}$$

Since $\hat{A}$ is deterministic given $\mathbf{z}_A$, and $\mathbf{z}_A \perp \mathbf{z}_X$, we can naturally say that $\hat{A} \perp \mathbf{z}_X$, that is

$$P(\mathbf{z}_X|\hat{A}) = P(\mathbf{z}_X). \tag{15}$$



(a) Data flow of invariant learning



(b) Data flow of correlation-aware learning

Figure 7. Illustration of the contradiction between fairness and performance goals in different methods. The methods of invariant learning and correlation-aware learning are in conflict when the feature to be reconstructed is not independent of the sensitive attribute. This can lead to a trade-off between fairness and performance, which can make it difficult to achieve both goals simultaneously.

When we multiply (13) and (15) on both sides, we have

$$\begin{aligned} & P(\hat{Y}|\mathbf{z}_X, \hat{A})P(\mathbf{z}_X|\hat{A}) && = P(\hat{Y}|\mathbf{z}_X)P(\mathbf{z}_X) \\ \Leftrightarrow \quad & \textstyle\int_{\mathbf{z}_X} P(\hat{Y}, \mathbf{z}_X|A) && = \displaystyle\int_{\mathbf{z}_X} P(\hat{Y}, \mathbf{z}_X) \\ \Leftrightarrow \quad & P(\hat{Y}|A) && = P(\hat{Y}), \end{aligned}$$

which implies that $\hat{Y}$ and A are independent, contradicting the fact that $Y \not\perp A$, since $\hat{Y} \approx Y$ and $\hat{A} \approx A$.

This establishes the theoretical foundation that the fairness and performance goals, which have been extensively applied in prior research, are inherently conflicting with each other.

9. Relation between conditional mutual information and conditional independence

Here, we prove zero conditional mutual information indicates conditional independence. Let us first define

$$F(x, y) := \frac{P(x|z)P(y|z)}{P(x, y|z)}. \tag{16}$$

Then we have

$$\sum_{x,y} P(x, y|z)F(x, y) = \sum_{x,y} P(x|z)P(y|z) = 1. \tag{17}$$

We can rewrite the conditional mutual information with $F(x, y)$ as

$$\begin{aligned} I(X; Y|Z) &= -\int_Z \sum_{x,y} P(x, y|z) \log F(x, y) \\ &= \int_Z \sum_{x,y} P(x, y|z)F(x, y) - 1 \\ &\quad - \sum_{x,y} P(x, y|z) \log F(x, y) \\ &= \int_Z \sum_{x,y} P(x, y|z)\big[F(x, y) - 1 - \log F(x, y)\big] \\ &= 0. \end{aligned} \tag{18}$$

Since $F(x, y) - 1 - \log F(x, y) \geq 0$ ($\log t \leq t - 1$ for all t), conditional mutual information is summing over the multiplication of two non-negative terms. Thus, if $I(X; Y|Z) = 0$, it naturally indicates $F(x, y) - 1 = \log F(x, y)$ and the equality holds only when $F(x, y) = 1$. This implies

$$F(x, y) = 1 \qquad \forall x, y, \tag{19}$$

which yields conditional independence between X and Y given Z_R.

Since $I(X; Y|Z)$ is non-negative and $P_{X,Y|Z}(x, y|z) \geq 0$ and $\log \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)} \geq 0$, minimizing CMI, *i.e.,* $I(X; Y|Z) = 0$, requires

$$\log \frac{P_{X,Y|Z}(x, y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)} = 0 \quad \forall x, y, \tag{20}$$

and this naturally indicates conditional independence

$$P_{X,Y|Z}(x, y|z) = P_{X|Z}(x|z)P_{Y|Z}(y|z). \tag{21}$$

Thus when CMI, $I_\theta(\hat{A}; \hat{Y}|\mathbf{z}_R)$, is minimized to zero, we can achieve the conditional independence between $\hat{Y}$ and $\hat{A}$ given Z_R, *i.e.,* $\hat{Y} \perp \hat{A}|\mathbf{z}_R$.

Figure 8. Feature translation of FADES on CelebA dataset

10. Detailed experimental setup: FADES

10.1. Description of disentangled subspace

A detailed explanation of the disentangled subspace can be found at the beginning of Section 5. To reiterate, we partition the latent representation into four subspaces: $\mathbf{z}_Y, \mathbf{z}_A, \mathbf{z}_R$, and $\mathbf{z}_X$. In this setup, $\mathbf{z}_Y$ and $\mathbf{z}_A$ account for the information solely related to Y and$A$, respectively.

For an example of image classification to predict smiling [25], Y-related $\mathbf{z}_Y$ may include semantic information related to attributes such as {high cheek, mouth slightly opened, narrow eyes, etc}, while A-related $\mathbf{z}_A$ may include semantics like {bald, facial structure, etc}. On the other hand, $\mathbf{z}_R$ is introduced to learn information relevant to both Y and A such as {wearing lipsticks, mustache, etc}, which could impact classifying both smiling and gender. Lastly, $\mathbf{z}_X$ is employed as an auxiliary space to guide the information that is irrelevant to both Y and A, e.g., a background of the image.

	Smiling			Blond Hair			Attractive		
	Acc ↑	EOD ↓	DP ↓	Acc ↑	EOD ↓	DP ↓	Acc ↑	EOD ↓	DP ↓
FADES	<u>0.918</u>	**0.032**	0.125	<u>0.931</u>	**0.119**	**0.158**	<u>0.764</u>	**0.292**	**0.318**
GVAE [10]	**0.919**	0.049	0.133	**0.944**	0.496	0.228	**0.780**	0.523	0.430
FFVAE [6]	0.891	0.075	<u>0.071</u>	0.924	<u>0.297</u>	0.171	0.746	<u>0.335</u>	<u>0.324</u>
ODVAE [47]	0.885	<u>0.038</u>	0.101	0.890	0.431	0.169	0.716	0.599	0.459
FairDisCo [31]	0.839	0.074	**0.051**	0.914	0.481	0.189	0.747	0.549	0.357
FairFactorVAE [32]	0.914	0.055	0.136	0.912	0.312	<u>0.166</u>	0.710	0.474	0.357

Table 3. Evaluation of downstream classification tasks on different target labels on CelebA dataset. FADES consistently exhibit better balance in fairness and accuracy.

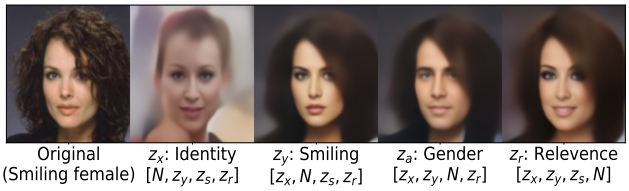


Figure 9. Reconstruction by replacing specific code with noise.

10.2. How to set subspace dimensions

In line with previous research, we opted for a single dimension in the case of the tabular dataset. For the image dataset, specifically the CelebA dataset, we maintained equal dimensional sizes for $\mathbf{z}_Y, \mathbf{z}_A, \mathbf{z}_R \in \mathbb{R}^d$, ensuring the dimensions are a power of 2. For the reported results, we assigned $d = 32$ empirically chosen by grid search. It is worth noting that the total latent dimension is fixed at $\mathbb{R}^{512}$ across all methods. Thus in our method we have $\mathbf{z}_X \in \mathbb{R}^{512-3\times32} = \mathbb{R}^{416}$.

Our empirical observations indicated that dimensions larger than 64 for d led to a decline in reconstruction performance. This degradation may be attributed to the excessive allocation of information to each subspace, ultimately compromising the reconstructive capacity.

11. Additional Experiments

11.1. Fair classification on CelebA

CelebA dataset [33] is widely employed to benchmark fair classification of vision models. On top of "Smiling" as the target label as reported in the main paper, we also conducted experiments considering other widely selected target labels, such as "Blond Hair" and "Attractiveness" under *gender* bias. Full comparison is summarized in Table 3. As in the results, FADES consistently demonstrates superior balance in fairness and accuracy across different target labels. Specifically, FADES achieves an accuracy level that is on par with the best-performing models while significantly improving fairness violations. This also validates the effectiveness of the proposed fair disentanglement learning method.

11.2. Counterfactual generation

In this subsection, we demonstrate the results of feature translation on CelebA dataset. In Figure 8, we display a grid of images generated by replacing certain latent codes. The first row is the direct reconstruction of the original input image in the last row. The intermediate rows are reconstructed images following the composition in the Y-axis. The superscript indicates the index of the image in each column. For instance, the second row, $[\mathbf{z}_X^{(0)}, \mathbf{z}_Y, \mathbf{z}_A, \mathbf{z}_R]$, generates images that have irrelevant features of the input image in the first column $\mathbf{z}_X^{(0)}$, such as blond hair, white background, etc, while following each column's $\mathbf{z}_Y$, $\mathbf{z}_A$, and $\mathbf{z}_R$. We observe that sensitive and target information is effectively translated to different images. Additionally, we can effectively translate both sensitive and target information simultaneously, allowing us more freedom in generating counterfactuals. For example, given a real image of a *smiling female*, we can generate {not smiling male, not smiling female, smiling male}. This enables us to improve individual fairness [14], which is to ensure individuals with similar features but a different sensitive attribute receive similar outcomes.

Furthermore, we visualize contribution of latent codes by replacing each code to normal noise in Figure 9. We observe specific feature changes by altering each code, e.g., $\mathbf{z}_A \sim \mathcal{N}$ in 4th image altered gender, which aligns with Figure 8.

To quantitatively analyze the results, we measure FID score [20] between the direct reconstruction of the input image and the reconstructed images with permuted sensitive code as

$$\Delta\text{FID} = \text{FID}(\hat{X}, \hat{X}_{perm}), \tag{22}$$

where $\hat{X}$ is reconstruction of X from VAE-based models and $\hat{X}_{perm}$ is reconstruction with randomly permuted sensitive code $\mathbf{z}_A$ within the evaluation set. The FID score difference, ΔFID, shows how natural the image translation is without image distortion or decrease in quality. As in Table 4, FADES achieves a significantly lower FID score difference compared to other methods for fair representation

learning, which indicates FADES renders counterfactuals with superior quality.

	FADES	FFVAE	GVAE	ODVAE
Δ FID $\downarrow$	1.166	3.712	1.407	15.08

Table 4. FID score difference between original reconstruction and random perturbation in sensitive code.

Figure 10. Reconstruction on an unbiased set of C-MNIST. Our method is the only one that correctly recovers both colors and digits, demonstrating its effectiveness in achieving disentanglement of the information.

11.3. Fair image reconstruction

To further qualitatively assess vision tasks, we reconstruct samples from the unbiased test set (*i.e.,* all colors are uniformly distributed over the digits) of the C-MNIST dataset. The effectiveness of disentangled methods can be measured by their ability to accurately reconstruct both the color and digit. The results are shown in Figure 10, where we observe that only FADES correctly recovers both color and digit, indicating its ability to effectively disentangle color and digit information in the latent space.

In addition, we illustrate t-SNE [48] visualization of the target code from the test set generated by each method, as in Figure 12A$\sim$12H to better understand the distribution and disentanglement of the learned representation. Each figure has two subfigures, the first is colored based on *Digit* and the second is colored based on *Color*. Since the goal is to filter out the sensitive information (color), we expect the distribution to be clustered by digit. However, most of the

(a) White Female (b) StyleCLIP (c) StyleCLIP + FADES

Figure 11. Style transfer by StyleCLIP and FADES extension. The example is to alter the source image (a) to "A doctor with curly hair". Fair text-to-image modification should be gender-invariant.

	Acc (Digit)$\uparrow$	Acc (Color)$\uparrow$
FADES (Ours)	**74.42**	<u>86.76</u>
GVAE [10]	62.24	67.83
FFVAE [6]	60.18	80.70
ODVAE [47]	55.59	64.26
FairDisCo [31]	54.54	**98.79**
FairFactorVAE [32]	<u>66.47</u>	64.92

Table 5. Digit and color recovery on unbiased C-MNIST trained on 95% color bias.

comparing methods have weak separation with respect to the digit information. Moreover, ODVAE (Figure 12B) and β-VAE (Figure 12H) have stronger corrlation with *color* than *digit*. In contrast, FADES achieve better separation with respect to digits, while different colors are uniformly distributed among the clusters. This also aligns with the results in Table 2 of the main paper.

Furthermore, we conducted experiments on the extreme setting (95% bias) with severe corruption ($\sigma = 0.005$) [41] as in Table 5. FADES significantly outperforms other methods on recovering both digit and color information. Note that FairDisCo requires color (sensitive) information as the input of decoder. This also validates that FADES better disentangle the sensitive and target information even under severe bias, showcasing its robustness and effectiveness.

11.4. Text-to-image editing

To further validate the capability of FADES, we integrated it as an adaptor on top of a pre-trained, frozen CLIP [45] image encoder, and trained with Facet dataset [17] to augment fairness in vision-language tasks. Notably, VLM exhibits career-gender biases. Specifically, in Figure 11, StyleCLIP [43] exhibits gender bias despite its identity preservation loss. In contrast, FADES addresses the bias while preserving effective modification.

Further, we evaluate CLIP with FADES in Table 6 on Facet dataset, where FADES shows enhanced fairness without compromising performance compared to linear probing (ERM), suggesting its applicability to various VLM tasks such as search and image retrieval with fairness. Also, we present an ablation study regarding contribution of each loss

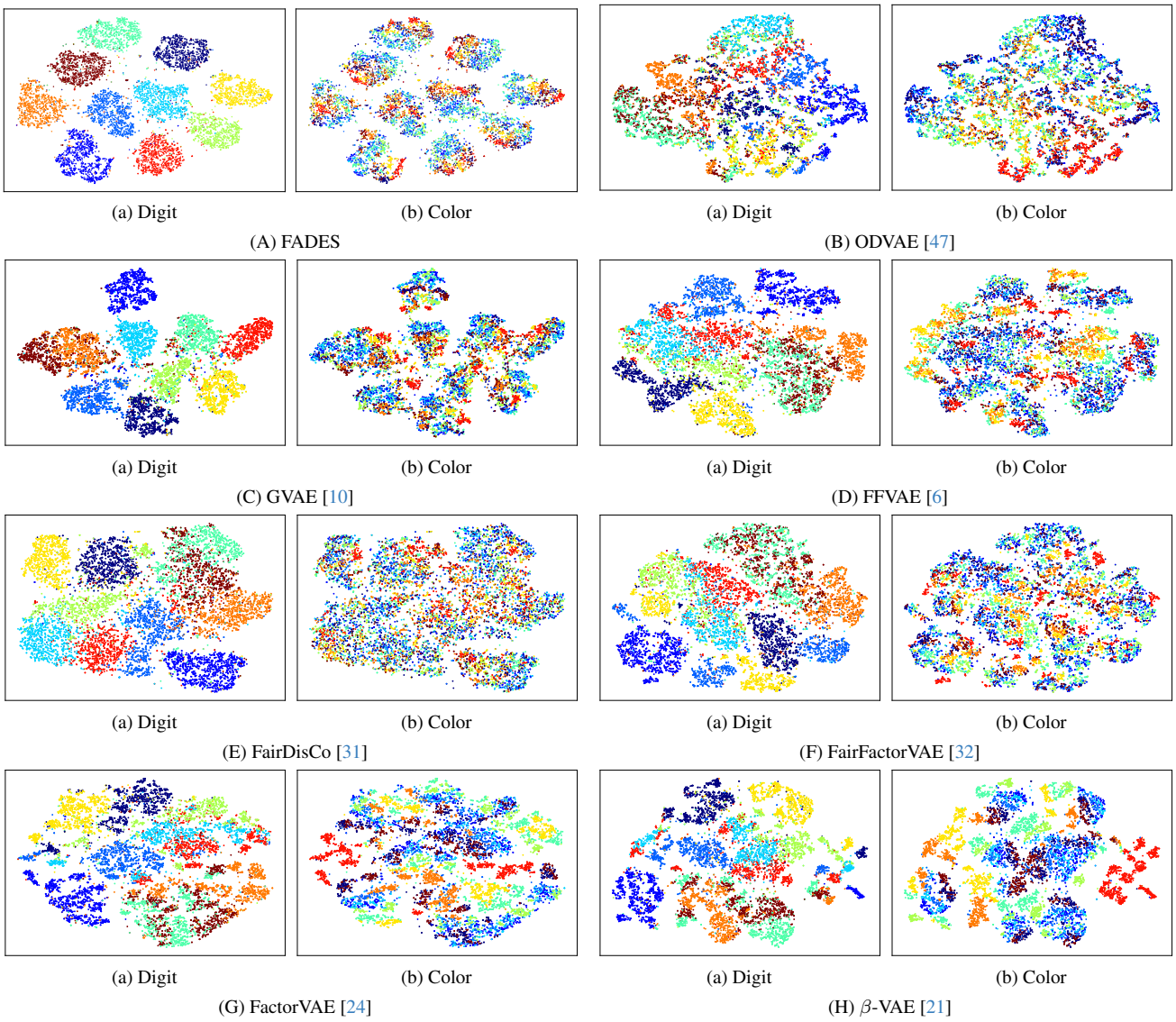


Figure 12. t-SNE visualization of learned representation corresponds to sensitive-irrelevant information on unbiased C-MNIST test set.

Method	**Top-1 Acc. (%)**			**Top-3 Acc. (%)**		
	WG	Avg	Gap	WG	Avg	Gap
Zero-shot	2.78	53.47	50.69	15.28	76.43	61.15
Linear prob	0.00	<u>65.67</u>	65.65	0.00	**85.55**	85.55
FADES	**69.83**	**69.22**	**0.61**	**84.98**	<u>85.26</u>	**0.28**
(Abl.) $\mathcal{L}_{\text{ELBO}}$	16.34	17.27	0.98	37.34	37.34	1.47
(Abl.) w/o $\mathcal{L}_{\text{CMI}}$	65.87	67.58	1.71	81.73	82.19	0.46
(Abl.) w/o $\mathcal{L}_{\text{reg}}$	30.51	34.15	3.63	59.11	60.05	0.94

Table 6. Performance of CLIP (ViT-B/32) [45] on facet dataset [17]. WG: Worst Group, Gap: Difference between WG and Avg.

terms in the last 3 rows. Each row corresponds to training with only $\mathcal{L}_{\text{ELBO}}$, without $\mathcal{L}_{\text{CMI}}$, and without $\mathcal{L}_{\text{reg}}$. It shows that FADES achieves optimal fairness while retaining accuracy.

12. Limitation and Discussion

The limitations of our work can be summarized in two folds. Firstly, while we address the contradiction between fairness and performance in previous works under certain data bias, there are many different sources that cause fairness violations in algorithms. FADES covers one of the main causes, *i.e.,* unwanted correlation between sensitive and target information, however, there are some other well-known causes, such as under-representation, model bias, missing or noisy features, etc. In real-world scenarios, these do not usually occur alone. As future work, it would be interesting to theoretically extend our approach to mitigate fairness violations under different causes or multiple sources of bias. Secondly, our method requires both target and sensitive information during the training phase, which may not be always available, since labeling is demanding or sometimes inaccessible due to laws and regulations. Moreover, one could consider extending our concept to an unsupervised manner in order to learn fair disentangled representation without the need for labels.