Fair Federated Learning with Biased Vision-Language Models

Huimin Zeng Zhenrui Yue Yang Zhang Lanyu Shang Dong Wang
University of Illinois at Urbana-Champaign
{huiminz3, zhenrui3, yzhangnd, lshang3, dwang24}@illinois.edu

Abstract

Existing literature that integrates CLIP into federated learning (FL) largely ignores the inherent group unfairness within CLIP and its ethical implications on FL applications. Furthermore, such CLIP bias may be amplified in FL, due to the unique issue of data heterogeneity across clients. However, in identity-sensitive FL applications, model fairness (i.e., group fairness) is imperative for model development. Therefore, this work explores a critical question ignored by the existing literature: how can we build a fair FL framework using biased pre-trained VLMs (e.g., CLIP)? To address this problem, we propose a fairness-aware adaptation framework tailored for VLM (e.g., CLIP) in the context of FL, named **F**air **F**ederated **D**eep **V**isiual **P**rompting or **FF-DVP**. As implied by its name, FF-DVP trains a fair FL model with fairness-aware deep visual prompting (DVP). Moreover, FF-DVP incorporates modality-fused classification heads to learn client-specific knowledge and fairness constraints. These modules explicitly address a unique kind of bias in FL, namely the bias triggered by data heterogeneity. We show that FF-DVP can be readily extended to prevailing parameter-efficient fine-tuning methods (e.g., adapter or LoRA) for debiasing purposes. To the best of our knowledge, FF-DVP is the first to leverage biased VLMs for building fair FL frameworks. Extensive results on human face attribute recognition (FAR) applications suggest that FF-DVP effectively improves model fairness and training convergence, outperforming state-of-the-art baselines.

1 Introduction

Federated learning (FL) emerges as a novel machine learning (ML) paradigm wherein ML models are trained from distributed data sources (McMahan et al., 2017). In FL, a central server stores a global model, while multiple local clients participate in the collaborative model training without sharing their private data. Such a decentralized design of FL makes it a promising solution for privacy-sensitive applications, like online facial services or medical diagnosis (McMahan et al., 2017). However, FL models commonly suffer from scalability issues. That is, traditional FL models struggle to achieve training convergence under high data complexity (e.g., data amount, data dimensionality) and high data heterogeneity (e.g., non-i.i.d. data) (Zhou et al., 2022). Therefore, we aim to design a scalable fair FL framework that withstands high data complexity and heterogeneity.

On the other hand, large foundation models become increasingly popular for ML tasks involving complex and large-scale data. These models, with billions parameters, are pre-trained using Internet-scale data (Chuang et al., 2023). They can extract domain-generalized features from inputs, and generalize to various downstream tasks across different domains (e.g., sentiment analysis, image classification (Bommasani et al., 2021)).

Recently, exploiting large vision-language-models (VLMs) for federated learning (FL) has gained increased attention (Lu et al., 2023; Yang et al., 2023; Guo et al., 2023a; Chen et al., 2023). The strong generalization ability of VLMs has empowered FL models to overcome the data complexity and data heterogeneity across clients, leading to better personalization and generalization (Lu et al., 2023). However, existing literature that combines VLMs (e.g., CLIP) and FL largely neglects the inherent bias within such VLMs and its ethical implications on FL applications. The unique issue of data heterogeneity in FL would exacerbate CLIP bias, severely compromising the fairness of the aggregated FL model (Chang and Shokri, 2023). On the other hand, in privacy- and identity-sensitive applications, it is imperative to develop FL models that are both accurate and fair. For instance, in an FAR-based criminal detection system, if the model consistently makes false-positive predictions on a specific demographic group, great ethical concerns

Findings of the Association for Computational Linguistics ACL 2024, pages 10002–10017
August 11-16, 2024 ©2024 Association for Computational Linguistics

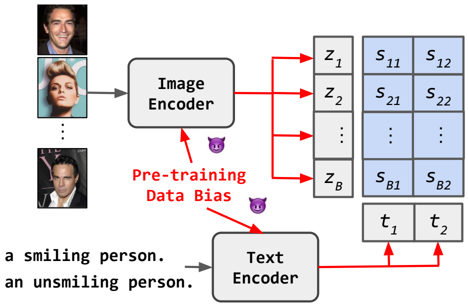


Figure 1: CLIP for Smiling Detection. For a query image, CLIP picks the prompt that depicts the image the base as the prediction. However, CLIP encodes bias in the pre-training data. In this work, we present a framework to debias FL model using biased CLIP.

would be raised (Najibi, 2022).

Unfortunately, as shown in Figure 1, pre-trained VLMs indeed encode the bias of their pre-training data, and may perpetuate such bias in downstream tasks (Chuang et al., 2023). Additionally, in FL applications, to transfer a pre-trained VLM to the domain-specific data, it is common to fine-tune the model on the clients' data and then update the global model via aggregation. In this case, if the local data distributions are biased and heterogeneous, fine-tuning a biased VLM on such local data would further amplify the bias of the global model. The bias triggered by data heterogeneity in FL represents a unique challenge, whereas there exist no fair methods tailored for VLM-based FL frameworks. Therefore, this work studies a critical question: how can we develop a fair FL framework using the biased pre-trained VLMs?

In this work, we focus on the fairness of the FL model w.r.t. **the unknown global data distribution**. Ideally, a fair FL model shall perform non-discriminatively against any demographic group (i.g., group fairness) while achieving satisfactory overall performance. To achieve this, we propose a fairness-aware adaptation framework tailored for VLMs in the context of FL: **F**air **F**ederated **D**eep **V**isiual **P**rompting or **FF-DVP**. We highlight that FF-DVP is specifically designed to mitigate the bias in FL settings: 1) FF-DVP minimizes the communication burden between the server and clients: given the large size of CLIP, FF-DVP is lightweighted in terms of communication costs in FL; 2) FF-DVP performs demographic-agnostic and domain-generalized feature extraction: it removes sensitive demographic information from CLIP features while retaining domain-generalized ones to counter data-heterogeneity-triggered bias in FL; 3) FF-DVP learns client-specific knowledge and fairness constraints: FF-DVP adapts CLIP to learn client-specific knowledge for targeted tasks (e.g., FAR applications) and meets client-specific fairness constraints (e.g., demographic parity).

To this end, FF-DVP consists of two debiasing modules. Firstly, FF-DVP debiases the CLIP features through fairness-aware deep visual prompting (DVP). The fairness-aware DVP is a sequence of learnable parameters prepended to the visual tokens. This DVP is designed to remove the demographic-related information from CLIP features while preserving the domain-generalized information. The extracted fair, domain-generalized feature overcomes the unique data-heterogeneity-trigger bias in FL. Secondly, FF-DVP learns client-specific knowledge and fairness constraints with modality-fused classification heads. The classification heads also contribute to the learning of fair and robust representations. With both modules, the aggregated global model achieves desired group fairness despite the data heterogeneity in local clients. Finally, we show that FF-DVP is flexible and can be extended to other parameter-efficient fine-tuning (PEFT) methods, such as adapter-style tuning (Houlsby et al., 2019) or low-rank adaptation (LoRA) (Hu et al., 2021). We summarize the contributions of our paper as follows[1]:

1. To the best of our knowledge, this work is first to study inherent bias of the pretrained VLMs in FL applications. Compared to centralized debiasing methods, we focus on the unique data-heterogeneity-triggered bias in FL, and proposed a scalable fair VLM-based FL framework.

2. Technically, FF-DVP debiases the pre-trained VLM (e.g., CLIP) through a novel fairness-aware deep visual prompting. Moreover, as a parameter-efficient adaptation method, we show that our method could be easily extended to other PEFT schemes.

3. We evaluate our method on federated face attribute recognition (FAR) for its privacy and ethical implications. On different FAR applications, experimental results suggest that FF-DVP can effectively improve fairness of FL models, outperforming the state-of-the-art baselines by a significant margin.

[1] We adopt publicly available datasets and will release the code upon acceptance.

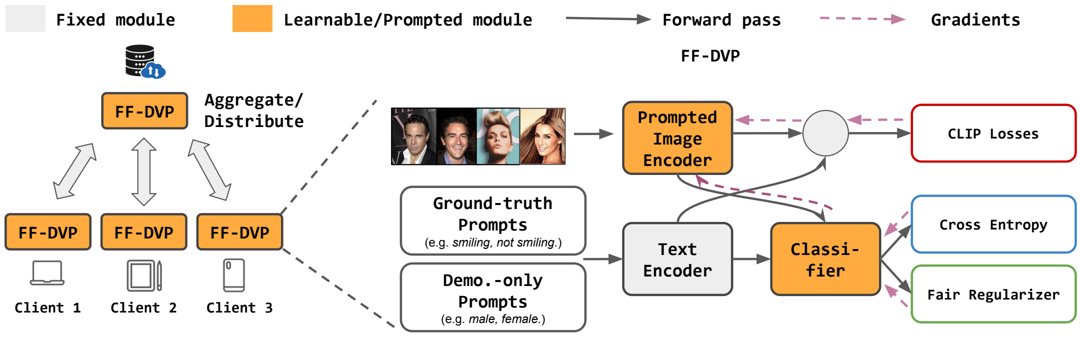


Figure 2: Overview of FF-DVP. On each client, FF-DVP adapts CLIP to application-specific data with fairness-aware DVP and client-specific classification heads. Then, the server aggregates the learnable modules using any existing aggregation protocol.

2 Related Work

Foundation Models and Federated Learning. Large foundation models (e.g., GPT family (OpenAI, 2023), LlaMA (Touvron et al., 2023), CLIP (Radford et al., 2021)) exhibit robust generalization capabilities across a wide spectrum of tasks. Recently, efforts have been made to integrate foundation models (e.g., CLIP) into FL frameworks for better personalization and generalization (Lu et al., 2023; Yang et al., 2023; Guo et al., 2023a; Chen et al., 2023). For instance, Guo et al. (2023a,b) focus on learning text prompts to personalize CLIP on client data, whereas Li et al. (2023) leverage visual prompts for the same goal. In addition to prompt learning, Lu et al. (2023); Chen et al. (2023) fine-tune CLIP with light-weighted adapters to adapt CLIP to the personalized data on clients. However, existing literature that combines CLIP and FL merely focuses on accuracy, largely overlooking the inherent bias of CLIP and its impact on fair FL applications. In contrast, our work is the first to consider the bias issue of CLIP in FL and leverage biased CLIP to build fair FL frameworks.

Group Fairness in Federated Learning. For identity-sensitive applications, concerns have been raised about the fairness of the FL models (i.e. group fairness): they could perform discriminatively against under-represented demographic groups (Ezzeldin et al., 2023). In terms of group fairness in FL, Cui et al. (2021) strive to satisfy the local personalized fairness on each client. In comparison, a larger amount of studies (including our work) focus on a more general notion of **global fairness**, for its profound significance in real-world applications Mohri et al. (2019); Du et al. (2021); Ezzeldin et al. (2023); Hong et al. (2021). For instance, agnostic federated learning (Mohri et al., 2019) achieves the good-intent fairness that protects the worst-case performance on any client. A fairness-aware aggregation protocol is introduced in (Ezzeldin et al., 2023) to obtain a fair global model. However, above methods usually suffer from the issue of scalability: in the era of large foundation models, the performance of traditional fair FL methods might degrade significantly when the data is complex and heterogeneous (Zhou et al., 2022). In this work, we present the very first framework that leverages large foundation models (e.g., CLIP) to address the bias issue in FL.

3 Preliminaries

Federated Learning. Assume there are K clients in an FL application. For all datasets, each data point consists of an input feature $x \sim \mathcal{X}$, a demographic attribute $a \sim \mathcal{A}$ and a label $y \sim \mathcal{Y}$. A local dataset of client k is denoted as $\mathcal{D}^{(k)} = \{(x_1^{(k)}, a_1^{(k)}, y_1^{(k)}), ...\}$. **For simplicity, if not specified, we use the notations without client index k to represent the data of an arbitrary client.**

To find the optimal global model f_θ^* in an FL application, McMahan et al. (2017) proposed FedAvg (Appendix A). Specifically, at each round, each local client trains its local model with its own data. Then, clients send the trained local model weights to the central server for aggregation. On the central server, the global model will be updated using a weighted-average of the received weights. However, if the local data distributions are imbalanced, the locally trained models are also biased. This eventually leads to an unfair global model after the aggregation (Ezzeldin et al., 2023).

Fairness Notions. To measure the model fairness and performance, we adopt the commonly used demographic parity Φ_{demo}, equalized odds Φ_{eq} (Hardt et al., 2016), accuracy parity Φ_{A} (Makhlouf et al., 2021) and balanced accuracy $\mathcal{A}_B$ (Brodersen et al., 2010). Due to space limit, we have summarized the formal definition of fairness notions (i.e., Φ_{demo}, Φ_{eq}, Φ_{A} and $\mathcal{A}_B$) in Appendix A.

CLIP. CLIP predicts which images are paired with which text prompts. We use f_I to denote the CLIP image encoder, and f_T for the CLIP text encoder. For a query image x and a list of candidate prompts for $|\mathcal{Y}|$ classes (e.g., `a photo of [class c]`), CLIP selects the candidate prompt with the highest normalized cosine similarity to x as the predicted class (more details in Appendix A):

$$\begin{aligned}&\hat{y} = \arg\max_{c} \frac{\exp\big(\cos(z, t_{candidate_c})/\tau\big)}{\sum_{c'} \exp\big(\cos(z, t_{candidate_{c'}})/\tau\big)},\\ &\text{where } z = f_I(x),\\ &t_{candidate_c} = f_T(\texttt{a photo of [class c]}),\\ &c \in \{1, 2, ..., |\mathcal{Y}|\}.\end{aligned} \tag{1}$$

For instance, in smiling detection (Figure 1), for a query image, CLIP selects either “a smiling person” or "an unsmiling person" as its prediction based on similarity. However, due to the inhere bias of CLIP (Chuang et al., 2023), directly integrating CLIP into FL results in a biased FL model.

4 Algorithm

4.1 Fairness-aware Deep Visual Prompting

To exploit CLIP for a fair FL model, we propose to firstly debias CLIP on each client. Specifically, we present fairness-aware deep visual prompting (DVP) that suppresses the demographic-related signals in CLIP features and preserves domain-generalized ones. For a better understanding of our fairness-aware DVP, we expand the image encoder f_I into L layers (Figure 3 (left)). As in (Jia et al., 2022), at the input layer, CLIP divides a query image x into J fixed-sized image patches $\{I_1, I_2, ...I_j, ..., I_J | I_j \in \mathbb{R}^{3\times h\times w}\}$, where the size of each image patch is $h \times w$. Note that the index of image patches j is different from the index of samples i defined in Section 3. The image patches are then embedded into embeddings using the embedding layer (i.e., layer 0) of f_I:

$$e_{0,j} = f_{I,\text{Embed}}(I_j),\ e_{0,j} \in \mathbb{R}^d,\ j \in \{1, 2, ..., J\} \tag{2}$$

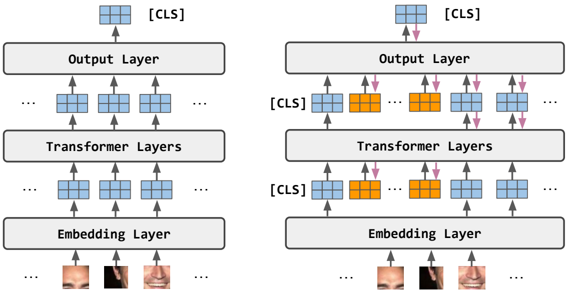


Figure 3: Fairness-aware Deep Visual Prompt.

After the embedding layer, the transformer layers of f_I then extract visual features from the image embeddings using different transformer layers. In particular, at the l-th transformer layer of f_I:

$$\begin{aligned}&[e_{l,0}, e_{l,1}, ..., e_{l,J}]\\ &\quad = f_{I,\text{Transformer}_l}([e_{l-1,0}, e_{l-1,1}, ..., e_{l-1,J}]),\end{aligned} \tag{3}$$

where $l \in \{1, 2, ..., L\}$ and $[\cdot, .., \cdot]$ represents the concatenation operation along the dimension of sequence length. For simplicity, we use E_l to denote $[e_{l,1}, ..., e_{l,J}]$. Thus, Equation 3 is simplified as:

$$[e_{l,0}, E_l] = f_{I,\text{Transformer}_l}([e_{l-1,0}, E_{l-1}]). \tag{4}$$

Note that in Equation 4, there is usually a `[CLS]` token prepended to the visual tokens. Thus, at the output layer, the first feature (corresponding to `[CLS]`) is used as the classification token to compute the cosine similarity: $z = e_{L,0}$.

To address the bias issue of CLIP, we insert fairness-aware visual prompts $V = \{p \in \mathbb{R}^d\}$ at the embedding layer and the transformer layers of f_I as in Figure 3 (right). At the embedding layer, the prompted image embeddings are formulated as:

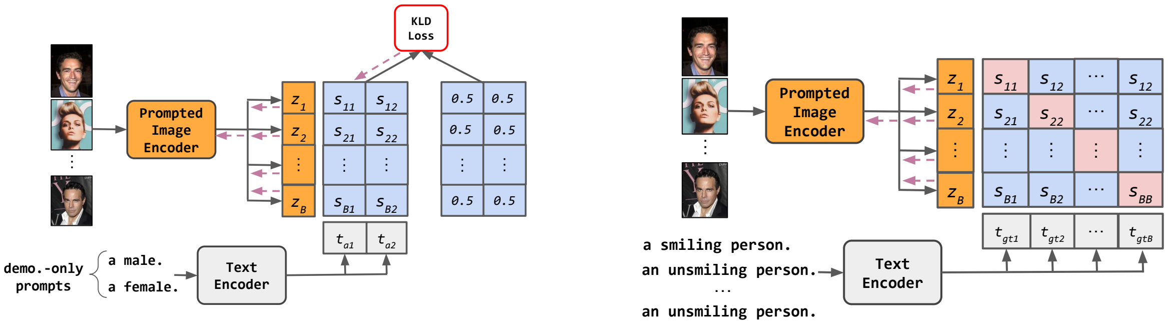
$$\Big[e_{0,0}, V_0, E_0\Big] = \Big[e_{0,0}, \underbrace{p_{0,1}, ..., p_{0,P}}_{\textbf{prompts}}, \underbrace{e_{0,1}, ..., e_{0,J}}_{\textbf{embeddings}}\Big], \tag{5}$$

where P is the length of visual prompts, and V_0 is the visual prompt for the embedding layer. Similarly, for the transformer layers, the forward pass of each layer is formulated as:

$$\begin{aligned}&[\tilde{e}_{1,0}, \_, \tilde{E}_1] = f_{I,\text{Transformer}_l}([e_{0,0}, V_0, E_0])\\ &[\tilde{e}_{l,0}, \_, \tilde{E}_l] = f_{I,\text{Transformer}_l}([\tilde{e}_{l-1,0}, V_{l-1}, \tilde{E}_{l-1}]),\end{aligned} \tag{6}$$

where $\tilde{e}$ and $\tilde{E}$ are prompted representations, and $l \in \{2, ..., L\}$. Therefore, at the output layer of the prompted f_I, the extract visual representation z is also prompted:

$$\tilde{z} = \tilde{e}_{L,0}. \tag{7}$$



(a) Fairness Branch.

(b) Adapting CLIP to FAR data.

Figure 4: Training the fairness-aware DVP using the fairness branch (left) and CLIP contrastive loss (right).

The intuition of debiasing CLIP is to mitigate the demographic-related information within $\tilde{z}$. For instance, as in Figure 1, if the goal is to debias CLIP for smiling detection w.r.t. gender, $\tilde{z}$ should not encompass information that could infer the gender of the query image.

To debias CLIP w.r.t. the demographic attribute a, we propose to construct a set of **demographic-only prompts** that depict the demographic information of the input images (Figure 4a). As before, CLIP encodes the demographic-only prompts into representations: $\{t_{a_1}, t_{a_2}, ..., t_{a_{|\mathcal{A}|}}\}$. With encoded demographic-only text prompts and the prompted visual representation $\tilde{z}$, we then propose a fairness-loss to debias the CLIP feature for an input x:

$$l_{fair}(x) = \mathrm{KL}(\hat{Pr}(A)||\mathcal{U}(1, |\mathcal{A}|)),$$

where

$$\hat{Pr}(A) = \mathrm{Softmax}(\frac{\cos(\tilde{z}, t_{a_1})}{\tau}, ..., \frac{\cos(\tilde{z}, t_{a_{|\mathcal{A}|}})}{\tau}),$$

$$\mathcal{U}(1, |\mathcal{A}|)) = [\frac{1}{|\mathcal{A}|}, ..., \frac{1}{|\mathcal{A}|}]. \tag{8}$$

Equation 8 computes the KL-divergence between the normalized cosine similarities $\hat{Pr}(A)$ and a uniform distribution $\mathcal{U}(1, |\mathcal{A}|)$. Each element in $\hat{Pr}(A)$ measures the relevance between the prompted visual representation $\tilde{z}$ and a demographic-only text prompt (e.g., `a photo of a [male].`). By minimizing Equation 8, all demographic-only text prompts become equally relevant to $\tilde{z}$. This indicates that CLIP can no longer distinguish which demographic group is more related to $\tilde{z}$ than others, thereby, debiasing CLIP.

In addition to fairness, we also propose to optimize a contrastive loss using the **ground-truth prompts** to maintain an overall high performance (Figure 4b). Formally, for training dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}$, we construct the ground truth prompts for x_i that contains textual description of y_i. Then, we compute the CLIP contrastive loss over the prompted visual representation and the prompts as in (Radford et al., 2021):

$$\mathcal{L}_{CLIP} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} - \log \frac{e^{\tilde{z}_i \cdot t_{gt_i}}}{\sum_{j=1}^{|\mathcal{D}|} e^{\tilde{z}_i \cdot t_{gt_j}}} + \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} - \log \frac{e^{\tilde{z}_i \cdot t_{gt_i}}}{\sum_{j=1}^{|\mathcal{D}|} e^{\tilde{z}_j \cdot t_{gt_i}}}, \tag{9}$$

where $\tilde{z}_i$ is the prompted visual representation (i.e., Equation 7) and $t_{gt_i} = f_T($`ground-truth prompt of` $x_i)$ is the encoded ground-truth prompt. To adapt CLIP into task-specific data and suppress the inhere bias of CLIP, Equation 9 and Equation 8 are optimized jointly, balanced via a non-negative factor λ_1:

$$\mathcal{L}_{CLIP} + \lambda_1 * \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} l_{fair}(x_i) \tag{10}$$

Finally, we highlight that FF-DVP is adjustable, allowing users to determine which transformer layers to prompt or not (i.e., not necessarily prompting all layers). We experimented this design by intervening only the first layer, first half layers and all layers of the CLIP image encoder in Section 5.3. We found that more intervened layers would generally improve the performance.

4.2 Client-Specific knowledge and Fairness

In FL, it is not always feasible to collect Internet-scale image data and text data on each client to finetune CLIP. Such data sparsity leads to training instability and performance degradation. Therefore, we further propose to build a shared light-weighted modality-fused classifier f_{cls} (i.e., a two-layer fully connected network) to stabilize the training of CLIP on clients and help CLIP learn robust latent representations (Appendix B).

In our FL setting, each client fine-tunes CLIP as well as trains this classifier to the FAR application using the client's local data. The classifier takes the prompted visual representation and the text representation as input. It then fuses the features from both data modalities:

$$\hat{y} = f_{cls}([\Pi(\tilde{z}), \Pi(t_{gt})]), \tag{11}$$

where $\tilde{z}$ is the prompted visual representation of x, and $t_{gt} = f_T(\texttt{ground-truth prompt of } x)$ is the encoded ground-truth prompts. $[\cdot,\cdot]$ represents the concatenation operation. In Equation 11, we introduce a learnable projection matrix Π to reduce the dimensionality of $\tilde{z}$ and t_{gt}. To train f_{cls}, we optimize the cross entropy loss as well as a fairness regularizer $\hat{\Phi}_{.}$ over the training set on each client:

$$\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} l_{ce}(\hat{y}_i, y_i) + \lambda_2 \cdot \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \hat{\Phi}_{.}(x_i, y_i, a_i). \tag{12}$$

The rationale behind this design is that: despite the debiasing efforts on CLIP features, the ultimate classification performance of the model does not necessarily fulfill the fairness constraints on local clients. Therefore, training f_{cls} on local data contributes to learning client-specific knowledge and fairness constraints. This design improves the fairness of the model in terms of the ultimate prediction performance after the global model aggregation. The fairness regularizers $\hat{\Phi}_{.}$ could be implemented flexibly using arbitrary differentiable version of fairness notion defined in Section 3.

4.3 Overall Framework

Training. The fairness-aware DVPs and client-specific classifiers are jointly trained in a standard FL fashion (i.e., local training and global aggregation). In our implementation, FedAvg is used for aggregation. The pseudocode of the training pipeline is summarized in Algorithm 1 (Appendix C). Finally, we highlight that the weights of CLIP encoders are not updated and exchanged. FF-DVP is a parameter-efficient method.

Inference. During the inference phase, we only use the prompted CLIP image encoder $\tilde{f_I}$ and the original text encoder f_T to perform inference for the query images. f_{cls} is dropped on purpose because f_{cls} is only designed to help CLIP learn client-specific knowledge and fairness constraints during training. Similar to Equation 1, the cosine similarity between **prompted** visual representations $\tilde{z}$ and candidate prompts is computed for making predictions.

5 Experiments

5.1 Experimental Setup

Dataset. We use CelebA (Liu et al., 2015) and FairFace (Karkkainen and Joo, 2021) to study different FAR applications in the context of FL. Due to space limit, we chose *smiling* and *age* as our predictive face attributes. As mentioned in (Shen et al., 2017), smiling detection is objective since smiling or not is easy to judge. In comparison, age detection is more challenging: it is formulated as a binary task of classifying 'young' and 'old', but both age groups exhibit a broad age range, causing a vague and hard-to-learn boundary. Finally, the age label is the only shared label in both datasets, which help us to test the generality of our method. Without loss of generality, we choose gender as the demographic attribute.

FL setup. During experiments, the training of some baseline methods could not converge under the high data complexity and data heterogeneity of FAR applications. Therefore, for fair comparison, we compare all methods under a setting of 5 clients, where all baseline methods could converge. Moreover, for training convergence and computational efficiency, we downsample 20000 images from both datasets and distribute the samples images to the 5 clients. We explicitly control population shifts for all clients, so that the local training data distributions are imbalanced and non-i.i.d. Finally, to eliminate the potential bias in the test data distribution that could affect the fairness evaluation, we sample a balanced test set of size 5000 to evaluate the FL model. More implementation details (i.e., local data distribution configuration, prompt design, hyperparameters, CLIP version) are summarized in Appendix D.

Baseline algorithms and models. Due to the lack of baselines that explore pre-trained foundation models for **fair FL**, we select FedAvg (McMahan et al., 2017), AFL (Mohri et al., 2019), FairFed (Ezzeldin et al., 2023), and FADE (Hong et al., 2021) as the baselines for comparison. Moreover, to better demonstrate the necessity of using pre-trained foundation models, we adopt a relatively

Face Application	**Metrics**	CLIP zero-shot	FedAvg	AFL	FairFed	FADE	**FF-DVP (Ours)**
Smiling Detection (CelebA)	$\mathcal{A}_B \uparrow$	0.848	$\underline{0.903_{\pm 0.009}}$	$0.899_{\pm 0.008}$	$0.900_{\pm 0.011}$	$0.866_{\pm 0.004}$	$\mathbf{0.905_{\pm 0.005}}$
	$\Phi_A \downarrow$	0.422	$\underline{0.191_{\pm 0.123}}$	$0.266_{\pm 0.103}$	$0.239_{\pm 0.113}$	$0.263_{\pm 0.100}$	$\mathbf{0.158_{\pm 0.043}}$
	$\Phi_{\text{demo}} \downarrow$	0.106	$0.012_{\pm 0.004}$	$0.011_{\pm 0.015}$	$\underline{0.011_{\pm 0.006}}$	$0.021_{\pm 0.002}$	$\mathbf{0.010_{\pm 0.011}}$
	$\Phi_{\text{eq}} \downarrow$	0.211	$0.037_{\pm 0.001}$	$0.042_{\pm 0.016}$	$\underline{0.033_{\pm 0.008}}$	$0.051_{\pm 0.006}$	$\mathbf{0.028_{\pm 0.016}}$
Age Detection (CelebA)	$\mathcal{A}_B \uparrow$	0.601	$0.534_{\pm 0.027}$	$0.622_{\pm 0.099}$	$0.584_{\pm 0.119}$	$\underline{0.773_{\pm 0.007}}$	$\mathbf{0.839_{\pm 0.009}}$
	$\Phi_A \downarrow$	1.829	$1.898_{\pm 0.073}$	$1.515_{\pm 0.410}$	$1.602_{\pm 0.563}$	$\underline{0.428_{\pm 0.055}}$	$\mathbf{0.284_{\pm 0.203}}$
	$\Phi_{\text{demo}} \downarrow$	0.281	$0.043_{\pm 0.030}$	$0.062_{\pm 0.060}$	$\underline{0.040_{\pm 0.056}}$	$0.105_{\pm 0.012}$	$\mathbf{0.026_{\pm 0.020}}$
	$\Phi_{\text{eq}} \downarrow$	0.562	$0.085_{\pm 0.060}$	$0.128_{\pm 0.120}$	$\underline{0.079_{\pm 0.112}}$	$0.210_{\pm 0.023}$	$\mathbf{0.053_{\pm 0.039}}$
Age Detection (FairFace)	$\mathcal{A}_B \uparrow$	0.544	$0.526_{\pm 0.036}$	$0.545_{\pm 0.054}$	$0.546_{\pm 0.048}$	$\underline{0.728_{\pm 0.017}}$	$\mathbf{0.848_{\pm 0.032}}$
	$\Phi_A \downarrow$	1.738	$1.926_{\pm 0.104}$	$1.863_{\pm 0.159}$	$1.212_{\pm 0.722}$	$\underline{0.910_{\pm 0.152}}$	$\mathbf{0.338_{\pm 0.265}}$
	$\Phi_{\text{demo}} \downarrow$	0.024	$\underline{0.028_{\pm 0.040}}$	$0.052_{\pm 0.064}$	$0.029_{\pm 0.026}$	$0.222_{\pm 0.035}$	$\mathbf{0.025_{\pm 0.011}}$
	$\Phi_{\text{eq}} \downarrow$	0.234	$\underline{0.057_{\pm 0.080}}$	$0.103_{\pm 0.128}$	$0.059_{\pm 0.053}$	$0.445_{\pm 0.071}$	$\mathbf{0.053_{\pm 0.019}}$

Table 1: Improving model **fairness** and **accuracy** with different schemes. The mean and standard deviation are reported. The best results are highlighted in bold and the second best results are highlighted with underline.

FAR	**Metrics**	**FF-DVP**	**FF-ADP**	**FF-LoRA**
Smiling (CelebA)	$\mathcal{A}_B \uparrow$	$0.905_{\pm 0.005}$	$0.877_{\pm 0.001}$	$0.907_{\pm 0.001}$
	$\Phi_A \downarrow$	$0.158_{\pm 0.043}$	$0.299_{\pm 0.028}$	$0.079_{\pm 0.009}$
	$\Phi_{\text{demo}} \downarrow$	$0.010_{\pm 0.011}$	$0.075_{\pm 0.002}$	$0.014_{\pm 0.006}$
	$\Phi_{\text{eq}} \downarrow$	$0.028_{\pm 0.016}$	$0.150_{\pm 0.014}$	$0.038_{\pm 0.019}$
Age (CelebA)	$\mathcal{A}_B \uparrow$	$0.839_{\pm 0.009}$	$0.806_{\pm 0.001}$	$0.852_{\pm 0.001}$
	$\Phi_A \downarrow$	$0.284_{\pm 0.203}$	$0.371_{\pm 0.036}$	$0.174_{\pm 0.018}$
	$\Phi_{\text{demo}} \downarrow$	$0.026_{\pm 0.020}$	$0.073_{\pm 0.007}$	$0.014_{\pm 0.010}$
	$\Phi_{\text{eq}} \downarrow$	$0.053_{\pm 0.039}$	$0.145_{\pm 0.013}$	$0.029_{\pm 0.0019}$
Age (FairFace)	$\mathcal{A}_B \uparrow$	$0.848_{\pm 0.032}$	$0.834_{\pm 0.001}$	$0.873_{\pm 0.002}$
	$\Phi_A \downarrow$	$0.338_{\pm 0.265}$	$0.135_{\pm 0.010}$	$0.173_{\pm 0.016}$
	$\Phi_{\text{demo}} \downarrow$	$0.025_{\pm 0.011}$	$0.034_{\pm 0.003}$	$0.041_{\pm 0.006}$
	$\Phi_{\text{eq}} \downarrow$	$0.053_{\pm 0.019}$	$0.067_{\pm 0.005}$	$0.083_{\pm 0.013}$

Table 2: Extending FF-DVP to FF-ADP and FF-LoRA.

smaller model, ResNet-18 (pre-trained on ImageNet) to run the selected baseline methods.

5.2 Evaluation: Fairness and Accuracy

Firstly, we compare FF-DVP with the baselines methods and present evaluation results in Table 1. Following (Lu et al., 2023), the experiments are repeated for 3 times. We observe: (1) the proposed FF-DVP effectively reduces the demographic bias of the FL model compared to all baseline fair FL methods. For instance, compared to CLIP (zero-shot), the model bias is reduced by approximately 87% w.r.t. Φ_{eq} on smiling detection; (2) FF-DVP achieves better fairness without necessarily sacrificing the accuracy. This is expected because our test dataset is subsampled from the original dataset and is balanced. After deploying FF-DVP, the FL model's performance on the minority group would be improved, leading to a natural increase in overall accuracy with the balanced test dataset. This indicates that the fairness-accuracy trade-off of FF-DVP is less pronounced. (3) FF-DVP achieves better training convergence than the baseline methods. Compared to baselines, FF-DVP achieves performance improvements in terms of both accuracy and fairness, whereas barely can traditional FL methods converge. We attribute the success of FF-DVP to the pre-trained foundation model as well as our novel fairness-aware adaptation strategy.

5.3 Fairness-aware PEFT

Next, we show that FF-DVP can generalize to other parameter-efficient fine-tuning (PEFT) methods, including adapter-style fine-tuning and Low-Rank Adaptation (i.e., LoRA). For the adapter-style fine-tuning, we combine FF-DVP with the attention-based adapter proposed in (Lu et al., 2023) (abbreviated as FF-ADP). As for LoRA, we combine FF-DVP with LoRA (abbreviated as FF-LoRA), and set the LoRA rank as 8, considering the communication cost of FL. As shown in Table 2, under the same FL setup, our debiasing strategy could still be effective in terms of debiasing FL models for both adapter-style fine-tuning and LoRA. Furthermore, we note that the number of intervened layers can influence the debiasing performance. For instance, FF-ADP only intervenes the output layer of the image encoder, which makes FF-ADP perform worse than FF-DVP and FF-LoRA. In comparison, FF-LoRA intervenes in all layers of the image encoder, and FF-LoRA generally achieves better performance than the others. As such, we highlight that our method is adjustable: FF-DVP allows the users to specify which transformer layers to debias, taking into account application-specific factors such as communication cost or dataset size.

FAR	Metrics	FF-DVP	w/o Contras. L.	w/o Classi. L.
Smiling (CelebA)	$\mathcal{A}_B \uparrow$	$0.905_{\pm 0.005}$	$0.435_{\pm 0.096}$	$0.894_{\pm 0.025}$
	$\Phi_A \downarrow$	$0.158_{\pm 0.043}$	$1.099_{\pm 0.300}$	$0.253_{\pm 0.228}$
	$\Phi_{\text{demo}} \downarrow$	$0.010_{\pm 0.011}$	$0.069_{\pm 0.029}$	$0.017_{\pm 0.009}$
	$\Phi_{\text{eq}} \downarrow$	$0.028_{\pm 0.016}$	$0.160_{\pm 0.088}$	$0.037_{\pm 0.019}$
Age (CelebA)	$\mathcal{A}_B \uparrow$	$0.839_{\pm 0.009}$	$0.513_{\pm 0.230}$	$0.838_{\pm 0.007}$
	$\Phi_A \downarrow$	$0.284_{\pm 0.203}$	$1.179_{\pm 0.589}$	$0.448_{\pm 0.152}$
	$\Phi_{\text{demo}} \downarrow$	$0.026_{\pm 0.020}$	$0.102_{\pm 0.072}$	$0.112_{\pm 0.038}$
	$\Phi_{\text{eq}} \downarrow$	$0.053_{\pm 0.039}$	$0.204_{\pm 0.142}$	$0.224_{\pm 0.076}$
Age (FairFace)	$\mathcal{A}_B \uparrow$	$0.848_{\pm 0.032}$	$0.538_{\pm 0.250}$	$0.868_{\pm 0.003}$
	$\Phi_A \downarrow$	$0.338_{\pm 0.265}$	$1.052_{\pm 0.706}$	$0.268_{\pm 0.058}$
	$\Phi_{\text{demo}} \downarrow$	$0.025_{\pm 0.011}$	$0.045_{\pm 0.046}$	$0.061_{\pm 0.011}$
	$\Phi_{\text{eq}} \downarrow$	$0.053_{\pm 0.019}$	$0.091_{\pm 0.001}$	$0.122_{\pm 0.021}$

Table 3: Ablation Study.

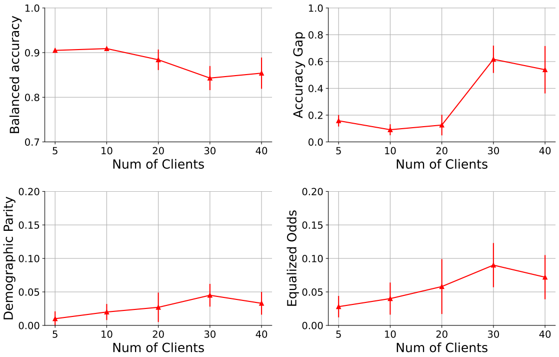


Figure 5: Scalability study on Smiling: we scaling up the number of clients to 40.

FAR	# Tkns	$\mathcal{A}_B \uparrow$	$\Phi_A \downarrow$	$\Phi_{\text{demo}} \downarrow$	$\Phi_{\text{eq}} \downarrow$
Smiling (CelebA)	10	$0.708_{\pm 0.162}$	$1.115_{\pm 0.719}$	$0.014_{\pm 0.010}$	$0.028_{\pm 0.021}$
	20	$0.905_{\pm 0.005}$	$0.158_{\pm 0.043}$	$0.010_{\pm 0.011}$	$0.028_{\pm 0.016}$
	30	$0.843_{\pm 0.092}$	$0.494_{\pm 0.479}$	$0.012_{\pm 0.008}$	$0.029_{\pm 0.009}$
	40	$0.894_{\pm 0.010}$	$0.323_{\pm 0.094}$	$0.044_{\pm 0.004}$	$0.087_{\pm 0.007}$
	50	$0.867_{\pm 0.026}$	$0.409_{\pm 0.218}$	$0.019_{\pm 0.017}$	$0.039_{\pm 0.033}$
Age (FairFace)	10	$0.838_{\pm 0.020}$	$0.466_{\pm 0.287}$	$0.059_{\pm 0.054}$	$0.131_{\pm 0.094}$
	20	$0.848_{\pm 0.032}$	$0.338_{\pm 0.265}$	$0.025_{\pm 0.011}$	$0.053_{\pm 0.019}$
	30	$0.843_{\pm 0.040}$	$0.416_{\pm 0.336}$	$0.032_{\pm 0.009}$	$0.066_{\pm 0.018}$
	40	$0.831_{\pm 0.034}$	$0.526_{\pm 0.319}$	$0.067_{\pm 0.044}$	$0.135_{\pm 0.087}$
	50	$0.785_{\pm 0.034}$	$1.067_{\pm 0.297}$	$0.217_{\pm 0.103}$	$0.434_{\pm 0.205}$

Table 4: Robustness w.r.t. number of tunnable tokens.

5.4 Ablation Study

We conduct an ablation study to evaluate the contribution of each key module of FF-DVP, namely the fairness-aware DVP and client-specific classifiers. The results are reported in Table 3. As expected, we observe that it is necessary to use contrastive loss to fine-tune the deep visual prompts. For instance, without the contrastive loss (as well as $\mathcal{L}_{fair}$), the adaptation basically fails. In comparison, f_{cls} indeed contributes to the fair representation learning of the model. For instance, the model fairness w.r.t demographic parity and equalized odds increase without using f_{cls}, indicating the contribution of f_{cls} to fair representation learning.

5.5 Scalability Study

We further study the scalability of FF-DVP w.r.t. the number of clients. We exclude the results of baseline methods in Figure 5, because baseline methods could barely converge. They make predictions randomly and achieve almost perfect but trivial fairness. We keep scaling up the number of clients until the training time exceeds the limit of a week. Moreover, we increase the number of clients but still control the group shifts to simulate non-i.i.d. heterogeneous and imbalanced data distributions (Appendix D). The results on smiling detection are visualized in Figure 5. We observe that, under a larger number of clients, our method could still converge and achieve similar or better fairness than the CLIP zero-shot performance.

5.6 Robustness Study

We finally study the relationship between the performance of FF-DVP and the length of visual prompts P. This hyperparameter is directly related to the communication cost in FL, as the visual prompts are shared and updated during the training process. In Table 4, we observe that there exists an optimal length of 20 to use the deep visual prompting. With shorter prompts, the expressive power of the model is reduced, indicating both debiasing and adaptation process is under-fitting the data. In comparison, with more tunable tokens, the training becomes slower and unstable. The instability is caused by the scarcity of the textual modality: there is insufficient amount of text data to train such a large model. For instance, in FairFace age detection, the FL model only achieves 78.5% accuracy with even 50 visual tokens. This observation indicates that more visual tokens will not necessarily improve the performance.

6 Conclusion

This work presents a novel fair FL framework using biased vision language models. FF-DVP provides an effective solution to develop fair ML models while protecting data privacy on users' end. We highlight the significance of this work: despite the promising integration of large foundation models with FL applications, existing literature largely overlooks the inherent bias of these foundation models. In contrast, our work is the first to address the inherent bias of these foundation models and their bias implication on FL applications. Finally, in addition to the performance of FF-DVP, we show that our method could easily extended to other PEFT methods for the adaptation of VLMs.

7 Limitations

One limitation of this work is that our method introduces extra hyperparameters. For different applications, one might need to finetune these hyperparameters, which brings extra computational cost, such as the trade-off factor λ and number of tunnable visual tokens. As for the actually trainable modules, there is only a small two-layer network and light-weight perturbations. Another limitation of this work is that our method only focuses on the bias, whereas CLIP has encoded other ethics-related issues (e.g., stereotypical data, racism and hate speech). Such malicious contents could have negative ethical implications on downstream FL applications as well. Therefore, a future research direction is to develop a benign, fair FL framework.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Hongyan Chang and Reza Shokri. 2023. Bias propagation in federated learning. *arXiv preprint arXiv:2309.02160*.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. 2023. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. *arXiv preprint arXiv:2308.12305*.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. 2021. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502.

Tao Guo, Song Guo, and Junxiao Wang. 2023a. pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. 2023b. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Junyuan Hong, Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H Dodge, and Jiayu Zhou. 2021. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 617–627.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Guanghao Li, Wansen Wu, Yan Sun, Li Shen, Baoyuan Wu, and Dacheng Tao. 2023. Visual prompt based personalized federated learning. *arXiv preprint arXiv:2303.08678*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. 2023. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR.

Alex Najibi. 2022. Racial discrimination in face recognition technology.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sijie Shen, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Fooling neural networks in face attractiveness evaluation: Adversarial examples with high attractiveness score but low subjective score. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 66–69. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. 2023. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19159–19168.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. *arXiv preprint arXiv:2208.09578*.

Hanhan Zhou, Tian Lan, Guru Venkataramani, and Wenbo Ding. 2022. On the convergence of heterogeneous federated learning with arbitrary adaptive online model pruning. *arXiv preprint arXiv:2201.11803*.

A Additional Prelinminaries

A.1 Federated Averaging

To find the optimal global model f_θ^* in an FL application, McMahan et al. (2017) proposed Federated Averaging (FedAvg). Specifically, at each round, each local client firstly receives a copy of the global model f_θ from the central server and trains the model with its own data. All clients will then obtain different local models $(f_{\theta^{(1)}}, f_{\theta^{(2)}}, ...f_{\theta^{(K)}})$, and send the trained model weights to the central server. Next, on the central server, the global model will be updated using a weighted-average of the received local model weights. Formally, FedAvg operates as follows:

$$\begin{aligned} \theta^* &= \frac{1}{\sum_k |\mathcal{D}^{(k)}|} \sum_{k=1}^{K} |\mathcal{D}^{(k)}| \cdot \theta^{(k)} \\ \text{s.t.} \quad \theta^{(k)} &= \arg\min_\theta \frac{1}{\sum_k |\mathcal{D}^{(k)}|} \sum_{i=1}^{|\mathcal{D}^{(k)}|} l(f_\theta(x_i^{(k)}, y_i^{(k)}) \\ k &\in \{1, ..., K\}. \end{aligned} \tag{13}$$

A.2 Fairness Notions

Definition 1 (Demographic Parity $\Phi_{\textbf{demo}}$ (binary case)). *For a classifier f_θ, demographic parity Φ_{demo} is defined as:*

$$\Phi_{demo}(\cdot) = |Pr(f_\theta(X) = 1|A = 0) - Pr(f_\theta(X) = 1|A = 1)|. \tag{14}$$

Furthermore, we also use fairness-aware accuracy metrics to measure the model's performance as in Yue et al. (2022), namely sub-group accuracy gap and balanced accuracy.

Definition 2 (Sub-group Accuracy $\mathcal{A}_{\text{sub}}$). *For a classifier f_θ, the sub-group accuracy $\mathcal{A}_{sub}$ is the accuracy on a specific demographic group characterized by A and Y.*

$$\mathcal{A}_{sub}(\cdot) = Pr(f_\theta(X) = Y|A = a) \tag{15}$$

Definition 3 (Accuracy Parity Φ_A). *For a classifier f_θ, the accuracy parity Φ_A sums up the absolute error among all demographic groups.*

$$\Phi_A(\cdot) = \sum_a \sum_y \sum_{a'} \sum_{y'} |\mathcal{A}_{sub}(a, y) - \mathcal{A}_{sub}(a', y')| \tag{16}$$

Definition 4 (Balanced Accuracy $\mathcal{A}_B$). *For a given classifier f_θ, the balanced accuracy computes the averaged sub-group accuracy for all demographic groups.*

$$\mathcal{A}_B(\cdot) = \frac{\sum_a \sum_y \mathcal{A}_{sub}(a, y)}{|\mathcal{A}| \cdot |\mathcal{Y}|}. \tag{17}$$

Since the model parameter is learned based on the training data distribution, local fairness could be highly heterogeneous across different local clients due to the discrepancy across different local data distributions.

A.3 CLIP Inference

CLIP is a large foundation model pre-trained with 400 million image-caption pairs. During pre-training, CLIP is trained to predict which images are paired with which texts. With such scale of pre-training data, CLIP is a powerful zero-shot image classifier and generalizes to different image classification tasks. To perform image classification, CLIP firstly encodes a query image and a set of text descriptions into latent representations. Next, CLIP computes the cosine similarity between the image representations and text representations. To produce the final prediction, CLIP selects the text with highest cosine similarity among all texts as the final prediction.

Formally, we use f_I to denote the CLIP image encoder and f_T to denote the CLIP text encoder. For a query image x and $|\mathcal{Y}|$ classes, we firstly craft a set of **candidate prompts** that contain class information (e.g., { `a photo of [class 1]`, `a photo of [class 2]...` }). Then, CLIP encodes x into z, and encodes the candidate prompts into representations $\{t_{candidate_1}, t_{candidate_2}, ..., t_{candidate_{|\mathcal{Y}|}}\}$, respectively. After computing cosine similarity between the image representation z and candidate prompt representations, CLIP selects the text with highest cosine similarity as the final prediction:

$$\begin{aligned} \hat{y} = \arg\max_c & \frac{\cos(z, t_{candidate_c})}{\sum_{c'} \cos(z, t_{candidate_{c'}})}, \\ \text{where} \quad & z = f_I(x), \\ & t_{candidate_c} = f_T(\text{"}\texttt{a photo of [class c]}.\text{"}), \\ & c \in \{1, 2, ..., |\mathcal{Y}|\}. \end{aligned} \tag{18}$$

B Modality-fused Classifier

The modality-fused classifier takes the prompted visual representation and the text representation as input. It then fuses the features from both data modalities:

$$\hat{y} = f_{cls}([\Pi(\tilde{z}), \Pi(t_{gt})]), \tag{19}$$

where $\tilde{z}$ is the prompted visual representation of x, and $t_{gt} = f_T(\texttt{ground-truth prompt of } x)$ is the encoded ground-truth prompts. $[\cdot, \cdot]$ represents the concatenation operation.

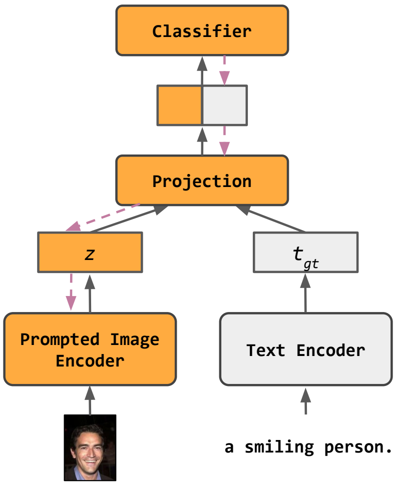


Figure 6: The classifier takes the debiased prompted visual representation and the text representation as input, and fuses the features from both data modalities.

C Algorithm in Pseudocode

Algorithm 1: FF-DVP

```
Input CLIP image encoder $f_I$, CLIP text encoder $f_T$, classifier $f_{cls}$ and projection matrix $\Pi$,
  datasets of local clients $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K$, task prompts $T = \{T_1, T_2, ..., T_C\}$, demographic-only
  prompts $T_A = \{T_{a_1}, T_{a_2}, .., T_{a_{|\mathcal{A}|}}\}$;
Hyperparameters learning rate $\eta$, trade-off factor $\lambda_1$ and $\lambda_2$, length of visual prompts $P$;
Initialize visual prompts $V$ of length $P$ at the central server ;
Clients download $f_I$ and $f_T$ ;
for global epochs do
    for k=1,2,...,K do
        Receive trainable models: $f_{cls}^{(k)} = f_{cls}$, $V^{(k)} = V$ ;
        for local epochs do
            compute $\mathcal{L}_1 = \mathcal{L}_{CLIP} + \lambda_1 \cdot \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} l_{fair}(x_i)$ ;
            compute $\mathcal{L}_2 = \mathcal{L}_{cls} + \lambda_2 \cdot \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \hat{\Phi}_{\cdot}(x_i)$;
            Update $V^{(k)}$ and $f_{cls}^{(k)}$ with gradient descent;
        end
        Send $V^{(k)}$ and $f_{cls}^{(k)}$ to the central server ;
    end
    Aggregate the received $V^{(k)}$s and $f_{cls}^{(k)}$s ;
end
Output the prompted CLIP image encode $\tilde{f_I}$ ;
```

D Implementation Details

D.1 Non-i.i.d. Local Data Distributions

We explicitly control population shifts for all clients, so that the local training data distributions are imbalanced and non-i.i.d. Below, we visualize the training data distribution of local clients.

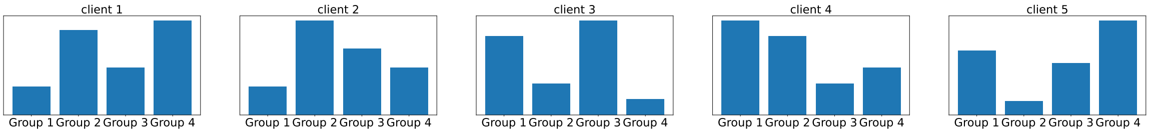


Figure 7: Client data distribution under 5 clients.

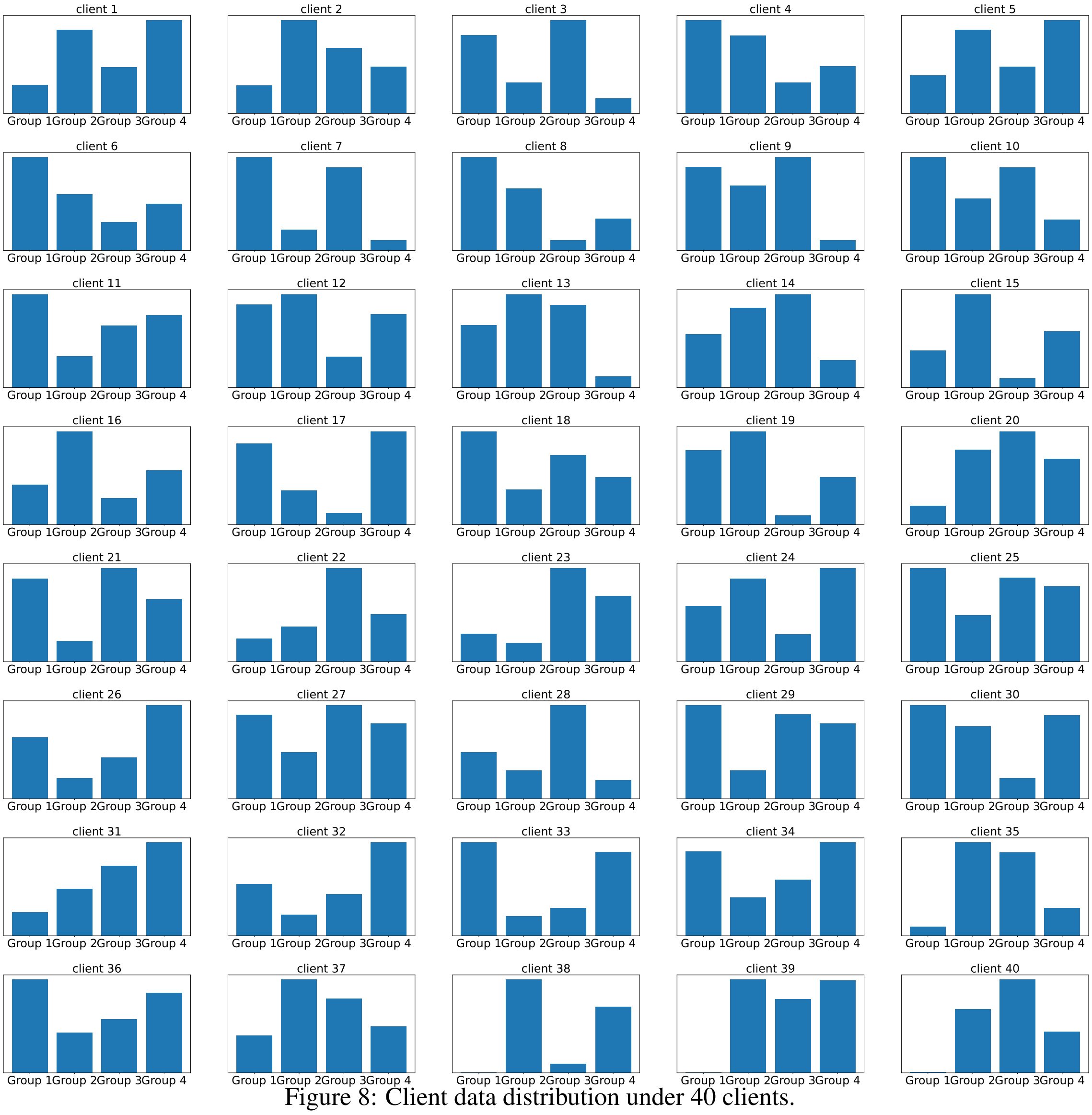


Figure 8: Client data distribution under 40 clients.

D.2 Hyperparameters and CLIP Configuration

For FF-DVP, FF-ADP, FF-LoRA, we select CLIP with configuration of ViT-L/14@336px to run experiments, and the learning rate is initialized as 1e-5. The models are optimized via AdamW. The local training epoch is 2 and the global epoch is also 2. For other baselines, including FedAvg, AFL, FairFed, FADE, we train the models by starting with the default training configuration and recommendations. However, we observed training divergence, and therefore, increased both local epochs (up to 30) and global epochs (up to 30). For all methods with key hyperparameters, we firstly performed grid search with the resolution of 0.1 until find the best performance. Based on that, we further reduce the search resolution to 0.01 until find best performance. Our hardware is NVIDIA A40.

D.3 Prompt Engineering

As a necessary input for CLIP, we manually design sets of prompts for different FAR applications. To determine whether the prompts are informative, we firstly compare the zero-shot performance of CLIP on FAR applications with biased or unbiased prompts. A biased prompt contains sensitive demographic information in addition to the class information. On the contrary, an unbiased prompt only contains class information. To enable FF-DVP, we also craft the demographic-only prompts defined in Section 4.1. For better understanding, we provide exemplar prompts for the FAR applications in Table 5 and Table 6.

In our experiments, the biased prompts generally achieve better accuracy because they impose demographic information on the face image, which potentially helps CLIP to recognize the face image. Given the zero-shot performance, we use biased prompts in experiments. Therefore, in our experiments, the biased prompts are used as the textual input for CLIP. Note that even if the biased prompts contain demographic information, FF-DVP can still effectively debias CLIP.

Type	Prompts

biased prompts (used)	`A photo of a male, and he is [smiling].`
	`A photo of a female, and she is [smiling].`
	`A photo of a male, and he is [not smiling].`
	`A photo of a female, and she is [not smiling].`

unbiased prompts	`A photo of a [smiling] person.`
	`A photo of a [not smiling] person.`

demographic-only prompts (used)	`A photo of a male.`
	`A photo of a female.`

Table 5: Crafted prompts for the CelebA smiling detection application. With biased prompts, CLIP achieves zero-shot accuracy of 84.8% whereas with unbiased prompts, the zero-shot accuracy is only 74.8%. Therefore, we choose to use the biased prompts in our experiments.

Type	Prompts

biased prompts (used)	`A photo of a male, and he is [young].`
	`A photo of a female, and she is [young].`
	`A photo of a male, and he is [not young].`
	`A photo of a female, and she is [not young].`

unbiased prompts	`A photo of a [young] person.`
	`A photo of a [not young] person.`

demo.-only prompts (used)	`A photo of a male.`
	`A photo of a female.`

Table 6: Crafted prompts for the CelebA age detection and FairFace age detection. To be consistent with smiling detection and considering the zero-shot accuracy, we also use biased prompts for age detection in our experiments.