LATENT FEATURE ALIGNMENT: DISCOVERING BIASED AND INTERPRETABLE SUBPOPULATIONS IN FACE RECOGNITION MODELS

Anonymous authorsPaper under double-blind review

ABSTRACT

Modern face recognition models achieve high overall accuracy but continue to exhibit systematic biases that disproportionately affect certain subpopulations. Conventional bias evaluation frameworks rely on labeled attributes to form subpopulations, which are expensive to obtain and limited to predefined categories. We introduce Latent Feature Alignment (LFA), an attribute-label-free algorithm that uses latent directions to identify subpopulations. This yields two main benefits over standard clustering: (i) semantically coherent grouping, where faces sharing common attributes are grouped together more reliably than by proximitybased methods, and (ii) discovery of interpretable directions, which correspond to semantic attributes such as age, ethnicity, or attire. Across four state-ofthe-art recognition models (ArcFace, CosFace, ElasticFace, PartialFC) and two benchmarks (RFW, CelebA), LFA consistently outperforms k-means and nearestneighbor search in intra-group semantic coherence, while uncovering interpretable latent directions aligned with demographic and contextual attributes. These results position LFA as a practical method for representation auditing of face recognition models, enabling practitioners to identify and interpret biased subpopulations without predefined attribute annotations.

1 Introduction

Face recognition technology (FRT) has seen widespread adoption in applications ranging from unlocking smartphones and securing sensitive facilities to assisting law enforcement and aiding in search operations. Rapid advancements in FRT have driven accuracy to levels that exceed human performance (Deng et al., 2019; Liu et al., 2017; Phillips et al., 2018; Schroff et al., 2015), promising enhanced efficiency and error reduction. Yet, as these systems proliferate, so do the risks associated with their misuse—ranging from privacy breaches to civil rights violations (Castelvecchi, 2020; Van Noorden, 2020; U.S. Commission on Civil Rights, 2024). Numerous studies have highlighted systematic biases in these systems (Cavazos et al., 2020; Drozdowski et al., 2020; Garcia et al., 2019; Wang & Deng, 2020; Gong et al., 2021; Robinson et al., 2020; Serna et al., 2022; Terhörst et al., 2021; Yucer et al., 2024). These biases can result in unfair treatment, exacerbating societal inequalities and undermining trust in their use (Garvie et al., 2016; Clayton, 2024).

Bias in face recognition algorithms refers to systematic differences in performance that affect certain groups of people based on characteristics such as race, gender, age, or other demographic factors. These biases creep in at various stages of the algorithm's lifecycle, from data collection and preprocessing to model design.

Traditionally, bias identification and mitigation rely on supervised approaches, where labeled datasets are used to define and evaluate group-specific performance gaps. However, such methods depend on explicit annotations (Terhörst et al., 2021), which may be unavailable, costly, or prone to unintended biases introduced by human annotators and unnoticed confounders (Balakrishnan et al., 2021). While synthetic datasets have been proposed as an alternative (Liang et al., 2023), they still only account for predefined attributes, leaving room for unconsidered biases to persist.

The challenge of identifying and analyzing bias in face recognition models has thus been largely constrained to known, predefined attributes. We address this limitation by developing a *label-free*

method that autonomously uncovers subpopulations correlated with performance disparities, without requiring attribute annotations or synthetic manipulations.

We introduce **Latent Feature Alignment** (**LFA**), an algorithm that uses latent directions in the embedding space to discover subpopulations and reveal interpretable structure in face recognition models. Using latent directions provides two key benefits over standard clustering: (i) **semantically coherent grouping**, where samples sharing common attributes are grouped more reliably than by distance-based methods, and (ii) **discovery of interpretable directions**, which align with semantic attributes such as age, ethnicity, or attire and expose bias-related subpopulations.

Our contributions are:

- Latent Feature Alignment (LFA): A novel attribute-label-free algorithm that discovers semantically coherent subpopulations in face recognition embeddings by aligning samples along latent directions.
- **Semantic coherence:** LFA consistently outperforms nearest-neighbor search and *k*-means clustering in grouping semantically related samples across multiple datasets (RFW, CelebA) and models (ArcFace, CosFace, ElasticFace, PartialFC).
- **Interpretable latent directions:** Directions discovered in the latent space by LFA correspond to human-interpretable attributes, validated both quantitatively and qualitatively.
- Bias discovery without annotations: LFA identifies subpopulations that exhibit systematic disparities in recognition performance, without requiring attribute labels.
- Resources: We validate vision—language model annotations as a reliable large-scale evaluation tool and release attribute annotations for RFW to support future bias auditing research.

2 Related Work

Our work intersects two central areas of research: (1) the discovery and interpretation of latent representations in computer vision models, and (2) the detection of bias in deep learning models. Prior work in these domains can be broadly classified by their supervision paradigm (supervised vs. unsupervised) and their application focus (interpretability or bias mitigation).

2.1 Supervised Approaches

Latent Concept Discovery in Vision Models. Early efforts to interpret neural networks revealed that human-defined concepts can often be represented as directions in latent space rather than as isolated neurons. Notably, *Concept Activation Vectors* (CAVs) (Kim et al., 2018) are learned by training a linear classifier to distinguish between concept-specific activations and random activations. Subsequent work has extended this idea to semi-supervised settings, using latent space decomposition to uncover concept vectors with partial labels (Graziani et al., 2023; O'Mahony et al., 2023).

Interpretable Feature Directions in Face Recognition. Plesh et al. (2024) explored interpretability in face recognition by identifying semantically meaningful directions in embedding spaces. Their method relied on annotated facial attributes and blended-image techniques to recover latent directions, which were then traversed for qualitative interpretation.

Synthetic Image Generation for Bias Detection. Counterfactual data augmentation has been proposed to diagnose bias by generating synthetic faces that selectively vary target attributes (Balakrishnan et al., 2021; Hutchinson et al., 2019; Joo & Kärkkäinen, 2020; Liang et al., 2023). While these methods support causal analyses, they are restricted to attributes chosen a priori, limiting their ability to uncover unexpected sources of bias.

2.2 Unsupervised Approaches

Discovery of Latent Semantics in Generative Models. Advances in Generative Adversarial Networks (GANs) have significantly enhanced our ability to interpret and manipulate latent spaces, leading to controlled and meaningful image generation (Härkönen et al., 2020). For example, SeFa

(Shen & Zhou, 2021) introduced a closed-form factorization for uncovering latent semantic directions without additional training or sampling. Other approaches enforce orthogonality and distinctiveness of latent directions (Song et al., 2023) or use contrastive objectives (Yüksel et al., 2021). These approaches demonstrate the promise of latent direction analysis, though they are generally applied to generative rather than discriminative models.

Unsupervised Bias Discovery via Clustering. Clustering-based methods have been proposed to detect bias in the absence of labels. For example, Krishnakumar et al. (2021) used hierarchical clustering of embeddings to identify subgroups with degraded classification performance in models trained for image classification. Glüge et al. (2020) showed that simple clustering-based measures are not reliable indicators of bias in face recognition models, arguing for more nuanced approaches.

These findings highlight both the promise and the limitations of unsupervised grouping for bias auditing: while clustering can reveal potential subgroups of concern, naïve distance-based methods may fail to align with semantically meaningful attributes or with actual performance disparities. Our method addresses these challenges by introducing *Latent Feature Alignment (LFA)*, which forms groups along latent directions in the embedding space. This provides higher semantic coherence than local similarity-based clustering, while simultaneously uncovering interpretable directions that correspond to meaningful, previously unlabeled factors of variation.

3 METHODOLOGY

3.1 LATENT FEATURE ALIGNMENT (LFA)

This work presents a novel framework that operates in the embedding space of face recognition models to discover subpopulations with distinct interpretable characteristics useful for bias auditing. At the core of our approach lies *Latent Feature Alignment (LFA)*, an algorithm that uses latent directions to identify semantically similar face images.

The algorithm computes a latent direction \vec{v} from a group of image embeddings. Specifically, the latent direction is obtained by taking the normalized average of the group's embeddings (see Algorithm 1). Next, among all face embeddings in the dataset, the algorithm identifies the one most aligned with this latent direction and adds it to the group (see Algorithm 2). Alignment is measured by projecting each embedding onto \vec{v} and selecting the embedding with the highest projection value.

This process repeats until the projection of the most aligned face embedding falls below a predefined threshold (τ) . By projecting the embeddings onto the latent vector, we capture the extent to which that direction is represented in each image. Note that all embeddings are normalized, so the embedding with the largest projection corresponds to the one with the greatest component of \vec{v} in its direction.

We now present the mathematical formulation of the method. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a dataset of N face images from which we aim extract a semantically consistent group. Using a deep feature extractor $\phi(\cdot)$, each image is mapped to a face embedding:

$$x_i \mapsto \phi(x_i) = \vec{e}_i \in \mathbb{R}^d$$

where d denotes the dimensionality of the embedding space. Let $\ell_i \in \{1, \dots, C_{\max}\}$ denote the identity label of x_i .

Given a current subset of indices $S \subset \{1, \dots, N\}$ of size n, define the count of samples with identity ℓ and the number of unique identities in S as:

$$c_{\ell} = |\{i \in \mathcal{S} : \ell_i = \ell\}|, \qquad C = |\{\ell_i : i \in \mathcal{S}\}|$$

The algorithm then iteratively performs the following steps:

1. Compute the latent direction of the subset:

$$\vec{v} = \frac{1}{C} \sum_{j \in \mathcal{S}} w_j \vec{e}_j$$
 with $w_j = \frac{1}{c_{\ell_j}}$ (weight of sample j) (1)

Algorithm 1 GETLATENTDIRECTION

```
Require: Embeddings \{\vec{e}_j\}_{j=1}^n and identity labels \{\ell_j\}_{j=1}^nEnsure: Latent direction \vec{v}\triangleright Set of unique identities1: \mathcal{I} \leftarrow \{\ell_j : j=1,\ldots,n\}\triangleright Set of unique identities2: for each identity i \in \mathcal{I} do\triangleright Count samples per identity3: c_i \leftarrow |\{j : \ell_j = i\}|\triangleright Count samples per identity4: end for\triangleright Inverse frequency weighting7: end for\triangleright Inverse frequency weighting8: \vec{v} \leftarrow \sum_{j=1}^n w_j \vec{e}_j\triangleright Weighted sum9: return \vec{v}
```

Algorithm 2 LATENT FEATURE ALIGNMENT

```
Require: Dataset embeddings E, initial group S
 1: while \max\{\boldsymbol{p}\} \geq \tau do
           \vec{v} \leftarrow \text{GetLatentDirection}(S)
 3:
                                                                                                ▶ Project embeddings onto direction
           m{p} \leftarrow m{E} \cdot m{ec{v}} \ / \ \|m{v}\|_2
 4:
           i \leftarrow \arg\max\{\boldsymbol{p}\}
                                                                                                               ▶ Most aligned embedding
 5:
           oldsymbol{S} \leftarrow oldsymbol{S} \cup \{ ec{oldsymbol{e}}_i \}
                                                                                                                                ▶ Add to group
           m{E} \leftarrow m{E} \setminus \{ ec{e}_i \}
 6:
                                                                                                                                        ▶ Remove
 7: end while
 8: return S, \vec{v}
```

2. Find the index of the most aligned embedding not already in S:

$$i^* = \underset{k \notin \mathcal{S}}{\operatorname{argmax}} \frac{\langle \vec{e}_k, \vec{v} \rangle}{||\vec{v}||_2}$$
 (2)

3. Check the stopping condition:

$$\frac{\langle \vec{e}_{i^*}, \vec{v} \rangle}{\|\vec{v}\|_2} < \tau \tag{3}$$

4. Update the subset by adding the index:

$$S = S \cup \{i^*\} \tag{4}$$

3.2 AUTOMATED DISCOVERY OF INTERPRETABLE AND BIASED GROUPS

Latent Feature Alignment relies on the existence of a coherent underlying factor shared across a group of images. When such a factor is present, the latent direction aggregates and reflects the common representation, yielding a semantically interpretable vector. Conversely, if the images are unrelated, the latent direction captures inconsistent signals, resulting in noise and an incoherent mixture of faces without a distinct representation. In this case, projections become uninformative, and the resulting groups lose interpretability.

The effectiveness of the method therefore depends critically on the initial group of images. Since no prior knowledge about groups is assumed, two strategies are possible: (1) exhaustive enumeration of small subsets (e.g., groups of size k < 10), or (2) initialization through pre-clustering. We adopt the latter, which scales to large datasets and provides semantically consistent starting points for *LFA*. Specifically, we construct an undirected similarity graph where each node corresponds to an embedding and edges are formed between embeddings exceeding a cosine similarity threshold of 0.5. Each connected component defines an initial group, while isolated nodes form singleton groups.

To compute the latent direction of a group, we average the embeddings. However, datasets often contain unbalanced numbers of images per identity, which can bias the mean toward dominant identities. To mitigate this, we use a weighted average that equalizes the contribution of each identity, ensuring that the latent direction reflects group-level semantic factors rather than identity frequency (Algorithm 1).

Table 1: Semantic coherence (\downarrow average attribute distance) across CelebA and RFW datasets. LFA thresholds and k values were chosen to yield comparable group sizes ($N \approx 100$). Across both datasets, LFA consistently produces more semantically coherent groups than nearest-neighbor search (NNS) and k-means.

	CelebA				RFW			
Method	ArcFace	CosFace	ElasticFace	PartialFC	ArcFace	CosFace	ElasticFace	PartialFC
NNS	8.61	8.59	8.60	8.52	3.13	3.04	3.13	3.10
k-means	7.65	7.61	7.55	7.50	2.88	2.66	2.86	2.86
LFA (ours)	6.42	6.41	6.57	6.28	1.86	1.52	1.84	1.55

After discovering semantically coherent groups via *LFA*, we evaluate their bias characteristics by analyzing impostor similarity scores within each group. Specifically, we compute the cosine similarity between embeddings belonging to different identities within the group. A high average impostor similarity indicates that the group is a biased subpopulation: the impostor score distribution shifts rightward, leading to elevated error rates in face recognition systems (Serna et al., 2022).

4 EXPERIMENTAL EVALUATION

We assess the effectiveness of our proposed method, Latent Feature Alignment (*LFA*), in identifying semantically meaningful and potentially biased subpopulations within face recognition datasets. Our evaluation focuses on three key aspects: (1) the semantic coherence of discovered clusters, (2) the interpretability of latent feature directions, and (3) the relationship between discovered groups and recognition performance bias. We conduct experiments on two widely-used face datasets: Racial Faces in the Wild (RFW) (Wang et al., 2019) and CelebA (Liu et al., 2015), and evaluate four state-of-the-art face recognition models.

4.1 Datasets and Models

We conduct our experiments on the following datasets:

RFW (Wang et al., 2019): A benchmark designed to measure racial bias in face recognition systems. It contains four subsets—Caucasian, African, Asian, and Indian—each with approximately 3,000 identities and 10,000 images. Its primarily source was the MS-Celeb-1M Guo et al. (2016) dataset.

CelebA (Liu et al., 2015): A large-scale dataset consisting of over 200,000 images of more than 10,000 celebrities, annotated with 40 binary attributes. It serves as a benchmark for attribute-based classification.

We annotate and validate RFW using state-of-the-art open-source Vision-Language Models (VLMs), including InternVL3 (Chen et al., 2024), Ola (Liu et al., 2025), Ovis2 (Lu et al., 2024), Qwen2.5-VL (Bai et al., 2025), and SAIL-VL (Dong et al., 2025). These annotations span 10 demographic and appearance attributes, enabling quantitative evaluation of the semantic coherence of discovered clusters. (Details of the annotation and validation in Appendix A.)

For our face embeddings, we use four widely adopted face recognition models: ArcFace (Deng et al., 2019), CosFace (Wang et al., 2018), Partial FC (An et al., 2021), and ElasticFace (Boutros et al., 2022).

4.2 SEMANTIC COHERENCE OF GROUPS

We measure intra-group attribute consistency using average Hamming distance between binary facial attributes (lower is better), which reflects the semantic homogeneity of each group. This metric quantifies the number of attribute differences between all pairs of images in the group, with the result representing the average number of changing attributes.

We compare against *k-means* clustering, a standard global partitioning method. Unlike classical clustering, we do not optimize for total coverage or disjointness; groups may overlap, and low-quality or ambiguous samples may remain ungrouped. To ensure fairness, we match the average

275

277

281

283 284

285

286

287

288

289 290 291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309 310

311 312

313

314

315

316 317

318

319

320

321

322

323

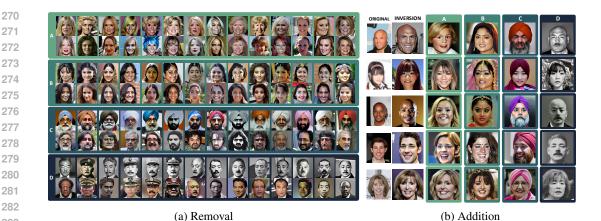


Figure 1: Left: Removal of semantic attributes from face images. For each group (A, B, C and D) the top image shows the inversion (embedding decoded) of the original image and the bottom one the original image after traversing back the embedding along the latent direction of the group. Right: Addition of semantic attributes. First column shows the original image, second column the inversion, and the rest of the columns (A-D) the original image decoded after traversing it along the latent direction of groups A-D.

group size between LFA and k-means: we run LFA with a threshold that yields groups of approximately n samples and set k such that the average cluster size is also $\sim n$. We note that optimizing k with standard clustering indices (e.g., Silhouette, Calinski-Harabasz) recovers the number of identities, which is unsuitable for our goal of discovering semantic attributes rather than identity partitions. We also compare against Nearest Neighbor Search (NNS), that is similar to LFA in the sense that it forms groups that can overlap and are not necessarily disjoint. For fairness we set k to n. This way, all compared algorithms posses an average group size of n.

Table 1 shows the comparison for n = 100. In the RFW dataset the average distance for standard clustering is a around 3 attributes when using all models, while for our algorithm is almost half. In the CelebA dataset our algorithm yields also a better semantic clustering. The reason behind the CelebA results being more than to times worse is twofold, first CelebA has 40 binary attributes in contrast to the 10 classes of the RFW annotations, and second, it has significant label noise Hand et al. (2018); Lingenfelter et al. (2022); Wu et al. (2023). (see Appendix A for details on robust annotation and validation, and Appendix E for quantitative ablations.)

We observe that, in contrast to the standard clustering baselines, LFA groups tend to be more cohesive and semantically similar. The advantage stems from LFA's ability to exploit the structure of the latent space, rather than relying purely on distance-based procedures that cannot capture broader semantic coherence.

INTERPRETABILITY OF LATENT DIRECTIONS 4.3

We evaluate qualitatively and quantiatively whether latent directions discovered by LFA correspond to meaningful attributes. For that, we decode embeddings as thay are traversed along each latent direction. To decode an embedding we use the identity-preserving generative model arc2face Papantoniou et al. (2024), that generates high-quality images of any subject given only its ArcFace embedding.

We selected four groups formed with our method in RFW database and tested the latent direction of each one. We traversed the directions to add or remove the common attributes. We removed the attributes of the face images in the selected groups by moving along its negative direction, and we added the attributes to random face images from CelebA dataset, similarly moving along the positive direction. The qualitative visual results are shown in Figures 1a and 1b. They reveal that latent directions discovered by LFA correspond to interpretable concepts such as blonde female, indian female, turban, or gray images of chinese men. The group A is composed of white blonde females, group B are young Indian women, group C Indian males with turbans, and group D Chinese

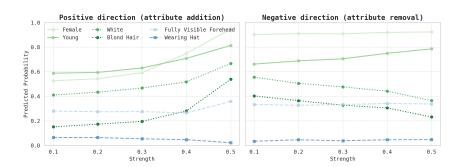


Figure 2: Attribute probability vs traversal strength. Predicted probabilities from an attribute classifier for images traversed along the latent direction of Group A (young, white, blonde, female). Relevant attributes (e.g., *Blond Hair, Young, Female, White*) change monotonically with traversal strength, while unrelated attributes (e.g., *Wearing Hat, Fully Visible Forehead*) remain stable.

men with mustaches in black-and-white.(Appendix D provides additional traversal visualizations and examples across multiple groups.)

Note that the directions were discovered using images from one database (RFW) but applied to the images of another database (CelebA), we did this to demonstrate the latent direction is a property of the latent space produced by the model and not tied to a particular database.

Also, note that traversing any direction in an arbitrary manner results in the decoding of a meaningless or generic image that has nothing to do with the original identity. The results are suprising since our algorithm operates on discriminative face recognition embeddings (e.g., ArcFace, CosFace), which are optimized for verification accuracy and not trained for semantic editing like generative models.

It can be seen that the traverse does not work equally well for every image or identity. This can be especially seen in group D, where some pictures has the mustache removed and others do not. Two key factors were identified as determinants: one, the image quality, and two, the identity. (More details on how traverse the embedding space in Appendix C.)

To quantify whether dominant directions discovered by LFA correspond to meaningful attributes we train an attribute classifier on CelebA following Lingenfelter & Hand (2021) and measure predicted attribute probabilities as embeddings are traversed and decoded.

Like for the qualitative analysis we used the latent directions identified by LFA on RFW onto CelebA embeddings. We traversed embeddings in both the positive (attribute addition on 1,000 random CelebA samples) and negative (attribute removal; 100 RFW samples aligned with the group's attributes) directions, measuring the predicted probabilities of multiple facial attributes at different interpolation strengths.

Figure 2 reports results for Group A (young, white, blonde, female). Traversals along the latent direction reveal meaningful and monotonic changes in relevant attributes (e.g., Blond Hair, Young, Female, White) while unrelated attributes (e.g., Wearing Hat, Fully Visible Forehead) remain stable.

Interestingly, we also observe some exceptions. For example, *Young* increases in both positive and negative traversals, and *Female* remains flat in the negative direction despite increasing in the positive one. This is consistent with what we observe in the qualitative results: sometimes, performing addition affects gender, but removal does not. Specifically, latent directions of female groups (A and B) change the gender of males to female, but latent directions of male groups (C and D) do not change it for females 1b.

While we do not yet have a definitive explanation, these effects may reflect non-linear structure in the embedding space, an underfitted representation, or entanglement of certain attributes. Such observations highlight both the interpretability and the complexity of the latent space learned by face recognition models.

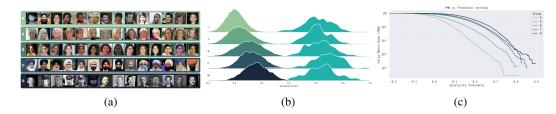


Figure 3: Qualitative and quantitative evidence of bias discovered by LFA on ArcFace. Left: Example images from different groups (R: random sample, A: Caucasian blonde females, B: young Indian women, C: Indian males with turbans, D: Chinese men with mustaches in black-and-white photographs). Center: Similarity score distributions for groups R–D. The right distribution of each group corresponds to genuine pairs (higher similarity), while the left one represents impostor pairs (lower similarity). Darker colors indicate higher error rates. Right: False Match Rate (FMR) curves for groups R–D, computed from genuine and impostor distributions (center). The extent of bias is indicated by the hue of the colour used; darker hues denote a higher degree of bias.

Table 2: Performance comparison of models under different training schemes. Comparison of False Match Rate (FMR) at fixed threshold (t = 0.2) between our method and attribute anotations. The four models are evaluated on four discovered groups plus a random group. Each subset includes the number of images and comparisons.

Model	Random	A (1	30 / 8k)	B (143 / 10k)		C (96 / 4k)		D (100 / 5k)	
	(265/35k)	LFA	Annotated	LFA	Annotated	LFA	Annotated	LFA	Annotated
ArcFace	0.0057	0.052 (x9)	0.0156 (x3)	0.180 (x32)	0.052 (x9)	0.215 (x38)	0.152 (x27)	0.318 (x56)	0.145 (x25)
CosFace	0.0045	0.052 (x12)	0.0196 (x4)	0.245 (x54)	0.068 (x15)	0.196 (x44)	0.138 (x31)	0.279 (x62)	0.117 (x26)
ElasticFace	0.0043	0.057 (x13)	0.0196(x5)	0.167 (x39)	0.052 (x12)	0.285 (x66)	0.188 (x44)	0.436 (x101)	0.175 (x41)
PartialFC	0.0016	0.016 (x10)	0.0057(x4)	0.106 (x66)	0.024 (x15)	0.081 (x51)	0.064 (x40)	0.140 (x88)	0.053 (x33)

4.4 BIAS DISCOVERY WITH ATTRIBUTE ANNOTATIONS

We evaluate whether the subpopulations discovered by LFA correspond to groups with higher error rates, and whether these trends are consistent with groups defined using explicit attribute annotations. We use the RFW database given that was specifically designed to measure bias (race bias).

Figure 3a displays example faces for different groups (R: random sample, A: Caucasian blonde females, B: young Indian women, C: Indian males with turbans, D: Chinese men with mustaches in black-and-white photographs). These exemplars illustrate the semantic coherence discovered by our method.

Figure 3b shows the distribution of cosine-similarity scores for genuine and imposter pairs within each group for the Arcface model. Genuine pairs are pairs of images belonging to the same person and imposter pairs are formed by pairing images from different identities. Similarity scores are computed using the cosine similarity of the corresponding face embeddings. Higher similarity values correspond to genuine pairs (right distribution of each group), while lower values correspond to imposters (left distribution).

Error rates are computed from the impostor and geniune comparisons: EER (Equal Error Rate), FNMR (False Non-Match Rate), and FMR (False Match Rate). The FNMR comes from the genuine distribution, which is similar across groups, but the impostor distribution varies greatly, so we focus on the FMR. Figure 3c shows FMR curves for the different groups evaluating the ArcFace model. The FMR curves show a clear disparity across groups, which is the sign of bias in biometric systems. The rest of the models (CosFace, ElasticFace, and PartialFC) have a similar score distributions and thus similar error rate curves (details are provided in Appendix B).

Table 2 reports false match rates (FMR) at a fixed threshold (t=0.2) for the LFA-discovered groups (A–D), compared against groups defined using explicit attribute annotations. Across all four backbones, LFA groups exhibit markedly higher FMR than a random subset, in some cases exceeding the baseline by more than $50\times$. Interestingly, although the exact groups of LFA groups do not fully align with those formed via attribute annotations, due to limitations in VLM attribute coverage (e.g., no attribute for black-and-white photos), the trends remain consistent: our method

tends to uncover even more homogeneous groups, always resulting in higher FMR. (see Appendix B for full bias evaluation metrics and confidence intervals.)

5 CONCLUSION AND DISCUSSION

In this work, we introduced Latent Feature Alignment (*LFA*), a simple yet effective method for discovering semantically coherent and interpretable subpopulations directly from face recognition embeddings. Unlike traditional clustering methods that rely purely on local distance metrics, *LFA* exploits latent directions in the embedding space, leading to groups that are both more coherent and more interpretable. Our experimental evaluation across CelebA and RFW demonstrates three main contributions.

First, we showed that *LFA* consistently produces groups with higher semantic coherence than *k*-means or nearest-neighbor search, across four state-of-the-art recognition models (ArcFace, CosFace, ElasticFace, PartialFC) and two widely used benchmarks (RFW, CelebA) (Section 4.2 and Appendix E). Second, we demonstrated that the dominant directions discovered by *LFA* correspond to meaningful semantic attributes, as evidenced by both qualitative traversals using arc2face and quantitative attribute probability trends (Section 4.3 and Appendix D). These results highlight the interpretability of the latent space—even when derived from discriminative embeddings not trained for generative editing (such as GANs or autoencoders). Third, we validated that the discovered groups present systematic disparities in recognition performance. Specifically, subpopulations identified by *LFA* exhibit higher error rates consistent with trends observed using explicit attribute annotations, confirming its value as a attribute-label-free method for bias discovery (Section 4.4 and Appendix B).

Our findings open several avenues for future work. While *LFA* uncovers meaningful latent directions, we also observed non-linear effects or entanglement between attributes, suggesting that more refined methods could improve disentanglement. Moreover, although we validated robustness with VLM-assisted annotations, the absence of certain attributes (e.g., black-and-white images) limits the completeness of comparisons. Extending the annotation coverage and applying *LFA* to larger and more diverse datasets could provide a fuller picture of its potential. Finally, we emphasize that *LFA* is not limited to faces: the principle of aligning groups along latent directions could apply broadly to other domains where embeddings capture high-level semantics.

In summary, *LFA* offers a practical and interpretable tool for analyzing embedding spaces, providing insights into both semantic structure and hidden biases. By broadening the scope of fairness auditing to domains where labeled attributes are limited, costly, or incomplete, LFA facilitates new methodologies for bias detection in high-stakes applications. We hope this work contributes to the growing literature on transparent and accountable machine learning.

6 LIMITATIONS

Initialization. The quality of *LFA* groups depends on initialization. Future work could investigate exhaustive exploration of initial combinations can yield more precise groups (Appendix D).

Observational Nature and Confounders. Our method is observational, relying solely on naturally occurring data. As such, it cannot disentangle semantic attributes from nuisance factors such as illumination, pose, or resolution. This limitation is inherent to any bias analysis performed without controlled interventions.

Linear Approximation of the Latent Space. *LFA* assumes locally linear directions in a highly dimensional embedding space. This linearization may overlook non-linear or entangled trajectories in the manifold, leading to partial or imperfect attribute respresentation.

REFERENCES

- Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1445–1449, October 2021.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv* preprint arXiv:2502.13923, 2025.
 - Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pp. 327–359. Springer, 2021.
 - Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1578–1587, 2022.
 - Davide Castelvecchi. Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–350, 2020.
 - Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.
 - Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024.
 - James Clayton. I was misidentified as shoplifter by facial recognition tech, May 2024. URL https://www.bbc.com/news/technology-69055945. Accessed: May 8, 2025.
 - Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.
 - Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025.
 - P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2): 89–103, 2020.
 - Raul Vicente Garcia, Lukasz Wandzik, Louisa Grabner, and Joerg Krueger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics* (*ICB*), pp. 1–6. IEEE, 2019.
 - Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up. *Georgetown Law Center on Privacy & Technology*, 18, 2016. URL www.perpetuallineup.org.
 - Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. How (not) to measure bias in face recognition networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pp. 125–137. Springer, 2020.
 - Sixue Gong, Xiaoming Liu, and Anil Jain. Mitigating face recognition bias via group adaptive classifier. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, June 2021. IEEE.
 - Mara Graziani, Laura O' Mahony, An-Phi Nguyen, Henning Müller, and Vincent Andrearczyk. Uncovering unique concept vectors through latent space decomposition. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

- Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A Dataset
 and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision* (ECCV), pp. 87–102, Amsterdam, The Netherlands, 2016. Springer.
 - Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
 - Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.
 - Ben Hutchinson, Emily Denton, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. In *Proceedings of the Fairness, Accountability, Transparency and Ethics in Computer Vision Workshop*, volume 2, 2019.
 - Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pp. 1–5, 2020.
 - Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning (ICML)*, pp. 2668–2677. PMLR, 2018.
 - Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference* (*BMVC*), 2021.
 - Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4977–4987, 2023.
 - Bryson Lingenfelter and Emily M Hand. Improving evaluation of facial attribute prediction models. In *International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021.
 - Bryson Lingenfelter, Sara R Davis, and Emily M Hand. A quantitative analysis of labeling issues in the CelebA dataset. In *International Symposium on Visual Computing*, pp. 129–141. Springer, 2022.
 - Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.
 - Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
 - Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv* preprint arXiv:2502.04328, 2025.
 - Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv*:2405.20797, 2024.
 - Laura O'Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. Disentangling neuron representations with concept vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3770–3775, 2023.
 - Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2Face: A foundation model for ID-consistent human faces. In *European Conference on Computer Vision (ECCV)*, pp. 241–261, 2024.

- P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacque line G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Castillo
 Chen, Carlos D. Chellappa, Rama White, David O'Toole, and Alice J. Face recognition accuracy
 of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.
 - Richard Plesh, Janez Križaj, Keivan Bahmani, Mahesh Banavar, Vitomir Štruc, and Stephanie Schuckers. Discovering interpretable feature directions in the embedding space of face recognition models. In 2024 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE, 2024.
 - Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face Recognition: Too bias, or Not Too Bias? In Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Virtual, 2020. IEEE.
 - Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.
 - Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, 2022.
 - Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings* of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1532–1540, 2021.
 - Ken Shoemake. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3): 245–254, July 1985.
 - Yue Song, Jichao Zhang, Nicu Sebe, and Wei Wang. Householder projector for unsupervised latent semantics discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7712–7722, 2023.
 - Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1): 16–30, 2021.
 - U.S. Commission on Civil Rights. The civil rights implications of the federal use of facial recognition technology. Technical report, U.S. Commission on Civil Rights, September 2024. URL https://www.usccr.gov/reports/2024/civil-rights-implications-federal-use-facial-recognition-technology.
 - Richard Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587 (7834):354–359, 2020.
 - Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.
 - Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9319–9328, Seattle, Washington, USA, 2020. IEEE.
 - Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 692–702, 2019.
- Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C King, and Kevin W Bowyer.
 Consistency and accuracy of CelebA attribute values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3258–3266, 2023.
 - Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. Racial bias within face recognition: A survey. *ACM Computing Surveys*, 57(4), 2024.

 Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14263–14272, 2021.

A RFW ANNOTATION AND VALIDATION

Table 3: OpenCompass VLM evaluation results of our selected Visual Language Models (VLMs) for annotating RFW. * Denotes the score of the unquantized version.

Method	Params (B)	Language Model	Vision Model	Avg. Score
InternVL3-38B-AWO	38.4	Owen2.5-32B	InternViT-6B-v2.5	77.8*
Ovis2-34B-GPTQ-Int8	34.9	Qwen2.5-32B	AIMv2-1B	76.5*
SAIL-VL-1.6-8B	8.33	Qwen2.5-7B	AIMv2 Huge	73.6
Ola-7b	8.88	Qwen2.5-7B	Oryx-ViT	72.6
Qwen2.5-VL-32B-AWQ	32.5	Qwen2.5-32B	QwenViT	_

Given the noise of CelebA attribute annotations, we decided to provide more robust evidence of the better clustering. For that we annotated RFW with five state of the art VLMs: InternVL3-38B-AWQ Chen et al. (2024), Ola-7b Liu et al. (2025), Ovis2-34B-GPTQ-Int8 Lu et al. (2024), Qwen2.5-VL-32B-Instruct-AWQ Bai et al. (2025), and SAIL-VL-1d6-8B Dong et al. (2025).

The models were chosen using the benchmark on the OpenCompass VLM leaderboard¹. The top 20 open source models belong to the 5 VLM families, and they vary in size. We chose one model from each family, 3 being of size \sim 32B, and 2 of size 7B. The 32B models are quantized versions of the models (to fit a GPU). Table 3 shows the models' vision and language backbones, number of parameters, and average score. The Qwen2.5-VL-32B model does not appear in the leadeboard.

The following prompt was used to elicit facial attribute classification from the VLMs:

```
You are an image analysis tool specialized in facial attribute classification.
For the provided face image, output a JSON object with the following attributes:
 "bangs": ["yes", "no"],
"baid": ["yes", "no"],
"beard": ["no", "mustache", "stubble", "full"],
"glasses": ["no", "regular", "sun"],
 "headwear": ["no", "beanie", "cap", "hat", "headband", "hijab", "helmet", "turban"]
Ensure the labeling is based on visible evidence only. If an attribute is unclear,
return "unknown".
Only output the JSON without any additional explanation or text.
Example JSON output:
 "gender": "female".
 "age": "middle-aged"
 "skin_color": "light",
 "ancestry": "asian",
 "hair_color": "black",
 "bangs": "no",
 "bald": "no",
 "beard": "no",
 "glasses": "sun",
 "headwear": "beanie"
```

https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Table 4 breaks down each attribute annotation showing how many samples fall into each class, both as a raw count and as a percentage of the total dataset. There is also an "Unknown" category for each attribute, which indicates how many images had no consensus among the models for that particular attribute. We quantify labeling bias computing compute agreement scores. These scores reflect the proportion of VLMs that agreed on the final label. For example, for the final label "mustache" in the 'beard' attribute of an image, if 3 VLMs say 'mustache', 1 says 'stubble' and 1 says 'no', the agreement ratio for that image and that attribute is 3/5 = 0.6. Table 4 presents mean agreement and standard deviation by attribute. Most classes show a high degree of agreement. Only *Hair Color* and *Ancestry* show less agreement for less frequent classes like *Latino/Hispanic*, *Indigenous*, or *Other*.

Table 4: Attribute annotations and agreement scores (mean and standard deviation) of the RFW dataset using five state-of-the-art Visual Language Models. *Unknown* class means there was no consensus among models. Total samples: 40,607.

Category	Class	Count	%	Mean Agr.	Std Agr.
	Middle-aged	19,072	47.0	0.88	0.15
Λα	Young	16,144	39.8	0.90	0.15
Age	Senior	5,312	13.1	0.92	0.14
	Unknown	79	0.2		
	Black	10,396	25.6	0.99	0.04
	White	10,161	25.0	0.95	0.11
	Asian	9,547	23.5	0.98	0.07
Ancestry	South Asian	9,191	22.6	0.93	0.12
,	Latino/Hispanic Middle Eastern	379 146	0.9 0.4	0.67 0.70	0.10 0.12
	Indigenous	146	0.4	0.70	0.12
	Unknown	769	1.9	0.61	0.16
	No	37,553	92.5	0.98	0.07
Bald	Yes	3,047	7.5	0.84	0.15
	Unknown	7	0.0		
	No	35,280	86.9	0.94	0.11
Bangs	Yes	5,297	13.0	0.86	0.16
	Unknown	30	0.1		
	No	30,356	74.8	0.91	0.12
ъ .	Stubble	3,940	9.7	0.74	0.15
Beard	Mustache Full	3,677 1,263	9.1 3.1	0.89 0.77	0.16 0.16
	Unknown	1,203	3.4	0.77	0.10
	Male	30,191	74.3	1.00	0.04
Gender	Female	10,400	25.6	0.99	0.04
	Unknown	16	0.0		
	No	34,589	85.2	1.00	0.03
Glasses	Regular	4,914	12.1	0.98	0.08
Giasses	Sun	1,080	2.7	0.94	0.12
	Unknown	24	0.1		
	Black	26,114	64.3	0.96	0.11
	Brown	5,795	14.3	0.82	0.15
II : G 1	Gray	5,529	13.6	0.91	0.14
Hair Color	Blonde Red	1,780 935	4.4 0.8	0.92 0.84	0.14 0.16
	Other	129	0.3	0.65	0.16
	Unknown	942	2.3	0.05	0.00
	No	35,543	87.5	0.99	0.04
	Cap	1,991	4.9	0.92	0.14
	Hat	1,089	2.7	0.87	0.17
	Headband	513	1.3	0.81	0.16
Headwear	Beanie	304	0.7	0.89	0.16
	Turban	262	0.6	0.94	0.13
	Helmet Hijab	255 199	0.6 0.5	0.89 0.91	0.14 0.15
	Unknown	451	1.1	0.71	0.13
	Medium	15,499	38.2	0.76	0.12
Skin Tone	Light	13,701	33.7	0.91	0.15
Skin Ione	Dark	10,952	27.0	0.97	0.09
	Unknown	455	1.1		

ANNOTATION MERGING VIA MAJORITY VOTING

To merge the predictions of the five VLMs into a single robust annotation per attribute, we used a majority voting scheme that explicitly ignores unknown values. For each image and attribute, we collected the predictions from all five models and discarded any labeled as unknown. The final consensus label was defined as the label that received strictly more than half of the valid (i.e., non-unknown) predictions. If no such label existed—either due to a tie or insufficient agreement—the annotation was marked as unknown.

Formally, let A_i^j be the label predicted by model i for attribute j, and define $V = \{A_i^j \mid A_i^j \neq \text{unknown}\}$ as the set of valid predictions for that attribute. Let n = |V| denote the number of valid votes. The consensus label C^j is then given by:

$$C^{j} = \begin{cases} \ell \in V, & \text{if } \operatorname{count}(\ell) > \frac{n}{2} \\ \operatorname{unknown}, & \text{otherwise} \end{cases}$$

This procedure ensures that only strong and unambiguous agreement among the models leads to an assigned label, while also allowing for occasional uncertainty (expressed via unknown) improving the reliability of the annotations used in our clustering analysis. We emphasize that the labels obtained via majority voting among VLMs serve as weak annotations and are not ground truth. In rare cases (fewer than 0.1% of samples), images with multiple faces were manually disambiguated by selecting the most centrally located face.

B BIAS EVALUATION

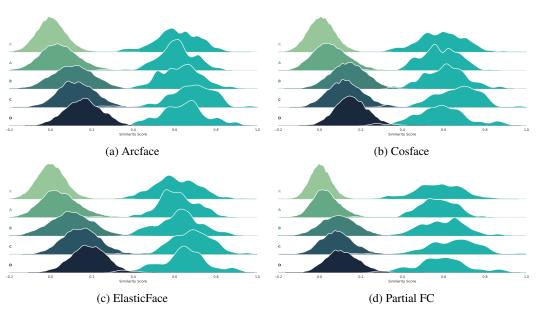


Figure 4: Distribution of cosine similarity scores for impostor and genuine comparisons within the same group.

We evaluate face recognition performance across some discovered groups (A-D) by forming genuine and imposter pairs exclusively within each group. Genuine pairs are pairs of images belonging to the same person and imposter pairs are formed by pairing images from different identities. Similarity scores are computed using the cosine similarity of the corresponding face embeddings.

Figure 4 shows the distribution of cosine similarity scores per model for genuine and imposter pairs within each group. The right distributions (gray) corresponds to genuine pairs (higher similarity values), while the left distribution (colored) corresponds to impostor pairs (lower similarity values).

Table 5: Performance (EER and FNMR) of the face recognition models across the discovered groups (A-D), with standard deviation across groups (lower means fairer, less biased). Each group includes the number of images and comparisons.

	EER				
Model	Group A (130 / 8k)	Group B (143 / 10k)	Group C (96 / 4k)	Group D (100 / 5k)	σ
ArcFace	8.70×10^{-3}	4.00×10^{-4}	2.42×10^{-2}	1.00×10^{-3}	0.0096
CosFace	5.10×10^{-3}	6.40×10^{-3}	1.83×10^{-2}	1.57×10^{-2}	0.0057
ElasticFace	8.20×10^{-3}	1.40×10^{-3}	1.22×10^{-2}	1.89×10^{-2}	0.0063
Partial FC	6.30×10^{-3}	1.20×10^{-3}	1.94×10^{-2}	1.78×10^{-2}	0.0077
		FNMR @	FMR=1%		
Model	Group A (130 / 8k)	Group B (143 / 10k)	Group C (96 / 4k)	Group D (100 / 5k)	σ
ArcFace	1.15×10^{-2}	0.00	2.08×10^{-2}	0.00	0.0087
CosFace	5.70×10^{-3}	0.00	3.12×10^{-2}	2.52×10^{-2}	0.0130
ElasticFace	1.15×10^{-2}	0.00	2.08×10^{-2}	3.36×10^{-2}	0.0123
Partial FC	5.70×10^{-3}	0.00	2.08×10^{-2}	1.68×10^{-2}	0.0083
		FNMR @	FMR=0.1%		
Model	Group A (130 / 8k)	Group B (143 / 10k)	Group C (96 / 4k)	Group D (100 / 5k)	σ
ArcFace	2.30×10^{-2}	0.00	3.12×10^{-2}	4.20×10^{-2}	0.0154
CosFace	1.72×10^{-2}	1.97×10^{-2}	4.17×10^{-2}	7.56×10^{-2}	0.0234
ElasticFace	2.30×10^{-2}	6.60×10^{-3}	3.12×10^{-2}	8.40×10^{-2}	0.0290
Partial FC	1.72×10^{-2}	6.60×10^{-3}	3.12×10^{-2}	8.40×10^{-2}	0.0297

Table 6: False Match Rates (FMR) with 95% confidence intervals at a fixed threshold (t=0.2) for each demographic group across four face recognition methods. Confidence intervals are computed via bootstrap resampling (1,000 iterations) at the image level.

Method	Group R	Group A	Group B	Group C	Group D
ArcFace	0.0057 ± 0.002	0.0522 ± 0.013	0.1798 ± 0.021	0.2167 ± 0.036	0.3181 ± 0.043
CosFace	0.0045 ± 0.002	0.0521 ± 0.014	0.2447 ± 0.027	0.1952 ± 0.038	0.2780 ± 0.036
ElasticFace	0.0043 ± 0.002	0.0569 ± 0.015	0.1676 ± 0.021	0.2861 ± 0.044	0.4365 ± 0.050
PartialFC	0.0016 ± 0.001	0.0161 ± 0.007	0.1060 ± 0.016	0.0809 ± 0.022	0.1400 ± 0.033

Darker colors indicate groups with higher error (imposter distributions shifted towards genuine), while lighter colors indicate groups with lower error.

From the resulting score distributions, we compute:

Equal Error Rate (EER): the operating point where false match rate (FMR) equals false non-match rate (FNMR).

FNMR at FMR=1% and 0.1%: the FNMR (False Non-Match rate) value when the decision threshold is set to yield an FMR (False Match Rate) of 1% and 0.1%.

Table 5 summarizes the verification metrics for all groups. Sample sizes vary per group and are reported alongside EER, FNMR@1%FMR, and FNMR@0.1%FMR. The standard deviation (σ) between A-D groups is a common intuitive proxy for bias in biometric systems as it indicates the statistical measure of variance across groups — higher σ means more disparity. FNMR@0.1%FMR means the false non-match rate at a fixed threshold for which the false match rate is 0.1%.

Although we observe that Groups C and D exhibit higher EER and FNMR than A and B, confirming the presence of bias in model performance for these groups, these groups lack the necessary number of comparisons for detailed measurements (notice the zeroes results for group B in FNMR @ FMR=1%).

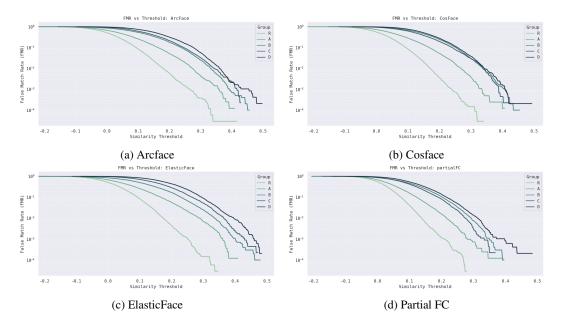


Figure 5: False Match Rate (FMR) curves for each face recognition model for all groups.

Although Groups C and D exhibit higher EER and FNMR compared to Groups A and B, indicating bias in the model's performance, they groups lack sufficient comparative data for detailed analysis at high error rates. For example, Group B has FNMR values of zero at FMR = 1%. To address this limitation, Table 6 reports false match rates (FMR) at a fixed threshold (t=0.2) with confidence intervals to assess statistical significance. Confidence intervals are computed via bootstrap resampling at the image level (1,000 iterations).

When evaluating face recognition systems at a fixed False Match Rate (FMR) of 0.1%, the decision threshold is chosen to allow only 1 in 1,000 impostor comparisons to be incorrectly accepted. The corresponding False Non-Match Rate (FNMR) is then computed over genuine comparisons using this threshold. However, when the number of genuine pairs is small—for example, approximately 120, like in our case—each false non-match has a disproportionately large effect on the FNMR estimate. Specifically, each false rejection increases the FNMR by roughly 0.83% (i.e., 1/120). This implies that the FNMR can only take on discrete values, such as 0%, 0.83%, 1.67%, etc., making it difficult to make fine-grained measurements. As a result, observed differences in FNMR at low FMRs may not be statistically significant or practically meaningful. Small improvements reported at this operating point could simply be due to random variation rather than true differences in system performance.

A more robust analysis of imposter distributions (in the order of thousands of scores) is to compute the FMR (False Match Rate) curve group-wise as a function of the threshold. Figure 5 shows the FMR curves for each face recognition model—ArcFace, CosFace, ElasticFace, and Partial FC—disaggregated by group. These curves illustrate how the False Match Rate varies across groups with the decision threshold, highlighting disparities between different groups.

C TRAVERSALS

To reconstruct images from facial embeddings, we use arc2face Papantoniou et al. (2024), which synthesizes high-quality, identity-preserving images given ArcFace Deng et al. (2019) embeddings.

Since ArcFace embeddings lie on the unit hypersphere, we cannot apply edits via direct vector addition. Instead, we use spherical linear interpolation (slerp) Shoemake (1985) to interpolate between embeddings. This approach respects the geometry of the hypersphere and provides fine-grained control over the edit intensity.

In contrast to linear interpolation or simple vector averaging—which may produce off-manifold results or require manual tuning of edit strength depending on the embedding and direction—slerp offers a principled mechanism for traversal along geodesics.

The slerp between two unit vectors p_0 and p_1 is defined as:

$$Slerp(p_0, p_1; t) = \frac{\sin((1-t)\theta)}{\sin \theta} p_0 + \frac{\sin(t\theta)}{\sin \theta} p_1$$

where $t \in [0, 1]$ is the interpolation parameter and $\theta = \cos^{-1}(p_0 \cdot p_1)$ is the angle between p_0 and p_1 . The interpolation smoothly transitions from p_0 to p_1 along the great arc on the hypersphere.

For the results shown in Figure 1a, we used a fixed interpolation strength of t=-0.5 to subtract and visualize the edit direction. For Figure 1b, the interpolation strength had to be adjusted between t=0.45 and t=0.5 depending on the embedding and direction. In a few rare cases, a value as low as t=0.4 was required to ensure plausible and visually coherent outputs.

D ABLATION STUDY – QUALITATIVE RESULTS

This section presents illustrative examples of high-quality groups formed using LFA on the CelebA dataset. All groups were constructed using embeddings extracted with Partial FC, and a threshold $\tau=0.35$ for LFA algorithm. The examples highlight attribute-based groups that exhibit strong semantic coherence, in part because all initial images belong to different identities.

Figures 6–10 display groups representing the following categories: males with turbans, males with baseball helmets, dark-skinned bald males, and light-skinned bald males. In each figure, the green box in the top-left corner indicates the initial group. Images are displayed in left-to-right, top-to-bottom sequence, reflecting the order in which they were added to the group.

We focus on CelebA due to its high number of images per identity, which increases the likelihood that grouping will be influenced by identity-related features, and makes it more complicated. When an initial group consists of images from the same identity, the latent direction in the embedding space may not yet be well defined. In such cases, expansion tends to favor look-alike identities rather than consistent attribute alignment.

To illustrate this, Figure 7 shows a group formed from four images of Asian females wearing tennis caps, all from different identities. The resulting group consistently includes images of individuals wearing (mostly white) tennis caps, indicating that the latent direction was successfully aligned with the intended visual attribute.

By contrast, Figures 11–13 show examples where the initial group consisted of a single image. In these cases, the grouping is driven primarily by identity similarity. While this is not inherently undesirable, it leads to groups where the dominant shared feature may be an identity-specific facial trait.

These examples validate the method's ability to discover visually meaningful groups when initialized with diverse identity faces. They also support our claim that exhaustive or informed initial selection improves grouping quality.

E ABLATION STUDY – QUANTITATIVE RESULTS

We evaluate the effectiveness of *LFA* versus *k-means* clustering on the CelebA dataset using four face recognition models: ArcFace, CosFace, ElasticFace, and PartialFC.

A fair comparison between *LFA* and *k-means* requires that both methods produce groups of similar size (i.e., similar average number of images per group). This is crucial because larger groups tend to exhibit greater variation, which increases the average intra-group attribute distance.

Therefore, instead of selecting a fixed value of k or threshold arbitrarily, we adjust k (for k-means) and τ (for LFA) to yield clusters with comparable average sizes. For example, a setting with k=2000 typically yields clusters with approximately 100 images. We adopt the same principle here and align group sizes across methods to allow meaningful comparison.

Table 7: Comparison of intra-group attribute distances between k-means and LFA at different levels of N (average number of samples). Thresholds τ for LFA and k for k-means are shown in parentheses. Bold values indicate better performance between the two methods under comparable group sizes.

Method	N (avg)	$\mathbf{ArcFace}\ (\tau\mid \mathbf{k})$	CosFace $(\tau \mid \mathbf{k})$	ElasticFace $(\tau \mid k)$	PartialFC $(\tau \mid k)$
K-Means	~50	6.58 (4000)	6.55 (4000)	6.04 (4000)	6.48 (4000)
LFA	\sim 50	6.09 (0.42)	6.08 (0.42)	6.21 (0.35)	6.13 (0.40)
K-Means	~100	7.65 (2000)	7.61 (2000)	7.55 (2000)	7.50 (2000)
LFA	~ 100	6.42 (0.40)	6.41 (0.40)	6.57 (0.33)	6.28 (0.39)
K-Means	~ 200	8.52 (1000)	8.49 (1000)	8.38 (1000)	8.28 (1000)
LFA	~ 200	6.87 (0.38)	6.95 (0.38)	7.03 (0.31)	6.64 (0.37)



Figure 6: CelebA images of males with turban. Initial group in green.

In Table 7, we report the average intra-group attribute distance (lower is better) for each method, at three different cluster size settings (50, 100, and 200 images per group). For each setting, we bold the lower (better) value between LFA and k-means for each model. Thresholds for LFA and values of k for k-means are shown in parentheses.

F IMPLEMENTATION DETAILS

All experiments were conducted on a single NVIDIA A100 GPU with 40GB of memory, an Intel Xeon IceLake-SP 8360Y CPU, and 512GB of RAM.

Processing Times Labeling the RFW dataset using visual-language models required several days. Extracting face embeddings with face recognition models took several hours.

Our method was evaluated on two datasets:

RFW (40k images): Runtime ranged from approximately 10 minutes (with a low threshold) to a few hours (with a high threshold).

CelebA (200k images): Runtime ranged from a couple of hours (low threshold) to several hours (high threshold).



Figure 7: CelebA images of asian females with a tennis cap. group seed in green.



Figure 8: CelebA images of males with baseball helmet. Initial group in green.



Figure 9: CelebA images of males with black skin and bald. Initial group in green.



Figure 10: CelebA images of males with white skin and bald. Initial group in green.



Figure 11: CelebA images of females. Single initial group in green.



Figure 12: CelebA images of females. Single initial group in green.



Figure 13: CelebA images of females. Single initial group in green.