Unified Spatio-Temporal Model for Fair Facial Attribute Classification

Madhurika Patil and Ajita Rattani*

Dept. of Computer Science and Engineering,
University of North Texas at Denton, Texas, USA
madhurikapatil@my.unt.edu; ajita.rattani@unt.edu

Abstract. Facial attribute classification models exhibit systematic demographic bias, leading to unequal performance across gender–racial subgroups. Existing mitigation strategies often impose a fairness–accuracy trade-off, yielding Pareto-inefficient outcomes - meaning any gain in fairness degrades accuracy and vice versa. Additionally, attributes such as gender and appearance traits exhibit high intra-class variation and inter-class similarity across demographics, necessitating fine-grained, robust representations. We propose a unified fairness-aware spatio-temporal framework that fuses CLIP-based semantic embeddings with geometric facial landmarks - a novel combination of geometric and semantic modalities for fairness-aware classification - via a learnable query-driven cross-attention mechanism for images UNIFY_SPA . The proposed framework is extended to video-based classification using spatio-temporal UNIFY_SPT attention with learnable positional embeddings to capture long-range inter-frame dependencies. Evaluation on five benchmarks, FairFace, UTKFace, LFWA+, CelebA, and CelebV-HQ, shows improved subgroup fairness over a ResNet-50 baseline, reducing the degree of bias from 4.2 to 1.15 and max-min ratio from 1.18 to 1.03, while maintaining a competitive overall accuracy of 94.47% - the best fairness metrics across all four published bias-mitigation baselines. For video-based classification on CelebV-HQ, our proposed bias mitigation model achieves only a 0.13% subgroup accuracy gap, indicating strong cross-group consistency and Pareto-superior performance over the ResNet-50 baseline.

Keywords: Fairness in AI · Facial Attribute Classification · Spatio temporal model.

1 Introduction

Automated facial analysis encompasses a broad spectrum of computer vision tasks like face detection, facial recognition and visual attribute classification such as gender, race, age and attractiveness prediction [14]. Among these, gender classification has attracted substantial research interest owing to its deployment across high-stakes real world domains including healthcare, retail analytics and identity verification [10]. Reliability of such systems is critically contingent on

* Corresponding author.

their ability to generalize equitably across diverse demographic populations. Empirical evidence shows that models trained on imbalanced or demographically skewed datasets develop systematic biases wherein spurious correlations between task-relevant attributes and demographic proxies such as race and age lead to biased predictions [22]. While gender classification remains the primary task, the proposed framework generalizes across age and other appearance-based gender-independent attributes.

Despite widespread commercial deployment, facial attribute classification models have pronounced intersectional performance disparities, especially at the intersection of gender and race. Female subgroups in underrepresented racial categories suffer disproportionately degraded accuracy [3,11,21]. These disparities reflect fundamental representational failure, in which models confuse demographic identity cues with task-discriminative facial features. Such failures have significant real-world consequences in high-stakes contexts like surveillance and identity verification where biased predictions affect marginalized populations.

Existing bias mitigation methods including self attention mechanism [15], adversarial debiasing [23], GAN based over sampling [27,20], multi task classification [5] and network pruning [13] all operate on the shared fundamental problem that their feature representations conflate the demographic appearance with task-relevant facial structure. Applying fairness correction on top of entangled representations causes a direct conflict, where reducing demographic bias hurts accuracy while improving accuracy reinforces bias [27]. This condition is known as *Pareto inefficiency* where improvement in fairness cannot be achieved without a corresponding loss in accuracy and no gain in accuracy can be made without the cost of fairness [27,23].

As these fairness violations stem from per-subgroup variance in entangled representations, the solution lies at feature level itself by building representations that are discriminative for the task but structurally resistant to demographic shortcuts. This motivates combining CLIP ViT-B/32 semantic embeddings which capture rich visual context, with 68-point facial landmark from Face Alignment Network (FAN) [2] which encodes geometric facial structure that is largely invariant to demographic appearance. A single learnable query, jointly optimized for classification accuracy and TPR parity [9], drives the cross attention mechanism—directing the model towards geometric, task relevant facial cues and making fairness a structural property of the representation rather than a post-hoc correction. We extend the model to fairness aware framework to video via temporal self attention with learnable positional embeddings and frame-level consistency regularization.

Contributions: 1) Unified Fair Facial Attribute for Spatial Model (UNIFY_SPA) cross attention framework which integrates frozen CLIP ViT-B/32 semantic embeddings with 68 point facial landmark geometry via learnable query driven multi-head cross attention mechanism which is optimized under TPR-parity fairness regularization. **2)** Extension of the spatial framework to Unified Fair Facial Attribute for Spatio Temporal Model (UNIFY_SPT) through temporal-self attention with learnable positional encoding and temporal consistency loss. **3)**

Extensive evaluation on FairFace, UTKFace, LFWA+, CelebA and CelebV-HQ datasets demonstrating Pareto-efficient fairness-accuracy trade-off achieving the lowest Degree of Bias (1.15) and tightest max-min ratio (1.03) compared to four published bias mitigation baselines.

2 Related Work

Prior bias mitigation strategies operate on entangled representations which conflate demographic appearance with task-relevant facial structure.

Bias Mitigation of Facial Attribute Classification. Demographic bias [3,11] of facial attribute classifier has been studied extensively. Dhar et al. [6] employed adversarial debiasing to suppress the gender information from pretrained facial embeddings, reducing gender bias in recognition systems. Ramaswamy et al. [22] uses latent space de-biasing to mitigate bias in facial attribute classification by targeting appearance based attributes like smiling, attractiveness and wavy hair which act as demographic proxies. Kisku and Patel [18] propose a framework which uses dual-attention on pretrained models with KL-divergence regularization to improve facial recognition and reduce demographic bias. Das et al. [5] address bias through multi-task joint classification of age, gender and race. Krishnan [12] and Padala and Gujar [17] improve fairness through semantic-preserving augmentation and Lagrangian-constrained loss function, respectively. Chuang and Mroueh [4] propose FairMixup which is a data interpolation strategy under the group fairness constraints, while Ramachandran [21] employ a self-supervised pipeline with meta-learning and pseudo-labeling for fair demographic classification. Zhang and Chunara [24] leverage the vision-language models to optimize the parity based fairness without using the group-labels. FineFace [16] and FairGRAPE [13] uses fairness aware gradient pruning which reduces demographic bias in facial attribute classification. In contrast, the proposed framework embeds fairness structurally at the representation level via complementary semantic and geometric modality fusion.

Spatial-temporal Models. Han et al. [8] employ CLIP-based facial component guidance to understand spatio-temporal cues for generalized deepfake detection. Hadid and Pietikäinen [7] demonstrate that appearance and motion cues via volumetric LBP descriptors improve gender recognition over image-based static classification models. Protsenko et al. [19] proposes a self-attention aggregation network which aggregates frame-level facial features from pretrained models while preserving the temporal order across video frames. Ali et al. [1] propose AgeFormer, extending spatial-temporal modeling to age classification through a two-stream transformer-EfficientNet architecture. To our knowledge, no prior work addresses demographic fairness in spatio-temporal video-based facial attribute classification models— a *gap* this work directly addresses.

3 Proposed Methodology

This section presents a unified fairness-aware spatio-temporal framework for facial attribute classification. The UNIFY_SPA for facial image based data is described in Section 3.1, The UNIFY_SPT for facial videos is described in Section 3.2 and

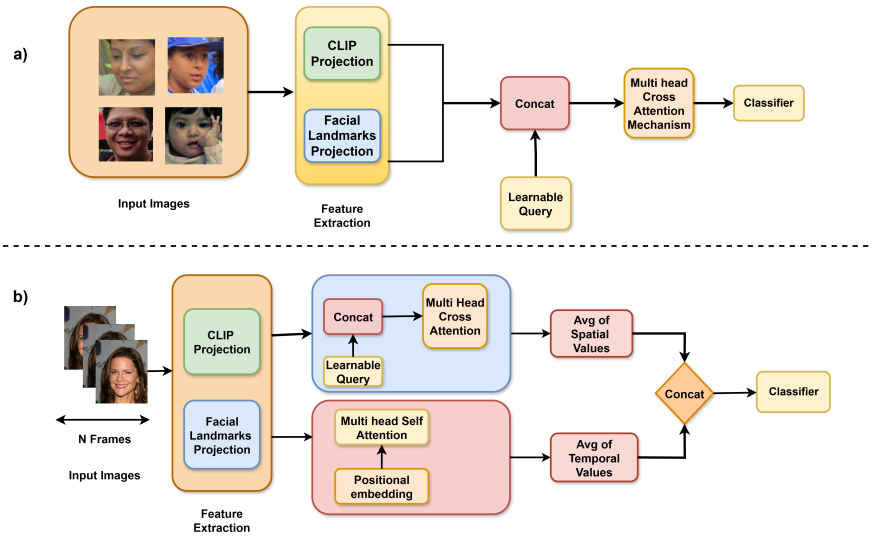

Fig. 1: Overview of the proposed unified fairness-aware spatio-temporal framework. **(a)** Spatial model (*UNIFY_SPA*): CLIP embeddings and facial landmark projections are concatenated with a learnable query and processed through a multi-head cross-attention mechanism with fairness-aware optimization for image-based spatial facial attribute classification. **(b)** Spatio-temporal model (*UNIFY_SPT*): CLIP embeddings and facial landmark features are extracted per frame and fused via a learnable query-driven multi-head cross-attention mechanism (Spatial Model, dashed border). Spatial and temporal features are averaged, concatenated, and passed to the classifier for video-level spatio-temporal facial attribute classification.

feature-level fusion mechanism is described in Section 3.3. The overview of the proposed framework is shown in Figure 1.

3.1 Spatial Architecture for Image Classification

The spatial model (UNIFY_SPA) model performs fairness-aware facial attribute classification which integrates high-level semantic embeddings with geometric facial structure through a cross-attention mechanism regularized via consistency and fairness aware optimization objectives.

3.1.1 Semantic and Geometric Facial Feature Extraction

CLIP based semantic embeddings: High dimensional semantic features using pretrained CLIP vision encoder with Vision Transformer (VIT-B/32) architecture, to obtain high dimension semantic representations of size 512 from the facial images. For each image, $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ the CLIP encoder computes $\mathbf{e} = f_{\text{CLIP}}(\mathbf{I}) \in \mathbb{R}^{512}$ where f_{CLIP} denotes CLIP visual encoder. The pretrained encoder weights

are held fixed to preserve cross-demographic generalization without fine-tuning. Freezing the encoder preserves the broad cross-demographic visual representations which are learned during large-scale pretraining, which, on fine-tuning, would risk overwriting.

Geometric facial landmark features: To incorporate structural facial geometry, we employ face alignment network (FAN)[2] to detect 68 facial landmarks from each facial image. These landmarks are represented as 2D coordinates $\ell_i(k) = (x_k, y_k) \in \mathbb{R}^2,\ k = 1, \ldots, 68$ and flattened into a 136-dimensional vector as $\mathbf{v}_i^{\text{landmark}} = [x_1, y_1, \ldots, x_{68}, y_{68}]^\mathsf{T} \in \mathbb{R}^{136}$.

3.1.2 Cross Attention Representation Encoder To fuse the complementary semantic and geometric modalities, a cross attention encoder is used which jointly processes CLIP embeddings and landmark features.

Linear feature projection: Since the dimensionality of both modalities reside in different dimension spaces, both are projected to a shared dimension space d (d = 256). The projections can be represented via linear transformations as $\mathbf{e}_{\text{proj}} = \mathbf{W}_e\mathbf{e} + \mathbf{b}_e, \quad \mathbf{e}_{\text{proj}} \in \mathbb{R}^{256}$ and $\mathbf{v}_{\text{proj}} = \mathbf{W}_v\mathbf{v}_{\text{landmark}} + \mathbf{b}_v, \quad \mathbf{v}_{\text{proj}} \in \mathbb{R}^{256}$ where, $\mathbf{W}_e \in \mathbb{R}^{256\times512}$ and $\mathbf{W}_v \in \mathbb{R}^{256\times136}$ are learnable weight matrices, and $\mathbf{b}_e, \mathbf{b}_v \in \mathbb{R}^{256}$ are bias terms. The bias term is set to 0 allowing the model to shift the weights flexibly.

Learnable query mechanism: The learnable query vector $q \in \mathbb{R}^{1\times d}$ serves as a trainable attention anchor which dynamically aggregates task-relevant information from both semantic and geometric modalities. The query is initialized as $q \sim \mathcal{N}(0, I)$ and functions as the query projection in cross-attention computation over concatenated features directing the attention of the model towards demographically invariant task-discriminative facial representations. The network parameters comprise $\Theta = \{\theta, q\}$ both the classification weights θ and learnable query q optimized via backpropagation under the objective :

$$\Theta^* = \arg\min_{\theta,q} \left[L_{\text{CE}}\big(f_\theta(q; e, v), y\big) + \lambda_{\text{fair}} \cdot L_{\text{fair}}\big(f_\theta(q; e, v)\big) \right]$$

where, f_θ is the classification network, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\mathcal{L}_{\text{fair}}$ is a fairness regularization term, and λ_{fair} controls the trade-off between accuracy and fairness.

Multi-head cross attention mechanism: To fuse semantic and geometric modalities the learnable query q attends over the concatenated projected features. Let $F = \text{Concat}\big(e^{\text{proj}}, v^{\text{proj}}\big) \in \mathbb{R}^{2\times d}$ which denotes the key-value source. Each attention head i, query, key and values can be computed as :

$$Q_i = qW_i^Q, \quad K_i = FW_i^K, \quad V_i = FW_i^V$$

where $W_i^Q \in \mathbb{R}^{d\times d_k}$ and $W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}$ are per-head projection matrices, with $d_k = d/h$ denoting the per-head dimension. Here W_i^K and W_i^V operate

independently on each of the two d-dimensional tokens in F. Each attention head computes: $\text{head}_i = \text{softmax}\Big(\frac{Q_i K_i^T}{\sqrt{d_k}}\Big) V_i$. The output is projected as :

$$\text{MultiHead}(q, F) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O$$

where, $W^O \in \mathbb{R}^{hd_k \times d}$ is the output projection matrix with q as the sole query source directing cross-modal feature retrieval towards task-discriminative facial representations.

Classification head: The attended feature representations are processed as a two-layer feedforward network comprising linear projections ($256 \to 256 \to 2$), RELU activation, dropouts ($p = 0.3$) with output class probabilities obtained via softmax.

3.1.3 Loss function and fairness-aware optimization

Supervised classification loss: The primary classification loss is standard binary cross entropy loss which is used to measure negative log-likelihood of the ground truth labels $y_i \in \{0, 1\}$ under the predicted probability $p_i \in (0, 1)$, over N samples as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \Big]$$

Consistency regularization via Jensen Shannon Divergence : In order to reduce sensitivity and improve robustness of the model to input perturbations independent of demographic labels, we use consistency regularization by injecting Gaussian noise on CLIP embeddings as follows:

$$\tilde{\mathbf{e}}_i = \mathbf{e}_i + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad \sigma = 0.005$$

Predictions are obtained for both original and perturbed embeddings as $\mathbf{p}_{\text{orig}} = \text{softmax}(f_\theta(\mathbf{e}_i, \mathbf{v}_i))$, $\mathbf{p}_{\text{aug}} = \text{softmax}(f_\theta(\tilde{\mathbf{e}}_i, \mathbf{v}_i))$. Consistency loss for JSD can be measured as :

$$\mathcal{L}_{\text{JSD}} = \frac{1}{2} \Big[\text{KL}(\mathbf{p}_{\text{orig}} \,\|\, \mathbf{m}) + \text{KL}(\mathbf{p}_{\text{aug}} \,\|\, \mathbf{m}) \Big], \quad \mathbf{m} = \frac{1}{2} \big(\mathbf{p}_{\text{orig}} + \mathbf{p}_{\text{aug}} \big)$$

where m is the mixture distributions; JSD is preferred for its symmetry and bounded gradients over alternative divergence measures.

Fairness-aware reward mechanism: To promote TPR (True Positive Rate) parity across demographic groups fairness reward is computed per batch as ratio of minimum to maximum group wise TPR as : $\text{TPR}_g = \frac{\sum_i \mathbb{1}[y_i = g \wedge \hat{y}_i = g]}{\sum_i \mathbb{1}[y_i = g]}$. $r_{\text{fair}} = \frac{\min(\text{TPR}_{\text{Female}}, \text{TPR}_{\text{Male}}) + \epsilon}{\max(\text{TPR}_{\text{Female}}, \text{TPR}_{\text{Male}}) + \epsilon}$, $\epsilon = 10^{-6}$,ϵ prevents numerical instability This reward scales the overall training loss:

$$\mathcal{L}_{\text{scaled}} = \big(1 - \lambda_f + \lambda_f \cdot r_{\text{fair}}\big) \mathcal{L}$$

where $\lambda_f = 0.5$ controls the strength of fairness enforcement. This helps in balancing TPR when loss is high in disparity. All hyperparameters are validated through ablation studies reported in the supplementary material.

3.2 Spatio Temporal Architecture for Image Classification

The spatio-temporal (UNIFY_SPT) attention model captures long range dependencies across video frames using multi-head self attention and learnable positional embeddings which aggregates per frame representations into video-level prediction for classification.

3.2.1 Model Architecture

Feature extraction : Let $\mathbf{X} \in \mathbb{R}^{B\times T\times D}$ denotes a batch of videos where B as the batch size, T as number of frames and d = 512, which is the dimension for CLIP embeddings extracted using pretrained ViT-B/32 encoder.

Positional Encoding: Self attention mechanism in the model is inherently permutation invariant due to which we add learnable positional embeddings which help in encoding temporal ordering information. The learnable positional encoding matrix can be defined as : $\mathbf{PE} \in \mathbb{R}^{T\times D}$, where each row $\mathbf{PE}_t$ corresponds to the positional encoding for frame position t. The positional encoded frames are computed as : $\mathbf{x}'_t = \mathbf{x}_t + \mathbf{PE}_t, \quad t \in \{1, 2, \ldots, T\}$ where, $\mathbf{x}'_t \in \mathbb{R}^D$ is a frame embedding which is augmented with temporal positional information. These positional embeddings are initialized from normal distribution scale by 0.001 to prevent overfitting in original frame embeddings.

Multi-head self attention : Temporal dependencies across the video frames are captured using multi-head self attention with number of heads = 8. For multi head self attention, we compute queries, keys and values based on positional encoding of frame features $\mathbf{X}' \in \mathbb{R}^{T\times D}$ as follows $\mathbf{Q} = \mathbf{X}'\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}'\mathbf{W}^K$, $\mathbf{V} = \mathbf{X}'\mathbf{W}^V$, where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{D\times D}$. Each attention head computes $\text{head}_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \quad$ where $d_k = D/H$ is the per-head dimension Multi-head attention applies this operation across $H = 8$ heads and combines the results:

$$\text{MultiHead}(\mathbf{X}') = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\mathbf{W}^O$$

where $W^O \in \mathbb{R}^{D\times D}$ is the output projection matrix.

Feature transformation and temporal pooling: From the output of multi-head self-attention model, we produce a sequence of frame level features which are refined by two layers of feed forward network which is applied independently to each frame. To reduce the feature dimensionality, Relu and layer normalization is used. The temporal average pooling of all the features is performed as:

$$z_{\text{temporal}} = \frac{1}{T}\sum_{t=1}^{T} h_t^{(2)} \in \mathbb{R}^{128}$$

where $h_t^{(2)} \in \mathbb{R}^{128}$ denotes the output of second feed-forward layer applied to frame t. The temporal forward pass can be expressed as :

$$z_{\text{temporal}} = \text{AvgPool}_T \circ \phi \circ \text{MultiHead} \circ \text{PE}(X) \in \mathbb{R}^{128}$$

where PE adds positional embeddings, MultiHead is self-attention, and ϕ is a 2-layer feed-forward network.

Temporal loss: To encourage temporal smoothness between attributes in each frame, we penalize the dissimilarity between each frame to maintain stability in each frame for video-level representation. To enforce temporal smoothness, we employ cosine similarity between each frame prediction.

$$\text{sim}(i,j) = \frac{(\mathbf{P}_i + \epsilon)^\top (\mathbf{P}_j + \epsilon)}{\|\mathbf{P}_i + \epsilon\|_2 \, \|\mathbf{P}_j + \epsilon\|_2}$$

Temporal consistency penalizes as follows :

$$\mathcal{L}_{\text{temporal}} = \frac{2}{T(T-1)} \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} (1 - \text{sim}(i,j))$$

where the normalization factor accounts for $\frac{T(T-1)}{2}$ unique frame pairs.

3.2.2 Feature Fusion Mechanism The model integrates spatial and temporal representations through concatenation using feature dimension. This simple concatenation method is used to preserve all information across both pathways and to learn optimal combination as $\mathbf{f}^{(st)} = [\,\mathbf{f}^{(spatial)}; \mathbf{f}^{(temporal)}\,] \in \mathbb{R}^{384}$. A progressive dimensionality reduction is implemented which uses a bottleneck architecture. Each layer applies linear transformation followed by normalization and non-linear activation with exception of final classification layer.

$$\mathbf{f}_{\text{fusion}} \in \mathbb{R}^{384} \xrightarrow{\text{Layer 1}} \mathbf{h}^{(1)} \in \mathbb{R}^{128} \xrightarrow{\text{Layer 2}} \mathbf{h}^{(2)} \in \mathbb{R}^{64} \xrightarrow{\text{Layer 3}} \boldsymbol{\ell} \in \mathbb{R}^2$$

Overall loss function: The loss function of the model balances multiple objectives like spatial robustness, temporal smoothness, fairness across the demographic groups and accurate classification making the model consistent and equitable. The total training loss combines the cross entropy loss of spatial temporal model, spatial loss, temporal loss and fairness weight as follows:

$$\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{CE}} + \lambda_s \, \mathcal{L}_{\text{spatial}} + \lambda_t \, \mathcal{L}_{\text{temporal}}) \times w_{\text{fairness}}$$

where $\lambda_s = 0.1$ and $\lambda_t = 0.05$ denote regularization coefficients, and w_{fairness} is a fairness weighting term that modulates the loss according to group-wise accuracy parity.

$$w_{\text{fairness}} = 1 - \alpha + \alpha \cdot r_{\text{fairness}},$$

$$r_{\text{fairness}} = \frac{\max_g \ \text{Acc}_g}{\min_g \ \text{Acc}_g}.$$

When group accuracies are equal $r_{\text{fairness}} = 1,\ w_{\text{fairness}} = 1$ and the loss is unmodified as the disparity grows $r_{\text{fairness}} > 1 \quad w_{\text{fairness}}$ increases which amplifies the overall loss so that optimization continues to push the model towards group parity.

4 Experimental Details

4.1 Datasets

We evaluate our proposed model on widely used benchmark datasets for fairness analysis of facial attribute classification which includes FairFace [10], UTK-Face [25], LFWA+ [14] and CelebA [14] datasets. All UNIFY_SPA model experiments use FairFace as primary dataset. LFWA+ and UTKFace provide race and gender annotations which enables intersectional fairness evaluation. CelebA lacks race labels therefore used for gender-based analysis and evaluation of 13 gender-independent attributes. These datasets and ImageNet pre-trained ResNet-50 baseline, fine tuned for cross entropy with no fairness intervention, are adopted in prior bias mitigation [16,22,12] studies making them well suitable for comparative evaluation. UTKFace and LFWA+ contain the races : White, Black, Indian, Asian and Others while FairFace dataset contains additional races like Southeast Asian, Middle Eastern and Latino Hispanic. For video-based evaluation of the proposed UNIFY_SPT model, CelebV-HQ [26] dataset is used which contains diverse facial attribute annotations spanning action, appearance and emotion based categories along with age, gender and 11 gender-independent attributes evaluated in our analysis. Additional details about the datasets are provided in the Table 1.

4.2 Evaluation Metrics

For fairness evaluation in gender classification, we use standard classification accuracy along with max-min accuracy ratio which measures gap between best and worse performing subgroups and Degree of Bias (DOB) [3] defined as standard deviation of subgroup accuracies [11]. A fair model is expected to have consistent accuracy across subgroups, with DOB close to 0 and max-min ratio close to 1 [9]. For gender-independent facial attributes we report True positive rate(TPR), Difference of Equal Opportunity(DEO) and Difference in Equalized Odds(DEOdds) [9]. Equal Opportunity requires TPR parity across subgroups while DEO quantifying violations for criterion [9]. DEOdds is used to measure correct positive predictions across groups. Maximum and minimum subgroup accuracies are also analyzed. An ideal classifier maintains high accuracy and TPR while maintaining low DEO and DEOdds thereby, achieving improved fairness [9].

5 Results

In section 5.1, UNIFY_SPA model on facial image dataset across gender-racial subgroups and gender independent attributes is evaluated. In section 5.2, the UNIFY_SPT model on facial videos for facial attribute analysis is evaluated.

Table 1: Dataset details including the number of images and demographic groups.

Dataset	Images/Videos	Demographic Groups
FairFace	108k	White, Black, Indian, Asian, Southeast Asian, Middle Eastern, Latino Hispanic
UTKFace	20k	White, Black, Indian, Asian, Others
LFWA+	13k	White, Black, Indian, Asian, Undefined
CelebA	202k	Only Gender Info
CelebV-HQ	36k	Only Gender Info

Table 2: Gender Classification Accuracy (%) on FairFace test set across demographics using ResNet-50 baseline and the proposed model. M: Male, F: Female. Best results in **bold**.

Model	Black		East Asian		Indian		Latino Hisp.		Middle Eastern		SE Asian		White		Max-Min Ratio	DOB	Overall
	M	F	M	F	M	F	M	F	M	F	M	F	M	F			
Baseline	94.4	82.4	97.2	88.9	98.1	93.3	95.6	92.2	97.8	92.4	94.4	91.5	96.5	89.9	1.18	4.20	93.20
UNIFY_SPA	92.59	91.32	92.38	95.41	92.35	94.53	94.01	95.24	97.72	94.66	93.23	94.06	95.94	95.81	**1.07**	**1.74**	**94.32**

5.1 Results for Spatial Classification Model

Intra-dataset Evaluation: Table 2 reports subgroup classification accuracy across gender-racial groups using our proposed UNIFY_SPA model against standard ResNet-50 baseline with comparison the published baselines in Table 5 and 7. Fairness is evaluated based on degree of bias (DOB) and max-min ratio. Baseline ResNet model achieves high overall accuracy (93.2%) but showcases demographic disparities with accuracy ranging from 82.4% in Black females to 98.1% in Indian males resulting in high max-min ratio (1.18) and DOB(4.2), indicating pronounced gender-race bias especially against female subgroups. In contrast, our proposed model achieves a high overall accuracy (94.32%) and reduces disparities with subgroup accuracies ranging from 91.32% to 97.72% which yields max-min ratio (1.07) and DOB (1.74). Female subgroups obtain comparable or higher accuracy than the male counterparts. Although CLIP ViT-B/32 provides a stronger backbone than ResNet-50, ablation study in supplementary confirm encoder capacity alone does not explain gains in fairness. OpenCLIP and DINOv2 under the same design yield DOB of 3.42 and 3.49 for OpenCLIP and DINOv2 respectively compared to 1.74 for CLIP.

The reduction in accuracy on LFWA+ is attributed to limited scale and demographic diversity relative to FairFace combined with fairness aware optimization moderating majority subgroup confidence to improve parity. The strong improvement on UTKFace reflects alignment between its demographic distribution and FairFace's training distribution which enables effective zero-shot generalization of CLIP's cross-demographic representations to under-performing subgroups. Fairness gains which are consistent across both datasets, confirm improvements which stem from proposed design.

Table 3: Cross-dataset Gender Classification Accuracy (%) on UTKFace and LFWA+ test sets across demographics. M: Male, F: Female. Best results in **bold**.

Dataset	Model	Asian		White		Black		Indian		Others		DOB	Max-Min Ratio
		M	F	M	F	M	F	M	F	M	F		
LFWA+	Baseline	96.5	78.3	96.9	89.1	98.7	78.8	97.9	95.0	96.8	91.1	7.70	1.26
	UNIFY_SPA	95.71	85.93	96.42	82.24	90.65	83.33	97.62	90.91	96.0	82.78	**6.16**	**1.18**
UTKFace	Baseline	88.2	65.7	90.2	72.2	94.6	67.6	95.1	74.9	89.1	78.8	11.10	1.45
	UNIFY_SPA	92.41	91.4	97.59	95.86	96.06	97.43	95.74	95.32	93.45	91.6	**2.30**	**1.06**

Table 4: Gender classification accuracy (%) on CelebA test set for ResNet-50 baseline and our proposed model. Best results in **bold**.

Model	M (%)	F (%)	Max–Min	Overall (%)	DOB (%)
Baseline	90.60	94.90	1.05	93.20	2.20
UNIFY_SPA	98.52	99.76	**1.01**	**99.14**	**0.87**

We evaluate the CelebA dataset for gender classification. The proposed model achieves substantially improved accuracy and fairness. A noticeable gap between male and female has been observed for the baseline model leading to DOB of 2.2 and max-min ratio of 1.05. Our model achieves parity performance across gender by reducing the DOB from 2.20 to 0.87 and reducing the max-min ratio from 1.05 to 1.01. Improvement in overall accuracy has been observed (from 93.20% to 99.14%) for our model. For CelebA, the model is trained and tested on CelebA using the standard train/test split making it a measure of classification performance within the same data distribution rather than cross dataset generalization.

Comparison to published work. Table 5 compares the performance of our model with the published bias mitigation techniques namely, Multi task learning [5], Adversarial debiasing [23], Consistency regularization [12] and deep generative views [20] across gender-racial subgroups. High overall accuracy has been observed in these methods while noticeable performance variability has been observed in the demographic subgroups resulting in higher degree of bias and max-min ratios. Lowest DOB is observed for our model (1.15) and lowest max-min ratio of (1.03) suggesting minimal disparity and bias between racial groups. Some methods like deep generative views achieve higher per group accuracies at the cost of increase in variance across demographic subgroups which results in higher DOB and max-min ratio. Our model achieves superior fairness-accuracy trade off as it emphasizes on demographic invariant cues while suppressing spurious correlations which lead to high accuracy across various subgroups while reducing disparity when compared to existing techniques.

Gender-independent facial attribute analysis. We evaluate the proposed model on CelebA dataset using 13 gender-independent attributes and compare it with the baseline as shown in the Table 6. The baseline achieves a high overall accuracy of 92.47% but exhibits substantial subgroup disparities with max-min accuracies ranging between 94.46% and 90.14% along with large TPR variation.

Table 5: Comparative Analysis with our model. A: Multi-Tasking, B: Adversarial debiasing, C: Consistency Regularization, D: Deep Generative Views. Best results in **bold**.

Method	Accuracy							Overall	DOB	Max-min
	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White			
A	91.26	94.45	95.05	95.19	97.35	94.20	94.96	94.64	1.81	1.067
B	87.66	91.93	93.67	93.80	95.96	91.81	93.96	92.69	2.62	1.095
C	90.83	93.60	94.48	94.48	95.94	93.64	94.57	94.00	1.59	1.056
D	91.64	95.29	95.38	95.32	97.11	97.11	94.92	94.72	1.72	1.060
UNIFY_SPA	91.91	94.00	93.53	94.66	96.59	93.64	95.88	94.47	**1.15**	**1.03**

Table 6: Facial attribute classification on CelebA (mean over 13 gender-independent attributes). Best results in **bold**.

Method	Acc	Max Acc	Min Acc	TPR	Max TPR	Min TPR	DEO	DEODD
Baseline	**92.47**	**94.46**	**90.14**	67.90	73.88	61.34	12.54	16.54
UNIFY_SPA	89.49	90.89	88.40	**90.55**	**92.66**	**84.23**	**8.41**	**9.11**

Table 7: Fairness methods on CelebA – mean scores over 13 gender-independent attributes. Best results in **bold**.

Metric	Weighting	Domain Indep.	Baseline Single	GAN Debias	Regularized	g-SMOTE Adap.	g-SMOTE	Baseline FairMixup	Fair Mixup	Our Model
Accuracy	91.45	91.24	92.47	92.12	91.05	92.56	92.64	**92.74**	88.46	89.49
Max grp acc	93.35	93.04	**94.46**	94.03	94.42	94.44	94.59	93.85	90.42	90.89
Min grp acc	89.06	88.93	88.93	89.85	87.86	90.36	90.35	**91.44**	86.36	88.40
TPR	64.02	70.74	67.90	66.13	54.20	67.11	66.14	79.13	46.67	**90.55**
Max grp TPR	67.41	75.61	73.88	70.36	56.11	74.06	73.43	80.89	47.85	**92.66**
Min grp TPR	59.41	66.05	61.34	61.25	52.34	59.78	58.32	72.92	44.27	**84.23**
DEO	7.67	9.56	12.54	9.11	3.77	14.28	15.11	7.97	**3.58**	8.41
DEODD	9.00	13.29	6.54	12.04	5.06	19.30	19.32	10.06	**4.29**	9.11

These imbalances cause high DEO (12.54) and DEODD (16.54) which indicates fairness violations. Our model reduces accuracy gap by 2.49% and improves min group TPR to 84.23% which eventually increases overall TPR (90.55%). Corresponding DEO and DEODD is reduced to 8.41 and 9.11 respectively, which demonstrates effective bias mitigation at a modest cost of overall accuracy. This trade-off is justified as fairness violations are reduced and strong overall classification performance is maintained.

Table 7 reports the performance of proposed model and compares against existing bias mitigation methods like re-weighting, domain independent learning, GAN based debiasing [20], regularization [12], data level balancing methods (g-SMOTE) and mix-up based approaches [4]. Some approaches (baseline FairMixup) achieve a higher overall accuracy but exhibit subgroup disparities particularly for TPR values with values ranging from 44.27% to 66.05% which leads to DEO and DEODD violations. Our model achieves higher TPR value (90.55%) and high minimum subgroup TPR value of 84.23% which showcases sensitivity

Table 8: Classification on CelebV-HQ test set for Gender and Age attributes. Age threshold: below 30 = young, above 30 = old.

Attribute	Overall Acc	Min Acc	Max Acc	TPR	Min TPR	Max TPR	DEO	DEODD
Gender	99.94	99.84	99.97	99.94	99.69	99.97	0.0031	0.0038
Age	92.20	87.83	94.69	84.62	81.44	87.71	0.06	0.11

among demographic subgroups. The accuracy gap between best and worst group is controlled (90.89% vs 88.4%). This showcases the model learns robust demographic features while maintaining TPR values and reducing DEO and DEODD violations. Low DEO and DEODD of certain competing methods reflect different fairness prioritization. These existing methods sacrifice TPR substantially (minimum subgroup TPR 44.27%–72.92%) whereas the proposed model achieves the highest minimum subgroup TPR (84.23%) by design.

5.2 Results for Spatio Temporal Classification Model

Intra-dataset evaluation: Table 8 reports the performance of the proposed spatio-temporal model (UNIFY_SPT) for gender attribute classification from facial videos, disaggregated by sensitive attributes—gender and age as no prior facial-aware video-based baselines exist for comparison, results are evaluated against fairness criteria in Section 4.2. For the gender attribute, the model achieves strong overall performance, with an accuracy of approximately 99.4%. The minimum and maximum subgroup accuracies are tightly clustered between 99.07% and 99.84%, demonstrating highly consistent performance across gender-groups. Similarly, the true positive rate (TPR) remains stable, ranging from 99.69% to 99.97%. The low disparity metrics, reflected by DEO (0.0031) and DEODD (0.0038), indicate minimal performance imbalance and confirm that the model satisfies fairness criteria while maintaining high predictive accuracy.

In comparison, greater variation is observed across age groups than across gender. The overall accuracy for age-based disaggregation is 92.2%, with subgroup accuracies ranging from 87.83% to 94.69%, indicating modest performance differences across age cohorts. Nevertheless, the observed disparities remain limited, suggesting that the proposed model generalizes effectively across both gender and age attributes while preserving fairness and robustness.

Supplementary material provides training details and hyperparameter justification, abbreviations, ablation studies validating CLIP over OpenCLIP and DINOV2 and JSD over KL-divergence, L2 and cosine similarity and gender-independent evaluations for both models on CelebA and CelebV-HQ datasets.

6 Conclusion

We present a fairness-aware spatio-temporal framework for facial attribute classification which addresses the demographic bias at the representation level. By integrating CLIP ViT-B/32 semantic embeddings with 68-point facial landmarks through a learnable query cross attention mechanism, jointly optimized under

TPR-parity regularization and JSD consistency loss- the UNIFY_SPA produced representations that are discriminative yet structurally resistant to demographic shortcuts. The framework is extended to video via temporal self-attention with learnable positional embeddings. UNIFY_SPT achieves a 0.13% max-min subgroup accuracy gap on CelebV-HQ which demonstrates near perfect cross-group consistency. Across the benchmark models, the proposed approach reduces the Degree of Bias from 4.2 to 1.15, achieving highest minimum subgroup TPR (84.23%) among all the compared methods, and improves the overall accuracy from 93.2% to 94.47% simultaneously, establishing a Pareto superior performance over published bias-mitigation baselines and suggesting that representation-level bias mitigation can reduce accuracy-fairness tension without explicit trade-off. Future work will evaluate the model behavior under degraded image conditions like heavy occlusion and low resolution, extend the fairness objective to multiple sensitive attributes including gender, age and race intersections, explore label-free formulations which annotate subgroups directly from data and broaden comparative evaluations to include baselines beyond ResNet-50.

References

1. Ali, A., Marisetty, A., Bremond, F.: P-age: Pexels dataset for robust spatio-temporal apparent age classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8606–8615 (2024)
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017)
3. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
4. Chuang, C.Y., Mroueh, Y.: Fair mixup: Fairness via interpolation. arXiv preprint arXiv:2103.06503 (2021)
5. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: Proceedings of the european conference on computer vision (eccv) workshops. pp. 0–0 (2018)
6. Dhar, P., Gleason, J., Souri, H., Castillo, C.D., Chellappa, R.: Towards gender-neutral face descriptors for mitigating bias in face recognition. arXiv preprint arXiv:2006.07845 (2020)
7. Hadid, A., Pietikäinen, M.: Combining appearance and motion for face and gender recognition from videos. Pattern Recognition **42**(11), 2818–2827 (2009)
8. Han, Y.H., Huang, T.M., Hua, K.L., Chen, J.C.: Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22995–23005 (2025)
9. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
10. Karkkainen, K., Joo, J.: 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. pp. 1548–1558 (2021)
11. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: 2020 19th IEEE international

conference on machine learning and applications (ICMLA). pp. 1028–1035. IEEE (2020)
12. Krishnan, A., Rattani, A.: A novel approach for bias mitigation of gender classification algorithms using consistency regularization. Image and Vision Computing **137**, 104793 (2023)
13. Lin, X., Kim, S., Joo, J.: Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In: European Conference on Computer Vision. pp. 414–432. Springer (2022)
14. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
15. Majumdar, P., Singh, R., Vatsa, M.: Attention aware debiasing for unbiased model prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4133–4141 (2021)
16. Manzoor, A., Rattani, A.: Fineface: Fair facial attribute classification leveraging fine-grained features. In: International Conference on Pattern Recognition. pp. 81–97. Springer (2024)
17. Padala, M., Gujar, S.: Fnnc: Achieving fairness through neural networks. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI. https://www. ijcai. org/proceedings/2020/0315. pdf Go to original source (2020)
18. Patel, S., Kisku, D.R.: Improving bias in facial attribute classification: A combined impact of kl divergence-induced loss function and dual attention. In: International Conference on Pattern Recognition. pp. 383–397. Springer (2024)
19. Protsenko, I., Lehinevych, T., Voitekh, D., Kroosh, I., Hasty, N., Johnson, A.: Self-attention aggregation network for video face representation and recognition. arXiv preprint arXiv:2010.05340 (2020)
20. Ramachandran, S., Rattani, A.: Deep generative views to mitigate gender classification bias across gender-race groups. In: International Conference on Pattern Recognition. pp. 551–569. Springer (2022)
21. Ramachandran, S., Rattani, A.: A self-supervised learning pipeline for demographically fair facial attribute classification. In: 2024 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2024)
22. Ramaswamy, V.V., Kim, S.S., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9301–9310 (2021)
23. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
24. Zhang, M., Chunara, R.: Leveraging vision-language models for fair facial attribute classification. arXiv preprint arXiv:2403.10624 (2024)
25. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)
26. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: CelebV-HQ: A large-scale video facial attributes dataset. In: ECCV (2022)
27. Zietlow, D., Lohaus, M., Balakrishnan, G., Kleindessner, M., Locatello, F., Schölkopf, B., Russell, C.: Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10410–10421 (2022)