How reaching a Reliable Ethnicity and Gender Inference (REGI)?

Anonymous authors
Paper under double-blind review

Abstract

Fairness evaluation in face analysis systems (FAS) uses automatic demographic attribute inference (DAI), which in turn relies on predefined demographic segmentation. However, the validity of fairness auditing hinges on the reliability of the DAI process. We revisit this process by proposing a fully reproducible ethnicity+gender prediction pipeline (REGI) that includes curation of diversified training and test datasets and a transfer-learning-inspired model training strategy. We ground our approach in a per-identity bound on the probability that an image is assigned a different segment than its identity's mean embedding. We show that the decision class this bounds covers, log-likelihood ratios of von Mises-Fisher mixtures, is realized by an RBF-SVM head on a pretrained face recognition encoder. Its margins being driven by the coverage of the training set motivates further our diversified training set WEG^{tr}. We audit this pipeline across three dimensions: accuracy, fairness, and a newly introduced notion of robustness, defined via intra-identity consistency. We compare our approach with seven competitive baselines across multiple datasets and training setups, including pipelines trained specifically for attribute prediction and generalist VLMs. Our pipeline outperforms all baselines for ethnicity, the more challenging attribute, across the three audited dimensions. REGI also ensures an interesting performance-efficiency balance, beating all the dedicated attribute predictors even with smaller computational budgets. Importantly, the newly proposed attribute training dataset generalizes better than existing counterparts, while our new test dataset, more challenging, is composed more suitably for evaluating the three audited dimensions. To promote reproducibility and adoption, we release the dataset metadata, full codebase, pretrained models, and evaluation toolkit.

1 Introduction

The accuracy of automatic face analysis systems (FAS) has significantly progressed over the past decade, driven by increasingly sophisticated models (Deng et al., 2019; Kim et al., 2022) and the availability of large-scale training datasets (Zhu et al., 2023). FAS now powers critical applications, from smartphone authentication to national security infrastructure (Apple Inc., 2017; U.S. Government Accountability Office, 2021; Stodder & Warrick, 2022). Given this societal impact, it must be not only accurate but also fair to ensure responsible deployment and public trust (Raji et al., 2020). Numerous studies have shown that FAS performance can vary significantly across demographic groups (Buolamwini & Gebru, 2018; Fournier-Montgieux et al., 2025; Grother et al., 2019; Gwilliam et al., 2021; Kotwal & Marcel, 2025; Sarridis et al., 2023). These disparities are attributable to demographic imbalance in training data (Gwilliam et al., 2021; Wu & Bowyer, 2023; Neto et al., 2025), limited data diversity (Fournier-Montgieux et al., 2025), visual variability (Terhörst et al., 2021), or suboptimal training procedures (Krishnapriya et al., 2020; Deka & Sutherland, 2023; Kamishima et al., 2012). While these studies have advanced our understanding of algorithmic bias, they largely overlook a foundational assumption: the reliability of the demographic attribute inference (DAI) process itself. In practice, attribute inference typically relies on off-the-shelf tools (Kotwal & Marcel, 2025), such as the FairFace pipeline (Kärkkäinen & Joo, 2021), referred to as FFNet to avoid any confusion with the FairFace dataset, the DeepFace library (Serengil & Özpınar, 2024), or VLMs (Bai et al., 2025). Yet, as illustrated in Figure 1, these tools offer suboptimal performance and compromise the reliability of downstream attribute analysis. We thus revisit ethnicity and gender inference

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

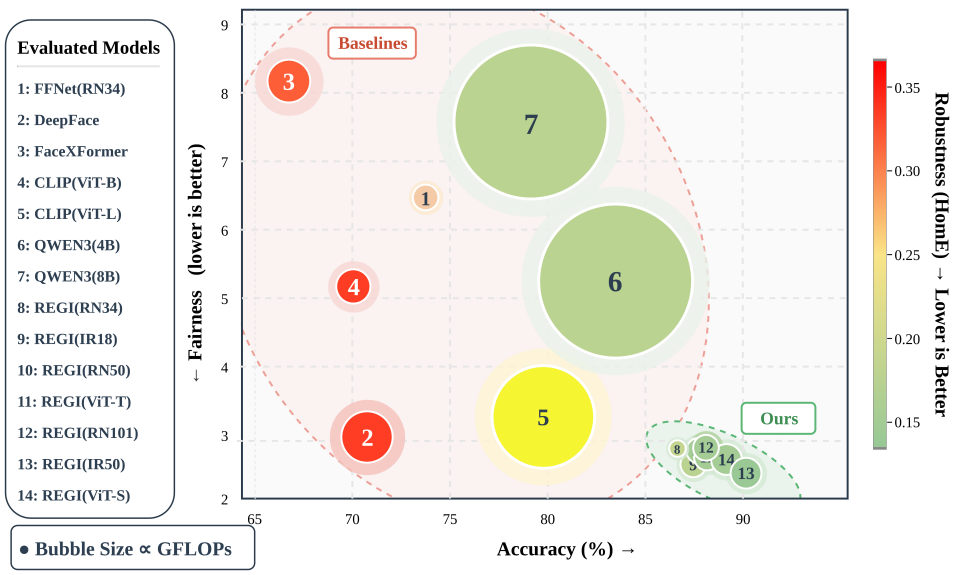


Figure 1: Ethnicity inference results (4 classes) audited using accuracy, fairness, and robustness dimensions on our proposed WEG^t dataset for baselines (FFNet (Kärkkäinen & Joo, 2021), the DeepFace Library (Serengil & Özpınar, 2024), FaceXFormer (Narayan et al., 2025a)), pretrained VLMs (Bai et al., 2025; Radford et al., 2021), and several instances of the proposed approach. Higher values are better for accuracy. **Lower** values are better for **fairness**. **Greener** is better for robustness. Circle sizes are proportional to GFLOPs per image, except for QWEN3, which would be too large to fit (we report the numerical value in Table 1). An ideal DAI system would be a collapsed circle located at the bottom-right corner of the plot.

and show that significant performance gains can be achieved by carefully curating DAI datasets and replacing the standard end-to-end learning strategy. We predict four coarse ethnicities (*African*, *Asian*, *Caucasian*, *Indian*) and two genders (*Female*, *Male*) for comparability with existing systems and datasets. A *segment* denotes one class of the resulting ethnicity $\times$ gender taxonomy.

First, we introduce reliable ethnicity and gender inference (**REGI**), which fits an RBF-SVM head on a frozen face recognition encoder pretrained with an angular-margin loss (Wang et al., 2018; Deng et al., 2019; Kim et al., 2022), instead of the end-to-end training or full fine-tuning used by existing tools. We show that this combination falls within a family of decision functions that offers a robustness guarantee. Furthermore, we show that on normalized embeddings, this family amounts to modeling each segment as a mixture of von Mises-Fisher (vMF) distributions, and the RBF-SVM head to a likelihood-ratio test between such mixtures centered on its support vectors.

Second, we propose **WEG**, a large curated **E**thnicity and **G**ender dataset that resamples **W**ebFace (Zhu et al., 2023) while emphasizing demographic diversity. The theory makes diversity an operative lever, since the margin of an identity is large only if the training set covers it. We aggregate FFNet (Kärkkäinen & Joo, 2021) predictions at the identity level to improve label reliability, boost diversity by matching WebFace identities to seed identities covering a wide range of ethnicities within each coarse category, and separate test and training identities.

Third, we advocate for conducting ethnicity and gender auditing across three complementary dimensions: **accuracy, fairness, and robustness**. Accuracy underpins correct interpretations; fairness ensures equitable error distribution across groups; robustness, an underexplored property, promotes intra-identity consistency. We introduce a robustness metric based on the simple yet sound hypothesis: *An ideal demographic classifier should predict the same segment for all images of the same person.* We formalize it through an entropy-based metric that generalizes across demographic taxonomies and is controlled by the bound above.

Across various backbones, attribute training datasets, pretraining datasets, and classification heads, REGI simultaneously improves accuracy, fairness, and robustness compared with widely used systems, namely FFNet (Kärkkäinen & Joo, 2021), DeepFace (Serengil & Özpınar, 2024), and FaceX-

Former (Narayan et al., 2025a), as well as large pretrained VLMs, such as CLIP (Radford et al., 2021) or Qwen3 (Bai et al., 2025), while also being more efficient. Its RBF-SVM head also outperforms linear and MLP heads, end-to-end training, and fine-tuning. We release the code, the dataset metadata, and the pretrained models under a restricted-use license limited to fairness auditing and related research (see section 7).

2 Related Work

Fairness evaluation (Raji & Fried, 2020; Sarridis et al., 2023) is needed to comprehensively audit AI-driven face analysis systems deployed in practical applications, including border control (Hidayat et al., 2024), consumer authentication (Teoh et al., 2021), public safety (Miethe et al., 2025), and face-based payment (Rija et al., 2022). These studies highlight significant performance variation across demographic segments, due to imbalance in face recognition training datasets (Gwilliam et al., 2021; Wu & Bowyer, 2023), insufficient data diversity (Fournier-Montgieux et al., 2025), or biases in the architectures or training strategies (Dooley et al., 2023; Nagpal et al., 2019; Rezgui et al., 2024). While understanding and mitigating biases is important, little effort is devoted to the reliability of the demographic attribute inference that enables their conclusions.

Training datasets are a core component of demographic classification. In supervised settings, their annotation follows a predefined segmentation, as in RFW (Wang et al., 2019), BFW (Robinson et al., 2020), FairFace (Kärkkäinen & Joo, 2021), and BUPT (Wang et al., 2021). All of these datasets are introducing demographic segment balancing, as it impacts the fairness of the models after training.

Training architecture and strategy. Ethnicity and gender auditing most often employs models trained specifically for the task. FFNet (Kärkkäinen & Joo, 2021) is the most widely used demographic classification tool (D'Incà et al., 2024; Deandres-Tame et al., 2024; Lee et al., 2024; Melzi et al., 2024) that jointly infers ethnicity, gender, and age. Also popular, FaceXFormer (Narayan et al., 2025a) trains a transformer on a merge of FairFace and UTKFace (Zhang et al., 2017), while the DeepFace library (Serengil & Özpınar, 2024) fully fine-tunes a pretrained model. Yet, these tools suffer performance limitations (Table 1) that impact any downstream fairness task. Beyond performance, they also expose a critical reproducibility gap, an issue already raised by the community, especially for FFNet: the training code and sufficient implementation details have not been released, despite numerous replication attempts, including ours.[1] General-purpose VLMs (Radford et al., 2021; Bai et al., 2025) can be applied zero-shot to DAI, but are far less efficient in compute (Table 1).

Hyperspherical geometry and robustness guarantees. Face recognition encoders trained with angular-margin losses (Wang et al., 2018; Deng et al., 2019; Kim et al., 2022) produce ℓ_2-normalized embeddings for which the von Mises-Fisher (vMF) distribution is the canonical directional model (Hasnat et al., 2017; Scott et al., 2021). Motivated by our newly introduced robustness audit dimension, we leverage these ingredients to derive a *per-identity* consistency bound over the natural variability of a person's images and for a class of decision functions defined as log-likelihood ratios of vMF mixtures. We further characterize the RBF kernel on normalized embeddings (Zhao & Song, 2018) to show that the RBF-SVM head yields a decision function that falls in this class.

3 Theoretical motivation

We motivate REGI by a guarantee that all images of an identity receive the same segment. Embeddings live on the unit hypersphere $\mathbb{S}^{d-1}$ ($d = 512$ in practice). The images of an identity id are mapped to i.i.d. embeddings $\mathbf{z} \sim Q_{id}$, where Q_{id} is an *arbitrary* distribution on $\mathbb{S}^{d-1}$ with mean direction $\boldsymbol{\mu}$. A C-class rule assigns $\hat{c}(\mathbf{z}) = \arg\max_c [\, g_c(\mathbf{z}) + \tau_c \,]$ for scores g_c and offsets τ_c, and the *flip probability* of identity id is

$$q_{id} \;:=\; \Pr_{\mathbf{z}\sim Q_{id}} \big(\hat{c}(\mathbf{z}) \neq \hat{c}(\boldsymbol{\mu})\big). \tag{1}$$

[1] `https://github.com/joojs/fairface/issues` (Accuracy only reached 75.06% on RFW after a 160-trial hyperparameter search, against 88.08% for the released weights).

The guarantee needs no distributional assumption; the vMF distribution (Mardia & Jupp, 2009) only motivates the decision class, since the log-likelihood ratio between mixtures of vMF components of shared concentration is a difference of log-sum-exp scores.

Definition 1 (The class $\mathcal{F}_{\kappa_b}$)**.** Fix $\kappa_b > 0$. $\mathcal{F}_{\kappa_b}$ is the set of C-class rules $\hat{c}(\mathbf{z}) = \arg\max_c [\, g_c(\mathbf{z}) + \tau_c \,]$ with offsets $\tau_c \in \mathbb{R}$ and log-sum-exp scores

$$g_c(\mathbf{z}) \;=\; \log \sum_{i=1}^{n_c} w_i^c \, e^{\kappa_b \, \boldsymbol{\mu}_i^{c\top} \mathbf{z}}, \tag{2}$$

with $n_c \geq 1$ component centers $\boldsymbol{\mu}_i^c \in \mathbb{S}^{d-1}$ and weights w^c in the probability simplex. Up to the shared normalizer, g_c is the log-density of $\sum_i w_i^c \, \mathrm{vMF}(\boldsymbol{\mu}_i^c, \kappa_b)$, and the offsets τ_c absorb priors and mass imbalance.

3.1 A DISTRIBUTION-FREE, PER-IDENTITY FLIP BOUND

Theorem 1 (Distribution-free flip bound)**.** *Let $\hat{c}$ be any C-class rule with scores $g_c + \tau_c$ (not necessarily in $\mathcal{F}_{\kappa_b}$), and let $\mathbf{z} \sim Q_{id}$ be arbitrary on $\mathbb{S}^{d-1}$. Write $c_0 := \hat{c}(\boldsymbol{\mu})$ and, for $c \neq c_0$, $F_c := g_{c_0} + \tau_{c_0} - g_c - \tau_c$, $m_c := \mathbb{E}_{Q_{id}}[F_c]$, and $s_c^2 := \mathrm{Var}_{Q_{id}}[F_c]$. If $m_c > 0$ for all $c \neq c_0$, then*

$$q_{id} \;\leq\; \sum_{c \neq c_0} \frac{s_c^2}{s_c^2 + m_c^2}. \tag{3}$$

If moreover each $F_c - m_c$ is sub-Gaussian with variance proxy σ_c^2, then $q_{id} \leq \sum_{c \neq c_0} \exp\big(-m_c^2/(2\sigma_c^2)\big)$.

The proof is in Appendix A.

Proposition 1 (Geometry of the score moments for $\mathcal{F}_{\kappa_b}$)**.** *For F_c built from equation 2, $\nabla F_c(\boldsymbol{\mu}) = \kappa_b\big(\bar{\mathbf{m}}_{c_0}(\boldsymbol{\mu}) - \bar{\mathbf{m}}_c(\boldsymbol{\mu})\big)$, with $\bar{\mathbf{m}}_c(\boldsymbol{\mu})$ the softmax-weighted component mean of class c at $\boldsymbol{\mu}$. Let Σ_{id} be the tangent covariance of Q_{id} at $\boldsymbol{\mu}$. To second order around $\boldsymbol{\mu}$, $m_c \approx F_c(\boldsymbol{\mu}) + \frac{1}{2}\,\mathrm{tr}\big(\nabla^2 F_c(\boldsymbol{\mu})\,\Sigma_{id}\big)$ and $s_c^2 \approx \nabla F_c(\boldsymbol{\mu})^\top \Sigma_{id}\, \nabla F_c(\boldsymbol{\mu})$, so*

$$\frac{m_c}{s_c} \;\approx\; \frac{\gamma_{\mathrm{lin}}}{\sigma_{id,\hat{\nabla}}}, \qquad \gamma_{\mathrm{lin}} := \frac{F_c(\boldsymbol{\mu})}{\|\nabla F_c(\boldsymbol{\mu})\|}, \qquad \sigma_{id,\hat{\nabla}}^2 := \hat{\nabla}^\top \Sigma_{id}\, \hat{\nabla}, \qquad \hat{\nabla} := \frac{\nabla F_c(\boldsymbol{\mu})}{\|\nabla F_c(\boldsymbol{\mu})\|}. \tag{4}$$

Under $Q_{id} = \mathrm{vMF}(\boldsymbol{\mu}, \kappa_{id})$, $\sigma_{id,\hat{\nabla}}^2 \approx 1/\kappa_{id}$ and $q_{id} \lesssim (C-1)\exp(-\kappa_{id}\,\gamma_{\mathrm{lin}}^2/2)$.

The SNR is the margin of the identity along the decision direction over its spread in that direction. The classical $\kappa_{id} \times (\mathrm{margin})^2$ exponent is its isotropic special case, which ties the guarantee to angular-margin pretraining that concentrates identities (see Appendix A).

The guarantee is a statement about training-set coverage. In any learned instantiation, the components $\boldsymbol{\mu}_i^c$ are (functions of) training points, so m_c is large iff a well-weighted same-class component sits near $\boldsymbol{\mu}$ while opposing components stay away. Empirically, γ_{lin} is rank-equivalent to the anchor gap $\delta_{\min}$, the cosine gap between the best-matching same-class component and the opposing ones (Spearman 0.997). The SNR adds the denominator: the guarantee also degrades where the identity's spread along the decision direction is large. The flip probabilities in turn control our robustness metric (Section A.5). We validate the bound and the SNR ranking out of sample on real face embeddings in Appendix C.

3.2 THE RBF-SVM HEAD REALIZES THE CLASS

On $\mathbb{S}^{d-1}$, $\|\mathbf{z}_i - \mathbf{z}\|_2^2 = 2 - 2\,\mathbf{z}_i^\top \mathbf{z}$, so the RBF kernel $K_{\mathrm{RBF}}(\mathbf{z}_i, \mathbf{z}) = e^{-\gamma_{\mathrm{RBF}}\|\mathbf{z}_i - \mathbf{z}\|_2^2} = e^{-\kappa_b} e^{\kappa_b \mathbf{z}_i^\top \mathbf{z}}$, with $\kappa_b = 2\gamma_{\mathrm{RBF}}$, is proportional to a vMF likelihood centered on $\mathbf{z}_i$. For a binary sub-problem with labels $y_i \in \{\pm 1\}$, dual coefficients $\alpha_i > 0$, and support vectors $\mathcal{S}$,

$$f(\mathbf{z}) = \sum_{i \in \mathcal{S}} \alpha_i y_i K_{\mathrm{RBF}}(\mathbf{z}_i, \mathbf{z}) + b = e^{-\kappa_b}\big[\beta_+ \, e^{g_+(\mathbf{z})} - \beta_- \, e^{g_-(\mathbf{z})}\big] + b, \tag{5}$$

where $g_\pm$ are scores (equation 2) on the support vectors of each class, with weights $w_i^\pm = \alpha_i/\beta_\pm$ and dual masses $\beta_\pm = \sum_{i \in \mathcal{S}:\, y_i = \pm 1} \alpha_i$. For $b = 0$, $\mathrm{sign}\, f$ is the rule $\arg\max_\pm [g_\pm + \tau_\pm]$ with

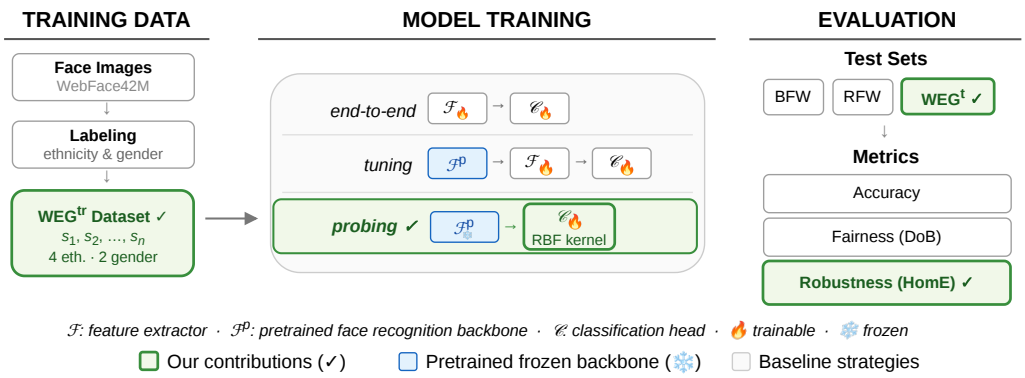


Figure 2: Pipeline for training and evaluating demographic attribute inference (s_i - ethnicity + gender segments ($n = 8$); $\mathcal{F}$ - feature extractor; $\mathcal{F}^p$ - pretrained $\mathcal{F}$, $\mathcal{C}$ - classification head). Our contributions are in green. We compare our probing strategy with end-to-end training (Kärkkäinen & Joo, 2021; Narayan et al., 2025a) and full fine-tuning (Serengil & Özpınar, 2024).

$\tau_\pm = \log \beta_\pm$: a log-likelihood ratio test between two vMF mixtures of support vectors. The decision statistic is therefore the score gap of Theorem 1: with $c_0 := \hat{c}(\boldsymbol{\mu})$ and c the opposing class, the flip event is exactly $\{F_c(\mathbf{z}) \leq 0\}$ for $F_c = g_{c_0} + \tau_{c_0} - g_c - \tau_c$, and the SVM boundary $\{f = 0\}$ is the zero-level set of F_c. A non-zero bias b shifts this comparison at the density level, $\beta_+ e^{g_+} - \beta_- e^{g_-} = -b\, e^{\kappa_b}$, rather than as a constant offset on the log-scores. Each binary head thus belongs to $\mathcal{F}_{\kappa_b}$, Theorem 1 applies to each of them. Since a multi-classification consists of the union of binary heads, the multi-class rule inherits the flip bound, and the coverage reading is literal (see Appendix B for details).

4 Ethnicity and Gender Attribute Inference

By Section 3.1, the margin of an identity is large only if a well-weighted same-segment component sits near its mean embedding. Since the components of the RBF-SVM head are training points (Section 3.2), the training set must cover each segment, including its fine-grained sub-populations. Figure 2 illustrates DAI pipelines and our contributions: **(1)** a diversified training set sampled from WebFace42M (Zhu et al., 2023) after reliable majority voting for demographic labeling, **(2)** a transfer-learning-inspired training that combines pretrained face recognition encoders and the RBF-SVM head, **(3)** a challenging test set also sampled from WebFace42M, **(4)** a robustness metric that quantifies intra-identity prediction consistency.

4.1 Ethnicity and Gender Datasets (WEG)

A curated dataset includes a set $S = \{s_1, \ldots, s_n\}$ of demographic segments (combination of Ethnicity+Gender). Ethnicity is defined using the predefined categories *African, Asian, Caucasian, Indian* (Kärkkäinen & Joo, 2021; Wang et al., 2021), used here for comparability with previous DAI systems (Kärkkäinen & Joo, 2021; Narayan et al., 2025a; Serengil & Özpınar, 2024). We justify this choice against the skin-color scheme in Section G.4. We encode gender using the *female* and *male* segments. Any such scheme reduces the diversity of human faces into a finite set of classes (Gwilliam et al., 2021; Khan & Fu, 2021; Popescu et al., 2022), raising ethical and technical challenges. Since labeling is expensive, we combine ethnicity and gender inference in WebFace42M (Zhu et al., 2023), leveraging intra-identity consistency to consolidate labels, and we diversify the dataset across fine-grained ethnicities within each coarse category, using an auxiliary seed dataset whose images are only used to drive diversification.

Seed dataset. We prompted GPT-5.1 to generate up to 100 male and 100 female names of living or 21st-century-active people for each of 270 fine-grained ethnicities, collected 60 images per name, and filtered them following Cao et al. (2018) (99.3% accuracy on 100 manually verified identities). Coarse ethnicity and gender are inferred per image by FFNet and aggregated at the identity level by

majority vote, which yields 244, 164, 241, and 211 fine categories for the African, Asian, Caucasian, and Indian ethnicities (details in Appendix F).

Identity-level labels. Assuming approximately independent errors across the N images of an identity, and retaining the majority label only when it gathers at least 60% of the votes, the label is correct with probability $\sum_{k=\lceil 0.6N \rceil}^{N} \binom{N}{k} p^k (1-p)^{N-k}$ for a per-image accuracy p. It exceeds 99.6% at $N = 20$ for $p = 0.83$ (FFNet accuracy on BUPT) well above the per-image accuracy, which breaks the potential circularity induced by the use of an automatic annotator (Section D.1).

WebFace resampling. We randomly split the 2M labeled WebFace42M identities (after removing the 28k identities used for training the backbones) into 80%, 10%, and 10% for training, validation, and test (WEG^{tr}, WEG^{v}, and WEG^{t}) before any labeling or sampling step, which ensures there is no identity contamination between the three subsets. We associate each WebFace identity with its closest seed identity in the embedding space of an AdaFace model (Kim et al., 2022), discard identities whose coarse ethnicity or gender differs from their proxy seed, and retain identities with at least 20 images whose gender and ethnicity consistency exceed 75% and 60%. These associations are only proxies for diversification and should not be regarded as ethnicity attributions. WEG^{v} and WEG^{t} are sampled uniformly among coarse ethnicities, then among their fine ethnicities. For WEG^{tr}, motivated by the training coverage discussion in Section 3.1, we sample a coarse group g uniformly, then a fine ethnicity e with probability proportional to $p^*(g|e)$, the Kullback–Leibler projection of the empirical conditional $p(g|e)$ onto the distributions whose group masses $\sum_e p^*(g|e)$ are equal, computed by iterative proportional fitting (Csiszár, 1975). This condition is necessary and sufficient for uniform fine-grained marginals (Appendix D), i.e., for covering every fine-grained subpopulation.

Dataset Resulting composition. WEG^{tr}, WEG^{v}, and WEG^{t} are respectively containing (20k identities, 40 images/identity), and (2400 identities, 40 images/identity) for both WEG^{v}, and WEG^{t}. While the main paper focuses on the gains for training on WEG^{tr}, we characterize the benefits of WEG^{t} in Appendix E

4.2 Model training (REGI)

Widely used DAI models, comprising a feature extractor $\mathcal{F}$ and a classifier $\mathcal{C}$, still rely on training from scratch (Kärkkäinen & Joo, 2021; Narayan et al., 2025a) or full fine-tuning (Serengil & Özpınar, 2024), despite the proven benefits of transfer learning (Sharif Razavian et al., 2014) in related face analysis tasks (Narayan et al., 2025b; Amato et al., 2019). Prior DAI works that employed pretraining (Hunanyan, 2022; Kalkatawi & Saeed, 2024) typically started from ImageNet models with linear heads, on the common assumption that the segments are linearly separable (Kärkkäinen & Joo, 2021; Sarridis et al., 2023), an assumption already questioned for age (Paplhám et al., 2024). We instead train $\mathcal{C}$ on top of a frozen face recognition backbone $\mathcal{F}^p$ pretrained with an angular-margin loss, which concentrates identities on the hypersphere, and use an RBF-SVM head, which realizes the class covered by the guarantee (Section 3.2).

4.3 Robustness as complementary evaluation dimension

DAI evaluation typically involves accuracy metrics, often complemented by fairness indicators (Sarridis et al., 2023; Fournier-Montgieux et al., 2025), and improvements in one may degrade the other (Zarei et al., 2025). We introduce robustness as a complementary dimension: a reliable attribute classifier should produce consistent predictions for different images of the same identity, despite variations in illumination, pose, and resolution (Merler et al., 2019; Thong et al., 2023). Consider D identities, identity id being represented by N_{id} images I_j^{id}, and a labeling function f with C classes. With $p_{id}^k = \frac{1}{N_{id}} \sum_j \delta_{f(I_j^{id})=k}$ the relative frequency of label k for identity id, the **homogeneity entropy** is

$$\mathrm{HomE} = \frac{1}{D} \sum_{id} \frac{1}{\log C} \sum_k -p_{id}^k \log p_{id}^k. \tag{6}$$

Lower values indicate higher robustness. HomE is **taxonomy-agnostic**, since it only requires identity labels, and it is controlled by the flip probabilities of Section 3.1 (Section A.5). Accuracy, fairness, and robustness form the three-axis protocol with which we audit DAI systems.

	Arch.	#GFLOP	Ethnicity								Gender					
			Accuracy ↑			Fairness ↓			Robust. ↓		Accuracy ↑		Fairness ↓		Robust. ↓	
			BFW	*RFW*	*WEG^t*	*BFW*	*RFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*
FFNet	RN34	1.94	87.9	88.1	74.4	3.50	4.10	6.44	0.19	0.29	96.3	91.4	2.01	0.9	0.13	0.28
DeepFace	VGG16	31.06	82.3	82.3	71.4	1.72	3.15	2.96	0.29	0.37	85.4	74.8	26.9	47.4	0.31	0.33
FaceXF.	Swin-B	30.92	84.7	83.4	67.2	3.54	3.11	8.25	0.23	0.33	95.1	89.3	0.97	4.64	0.12	0.21
CLIP	ViT-B	8.74	83.3	83.3	70.7	5.23	3.37	5.14	0.21	0.36	73.3	79.2	53.2	39.8	0.44	0.42
CLIP	ViT-L	155.4	87.6	89.2	80.8	2.83	1.60	3.20	0.20	0.25	68.0	75.1	63.9	48.3	0.38	0.39
QWEN3	4B	1101	93.6	92.8	83.9	2.28	2.40	5.15	0.11	0.19	98.2	95.4	0.87	2.10	0.07	0.16
QWEN3	8B	8767	89.7	90.9	80.3	4.82	3.81	7.45	0.16	0.18	**99.1**	**95.5**	1.65	2.30	**0.05**	**0.15**
REGI	RN18	0.98	94.0	93.4	86.2	1.34	1.93	2.93	0.11	0.22	96.9	91.9	**0.86**	0.82	0.11	0.26
REGI	IR50	12.64	**95.9**	**95.8**	**90.8**	**0.95**	**1.22**	**2.52**	**0.07**	**0.13**	97.9	94.7	1.14	**0.11**	0.07	0.19

Table 1: Comparison of ethnicity and gender inference based on accuracy, fairness, and robustness dimensions. We evaluate REGI (with RBF kernel) with two backbones and a fixed pretraining dataset (1.4M images), our diversified dataset (WEG^{tr}) against competitive baselines. **REGI**: ours. Best results in bold, second-best underlined; REGI with other backbones are compared in Table 7. Despite the gap in backbone size, REGI yields remarkably stable results on both IR50 and RN18, still showing competitive performance on the latter.

5 Experiments

5.1 Experimental setup

Baselines. We compare REGI with task-specific DAI pipelines and generalist VLMs. The first category includes: (1) FFNet (Kärkkäinen & Joo, 2021), which trains a ResNet34 (RN34) (He et al., 2016) end-to-end on a publicly released DAI subset combined with unavailable proprietary data; (2) the DeepFace library (Serengil & Özpınar, 2024)[2], which fully fine-tunes a VGGFace model (Parkhi et al., 2015) pretrained for face recognition on data from multiple DAI sources; and (3) FaceXFormer (Narayan et al., 2025a), which implements multi-task end-to-end training of a Swin-B model on UTKFace (Zhang et al., 2017) and FairFace. The second category includes CLIP (Radford et al., 2021) with ViT-B and ViT-L backbones and QWEN3-VL-Instruct (Bai et al., 2025) with 4B and 8B parameters.

Test datasets and metrics. BFW (Robinson et al., 2023) and WEG^t employ a common four-class ethnicity and two-gender segmentation and include multiple samples per identity, enabling accuracy, fairness, and robustness evaluation. RFW (Wang et al., 2019) includes per-sample ethnicity but not enough images per identity, so we use it only for the accuracy and fairness of ethnicity predictions. We report accuracy, the degree of bias (DoB), computed as the standard deviation of per-group accuracies, for fairness, and HomE (Equation 6) for robustness. Additional robustness measures are described in Appendix G.

Ethnicity and gender inference pipelines. In Table 1, we fix the pretraining dataset (28k identities and 1.4M WebFace42M samples balanced across ethnicity and gender), the classification head (RBF-SVM), and the attribute training dataset (WEG^{tr}), and report a compact (ResNet18, RN18 (He et al., 2016)) and our best (iResNet50, IR50 (Deng et al., 2019)) backbone, with the InsightFace implementation. We detail the experimental setup in Appendix F.

5.2 Main results

Table 1 shows that REGI variants improve ethnicity prediction across all datasets and evaluation dimensions, and are second-best after the much larger QWEN3 models for gender.

Ethnicity. REGI outperforms the baselines trained specifically for the attribute for all audited dimensions and datasets. Accuracy gains range from 7 to 16 points relative to FFNet, with larger gains for DeepFace and FaceXFormer, and fairness and robustness scores are at least halved in all three cases. The differences remain significant with the compact RN18 backbone, which requires half the GFLOPs of the RN34 used by FFNet. Among the generalist baselines, which employ much larger

[2]Not to be confounded with the original DeepFace publication (Taigman et al., 2014).

	WEG^{t}				BFW				RFW			
Training	**FC**	**LIN**	**MLP**	**RBF**	**FC**	**LIN**	**MLP**	**RBF**	**FC**	**LIN**	**MLP**	**RBF**
FairFace	70.1	75.2	71.7	*75.3*	79.2	83.7	82.6	*85.9*	80.1	85.3	82.4	*86.2*
BUPT	81.7	81.7	84.0	*87.9*	90.5	90.6	93.7	*95.6*	90.4	90.6	93.2	*95.2*
WEG^{tr}	**83.1**	**84.1**	**89.2**	***90.8***	**91.0**	**91.6**	**95.5**	***95.9***	**90.9**	**91.3**	**94.8**	***95.8***

Table 2: Impact of the classification head and of the ethnicity training dataset. *Best results for the classification head*. **Best results for the ethnicity training dataset.**

architectures and pretraining datasets, QWEN3-4B comes closest to REGI, but still lags behind by 2 to 7 accuracy points, doubles the fairness metric, is roughly 50% higher on robustness, and its inference is orders of magnitude slower. Performance differences between compact and larger REGI backbones remain modest (Section H.1).

Gender. This inference is simpler (Kärkkäinen & Joo, 2021), and REGI again outperforms the three task-specific baselines and the two CLIP versions. The QWEN3 models improve accuracy by 1 point compared with the best REGI versions and are more robust, but remain less fair, at the cost of much worse efficiency (GFLOP column of Table 1).

Test dataset. WEG^t is more challenging than BFW and RFW for both attributes and the three evaluation dimensions: the remaining error rate is roughly twice as high for ethnicity and nearly five times higher for gender, as an effect of the emphasis on diversification during its curation (Appendix E). Finally, REGI with RN18 outperforms FFNet with the larger RN34, which rules out that the gains of REGI come from backbone capacity alone.

5.3 Ablation study

We assess the classification head, the attribute training dataset, the pretraining size, and the training strategy for ethnicity, the more challenging attribute, with an IR50 backbone pretrained on 1.4M WebFace images unless indicated otherwise; further ablations are in Appendix H.

Classification head. In addition to the RBF-SVM, we test (1) FC, the standard fully-connected linear layer; (2) LIN, a linear SVM head; and (3) MLP, two 512-dimensional layers with ReLU, the head recommended for age prediction by Paplhám et al. (2024). The RBF head provides the best performance in Table 2, followed by the MLP, the linear SVM, and FC. This is consistent with Section 3.2: among the tested heads, only the RBF-SVM realizes the vMF-mixture class $\mathcal{F}_{\kappa_b}$, and the gap between MLP and RBF rules out that non-linearity alone explains the gains.

	90K	**180K**	**360K**	**700K**	**1.4M**	**4M**
WEG^t	82.3	85.4	87.6	89.4	90.8	**91.2**
BFW	91.5	93.6	94.6	95.2	95.9	**96.2**
RFW	91.0	92.9	94.1	95.2	**95.8**	95.6

(a) Impact of face recognition pretraining dataset size on ethnicity accuracy (%) with the RBF head trained on WEG^{tr}. Last column is the model trained by Kim et al. (2022) on WebFace4M.

	E2E	**FT**	**RBF**
WEG^t	78.1	89.6	**90.8**
BFW	87.0	95.8	**95.9**
RFW	87.7	95.2	**95.8**

(b) Impact of training strategies for ethnicity accuracy (%). FT and RBF pretraining uses 1.4M images. WEG^{tr} is used for training.

Ethnicity training dataset. With the same pretraining configuration, WEG^{tr} outperforms FairFace and BUPT for all test datasets and classification heads (Table 2), with a larger margin on WEG^t, the most challenging test dataset. WEG and BUPT have similar sizes and numbers of identities; the main difference lies in the emphasis of WEG on fine-grained ethnic diversity, i.e., on the coverage of each segment, which is the lever identified by Section 3.1. The size difference does not explain the gap with FairFace, since WEG^{tr} outperforms it even when smaller (Appendix H), and the gains observed for every head rule out that the benefit of WEG is specific to the RBF-SVM.

Pretraining size and training strategy. Accuracy increases with the pretraining size up to 1.4M images and saturates beyond (Table 3a), probably as an effect of the ethnicity imbalance of the

full WebFace dataset; this increase is consistent with the isotropic reading of Proposition 1, since stronger pretraining concentrates identities. Among the training strategies of Figure 2 (Table 3b), probing outperforms fine-tuning and end-to-end training, both carefully optimized with Optuna, in addition to being faster to train. Probing outperforms end-to-end training even with the smallest pretraining dataset (82.3 vs. 78.1 on WEG^t), which rules out that the gains of REGI arise solely from strong pretraining: face recognition is a suitable pretraining task for ethnicity and gender inference and, advantageously, replaces end-to-end training.

6 Limitations and Ethics

Looseness and estimation of the guarantee. Theorem 1 is first distribution-free bound wich we further refine under vMF assumption. It pays for it with a polynomial rate: the bound averages ≈ 0.5 against a flip rate ≈ 0.2 (Appendix C). Given the gap between the bound and the results obtained in practice, we leave for future work the search of a tighter bound associated to functions of the F_c family. The moments are estimated from at most 40 images per identity, hence the out-of-sample protocol, under which score residuals have heavier tails than in-sample; the sub-Gaussian variant is therefore only a remark. It is important to note that the guarantee assumes i.i.d. images given the identity, while the images of an identity may share capture conditions. Similarly, we simplify the theoritical framework by assuming that identity distribution can be modelized by isotropic vMFs with constant κ. While motivated by the angular loss formulations, such property may not hold in practice. Finally, the coverage reading is supported by correlations and by the training-dataset ablation, not by a precise causal intervention on the training set, which we leave for future work.

Model size, interpretability, and attribute completeness. We focus on compact models, but larger architectures and pretraining datasets bring modest gains, which calls for a performance-computation tradeoff, and the learned features remain opaque. We focus on ethnicity and gender because these attributes are stable at the identity level, so that labels can be aggregated and intra-identity consistency is meaningful; age varies within each identity, and Paplhám et al. (2024) recommend strong pretraining followed by non-linear probing for it, a finding that resonates with ours.

Ethical risks and potential misuse. The pipeline, designed for fairness auditing, could be repurposed for surveillance, profiling, or discrimination, and raises concerns around consent, privacy, and bias amplification; technical safeguards alone are insufficient (see the ethics statement).

7 Conclusion

We revisit the inference of ethnicity and gender and propose diverse training and test sets for these tasks. We implement a transfer-learning-based pipeline that combines a model pretrained for face recognition and an RBF-SVM head, motivated by a distribution-free, per-identity bound on intra-identity consistency that is measurable from the images of an identity, reads as a signal-to-noise ratio, and ties robustness to pretraining concentration and training-set coverage. We add robustness as a third audit dimension and demonstrate consistent performance for ethnicity and an interesting performance-efficiency balance for gender. Looking ahead, we will evaluate larger pretrained models while preserving the emphasis on dataset diversity, test the coverage reading by targeted interventions on the training set, and address age with automatically labeled, diversified datasets.

Ethics statement

This work predicts perceived ethnicity and gender from face images. We consider the risks of such technology serious and address them explicitly.

On the taxonomy. We predict four coarse ethnicity categories and two genders because they are the categories used by the benchmarks and systems our results must be comparable with (Kärkkäinen & Joo, 2021; Wang et al., 2019; 2021). We also leverage finer ethnicity categories to drive diversification within test and training sets. We do not claim these categories are natural, exhaustive, or an adequate description of human diversity: ethnicity is a social construct mixing cultural and origin-based dimensions (Ford & Kelly, 2005), individuals can legitimately be placed in different coarse categories, and gender is not binary. Any partition of human appearance into finitely many classes

is open to criticism (Khan & Fu, 2021; Kotwal & Marcel, 2025). Our position is methodological: whichever taxonomy an audit adopts, its adequacy should be measured rather than assumed, and the taxonomy-agnostic robustness metric of Section 4.3 permits such a comparison across label spaces (Section G.4).

On intended and forbidden use. Demographic inference on faces is a biometric categorization technology and is treated as sensitive under, among others, the EU AI Act. We release the DAI tool under a restricted-use license permitting *only* aggregate fairness measurement, bias diagnosis, and bias mitigation research. The license explicitly forbids individual-level profiling, identification, authentication, surveillance, targeting, and any decision-making about individuals. We are aware that a license is not an enforcement mechanism; we consider the alternative, leaving fairness auditors dependent on less accurate and less robust tools, to be worse.

On the data. WEG is resampled from WebFace42M (Zhu et al., 2023) and inherits its collection conditions; we release metadata and identity lists rather than redistributing images. Seed identities are public figures with a documented public presence. Demographic labels are algorithmically produced with a measured error rate and must not be interpreted as statements about any individual's actual identity, ancestry, or self-identification; the identity-to-seed matching of Section 4.1 is a diversification device only.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions in section 3 and Section 3.1; the proofs are given in Appendix A, and the derivation of the RBF-SVM as a member of $\mathcal{F}_{\kappa_b}$ in Appendix B. The validation protocol of the bound (rule, components, bandwidth, split, and out-of-sample protocol) is described in Appendix C. WEG construction is specified end to end in Section 4.1, Appendix D, and Appendix F: seed collection and filtering, the $80/10/10$ identity split preceding all labeling, the supermajority thresholds (75% gender, 60% ethnicity, ≥ 20 images per identity), and the iterative proportional fitting procedure. Training configurations, including the hyperparameter search of the end-to-end and fine-tuning baselines, are reported in Appendix F. We release the dataset metadata, the full training and inference code, the pretrained models, and the evaluation toolkit, under the restricted-use license described above.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

An LLM was used as a component of the research method. To build the seed dataset of Section 4.1, we prompted GPT-5.1 with a list of 270 fine-grained ethnicities and requested up to 100 male and 100 female names of living or 21st-century-active public figures per ethnicity. The model was used only as a name-retrieval device: every generated entry was subsequently filtered, and the collected face images were filtered by an embedding-based procedure following Cao et al. (2018) and manually verified on 100 identities (99.3% accuracy). The LLM produced no demographic label used in the paper: coarse ethnicity and gender labels come from the identity-level aggregation of Section 4.1. This step inherits the coverage bias of the LLM over public figures, which is one reason the seed set is used only as a diversification prior. LLMs were also used for editing assistance on the manuscript; all technical content, proofs, experiments, and claims are the authors'.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fabio Valerio Massoli, Nikolaos Passalis, Anastasios Tefas, Alessandro Trivilini, and Claudio Vairo. Face verification and recognition for digital forensics and information security. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–6. IEEE, 2019.

Apple Inc. Face ID Security. Technical White Paper, September 2017. URL `https://support.apple.com/en-us/HT208108`. Details the depth-sensing TrueDepth camera and neural-network matching that enable Face ID for high-assurance smartphone authentication.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pp. 77–91, 2018.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, 2018.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Ivan Deandres-Tame, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability. *IEEE Access*, 12:34390–34401, 2024.

Namrata Deka and Danica J Sutherland. Mmd-b-fair: Learning fair representations with statistical testing. In *International Conference on Artificial Intelligence and Statistics*, pp. 9564–9576. PMLR, 2023.

William Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

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.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12225–12235, 2024.

Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems*, 36:74366–74393, 2023.

Marvella E Ford and P Adam Kelly. Conceptualizing and categorizing race and ethnicity in health services research. *Health services research*, 40(5p2):1658–1675, 2005.

Alexandre Fournier-Montgieux, Michael Soumm, Adrian Popescu, Bertrand Luvison, and Hervé Le Borgne. Fairer analysis and demographically balanced face generation for fairer face verification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2788–2798. IEEE, 2025.

Sixue Gong, Xiaoming Liu, and A Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, pp. 330–347, 2020.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology, 2019.

Matthew Gwilliam, Srinidhi Hegde, Lade Tinubu, and Alex Hanson. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4123–4132, 2021.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. von Mises–Fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Fadhil Hidayat, Ulva Elviani, George Bryan Gabriel Situmorang, Muhammad Zaky Ramadhan, Figo Agil Alunjati, and Reza Fauzi Sucipto. Face recognition for automatic border control: a systematic literature review. *IEEE Access*, 12:37288–37309, 2024.

Arpi Hunanyan. Race from face. Technical report, American University of Armenia, 2022.

C. T. Ireland and Solomon Kullback. Contingency tables with given marginals. *Biometrika*, 55(1): 179–188, 1968.

Abdul-aziz Kalkatawi and Usman Saeed. Ethnicity classification based on facial images using deep learning approach. *International Journal of Advanced Computer Science & Applications*, 15(2), 2024.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 35–50. Springer, 2012.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1547–1557, 2021.

Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, pp. 587–597, 2021.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Ketan Kotwal and Sébastien Marcel. Review of demographic fairness in face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.

KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

Tony Lee, Haoqin Tu, Chi H Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin S Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666, 2024.

Li Lin, Santosh Santosh, Mingyang Wu, Xin Wang, and Shu Hu. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3503–3515, June 2025.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Pietro Melzi, Ruben Tolosana, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Ivan DeAndres-Tame, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, et al. Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Information Fusion*, 107:102322, 2024.

Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

Terance D Miethe, Tanya Dudinskaya, Christopher Forepaugh, and William H Sousa. Facial recognition technology in policing: A national survey of public support for this technology and privacy/safety concerns. *Crime & Delinquency*, 71(4):1025–1051, 2025.

Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. In *CVPR*, 2025a.

Kartik Narayan, Vibashan VS, and Vishal M Patel. Facexbench: Evaluating multimodal llms on face understanding. *arXiv preprint arXiv:2501.10360*, 2025b.

Pedro C Neto, Naser Damer, Jaime S Cardoso, and Ana F Sequeira. Balancing beyond discrete categories: Continuous demographic labels for fair face recognition. *arXiv preprint arXiv:2506.01532*, 2025.

Jakub Paplhám, Vojt Franc, et al. A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1196–1205, 2024.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Adrian Popescu, Liviu-Daniel Ștefan, Jérôme Deshayes-Chossart, and Bogdan Ionescu. Face verification with challenging imposters and diversified demographics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3357–3366, 2022.

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

Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. In *AAAI 2020 Workshop on AI Evaluation*, 2020. URL `https://arxiv.org/abs/2102.00813`.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Remi Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151, 2020.

Zohra Rezgui, Nicola Strisciuglio, and Raymond Veldhuis. Controllable privacy in face recognition: A filter-based approach. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9. IEEE, 2024.

Rasha Khudiar Rija, Ghaida Muttasher, and Ahmed Al-Araji. Payment systems based on face recognition: A survey. *Journal of Optoelectronics Laser*, 41(5):563–571, 2022.

Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pp. 0–1, 2020.

Joseph P. Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, 32: 4365–4377, 2023. doi: 10.1109/TIP.2023.3282837.

Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Towards fair face verification: An in-depth analysis of demographic biases. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 194–208. Springer, 2023.

Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10612–10622, 2021.

Sefik Serengil and Alper Özpınar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, 17(2):95–107, 2024.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Seth Stodder and Thomas Warrick. Biometrics at the Border: Balancing Security, Convenience, and Civil Liberties. Technical report, Atlantic Council, 2022. URL `https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/biometrics-at-the-border/`.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

KH Teoh, RC Ismail, SZM Naziri, R Hussin, MNM Isa, and MSSM Basir. Face recognition and identification using deep learning approach. In *Journal of Physics: Conference Series*, volume 1755. IOP Publishing, 2021.

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.

William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4903–4913, 2023.

U.S. Government Accountability Office. Facial Recognition Technology: Current and Planned Uses by Federal Agencies. Technical Report GAO-21-526, United States Government Accountability Office, August 2021. URL `https://www.gao.gov/assets/gao-21-526.pdf`.

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, 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 (ICCV)*, pp. 692–702, 2019.

Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021.

Haiyu Wu and Kevin Bowyer. What should be balanced in a "balanced" face recognition dataset? In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. URL https://papers.bmvc2023.org/0235.pdf.

Amir Zarei, Ahmad Hassanpour, and Kiran Raja. On privacy, accuracy, and fairness trade-offs in facial recognition. *IEEE Access*, 2025.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

Chenchao Zhao and Jun S Song. Exact heat kernel on a hypersphere and its applications in kernel svm. *Frontiers in Applied Mathematics and Statistics*, 4:1, 2018.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Dalong Du, Jiwen Lu, and Jie Zhou. Webface260m: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2627–2644, 2023. doi: 10.1109/TPAMI.2022.3169734.

A Proofs and complements for the theoretical motivation

A.1 Softmax gradient

Lemma 1 (Softmax gradient)**.** *For g of the form equation 2, $\nabla g(\mathbf{z}) = \kappa_b \sum_i \pi_i(\mathbf{z})\, \boldsymbol{\mu}_i$ with softmax weights $\pi_i(\mathbf{z}) \propto w_i\, e^{\kappa_b \boldsymbol{\mu}_i^\top \mathbf{z}}$; in particular $\|\nabla g\| \le \kappa_b$, so g is κ_b-Lipschitz.*

Proof. Direct differentiation of equation 2 gives $\nabla g(\mathbf{z}) = \kappa_b \sum_i \pi_i(\mathbf{z})\, \boldsymbol{\mu}_i$ with $\pi_i(\mathbf{z}) = w_i e^{\kappa_b \boldsymbol{\mu}_i^\top \mathbf{z}} / \sum_j w_j e^{\kappa_b \boldsymbol{\mu}_j^\top \mathbf{z}}$. A convex combination of unit vectors has norm ≤ 1, hence $\|\nabla g\| \le \kappa_b$. □

A.2 Proof of Theorem 1

Proof. $\{\hat{c}(\mathbf{z}) \neq c_0\} \subseteq \bigcup_{c \neq c_0} \{F_c(\mathbf{z}) \le 0\}$, and Cantelli's one-sided inequality (Boucheron et al., 2013) gives $\Pr(F_c \le 0) = \Pr(F_c - m_c \le -m_c) \le s_c^2/(s_c^2 + m_c^2)$; we conclude by the union bound over the $C-1$ opposing classes. The exponential version replaces Cantelli by the sub-Gaussian lower-tail bound. □

The price of distribution-freeness is a polynomial (Cantelli) rate; the exponential rate requires a sub-Gaussian score, which must be justified or measured (Appendix C).

A.3 Variant of Theorem 1 used in the validation

Let $F_{\min} := \min_{c \neq c_0} F_c$. Since $\{\hat{c}(\mathbf{z}) \neq c_0\} \subseteq \{F_{\min}(\mathbf{z}) \le 0\}$, the argument of Theorem 1 applied to the single scalar $F_{\min}$, with mean $m_{\min} > 0$ and variance $s_{\min}^2$ under Q_{id}, gives $q_{id} \le s_{\min}^2/(s_{\min}^2 + m_{\min}^2)$, without union bound. This is the per-identity min-class variant reported in Appendix C.

A.4 Proof of Proposition 1

Throughout, $P := I - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ denotes the orthogonal projector onto the tangent space of $\mathbb{S}^{d-1}$ at $\boldsymbol{\mu}$, and we use the exact tangent–radial decomposition

$$\mathbf{z} \;=\; t\,\boldsymbol{\mu} + \mathbf{v}, \qquad t := \boldsymbol{\mu}^\top \mathbf{z}, \qquad \mathbf{v} := P\mathbf{z}, \qquad \|\mathbf{v}\|^2 = 1 - t^2, \tag{7}$$

valid for every $\mathbf{z} \in \mathbb{S}^{d-1}$. Since $\boldsymbol{\mu}$ is the mean direction of Q_{id}, $\mathbb{E}[\mathbf{z}] \propto \boldsymbol{\mu}$ and therefore $\mathbb{E}[\mathbf{v}] = P\,\mathbb{E}[\mathbf{z}] = 0$ *exactly*; the tangent covariance $\Sigma_{id} := \mathbb{E}[\mathbf{v}\mathbf{v}^\top]$ is supported on the tangent space ($P\Sigma_{id}P = \Sigma_{id}$, $\Sigma_{id}\boldsymbol{\mu} = 0$). "To second order" means up to remainders controlled by third moments, $O(\mathbb{E}\|\mathbf{v}\|^3)$, with constants depending only on κ_b (Step 1).

Step 1: derivatives of the scores. By Lemma 1, $\nabla g_c(\mathbf{z}) = \kappa_b \sum_i \pi_i^c(\mathbf{z})\,\boldsymbol{\mu}_i^c =: \kappa_b\,\bar{\mathbf{m}}_c(\mathbf{z})$, the softmax-weighted component mean of class c. Since the offsets are constants, $F_c = g_{c_0} - g_c + \tau_{c_0} - \tau_c$ gives

$$\nabla F_c(\boldsymbol{\mu}) \;=\; \kappa_b\big(\bar{\mathbf{m}}_{c_0}(\boldsymbol{\mu}) - \bar{\mathbf{m}}_c(\boldsymbol{\mu})\big), \tag{8}$$

the stated gradient formula. Differentiating once more,

$$\nabla^2 g_c(\mathbf{z}) \;=\; \kappa_b^2\Big(\textstyle\sum_i \pi_i^c(\mathbf{z})\,\boldsymbol{\mu}_i^c\boldsymbol{\mu}_i^{c\top} - \bar{\mathbf{m}}_c(\mathbf{z})\bar{\mathbf{m}}_c(\mathbf{z})^\top\Big) \;=\; \kappa_b^2\,\mathrm{Cov}_{\pi^c(\mathbf{z})}\big(\boldsymbol{\mu}_i^c\big), \tag{9}$$

the covariance of the component centers under the softmax weights. In particular $0 \preceq \nabla^2 g_c \preceq \kappa_b^2 I$, so $\|\nabla^2 F_c\| \le \kappa_b^2$, and the same computation bounds all third derivatives of F_c by $O(\kappa_b^3)$ uniformly. F_c is thus a smooth function of the ambient variable with derivatives bounded in terms of κ_b only, which licenses the Taylor expansion below and quantifies its remainder.

Step 2: Taylor expansion along the sphere. From equation 7, $t = \sqrt{1 - \|\mathbf{v}\|^2} = 1 - \frac{1}{2}\|\mathbf{v}\|^2 + O(\|\mathbf{v}\|^4)$, so

$$\mathbf{z} - \boldsymbol{\mu} \;=\; \mathbf{v} - \tfrac{1}{2}\|\mathbf{v}\|^2\,\boldsymbol{\mu} + O(\|\mathbf{v}\|^4). \tag{10}$$

Substituting into the ambient second-order expansion of F_c at $\boldsymbol{\mu}$ and collecting terms by order in $\mathbf{v}$,

$$F_c(\mathbf{z}) \;=\; F_c(\boldsymbol{\mu}) + \nabla F_c(\boldsymbol{\mu})^\top \mathbf{v} + \tfrac{1}{2}\,\mathbf{v}^\top H_c\,\mathbf{v} + R(\mathbf{v}), \qquad |R(\mathbf{v})| \le C_{\kappa_b}\|\mathbf{v}\|^3, \tag{11}$$

where

$$H_c \;:=\; P\,\nabla^2 F_c(\boldsymbol{\mu})\,P \;-\; \big(\boldsymbol{\mu}^\top \nabla F_c(\boldsymbol{\mu})\big)\,P \tag{12}$$

is the *spherical (intrinsic) Hessian* of F_c at $\boldsymbol{\mu}$: the term $-\|\mathbf{v}\|^2\boldsymbol{\mu}^\top\nabla F_c(\boldsymbol{\mu})$, produced by the radial retraction $-\frac{1}{2}\|\mathbf{v}\|^2\boldsymbol{\mu}$ hitting the gradient, is precisely the curvature correction that turns the ambient Hessian into the second derivative of F_c along geodesics of $\mathbb{S}^{d-1}$. In the statement of Proposition 1, $\nabla^2 F_c(\boldsymbol{\mu})$ is to be read in this intrinsic sense; the difference between the two readings, $-\frac{1}{2}(\boldsymbol{\mu}^\top\nabla F_c(\boldsymbol{\mu}))\,\mathrm{tr}\,\Sigma_{id}$, is negligible in practice because $\nabla F_c(\boldsymbol{\mu})$ is essentially tangent (median tangent fraction 0.994, Appendix C).

Step 3: mean. Taking Q_{id}-expectations in equation 11: the linear term vanishes because $\mathbb{E}[\mathbf{v}] = 0$, and $\mathbb{E}[\mathbf{v}^\top H_c\mathbf{v}] = \mathrm{tr}(H_c\Sigma_{id})$, so

$$m_c \;=\; F_c(\boldsymbol{\mu}) + \tfrac{1}{2}\,\mathrm{tr}\big(H_c\,\Sigma_{id}\big) + O(\mathbb{E}\|\mathbf{v}\|^3), \tag{13}$$

which is the stated mean formula.

Step 4: variance. Subtracting the mean in equation 11, $F_c - m_c \;=\; \nabla F_c(\boldsymbol{\mu})^\top\mathbf{v} + \big(q(\mathbf{v}) - \mathbb{E}[q(\mathbf{v})]\big) + \big(R(\mathbf{v}) - \mathbb{E}[R(\mathbf{v})]\big)$ with $q(\mathbf{v}) := \frac{1}{2}\mathbf{v}^\top H_c\mathbf{v}$. Expanding the square, the leading term is $\nabla F_c(\boldsymbol{\mu})^\top\Sigma_{id}\nabla F_c(\boldsymbol{\mu})$; the cross term $2\,\mathbb{E}[(\nabla F_c^\top\mathbf{v})(q - \mathbb{E}q)]$ involves third moments of $\mathbf{v}$ and is $O(\mathbb{E}\|\mathbf{v}\|^3)$; $\mathrm{Var}[q] = O\big((\mathbb{E}\|\mathbf{v}\|^2)^2\big)$; and the remainder contributes $O(\mathbb{E}\|\mathbf{v}\|^3)$ as well. Hence

$$s_c^2 \;=\; \nabla F_c(\boldsymbol{\mu})^\top\Sigma_{id}\,\nabla F_c(\boldsymbol{\mu}) + O\big(\mathbb{E}\|\mathbf{v}\|^3 + (\mathrm{tr}\,\Sigma_{id})^2\big), \tag{14}$$

and since Σ_{id} is tangent-supported, only the tangent part of the gradient contributes: $\nabla F_c^\top\Sigma_{id}\nabla F_c = (P\nabla F_c)^\top\Sigma_{id}(P\nabla F_c)$.

Step 5: the SNR. When the second-order corrections are subleading ($\|\Sigma_{id}\|$ small), $m_c \approx F_c(\boldsymbol{\mu})$ and $s_c \approx \|\nabla F_c(\boldsymbol{\mu})\|\,\sigma_{id,\hat{\nabla}}$ with $\sigma^2_{id,\hat{\nabla}} = \hat{\nabla}^\top \Sigma_{id} \hat{\nabla}$; dividing yields equation 4. The SNR thus compares the linearized margin $\gamma_{\text{lin}} = F_c(\boldsymbol{\mu})/\|\nabla F_c(\boldsymbol{\mu})\|$ (the distance from $\boldsymbol{\mu}$ to the linearized decision boundary $\{F_c = 0\}$) to the identity's spread along the decision direction $\hat{\nabla}$.

Step 6: vMF specialization. Let $Q_{id} = \text{vMF}(\boldsymbol{\mu}, \kappa_{id})$ and use the tangent-normal decomposition $\mathbf{z} = t\boldsymbol{\mu} + \sqrt{1-t^2}\,\boldsymbol{\xi}$, where t has density $f(t) \propto (1-t^2)^{(d-3)/2} e^{\kappa_{id} t}$ on $[-1, 1]$ and $\boldsymbol{\xi}$ is uniform on the unit subsphere of the tangent space, independent of t (Mardia & Jupp, 2009). Isotropy of $\boldsymbol{\xi}$ gives $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = P/(d-1)$, hence

$$\Sigma_{id} \;=\; \frac{\mathbb{E}[1-t^2]}{d-1}\, P. \tag{15}$$

For the radial moment, substitute $u := \kappa_{id}(1-t)$: the density of u is proportional to $u^{(d-3)/2}\big(2 - u/\kappa_{id}\big)^{(d-3)/2} e^{-u}$, which converges (as $\kappa_{id} \to \infty$, up to normalization) to the $\text{Gamma}\big(\frac{d-1}{2}, 1\big)$ density, so $\mathbb{E}[1-t] \approx \frac{d-1}{2\kappa_{id}}$ and $\mathbb{E}[1-t^2] = \mathbb{E}[(1-t)(1+t)] \approx 2\,\mathbb{E}[1-t] \approx \frac{d-1}{\kappa_{id}}$. Therefore $\Sigma_{id} \approx \kappa_{id}^{-1} P$: the spread is $\sigma^2_{id,\mathbf{u}} \approx 1/\kappa_{id}$ along *every* tangent direction $\mathbf{u}$, in particular along $\hat{\nabla}$, and $\mathbb{E}\|\mathbf{v}\|^2 = \text{tr}\,\Sigma_{id} \approx (d-1)/\kappa_{id}$. Plugging into equation 4 gives $m_c/s_c \approx \sqrt{\kappa_{id}}\,\gamma_{\text{lin}}$. Finally, in the same high-concentration regime the rescaled tangent fluctuation $\sqrt{\kappa_{id}}\,\mathbf{v}$ is asymptotically $\mathcal{N}(0, P)$, so the linearized score $\nabla F_c(\boldsymbol{\mu})^\top \mathbf{v}$ is sub-Gaussian with variance proxy $\sigma_c^2 \approx s_c^2 \approx \|\nabla F_c(\boldsymbol{\mu})\|^2/\kappa_{id}$. The exponential branch of Theorem 1 then reads $\Pr(F_c \le 0) \lesssim \exp\big(-m_c^2/(2\sigma_c^2)\big) \approx \exp\big(-\kappa_{id}\,\gamma_{\text{lin}}^2/2\big)$, and the union bound over the $C-1$ opposing classes yields $q_{id} \lesssim (C-1)\exp\big(-\kappa_{id}\,\gamma_{\text{lin}}^2/2\big)$. The symbol $\lesssim$ records the second-order approximations of Steps 3–4 and the Gaussian tangent limit. □

A.5 From flip probabilities to robustness

The flip probability controls the homogeneity entropy (HomE, Equation 6), our robustness metric: with H_b the binary entropy,

$$\mathbb{E}[\text{HomE}] \;\le\; \frac{1}{D}\sum_{id} \frac{H_b(q_{id}) + q_{id}\log(C-1)}{\log C}, \tag{16}$$

and plugging equation 3, capped at $1 - 1/C$, yields a HomE certificate. HomE measures the dispersion around the center decision $c_0 = \hat{c}(\boldsymbol{\mu})$, not whether c_0 is correct; we therefore always report it with accuracy, and the pair (CMR, HomE$^+$) of Section G.3 makes this (c_0, q_{id}) split explicit.

Proof of equation 16. Fix an identity id and let P_{id} be the distribution of the label $\hat{c}(\mathbf{z})$, $\mathbf{z} \sim Q_{id}$, so that $P_{id}(c_0) = 1 - q_{id}$. The empirical label frequencies $\hat{p}_{id}$ over the N_{id} i.i.d. images of id satisfy $\mathbb{E}[\hat{p}_{id}] = P_{id}$, and the entropy H is concave, so by Jensen's inequality $\mathbb{E}[H(\hat{p}_{id})] \le H(P_{id})$. By the grouping property of entropy, with $P_{id}^{\neq}$ the distribution of the label conditioned on $\hat{c}(\mathbf{z}) \neq c_0$, supported on $C-1$ labels,

$$H(P_{id}) = H_b(q_{id}) + q_{id}\,H(P_{id}^{\neq}) \le H_b(q_{id}) + q_{id}\log(C-1) =: \varphi(q_{id}). \tag{17}$$

Averaging over identities and dividing by $\log C$ gives equation 16. Finally, $\varphi'(q) = \log\frac{1-q}{q} + \log(C-1) \ge 0$ if and only if $q \le 1 - 1/C$, so φ is increasing on $[0, 1-1/C]$, where it reaches its maximum $\log C$; replacing q_{id} by $\min(B_{id}, 1-1/C)$, with B_{id} the right-hand side of equation 3, therefore preserves the inequality and yields a HomE certificate computed from the score moments only.

B Complete derivation: the RBF-SVM on angular-margin embeddings realizes $\mathcal{F}_{\kappa_b}$

We detail the derivation summarized in Section 3.2. The individual components are established results from directional statistics (Mardia & Jupp, 2009) and metric learning (Deng et al., 2019; Scott et al., 2021); we connect them to the decision class $\mathcal{F}_{\kappa_b}$ of Definition 1, for which the guarantee of Section 3.1 is stated.

B.1 Background: angular-margin losses and hyperspherical geometry

Modern face recognition models employ angular-margin losses such as CosFace (Wang et al., 2018), ArcFace (Deng et al., 2019), or AdaFace (Kim et al., 2022), which operate on ℓ_2-normalized embeddings. Let $f : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ denote a deep neural network that maps an input face image to a d-dimensional embedding. The normalization constraint ensures that feature vectors lie on the unit hypersphere:

$$\mathbf{z} = \frac{f(\mathbf{x})}{\|f(\mathbf{x})\|_2} \in \mathbb{S}^{d-1}. \tag{18}$$

Angular losses for a training sample $(\mathbf{x}_i, y_i)$ with ground-truth identity $y_i \in \{1, \dots, K\}$ can be written as

$$\mathcal{L}_{\text{angular}} = -\log \frac{\exp(s\cos(\theta_{y_i} + m))}{\exp(s\cos(\theta_{y_i} + m)) + \sum_{j\neq y_i} \exp(s\cos\theta_j)}, \tag{19}$$

where $\theta_j = \arccos(\mathbf{w}_j^\top \mathbf{z}_i)$ is the angle between the embedding $\mathbf{z}_i$ and the j-th class center $\mathbf{w}_j \in \mathbb{S}^{d-1}$, s is a scale parameter, and m is the angular margin penalty.

B.2 Angular losses and von Mises–Fisher distributions

The von Mises–Fisher (vMF) distribution is the canonical probability distribution for directional data on the unit hypersphere (Mardia & Jupp, 2009). For a random unit vector $\mathbf{z} \in \mathbb{S}^{d-1}$, the vMF density is

$$p(\mathbf{z} \mid \boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{z}\right), \tag{20}$$

where $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is the mean direction, $\kappa \geq 0$ is the concentration parameter, and $C_d(\kappa) = \kappa^{d/2-1}/[(2\pi)^{d/2} I_{d/2-1}(\kappa)]$ is the normalization constant, with I_ν the modified Bessel function of the first kind. The softmax classifier with ℓ_2-normalized features and class centers is equivalent to a generative model where each class follows a vMF distribution (Mardia & Jupp, 2009; Scott et al., 2021): under uniform class priors,

$$P(y = c \mid \mathbf{z}) = \frac{p(\mathbf{z} \mid \mathbf{w}_c, s)}{\sum_{j=1}^K p(\mathbf{z} \mid \mathbf{w}_j, s)} = \frac{\exp(s\mathbf{w}_c^\top \mathbf{z})}{\sum_{j=1}^K \exp(s\mathbf{w}_j^\top \mathbf{z})}, \tag{21}$$

where each class c is modeled as a vMF distribution with mean direction $\mathbf{w}_c$ and shared concentration $\kappa = s$. The angular margin m decreases the likelihood of the correct class during training through $\cos(\theta_c) \to \cos(\theta_c + m)$, forcing the model to achieve higher concentration (smaller intra-class angles) to overcome the penalty. This results in compact, well-separated vMF-like identity clusters on $\mathbb{S}^{d-1}$, i.e., in a high identity concentration κ_{id} in the isotropic reading of Proposition 1.

B.3 RBF kernel as von Mises–Fisher similarity

The RBF kernel between two vectors is $K_{\text{RBF}}(\mathbf{z}_i, \mathbf{z}_j) = \exp(-\gamma_{\text{RBF}}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2)$. For $\mathbf{z}_i, \mathbf{z}_j \in \mathbb{S}^{d-1}$, $\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 = 2(1 - \mathbf{z}_i^\top \mathbf{z}_j)$, so that, with $\kappa_b := 2\gamma_{\text{RBF}}$,

$$K_{\text{RBF}}(\mathbf{z}_i, \mathbf{z}_j) = e^{-\kappa_b}\, e^{\kappa_b \mathbf{z}_i^\top \mathbf{z}_j} = \frac{e^{-\kappa_b}}{C_d(\kappa_b)}\, p(\mathbf{z}_j \mid \mathbf{z}_i, \kappa_b). \tag{22}$$

The RBF kernel on normalized features is thus proportional to the vMF likelihood centered on $\mathbf{z}_i$ (Mardia & Jupp, 2009), and the constant of proportionality does not depend on $\mathbf{z}_i$ or $\mathbf{z}_j$.

B.4 Decision boundaries as vMF mixture comparisons

Consider a binary sub-problem with $y_i \in \{-1, +1\}$. The SVM decision function is

$$f(\mathbf{z}) = \sum_{i\in\mathcal{S}} \alpha_i y_i K_{\text{RBF}}(\mathbf{z}_i, \mathbf{z}) + b, \tag{23}$$

where $\mathcal{S}$ is the set of support vectors and $\alpha_i > 0$ are the dual coefficients. Partitioning the support vectors by class, $\mathcal{S}_\pm := \{i \in \mathcal{S} : y_i = \pm 1\}$, and substituting equation 22,

$$f(\mathbf{z}) = \Lambda\Big[\sum_{i\in\mathcal{S}_+} \alpha_i\, p(\mathbf{z} \mid \mathbf{z}_i, \kappa_b) - \sum_{i\in\mathcal{S}_-} \alpha_i\, p(\mathbf{z} \mid \mathbf{z}_i, \kappa_b)\Big] + b, \qquad \Lambda := \frac{e^{-\kappa_b}}{C_d(\kappa_b)} > 0. \tag{24}$$

With the dual masses $\beta_\pm := \sum_{j\in\mathcal{S}_\pm} \alpha_j$ and the normalized weights $w_i^\pm := \alpha_i/\beta_\pm$, which lie in the probability simplex, each sum defines a vMF mixture density

$$p_\pm(\mathbf{z}) := \sum_{i\in\mathcal{S}_\pm} w_i^\pm \, p(\mathbf{z} \mid \mathbf{z}_i, \kappa_b), \qquad f(\mathbf{z}) = \Lambda\,[\beta_+ p_+(\mathbf{z}) - \beta_- p_-(\mathbf{z})] + b. \tag{25}$$

B.5 Membership in $\mathcal{F}_{\kappa_b}$

Up to the shared normalizer $C_d(\kappa_b)$, $\log p_\pm(\mathbf{z}) = g_\pm(\mathbf{z}) + \log C_d(\kappa_b)$ with

$$g_\pm(\mathbf{z}) = \log \sum_{i\in\mathcal{S}_\pm} w_i^\pm \, e^{\kappa_b \mathbf{z}_i^\top \mathbf{z}}, \tag{26}$$

which is exactly the log-sum-exp score equation 2 with components $\boldsymbol{\mu}_i^\pm = \mathbf{z}_i$ (the support vectors) and concentration $\kappa_b = 2\gamma_{\mathrm{RBF}}$. For $b = 0$, $\operatorname{sign} f(\mathbf{z}) = \operatorname{sign}\big[(g_+(\mathbf{z}) + \log\beta_+) - (g_-(\mathbf{z}) + \log\beta_-)\big]$, i.e., the rule $\arg\max_{c\in\{+,-\}}[g_c(\mathbf{z}) + \tau_c]$ with offsets $\tau_\pm = \log\beta_\pm$, a log-likelihood ratio test between the two vMF mixtures, with the dual masses acting as priors:

$$\hat{y} = \operatorname{sign}\left[\log \frac{\beta_+ \, p_+(\mathbf{z})}{\beta_- \, p_-(\mathbf{z})}\right]. \tag{27}$$

A non-zero bias b shifts the decision threshold of the mixture comparison, as a prior ratio would. The bandwidth parameter thus sets the component concentration κ_b, which can be adapted to the clustering scale induced during pretraining, and the support vectors, which are training points, are the components whose placement determines the margin $F_c(\boldsymbol{\mu})$ of each identity (coverage reading of Section 3.1). Unlike linear classifiers, which assume Euclidean separability, the RBF-SVM accommodates the curved geometry of $\mathbb{S}^{d-1}$; hyperspherical kernels have been studied for SVMs (Zhao & Song, 2018), and the connection above places the RBF-SVM head within the class covered by the geometric reading of the guarantee.

C Empirical validation of the flip bound: protocol and diagnostics

This appendix tests the two theoretical statements of Section 3.1 on real face embeddings: the distribution-free flip bound (Theorem 1) and its geometric reading (Proposition 1). Three questions structure the validation: (i) is the bound *valid*, i.e., does the measured flip rate stay below the certificate, including when the moments are estimated on images disjoint from those on which the flip rate is measured? (ii) is the SNR m_c/s_c a good *ranking* of identities by robustness? (iii) does the SNR behave as its geometric reading predicts?

The rule under test. We instantiate an explicit member of $\mathcal{F}_{\kappa_b}$ on AdaFace IR-50 embeddings ($d = 512$). Each of the four ethnicity class scores is the log-sum-exp score equation 2 over $n_c = 10{,}000$ components (40k in total): the components are embeddings drawn from WEG^{tr}, the weights are uniform, the offsets are zero, and the concentration is set to $\kappa_b = 736.6$, the estimated identity concentration $\hat{\kappa}_{id}$ of the encoder. We choose this rule rather than the trained RBF-SVM head because it makes every quantity of the theory explicit and directly inspectable (components, softmax weights $\pi_i^c(\boldsymbol{\mu})$, gradients $\nabla F_c(\boldsymbol{\mu})$). The price is that it is weaker than a trained head: its center accuracy (accuracy of $c_0 = \hat{c}(\boldsymbol{\mu})$) is 0.86 and its mean per-identity flip rate is 0.21. This is a feature for the validation: a stressed rule populates the whole range of flip rates, so the bound and the ranking are tested where they matter, not only in the easy regime.

Evaluation data and per-identity quantities. The rule is evaluated on 2,400 WEG^v identities with 20 images each; these identities are disjoint from the components. For each identity we compute: the center $\boldsymbol{\mu}$ (normalized mean of the 20 embeddings) and its label $c_0 = \hat{c}(\boldsymbol{\mu})$; the empirical flip rate $\hat{q}_{id}$ (fraction of the images with $\hat{c}(\mathbf{z}) \neq c_0$); and, for each opposing class $c \neq c_0$, the empirical mean m_c and unbiased standard deviation s_c of the scalar score F_c over the images. From the moments we form the two certificates: the union bound equation 3, summing $s_c^2/(s_c^2 + m_c^2)$ over the three opposing classes, and the min-class variant of Section A.3, which applies Cantelli to the single score $F_{\min} = \min_{c\neq c_0} F_c$. When a single scalar summary is needed, the SNR of an identity is the worst-case one, $\min_{c\neq c_0} m_c/s_c$.

Out-of-sample protocol. The moments and the flip rate are, in the first instance, computed from the same 20 images, so a valid bound could partly reflect this reuse: the certificate would be optimistic on fresh images of the same identity. To measure this effect, we repeat 20 random half-splits of the images of each identity: on half A (10 images) we estimate c_0, m_c, and s_c, hence the certificate; on the disjoint half B we measure the flip rate. This simulates the intended use case, in which an identity's certificate is computed once and must hold for its future images. The same splits also give the *reliability ceiling*: the Spearman correlation between the flip rates measured on the two halves is 0.73, and no predictor computed from half A can be expected to rank the half-B flip rates better than the half-A flip rate itself does.

Validity of the bound. Figure 3 summarizes the validation. In-sample (moments and flip rate on the same images), the bound holds for 100% of identities, for both variants. Out of sample, the union bound equation 3 holds for 99.8% of identities and the tighter min-class variant for 96.5%; the failures are identities whose moments happened to be favourably estimated on half A. Validity is therefore essentially preserved under the protocol.

Tightness. The certificate is conservative, as a distribution-free bound must be. Out of sample, the min-class bound averages ≈ 0.51 against a mean flip rate of ≈ 0.22, and the union version averages ≈ 0.72; the min-class bound is non-vacuous ($< 1/2$) for $\approx 56\%$ of identities, and 64% of identities satisfy the condition $m_c \geq s_c$ for all opposing classes (the regime in which Theorem 1 is informative). E therefore think that future work could tighten this bound if, beyond theoretical motivation, a tight guarantee is needed.

The SNR as a ranking. The SNR orders identities by robustness: Spearman correlation between the SNR and the flip rate of -0.94 in-sample, and -0.75 out of sample (above the reliability ceiling of 0.73, i.e., the SNR computed on half A ranks the half-B flip rates slightly better than the half-A flip rate itself). The ranking is also well calibrated in levels: across SNR bins from < 0 to ≥ 3, the mean flip rate falls monotonically from 0.74 to 0.00 (Figure 3a).

The SNR subsumes the bare margin. Proposition 1 says the relevant quantity is the margin *divided by the directional spread*, not the margin alone. The data agree: controlling for the SNR, the linearized margin carries no residual information about the flip rate (partial Spearman $\mathrm{partial}(\hat{q}_{id}, \gamma_{\mathrm{lin}} \mid \mathrm{SNR}) = -0.06$), whereas controlling for the margin leaves the SNR almost fully informative ($\mathrm{partial}(\hat{q}_{id}, \mathrm{SNR} \mid \gamma_{\mathrm{lin}}) = -0.89$). The denominator s_c (the identity's spread along the decision direction) is thus not a refinement but a necessary part of the predictor.

Fairness reading. The SNR ordering across segments coincides with the flip-rate ordering (median SNR / mean flip rate: African $1.78/0.15$, Asian $1.59/0.16$, Caucasian $1.27/0.22$, Indian $0.81/0.30$). Differences in robustness between segments are therefore explained by the same mechanism as differences between identities (coverage of the segment by the components and spread along the decision directions) which ties the fairness of robustness to segment-wise coverage.

Sub-Gaussian variant. Theorem 1 also offers an exponential bound if the centered score $F_c - m_c$ is sub-Gaussian. In-sample, the standardized residuals $(F_c - m_c)/s_c$ are Gaussian-like: pooled lower-tail sub-Gaussian ratio $r \approx 1.0$, skewness 0.07, excess kurtosis -0.37 (Figure 4, bottom left). Out of sample, however, the standardized tails are heavier, not because the scores change, but because standardizing by moments estimated from only 10 images inflates the tails. Since the exponential bound is exactly the one that is sensitive to the tails, we state the polynomial Cantelli bound as the main result and keep the Sub-Gaussian variant as a remark.

Geometric reading. Two predictions of Proposition 1 are checked directly. First, the gradient $\nabla F_c(\boldsymbol{\mu})$ should be essentially tangent to the sphere at $\boldsymbol{\mu}$; measured, the median tangent fraction of its norm is 0.994. Second, the measured SNR m_c/s_c should track its geometric approximation $\gamma_{\mathrm{lin}}/\sigma_{id,\hat{\nabla}}$; the two agree up to a Spearman correlation of 0.67 (Figure 4, bottom right). The dispersion around the trend is expected: it reflects the second-order (curvature) correction to m_c discussed in Appendix A.

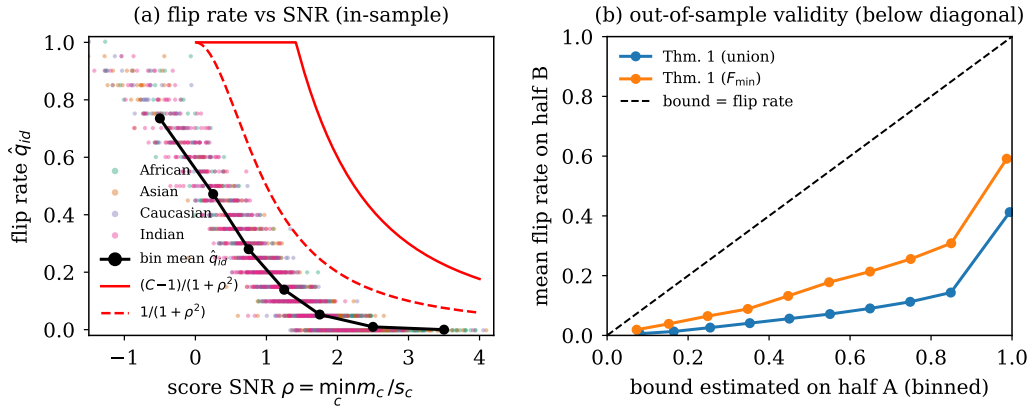


Figure 3: Validation of Theorem 1 (AdaFace IR-50, explicit member of $\mathcal{F}_{\kappa_b}$, 2,400 held-out WEG^v identities $\times$ 20 images). **(a)** Per-identity flip rate against the score SNR, with binned means and the Cantelli curves. **(b)** Out-of-sample validity: moments and c_0 estimated on half of the images, flip rate measured on the other half; the bound is valid below the diagonal. ($F_{\min}$: Cantelli applied to $\min_c F_c$, Section A.3.)

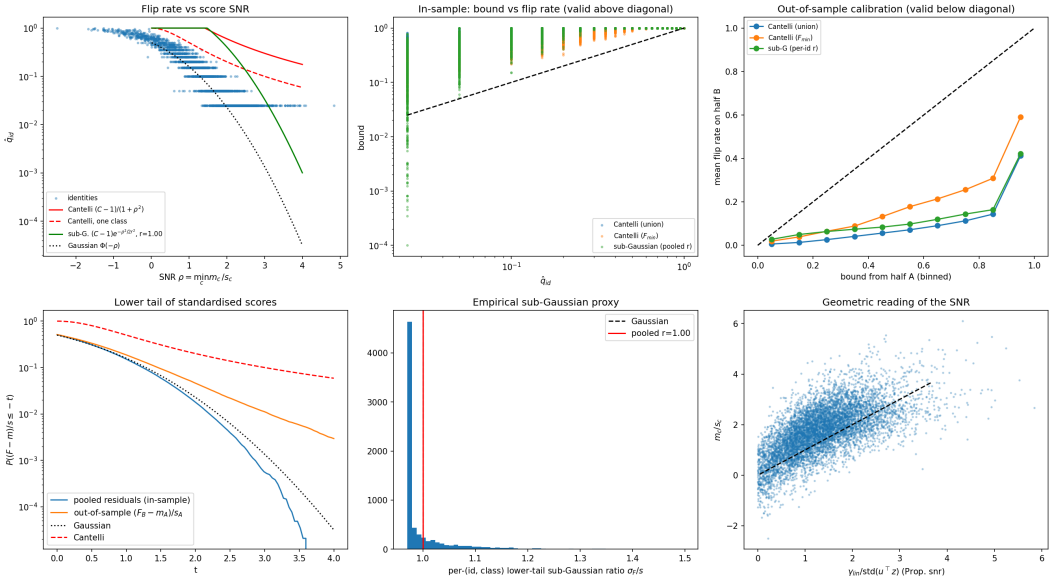


Figure 4: Full diagnostics of the validation of Theorem 1. **Top left:** per-identity flip rate against the SNR, with the Cantelli, sub-Gaussian, and Gaussian curves. **Top center:** in-sample validity (bound above the diagonal). **Top right:** out-of-sample calibration (valid below the diagonal). **Bottom left:** lower tail of the standardized scores $(F_c - m_c)/s_c$, in-sample and out of sample. **Bottom center:** per-(identity, class) lower-tail sub-Gaussian ratio. **Bottom right:** measured SNR m_c/s_c against its geometric reading $\gamma_{\text{lin}}/\sigma_{id,\hat{\nabla}}$ (Proposition 1).

D Balanced Sampling via Iterative Proportional Fitting

In this section, we formally describe the label inference and the sampling for creating WEG^{tr}, as mentioned in Section 4.1, and validate it experimentally. The motivation behind the sampling is to create a dataset that faithfully reproduces the original structure of fine ethnicities, which are not compartmentalized into the 4 coarse ethnicities, but can often overlap them. We identify a good approach, a sampling based on the conditional probability $p(g|e)$ of a coarse ethnicity g, given a fine ethnicity e.

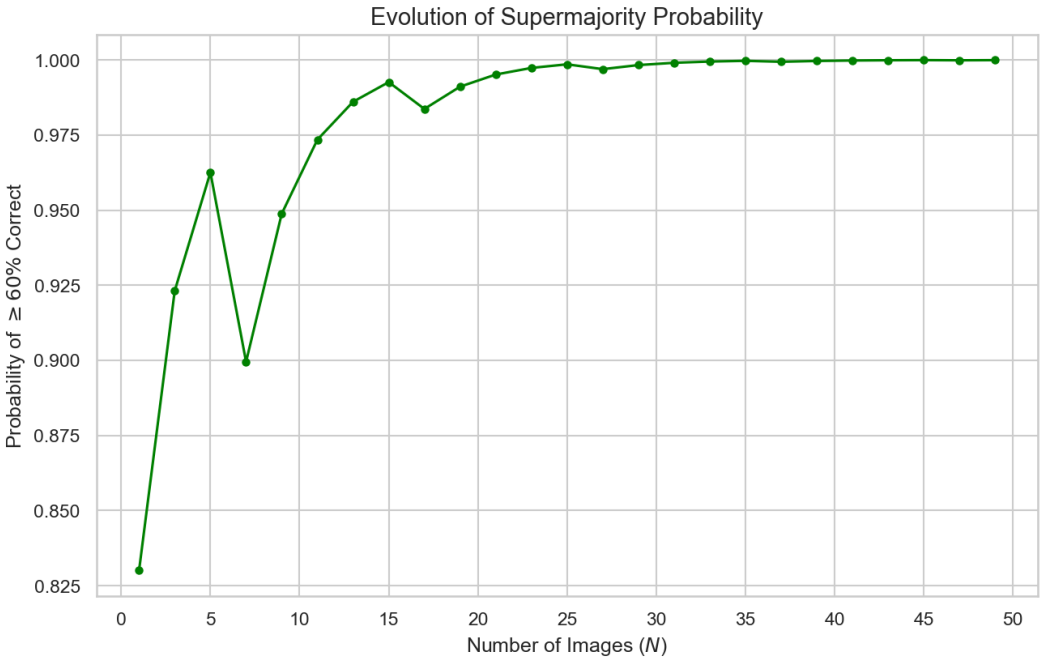


Figure 5: Theoretical evolution of the supermajority label being accurate depending on N, when k = 0.6 N (i.e., our threshold for WEG ethnicity estimation), and with p = 0.83

D.1 Theoretical and Experimental Guarantees for Label Inference Through Identity-Wise Majority Voting

Leveraging the predictions made by an ethnicity or gender classifier on several images of the same identity allows the aggregated label to substantially exceed the per-image accuracy of the underlying model. Formally, let p denote the probability that the classifier assigns the correct label to a single image. Assuming that prediction errors across different images of the same identity are approximately independent – a reasonable assumption given the diversity of pose, illumination, and expression within a WebFace42M identity – the number of correct predictions over N images follows a $\mathrm{Binomial}(N, p)$ distribution. We retain the majority label only when it represents at least 60% (the threshold we kept for ethnicity estimation) of the N predictions, i.e., when the number of correct votes $k \geq \lceil 0.6\, N \rceil$. The probability that this supermajority label is correct is therefore:

$$P(k \geq \lceil 0.6\, N \rceil) \;=\; \sum_{k=\lceil 0.6\, N \rceil}^{N} \binom{N}{k} p^k \, (1-p)^{N-k}. \tag{28}$$

As N increases, this probability converges rapidly to 1 for any $p > 0.6$. We provide the theoretical evolution of this probability for $p = 0.83$ (FFNet accuracy on the full BUPT dataset) in Figure 5. Even at $N = 5$ images, the supermajority probability already reaches 96.3%, and at $N = 20$ it exceeds 99.6%. In practice, our filtering step retains only identities whose per-image consensus exceeds a stricter threshold, further tightening this bound. The obtained label accuracy on BUPT with our thresholds reaches 95.5%, vastly outperforming the per-image baseline and reducing the error rate well below 5%, thereby breaking the potential circularity induced by the use of an automatic annotator.

As a final quality check, we manually verify the accuracy of ethnicity and gender labels for WEG^t. Confirming the robustness of the identity-level attribute labeling, fewer than 0.3% of samples are removed during manual verification.

D.2 Balanced Sampling Procedure

Consider a dataset with N groups indexed by $g \in \{1, \dots, N\}$ and M subgroups indexed by $e \in \{1, \dots, M\}$. Let $p(g|e)$ denote the conditional probability of group g given subgroup e, satisfying $\sum_{g=1}^{N} p(g|e) = 1$ for all e.

We propose the following two-stage sampling procedure:

1. Sample a group g uniformly with probability $p(g) = \frac{1}{N}$
2. Within group g, sample subgroup e with probability $p(e|g) = \frac{p(g|e)}{S_g}$

where $S_g = \sum_{e=1}^{M} p(g|e)$ denotes the total mass of group g across all subgroups.

Theorem 2 (Balanced Sampling Condition)**.** *The marginal sampling probability of subgroup e is*

$$p(e) = \frac{1}{N} \sum_{g=1}^{N} \frac{p(g|e)}{S_g} \tag{29}$$

When $p(g) = \frac{1}{N}$ the following equivalence yields:

$p(e) = \frac{1}{M}$ for all e if and only if $S_g = \frac{M}{N}$ for all $g \in \{1, \dots, N\}$. Thus, when sampling with regard to $p(g|e)$, ensuring that $S_g = \frac{M}{N}$ and sampling the coarse ethnicities uniformly enforce diversity with regard to the fine ethnicities.

This theoretical element motivates the use of this sampling.

Proof. By the law of total probability,

$$p(e) = \sum_{g=1}^{N} p(e|g)p(g) = \sum_{g=1}^{N} \frac{p(g|e)}{S_g} \cdot \frac{1}{N} = \frac{1}{N} \sum_{g=1}^{N} \frac{p(g|e)}{S_g}. \tag{30}$$

For the sufficiency, if $S_g = \frac{M}{N}$ for all g, then

$$p(e) = \frac{1}{N} \sum_{g=1}^{N} \frac{p(g|e)}{M/N} = \frac{N}{M} \cdot \frac{1}{N} \sum_{g=1}^{N} p(g|e) = \frac{1}{M}. \tag{31}$$

For the necessity proof, suppose $p(e) = \frac{1}{M}$ for all e. Note that

$$S_g = \sum_{e=1}^{M} p(g|e) \tag{32}$$

$$= \sum_{e=1}^{M} p(e|g) * \underbrace{p(g)}_{1/N} / \underbrace{p(e)}_{1/M} \tag{33}$$

$$= \underbrace{\sum_{e=1}^{M} p(e|g)}_{=1} * M/N = \frac{M}{N} \tag{34}$$

If the distribution is uniform over subgroups while maintaining uniform group sampling, the balance constraint $S_g = \frac{M}{N}$ must hold to ensure compatibility of the marginals. □

D.3 Enforcing Balance via Iterative Proportional Fitting

Given an empirical conditional probability matrix $\{\tilde{P}_e\}_{e=1}^{M} = \{\tilde{p}(g|e)\}_{e=1}^{M}$, we seek an adjusted matrix $\{P_e^*\}_{e=1}^{M} = \{p^*(g|e)\}_{e=1}^{M}$ that satisfies the balance constraints while minimally perturbing the original distribution. This is formalized as a constrained Kullback-Leibler divergence minimization problem (Csiszár, 1975; Ireland & Kullback, 1968).

Algorithm 1 Iterative Proportional Fitting (IPF)

Input: Empirical matrix $\tilde{P}$, tolerance ϵ
Initialize: $P^{(0)} \leftarrow \tilde{P}$, $t \leftarrow 0$
repeat
  // Row normalization
  for $g = 1$ to N **do**
    $r_g^{(t)} \leftarrow \sum_{e=1}^{M} p^{(t)}(g|e)$
    $\hat{p}^{(t)}(g|e) \leftarrow p^{(t)}(g|e) \cdot \frac{M/N}{r_g^{(t)}} \quad \forall e$
  end for
  // Column normalization
  for $e = 1$ to M **do**
    $c_e^{(t)} \leftarrow \sum_{g=1}^{N} \hat{p}^{(t)}(g|e)$
    $p^{(t+1)}(g|e) \leftarrow \hat{p}^{(t)}(g|e) \cdot \frac{1}{c_e^{(t)}} \quad \forall g$
  end for
  $t \leftarrow t + 1$
until $\max_g |r_g^{(t)} - M/N| < \epsilon$ and $\max_e |c_e^{(t)} - 1| < \epsilon$
Return: $P^* \leftarrow P^{(t)}$

Problem 1 (Constrained KL-Divergence Minimization)**.** *Find P^* that solves*

$$\begin{aligned} \min_{P_e} \sum_{e}^{M} \quad & D_{\mathrm{KL}}(P_e \| \tilde{P}_e) = \sum_{e=1}^{M} \sum_{g=1}^{N} p(g|e) \log \frac{p(g|e)}{\tilde{p}(g|e)} \\ \textit{s.t.} \quad & \sum_{e=1}^{M} p(g|e) = \frac{M}{N} \quad \forall g \in \{1, \dots, N\} \\ & \sum_{g=1}^{N} p(g|e) = 1 \quad \forall e \in \{1, \dots, M\} \\ & p(g|e) \geq 0 \quad \forall g, e \end{aligned} \tag{35}$$

The Iterative Proportional Fitting (IPF) algorithm (Deming & Stephan, 1940; Sinkhorn & Knopp, 1967) provides an efficient method to solve Problem 1. IPF alternates between normalizing rows and columns until convergence. We apply this algorithm to our problem, following the pseudo code of Algorithm 1.

D.4 Theoretical Guarantees

Algorithm 1 converges to the unique solution P^* of Problem 1. The solution has the multiplicative scaling form

$$p^*(g|e) = u_g^* v_e^* \tilde{p}(g|e) \tag{36}$$

where $u_g^*, v_e^* > 0$ are scaling factors satisfying the marginal constraints

$$\sum_{e=1}^{M} u_g^* v_e^* \tilde{p}(g|e) = \frac{M}{N}, \quad \sum_{g=1}^{N} u_g^* v_e^* \tilde{p}(g|e) = 1. \tag{37}$$

The convergence and optimality of IPF for KL-divergence minimization subject to marginal constraints is well-established in the literature (Csiszár, 1975; Sinkhorn & Knopp, 1967). The algorithm performs block coordinate ascent on the dual problem, where each row (column) normalization step corresponds to optimizing over one set of dual variables while fixing the other (Peyré & Cuturi, 2019). The objective function is strictly convex, ensuring a unique global optimum. By strong duality (Boyd & Vandenberghe, 2004), the primal solution obtained from the optimal dual variables $p^*(g|e) = u_g^* v_e^* \tilde{p}(g|e)$ is the unique minimizer of Problem 1. For a comprehensive treatment of I-projections and the information-geometric perspective, we refer to (Csiszár, 1975).

	Arch.	#GFLOP	Ethnicity								Gender					
			Accuracy ↑			Fairness ↓			Robust. ↓		Accuracy ↑		Fairness ↓		Robust. ↓	
			BFW	*RFW*	*WEG^t*	*BFW*	*RFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*
WEG IPF	IR50	12.64	**95.9**	**95.8**	90.8	**0.95**	1.22	**2.52**	0.07	**0.13**	**97.9**	94.7	1.14	**0.11**	**0.07**	0.19
WEG Cond	IR50	12.64	95.8	**95.8**	90.7	1.21	**1.18**	2.66	0.07	0.14	97.8	94.7	1.28	0.41	0.08	**0.18**
WEG Uniform	IR50	12.64	95.6	95.6	**91.0**	1.12	1.23	2.61	0.07	0.14	**97.9**	94.7	**1.08**	0.29	**0.07**	0.19

Table 4: Ethnicity inference accuracy, fairness (DoB $\times 100$), and robustness (HoE) for three training set variants of REGI (IR50, SVM-RBF). Gender columns are left blank. **Bold**: best; underline: second best.

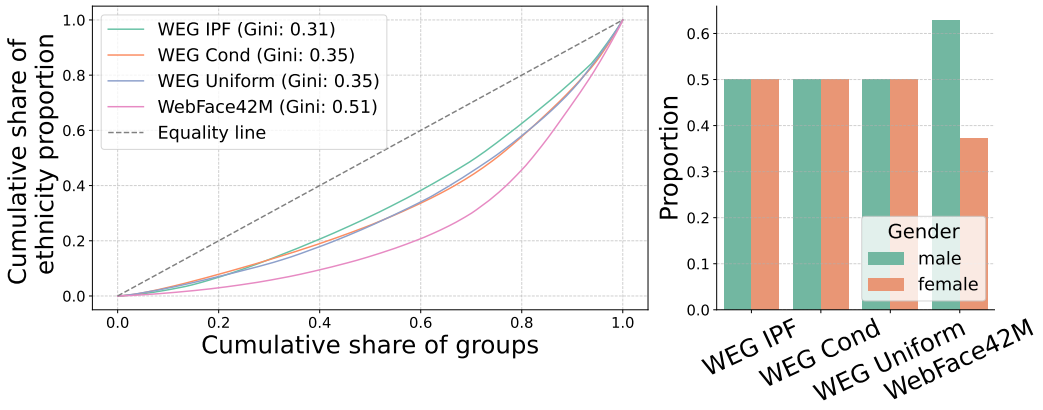


Figure 6: Fine ethnicities and gender proportion study for different balancing approaches.

The IPF solution P^* preserves the zero structure of $\tilde{P}$ (if $\tilde{p}(g|e) = 0$ then $p^*(g|e) = 0$) and minimally perturbs the original distribution in the sense of KL-divergence. This makes it particularly suitable for applications where the empirical associations $\tilde{p}(g|e)$ reflect meaningful semantic relationships that should be preserved while enforcing balance constraints.

D.5 Experimental Verification

We compare across 3 sampling strategies, the results obtained when training REGI on the resulting dataset. The 3 sampling strategies are as follow:

- **WEG Uniform** consists in sampling an ethnicity present in a coarse ethnicity g with probability $\frac{1}{M_g}$ with $M_g = \#\{e, e \in g\}$, the number of fine ethnicities present in g.
- **WEG Cond** consists in sampling with regard to $\tilde{p}(g|e)$.
- **WEG IPF** consists in sampling with regard to $p^*(g|e)$

We show in Table 4 that the WEG IPF sampling yields equivalent Accuracy (slightly better), but most importantly, yields fairer and more robust predictions than the other approaches. We think these results are a consequence of a more effective balancing (as shown by the Lorenz curve and the reported GINI coefficient in Figure 6) than other sampling strategies. Figure 6 Results are coherent with Theorem 2.

E Comparison of WEG^t with Existing Benchmarks

In this section, we highlight the specific characteristics of WEG^t compared with existing benchmark sets. These proper characteristics can be decomposed in three points: **Fine-grained demographic diversity.** WEG^t improves fine-grained demographic diversity over both BFW and RFW, as measured by the Gini coefficient (lower is better) computed across the fine-grained ethnicity sub-groups

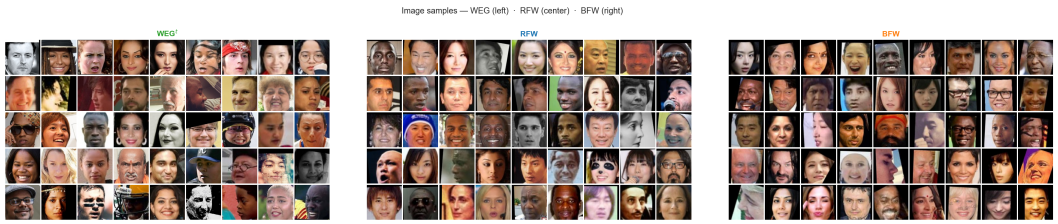


Figure 7: Samples drawn randomly from the different benchmark sets.

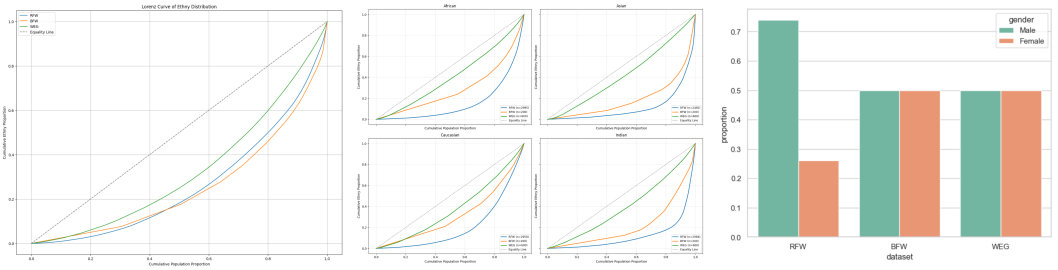


Figure 8: **Left** Lorenz curve for the overall distribution of fine ethnicities in different benchmark sets. The closer to the equality line, the more balanced.
Center: Lorenz curves for each coarse ethnicity. **Right**: Proportion occupied by each gender in each benchmark set.

within each coarse label. WEG^t achieves a Gini coefficient of 0.35, compared with 0.48 for RFW and 0.49 for BFW. This lower value reflects a more balanced representation of fine-grained subgroups within each coarse ethnic category, exposing identities that BFW and RFW systematically underrepresent. Concretely, the diversification procedure described in Section 3.1 (which samples uniformly over fine-grained ethnicities rather than only over the four coarse labels) directly reduces this imbalance and produces a test set that is more sensitive to intra-group performance variations.

Reduced accuracy saturation. BFW and RFW exhibit pronounced accuracy saturation for state-of-the-art models, which causes *fairness metrics to collapse*: when all groups are predicted with near-ceiling accuracy, the degree of bias (DoB) tends toward zero regardless of the model's actual discriminative properties, rendering the benchmark uninformative. As shown in Table 1, several baselines already reach accuracy plateaus on BFW and RFW, while WEG^t preserves a larger margin for discrimination across all evaluated models. This additional headroom makes WEG^t *more discriminative* for both current and future models, and ensures that fairness metrics remain meaningful as DAI performance continues to improve.

Stronger robustness evaluation. WEG^t contains approximately $\times 3$ more identities and $\times 2.5$ more images than BFW, providing the statistical depth required for reliable intra-identity consistency estimation via the HomE metric (Eq. 4). RFW, by contrast, does not provide identity-level annotations

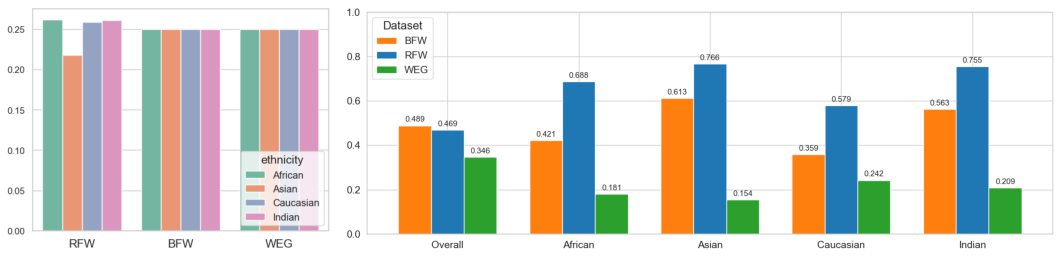


Figure 9: **Left:** Coarse ethnicities proportions for each benchmark set. **Right:** Gini coefficient for the fine ethnicities for each dataset. Overall stands for the Gini coefficient obtained on the whole dataset. We report the Gini coefficient of fine ethnicities for each coarse ethnicity.

and therefore cannot support the same identity-level robustness analysis; accuracy and fairness are the only dimensions assessable on RFW, as noted in Section 4.1. The larger number of identities in WEG^t also reduces the variance of the HomE estimator, yielding more stable robustness measurements across demographic groups and making cross-model comparisons more reliable.

F Implementation detailed description

F.1 Existing Datasets References

FairFace (Kärkkäinen & Joo, 2021) is described in detail in section 2.

BUPT (Wang et al., 2021) is a dataset of real images that can be found at this URL. We actually consider *BUPT-BalanceFace* that contains 1.3M images from 28k celebrities. Four labels are considered, namely *caucasian*, *indian*, *asian* and *african*. The dataset is balanced w.r.t the number of identity per class (7k per label) and approximately balanced in terms of number of images: 324k for *african*, 325k for *asian*, 326k for *caucasian* and 375k for *indian*.

WebFace (Zhu et al., 2023) is a dataset of real images, containing 4M identities represented by 260M face images, which a part (4M identities and 42M face images) has been curated to result into Webface42M.

F.2 Seed Images Collecion

The list of ethnicities covers the different regions of the world to optimize the diversity of the created datasets. The initial list was curated by prompting GPT-5.1 to provide up to 50 most populous ethnic groups including at least 1 million persons for major world regions (Africa, Asia, Europe, North America, South America, Oceania). In addition, we ran specific queries to get up to 20 ethnicities for large multi-ethnic countries, such as India or Indonesia, to improve their coverage. For instance, the list includes items such as *Javanese, Malay, Sundanese, Batak* for *Indonesia*. Since there is no one-to-one mapping between ethnicities and regions/countries, we deduplicated the items and obtained a list of 270 ethnicities.

We then populated the ethnicities by prompting GPT-5.1 to get 100 notable person names for each gender using ``List up to 100 notable living ETHNICITY GENDER. Diversify the professions and occupations. Format the output in JSON, providing the name and the profession. Avoid duplicate names.'' The list formatting was correct, but some duplicate names or name concatenations appeared in the results, and we filtered them out. After filtering, the list included nearly 30,000 unique names, with a 45% ratio of females.

Next, we collected 60 images for each name using a SERP API service over the Bing search engine and used RetinaFace (Deng et al., 2020) to detect faces in the images and kept those whose width was at least 96 pixels. We kept the identities that included at least 30 detected faces. Due to name ambiguity and retrieval imperfections, the search results contained noisy images and required filtering. We implemented an automatic filtering method inspired by (Cao et al., 2018). We extracted face embeddings using the AdaFace R100 model trained with WebFace12M [3]. Then we compute an identity prototype by averaging the embeddings of the first five images including a single detected face. We compared each face embedding with the prototype identity representation and retained only faces that are closer to the target prototype than to any other. Finally, we applied a face verification protocol to further ensure that all faces belonged to the same identity. Preselected images of an identity were kept if their distance to at least one other image of the target identity was below a verification threshold. This threshold was optimized on the FAVCI2D dataset (Popescu et al., 2022), a demographically diversified face verification dataset.

To verify the quality of the filtering, we manually annotated the images of 100 randomly-selected identities. The average accuracy of the face-identity attributions was 99.3%. This score confirms the robustness of the proposed seed identity filtering.

[3] https://github.com/mk-minchul/adaface

F.3 TRAINING AND INFERENCE DETAILS

For training the end-to-end (E2E) and finetuned (FT) classifier (c.f. Table 3b, we set the training duration to 30 epochs for E2E and 10 epochs for FT, with a batch size of 256. We use a warmup-cosine learning rate scheduler. The learning rate reaches its peak at 1e-3 for E2E training and 1e-4 for FT. The whole backbone is updated in both cases. Fine-tuning is performed after the same 30-epoch face-recognition pretraining as REGI. For this ablation, use the same iResNet50 backbone for fair comparison. We use a CrossEntropy loss and the augmentation described below in both cases.

For training the **face recognition classifiers** (REGI pretraining task), we employed the standard setup from AdaFace. We used the AdaFace loss with parameter h set to 0.333 and m set to 0.4. We applied the standard framework augmentations for face recognition training: random flipping, crop augmentation, low-resolution augmentation, and photometric augmentation. We trained the model for 30 epochs using the SGD optimizer with momentum and weight decay fixed at 5e-4. The scheduler was a step LR scheduler that divides the learning rate by 10 at epochs 12, 20, and 24. To ensure stability of the AdaFace loss when training the ResNets (torchvision implementation) 50 and 101, we add a fully connected layer to the models to project the original 2024 latent space to 512.

The SVM implementation is from scikit-learn (Pedregosa et al., 2011). To accelerate the computation of cross distances, we implemented a wrapper that performs this phase on a GPU. This allows us to drastically reduce both training and inference time (training time drops from 40 minutes to 2 minutes on a training set of 100k samples). For the RBF kernel, we use the scaling gamma defined as $\frac{1}{D \cdot \mathrm{Var}(X)}$, where D is the dimension of the feature vectors (512) and X are the input vectors. For the ablation on the datasets (Table 2), we found that, in the case of imbalanced training set with regard to the predicted label, a weighted hinge loss yields better performance than the unweighted one. Thus, we use, when training on FairFace, and when training on BUPT for gender inference, the weighting defined as: $\mathrm{Weight}(\mathrm{Class}_i) = \frac{0.5}{\mathrm{Frequency}(\mathrm{Class}_i)}$. To yield full compatibility with the hypothesis of Equation 27, our dataset is carefully balanced with regard to gender and ethnicity (c.f. Appendix D), and thus does not require such weighting. We use the squared hinge loss with L2 penalization and 1.0 as the regularization parameter C.

All trainings are performed on a compute device having 250 GB of RAM, 32 CPU cores and an NVIDIA H100 GPU.

As for the inference of the VLMs we used the following approaches: For the CLIP models, we leverage the common text/image encoding space: we encode each class name and compute the cosine distance between the input image embedding and each class embedding, assigning the class with the smallest distance.

For QWEN models, we use the following prompt for ethnicity prediction: ``Please analyze the image and determine the apparent ancestry of the person shown. Answer with one of the following: "Asian", "Caucasian", "African", "Indian. Return only the ancestry label in lowercase, nothing else.``

As for gender: ``Please analyze the image and determine the apparent gender of the person shown. Answer with one of the following: 'male', 'female'. Return only the gender label in lowercase, nothing else.``

F.4 ARCHITECTURE PARAMETERS

We report in Table 5 the parameters and GFLOPS associated with each architecture mentioned in the main paper. While allowing for significantly better performances, the backbones used by REGI-Face require lower parameter count and computational power.

Architecture(s)	#Parameters (M)	GFLOPS
RN18	11.2	0.98
ViT-T	19.1	2.76
RN34	21.3	1.94
IR18	24.0	5.24
RN50	24.6	2.18
IR50	43.5	12.64
RN101	42.5	4.04
ViT-S	76.0	11.00
Swin-B	109	30.92
VGG-Face	134	31.06
CLIP ViT-L	427	155.4
CLIP ViT-B	151	8.74
QWEN3 4B	4437	1101
QWEN3 8B	8767	1973

Table 5: Parameter count and FLOPS (computed with ptflops, except for QWEN, which required the use of deepspeed.profiling) to architecture mapping for DAI pipelines.

G Details on metrics

G.1 Performance Metrics

The annotation raw and group-wise performance of 1 is estimated with *Accuracy*, with the exception of the group-wise gender performance, which relies on the TPR, since it is a binary classification task, defined, with its counterpart FPR as:

- *True Match Rate (TMR)* or *True positive Rate (TPR)* or *Sensitivity* is the number of actual positive cases correctly identified over the actual number of positive samples in the dataset
- *False Match Rate (FMR)*, or *False Positive Rate (FPR)*, is the number of negative cases incorrectly identified as positive by the FAS over the actual number of negative samples.

G.2 Fairness Metrics

To estimate the fairness in the main paper, we considered the Degree of Bias (DoB) (Gong et al., 2020) for ethnicity inference, which is the standard deviation of the per-group accuracies. A lower standard deviation means that the performance varies less from one group to another, thus implying a fairer annotator.

Other aggregation metrics, such as *Demographic Parity Difference (DPD)* and *Demographic Parity Ratio (DPR)* (Agarwal et al., 2018; 2019) are derived from the same principle, aiming at achieving equal representation of the demographic groups.

Equalized Odds Difference (EOD) and *Equalized Odds Ratio (EOR)* (Agarwal et al., 2018; Hardt et al., 2016) require that the FAS's performance is independent of the demographic group, that is, the TPR or FPR are the same across all groups. The EOD is defined by the absolute difference between the smallest and largest values for the considered performance metric (TPR or FPR) across groups, while EOR is defined according to their ratio. Similarly to demographic parity, these metrics are in $[0, 1]$ a FAS is perfectly fair when its EOD is zero or its EOR is one. We used EOD for evaluating the fairness with regard to gender inference, which in a binary classification setup is equivalent as measuring the difference between the two TPR for Male and Female.

We report below additional plots to illustrate predictor fairness. Confirming the results reported in Table 1, Figure 10 reports the accuracy ethnicity-wise for the ethnicity prediction task, showing that REGI achieves stable prediction performance across all 4 segments. We also analyze the fairness of ethnicity prediction with regard to gender in Figure 11, highlighting the robustness of REGI.

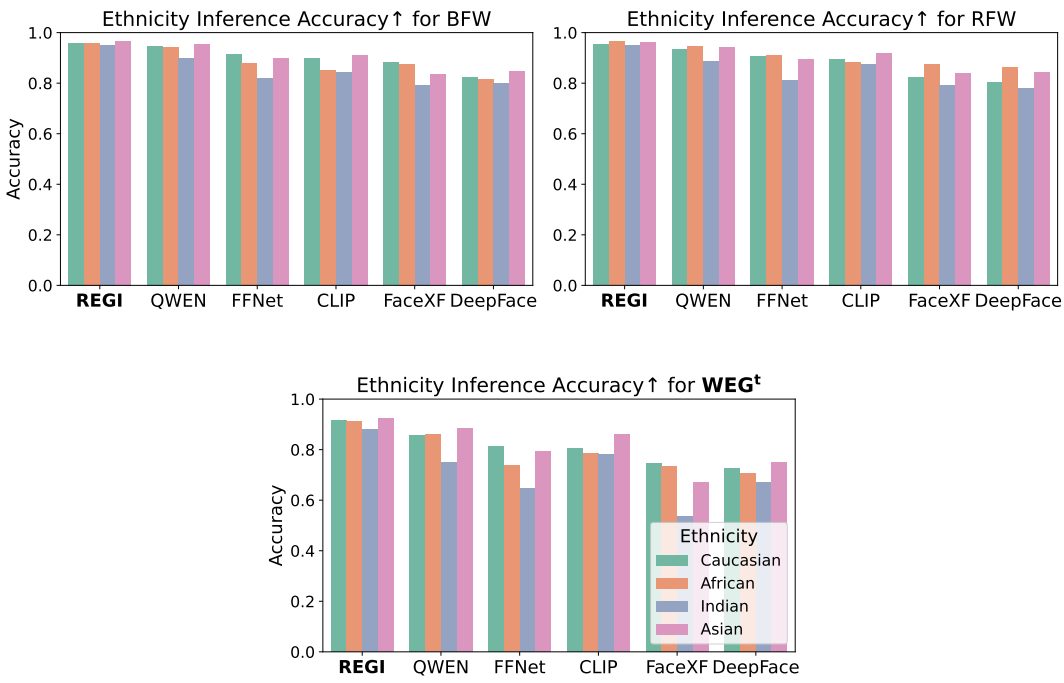


Figure 10: Accuracy per ethnicity group on ethnicity inference task.

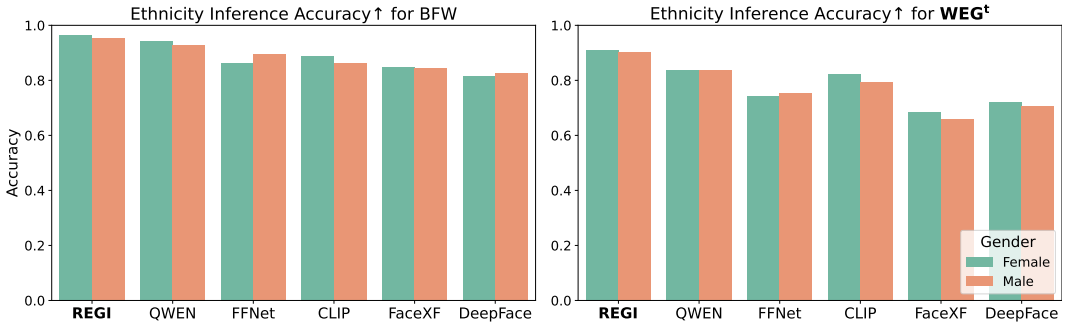


Figure 11: Accuracy per gender on ethnicity inference task.

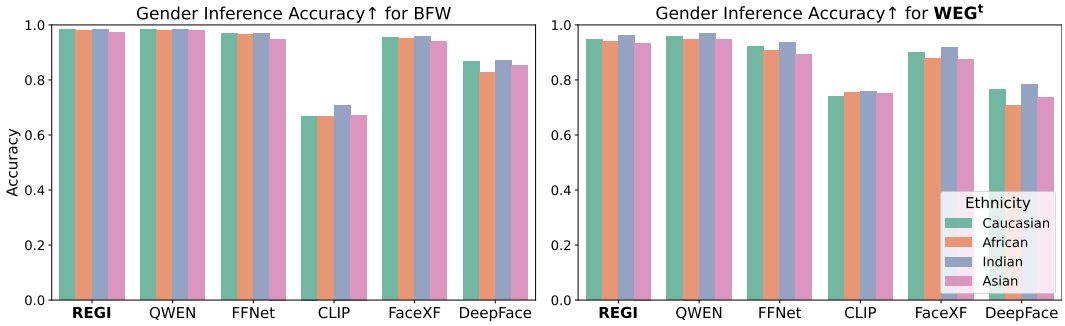


Figure 12: Accuracy per ethnicity on gender inference task.

G.3 Label Robustness Metric

Let us consider an image dataset $\mathcal{I} = \{I_j^{id}; id \in [\![1, D]\!], j \in [\![1, N_{id}]\!]\}$, with D identities, an identitity id being represented by N_{id} images. Such datasets are typically face recognition-like, containing images grouped by identity (BFW, WebFace, BUPT, CASIA, etc.). We consider a labeling function $f_{\mathcal{A}}^{\mathcal{C}}$ that attributes a label to each image, each label being an element of a finite set $\mathcal{C}$, and depends of a dataset $\mathcal{A}$ ($f_{\mathcal{A}}^{\mathcal{C}}$ is defined below). We propose to consider three metrics to reflect the robustness of the labeling system $(\mathcal{A}, \mathcal{C})$. The general idea is that a good labeling system should (at least) attribute as much as possible images of a given identity to the same label.

Macro majority accuracy. For each identity id, we consider the frequency of the majority label and compute the average over all labels of the dataset

$$MaMA = \frac{\sum_{id=1}^{D} \max_{k=1...N_{id}} \left(\sum_{j=1}^{N^{id}} \delta_{f_{\mathcal{A}}^{\mathcal{C}}(I_j^{id}) = f_{\mathcal{A}}^{\mathcal{C}}(I_k^{id})} \right)}{\sum_{id} N_{id}} \tag{38}$$

Micro majority accuracy For each identity, we consider the proportion of images labeled with the majority label and compute its average over all identities:

$$MiMA = \frac{1}{D} \sum_{id=1}^{D} \frac{1}{N_{id}} \max_{k=1...N_{id}} \left(\sum_{j=1}^{N^{id}} \delta_{f_{\mathcal{A}}^{\mathcal{C}}(I_j^{id}) = f_{\mathcal{A}}^{\mathcal{C}}(I_k^{id})} \right) \tag{39}$$

Last, let remind that in the main paper we consider, for each identity id, the relative frequency of each possible label $k \in \mathcal{C}$ noted as $p_{id}^k = \frac{1}{N_{id}} \sum_{j=1}^{N^{id}} \delta_{f_{\mathcal{A}}^{\mathcal{C}}(I_j^{id}) = f_{\mathcal{A}}^{\mathcal{C}}(I_k^{id})}$. We then compute the **homogeneity entropy** as the average entropy of p over all identities:

$$HomE = \frac{1}{D} \sum_{id} \frac{1}{\log(C)} \sum_{k} -p_{id}^k \log(p_{id}^k) \tag{6 revisited}$$

These metrics reflect the average variation of the attributions for each entity. Ideally, all images of a given entity are attributed to the same label. In this case, the *homogeneity entropy macro majority accuracy* and *micro majority accuracy* are respectively 0,1 and 1. In the worst case, the images of a given entity are uniformly distributed over all C labels, leading to the three metrics above being respectively 1, $\frac{1}{C}$ and $\frac{1}{C}$. We thus normalize MiMA and MaMA by multiplying by $\frac{C}{C-1}$ and removing $\frac{1}{C-1}$ such that the worst value is 0 and the best value is 1

Labelling function the function $f_{\mathcal{A}}^{\mathcal{C}}$ project an image onto a unique label in $\mathcal{C}$. It is defined according an image dataset $\mathcal{A}$. We consider two versions of $f_{\mathcal{A}}^{\mathcal{C}}$:

- a non-parametric function: each image of $\mathcal{A}$ is described with an image encoder that project it into an euclidean space. It then attributes a label to an image according to its nearest neighbors (1-NN or k-NN)
- a parametric function, such as a neural network, that is learnt on $\mathcal{A}$, with labels in $\mathcal{C}$, and a cross-entropy loss.

G.4 Robustness Results

Results of robustness for all the methods introduced in Table 1 are reported in Table 6, in terms of all three metrics. In this table, MiMA and MaMA show very similar results. This is not surprising. Their formulation implies different behavior under heterigeneous number of images per identity. To see their specific behavior, we reported robustness values on the IMDb dataset in Figure 16.

In this Figure, we compare REGI-Face against approaches based on skin color (CASCo, Beyond Skin Tone (Thong et al., 2023), and AI-Face hybrid annotation on IMDb (Lin et al., 2025)). Such a comparison between different taxonomies is not possible with traditional performance metrics, such as Accuracy, and is one of the strongest motivations for using our robustness metrics. While a perfect

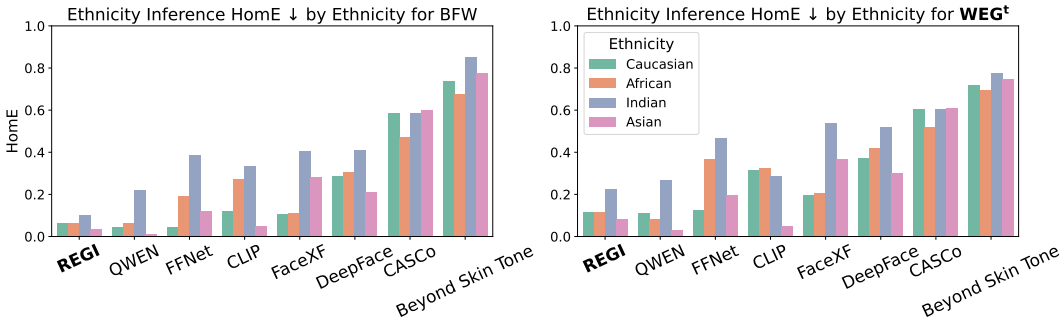


Figure 13: Per ethnicity segment robustness. We compare REGI with the baselines of Table 1 alongside two skin color based approaches.

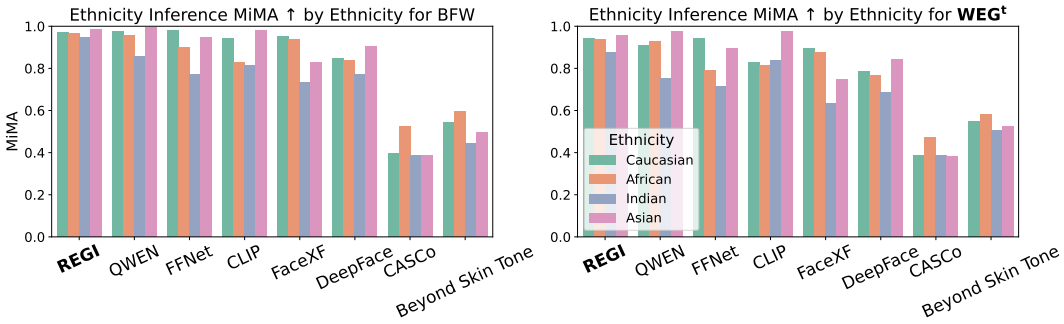


Figure 14: Per ethnicity segment robustness. We compare REGI with the baselines of Table 1 alongside two skin color based approaches.

robustness score is not sufficient to assess that an annotator is perfect, **poor robustness metrics imply an annotation scheme of poor quality**. The application of these robustness analyses on skin tone-based approaches (Figure 15, Figure 14, Figure 13, and Figure 16) allows us to conclude that demographic schemes based on skin color are unfortunately of poor quality compared to classic 4-ethnicity classification. This is especially highlighted by the comparison between REGI Face and the hybrid annotation performed by (Lin et al., 2025) in Figure 16.

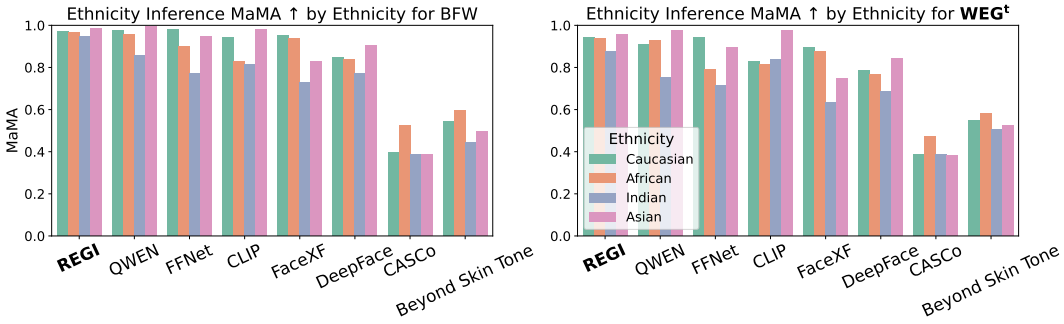


Figure 15: Per ethnicity segment robustness. We compare REGI with the baselines of Table 1 alongside two skin color based approaches.

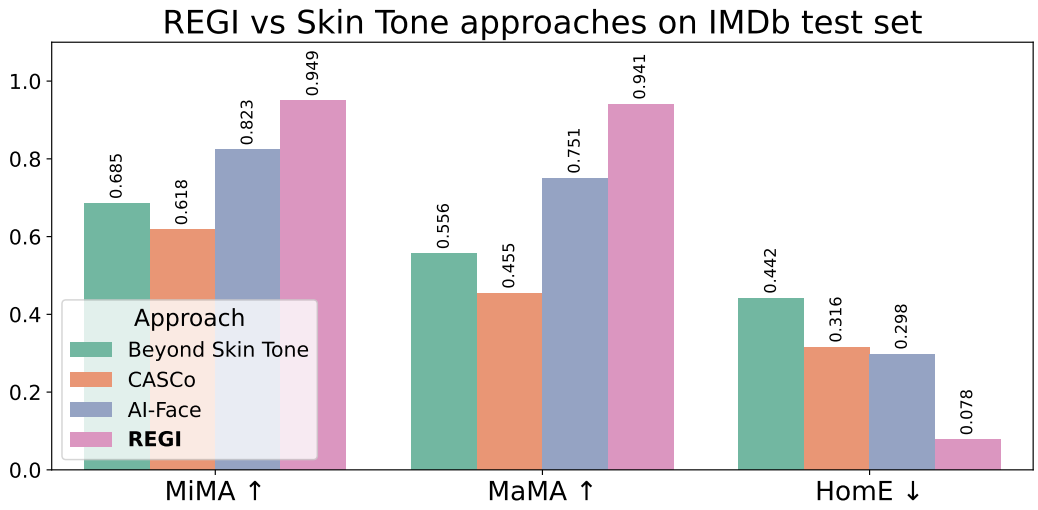


Figure 16: Robustness comparison between REGI and Skin Colors base approaches : Beyond Skin Tone (ICCV 2023), CASCo (Social Science Quarterly 2023) and the hybrid annotation of AI-Face (2025)

	Arch.	#GFLOP	Ethnicity						Gender					
			MaMA ↑		MiMA ↑		HomE ↓		MaMA ↑		MiMA ↑		HomE ↓	
			BFW	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*
FFNet (Kärkkäinen & Joo, 2021)	RN34	1.94	89.97	83.51	90.00	83.51	0.19	0.29	96.60	91.97	96.60	91.97	0.13	0.28
DeepFace (Serengil & Özpınar, 2024)	VGG16	31.06	83.15	79.12	83.16	79.12	0.30	0.37	90.00	89.41	90.01	89.41	0.31	0.33
FaceXF. (Narayan et al., 2025a)	Swin-B	30.92	86.18	78.93	86.19	78.93	0.23	0.33	95.64	90.50	95.63	90.50	0.12	0.21
CLIP (Radford et al., 2021)	ViT-B	8.74	88.29	78.88	88.34	78.88	0.21	0.36	82.93	84.90	82.99	84.90	0.44	0.42
CLIP (Radford et al., 2021)	ViT-L	155.4	89.17	87.88	89.16	88.13	0.20	0.21	86.29	86.32	86.34	86.32	0.39	0.39
QWEN3 (Bai et al., 2025)	4B	1101	94.60	89.44	94.59	89.44	0.11	0.19	98.22	**95.98**	98.22	**95.98**	0.07	**0.15**
QWEN3 (Bai et al., 2025)	8B	8767	91.54	89.93	91.58	89.93	0.16	0.18	**99.08**	**95.98**	**99.04**	**95.98**	**0.05**	**0.15**
REGI	RN18	0.98	94.48	88.34	94.48	88.34	0.11	0.22	97.25	92.42	97.25	92.42	0.11	0.26
REGI	RN34	1.94	92.85	85.09	92.85	85.09	0.15	0.27	96.11	90.61	96.12	90.61	0.15	0.31
REGI	IR18	5.24	95.31	89.93	95.31	89.93	0.10	0.19	97.61	93.85	97.62	93.85	0.10	0.22
REGI	RN50	2.18	96.02	90.60	96.01	90.60	0.08	0.17	98.11	94.17	98.12	94.17	0.08	0.21
REGI	ViT-T	2.76	96.00	90.75	96.01	90.75	0.09	0.17	97.95	94.36	97.97	94.36	0.08	0.20
REGI	RN101	4.04	96.07	91.02	96.06	91.02	0.08	0.17	98.21	94.39	98.22	94.39	0.08	0.20
REGI	IR50	12.64	**96.68**	**92.68**	**96.66**	**92.68**	**0.07**	**0.13**	98.18	94.84	98.18	94.84	0.07	0.19
REGI	ViT-S	11.00	96.63	91.69	96.63	91.69	**0.07**	0.15	98.17	94.80	98.18	94.80	0.08	0.19

Table 6: Comparison of ethnicity and gender inference based on MaMA, MiMA, and HomE dimensions. We evaluate REGI with different backbones and a fixed pretraining dataset (1.4M images), our diversified dataset (WEG^{tr}), and a classification head (SVM-RBF) against competitive baselines. Note that QWEN3's architecture column reflects parameter count (4B/8B) rather than a standard architecture acronym.

H Additional Abalation

H.1 Backbone architecture

Table 7 reports REGI with all tested backbones, from iResNet18 (IR18) (Deng et al., 2019) to ResNet101 (RN101) (He et al., 2016), as well as ViT-T and ViT-S. Performance differences between compact and larger backbones remain modest; we recommend RN50 or ViT-T for constrained environments and IR50 for optimal performance.

H.2 Impact of the Demographic Training set structure

We perform an ablation on the training set structure fed to the SVM with RBF kernel in Table 8. The results clearly show that while the amount of data indeed affects REGI accuracy, ethnicity

	Arch.	#GFLOP	Ethnicity								Gender					
			Accuracy ↑			Fairness ↓			Robust. ↓		Accuracy ↑		Fairness ↓		Robust. ↓	
			BFW	RFW	WEG^t	BFW	RFW	WEG^t	BFW	WEG^t	BFW	WEG^t	BFW	WEG^t	BFW	WEG^t
REGI	RN18	0.98	94.0	93.4	86.2	1.34	1.93	2.93	0.11	0.22	96.9	91.9	0.86	0.82	0.11	0.26
REGI	RN34	1.94	92.3	91.3	82.8	1.48	2.53	3.29	0.15	0.27	95.9	89.8	1.15	3.02	0.15	0.31
REGI	IR18	5.24	94.6	94.5	88.2	1.13	1.52	**2.51**	0.10	0.19	97.3	93.4	0.95	0.44	0.10	0.22
REGI	RN50	2.18	95.5	95.0	88.6	1.01	1.43	2.82	0.08	0.17	97.8	93.7	0.32	1.33	0.08	0.21
REGI	ViT-T	2.76	95.5	94.7	88.9	0.89	1.39	2.72	0.09	0.17	97.6	93.9	1.00	0.88	0.08	0.20
REGI	RN101	4.04	95.4	95.0	88.9	1.08	1.37	2.84	0.08	0.17	**97.9**	94.1	0.74	1.02	0.08	0.20
REGI	IR50	12.64	95.9	**95.8**	**90.8**	0.95	1.22	2.52	**0.07**	**0.13**	**97.9**	**94.7**	1.14	**0.11**	**0.07**	**0.19**
REGI	ViT-S	11.00	**96.2**	95.3	89.9	**0.82**	**1.20**	2.68	**0.07**	0.15	**97.9**	94.5	**0.31**	1.74	0.08	**0.19**

Table 7: Ablation on the backbone architecture of REGI (fixed 1.4M-image pretraining, WEG^{tr}, RBF-SVM head); baselines are reported in Table 1. Best results among REGI backbones in bold, second-best underlined.

	1	**2**	**5**	**10**	**15**	**20**	**40**
500	79.0	80.9	82.2	82.9	83.1	83.2	83.5
1000	82.0	83.1	83.8	84.2	84.5	84.6	85.3
2000	83.5	84.2	84.8	85.6	87.1	87.6	88.8
5000	84.8	85.1	88.5	89.9	90.2	90.4	—
10000	85.6	88.4	90.5	91.0	—	—	—
20000	88.7	90.5	91.1	—	—	—	—

Table 8: Impact of the structure of the ethnicity classification dataset. We vary the number of identities (rows) and the number of samples per identity (columns). 40 is the maximum number of samples per identity. We limit the total dataset size to 100k. We report the WEG^t accuracy obtained with the IR50 backbone pretrained on 1M WebFace images.

accuracy remains above several baselines, even under a scarce-data regime. Moreover, Table 8 shows that, when seeking more data, increasing the number of identities in the training set yields greater improvement than increasing the number of images per identity. This conclusion is similar to the one in Section H.5

H.3 Backbone Training impact on REGI

We further explore the impact of the face recognition backbone pretraining on REGI-Face.

More specifically, we plot the impact of the number of identities and images of the FR training set on the accuracy in Figure 17 and Figure 18. The figures show that, on both gender and ethnicity prediction tasks, a larger pretraining set yields better results, providing a **strong incentive for using the REGI approach**. While demographic annotations are limited and hard to obtain, face recognition data is abundant. These conclusions must, however, be tempered: The plot clearly shows a plateau effect. With the x scale in log2, doubling the dataset pretraining size indeed improves the final accuracy, but the benefits decrease as the training set size increases.

We extend the analysis of the backbone pretraining by exploring in Table 9 the impact of the balancing of the pretraining set. Surprisingly, a balanced training set seems to improve REGI-Face performance across all architectures.

Finally, to confirm the claim in Section 4.2 on the compatibility of RBF-SVM with angular losses, we inspect the results on several FR angular loss, such as AdaFace, ArcFace, CurricularFace, and MagFace in Table 10. Confirming our claim, the performance of AdaFace, CurricularFace, and ArcFace is similar. The results on MagFace, which is farther from our theoretical hypothesis, tell a different story. On WEG^t, the accuracy drops by almost 3 points, demonstrating that using a loss that closely matches our framework is important.

Arch.	BFW		RFW		WEG^t	
	Imb	Bal	Imb	Bal	Imbb	Bal
IR- 18	94.6	**94.8**	94.5	**94.6**	88.2	**88.3**
IR- 50	95.8	**95.9**	95.3	**95.8**	89.9	**90.8**
RN101	94.8	**95.4**	94.4	**95.0**	87.9	**88.9**
RN34	91.0	**92.3**	89.8	**91.3**	80.9	**82.8**
RN50	94.6	**95.5**	94.4	**95.0**	87.9	**88.6**
ViT- s	95.3	**96.2**	94.9	**95.3**	88.9	**89.9**
ViT- t	94.9	**95.5**	94.2	**94.7**	88.4	**88.9**

Table 9: BFW, RFW, and WEG^t accuracy when pretraining the IR50 backbone with Imbalanced (Imb) or balanced (Bal) WebFace subset comprising 1M images.

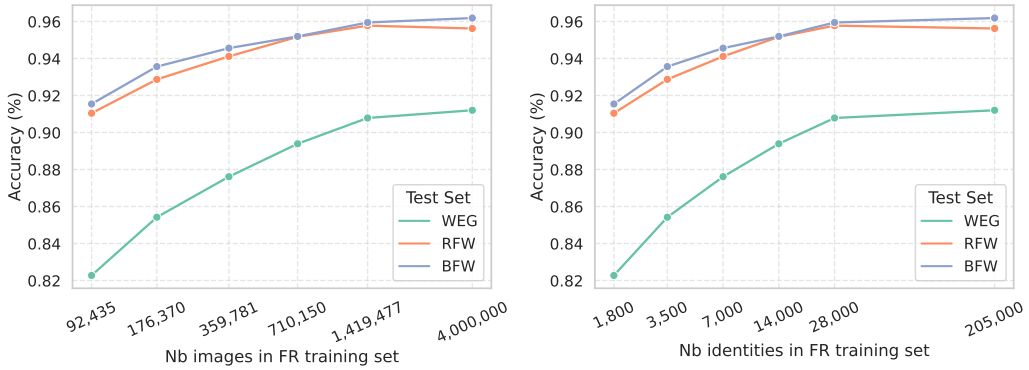


Figure 17: Impact of the size of the backbone's training set on REGI Face ethnicity accuracy. Results are plotted using an IR50 training on balanced subsets of WebFace. The predictor is following the configuration of REGI WEG^{tr} (except for the 4M checkpoint that is coming from AdaFace (Kim et al., 2022) repository)

H.4 Complementary results on Gender

We report in Table 11 the analysis on the impact of our training set WEG^{tr} and the SVM-RBFW head on the gender inference task. As stated in the previous section, we also report the impact of the number of identities in the pretraining set in Figure 18. We thus complete the results reported in Table 3a and Table 2. These results draw the same conclusion as the results on the ethnicity prediction task, confirming the superiority of REGI-Face, both in terms of the framework and the training set, over baselines.

H.5 Marginal Effects Analysis

To rigorously quantify the impact of training configuration choices on DAI accuracy performance, we conduct marginal effects analysis using ordinary least squares (OLS) regression. This approach allows us to estimate the average causal effect of each variable while controlling for confounders, providing a more robust understanding than simple bivariate comparisons.

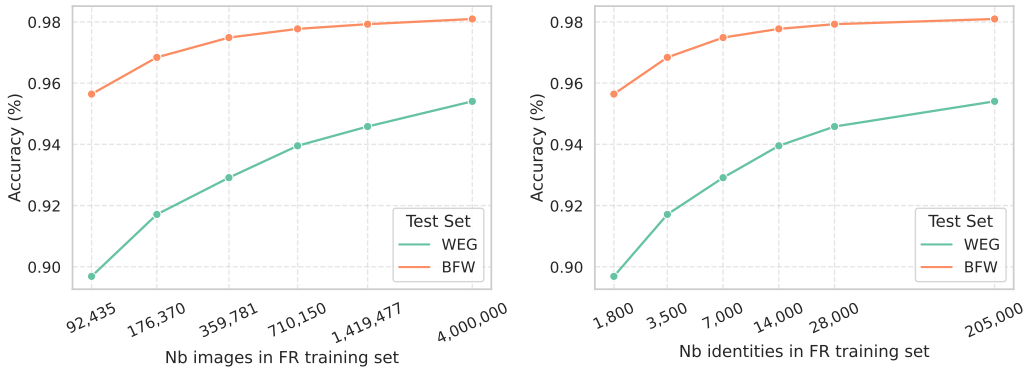


Figure 18: Impact of the size of the backbone's training set on REGI Face gender accuracy. Results are plotted using an IR50 training on balanced subsets of WebFace. The predictor is following the configuration of REGI WEG^{tr} (except for the 4M checkpoint that is coming from AdaFace (Kim et al., 2022) repository)

Loss	Ethnicity								Gender					
	Accuracy ↑			Fairness ↓			Robust. ↓		Accuracy ↑		Fairness ↓		Robust. ↓	
	BFW	*RFW*	*WEG^t*	*BFW*	*RFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*	*BFW*	*WEG^t*
AdaFace	95.94	95.77	90.78	0.95	1.22	2.52	0.07	0.13	97.92	94.65	1.14	0.11	0.07	0.19
ArcFace	96.37	95.80	90.96	0.97	1.17	2.51	0.06	0.13	98.20	95.03	0.61	1.13	0.06	0.17
CurricularFace	95.83	95.52	90.35	1.02	1.27	2.35	0.07	0.14	97.96	95.12	0.73	0.52	0.07	0.17
MagFace	94.87	94.37	87.64	1.05	1.56	2.55	0.10	0.19	97.32	92.98	0.77	1.01	0.10	0.23

Table 10: Comparison of different losses based on ethnicity and gender dimensions (accuracy, fairness, robustness).

H.5.1 METHODOLOGY

Model Specification We specify two related regression models to analyze different aspects of our experimental results:

Model 1: Best Configuration Analysis For the first analysis, examining the best performing configuration per architecture, we model the target DAI accuracy metric y_i (best RFW validation accuracy at optimal epoch) as:

$$y_i = \beta_0 + \beta_1 \log(\text{IDs}_i) + \beta_2 \log(\text{Samples}_i) + \\ \beta_3 \text{FV}_i + \beta_4 \text{Unbal}_i + \epsilon_i$$

where:

- IDs_i is the number of identities in the training set
- Samples_i is the number of samples per identity
- FV_i is the face verification accuracy of the base model
- Unbal_i is a binary indicator for unbalanced training (vs. balanced baseline)
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term

Model 2: Per-Epoch Analysis For the second analysis examining performance across training epochs for specific evaluation sets, we extend the model to:

	WEG^{t}				BFW			
Training	**FC**	**LIN**	**MLP**	**RBF**	**FC**	**LIN**	**MLP**	**RBF**
FairFace	85.5	87.6	87.1	*89.2*	91.4	92.7	93.4	*93.6*
BUPT	87.7	88.7	92.7	*93.6*	94.5	94.7	97.3	*97.8*
WEG^{tr}	**88.9**	**89.5**	**94.3**	***94.7***	**94.9**	**95.3**	**97.6**	***97.9***

Table 11: Impact of the classification head and of the gender training dataset. *Best results for the classification head*. **Best results for the gender training dataset.**

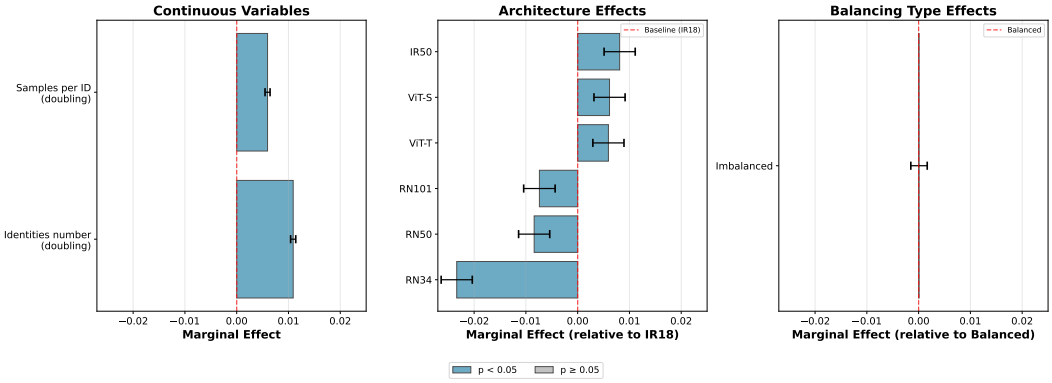


Figure 19: Marginal effects for OLS Model 2 on RFW

$$y_{iet} = \beta_0 + \beta_1 \log(\text{IDs}_i) + \beta_2 \log(\text{Samples}_i) + \beta_3 f(\text{Epoch}_t) + \beta_4 C_i + \beta_5 \text{Unbal}_i + \epsilon_{iet}$$

where y_{iet} is the accuracy for configuration i, evaluation set e, at epoch t, and C_i represents either:

- A continuous control variable (FV accuracy on a specific benchmark), or
- A set of architecture dummy variables: $\sum_{k=2}^{K} \gamma_k \mathbb{1}(\text{Arch}_i = k)$

The function $f(\cdot)$ applies logarithmic transformation if epochs span more than one order of magnitude, otherwise uses linear scaling.

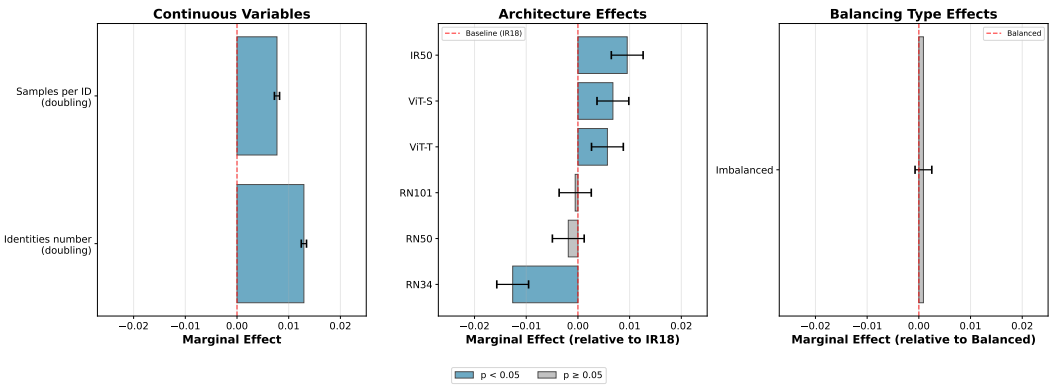


Figure 20: Marginal effects for OLS Model 2 on BFW

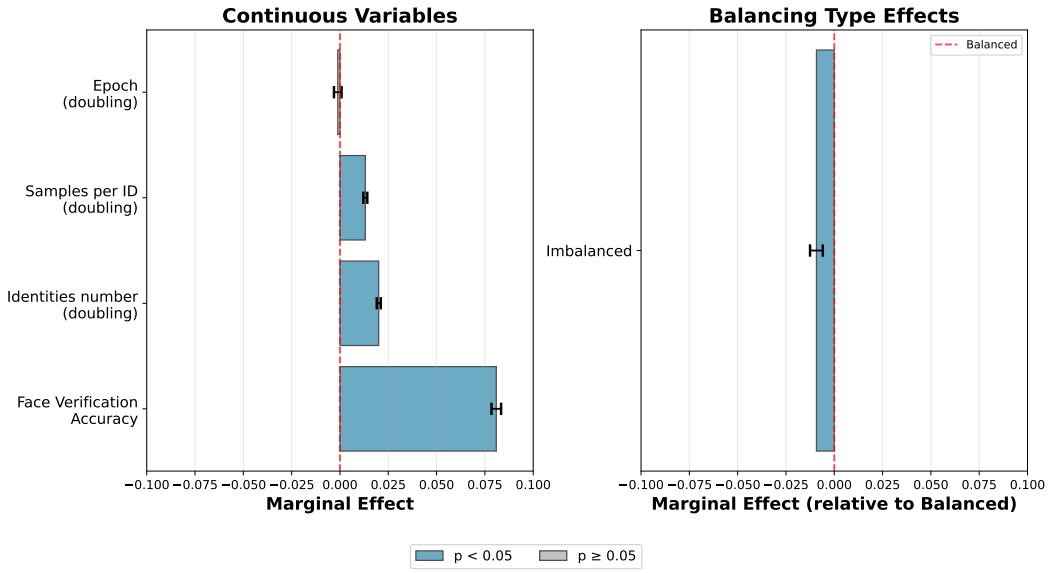


Figure 21: Marginal effects for OLS Model 1 on RFW

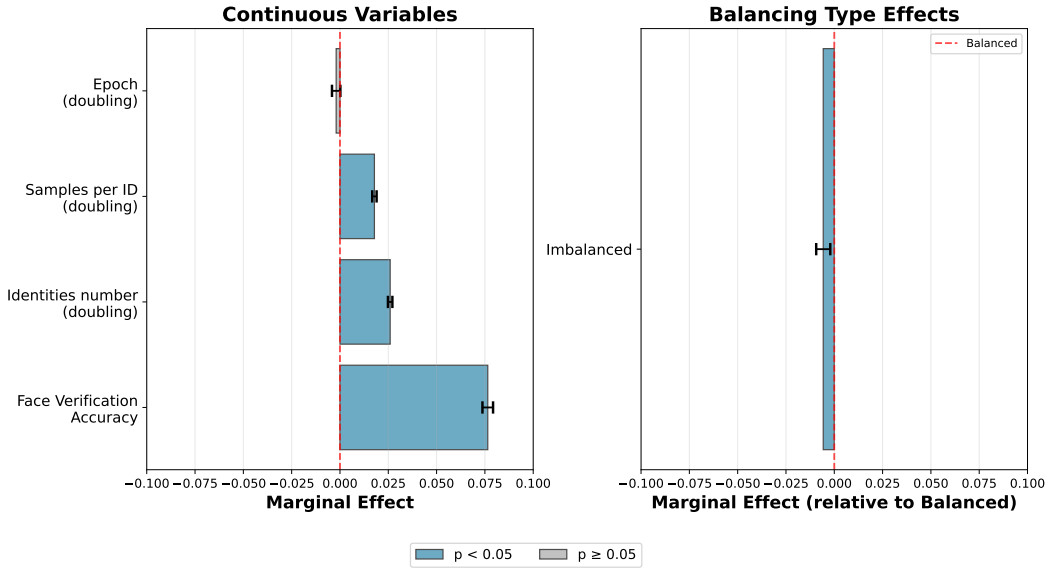


Figure 22: Marginal effects for OLS Model 1 on BFW

Variable Transformations To ensure proper model specification and interpretability, we apply the following transformations:

Log Transformations Variables spanning multiple orders of magnitude (number of training identities: 500–20,000; samples per identity: 1–40) are log-transformed. This addresses two concerns:

1. **Heteroskedasticity**: Raw values exhibit increasing variance with scale
2. **Interpretability**: Coefficients represent *elasticities* (percentage changes) rather than absolute changes

For a log-transformed variable, the marginal effect of doubling is computed as:

$$\Delta y_{\text{double}} = \beta_{\log(x)} \times \log(2) \approx 0.693 \beta_{\log(x)} \tag{40}$$

Categorical Variables Categorical variables (architecture, balancing type) are dummy-coded with one category as the baseline. The coefficient γ_k represents the average difference in outcome between category k and the baseline, holding all other variables constant.

Why Control for Face Verification Accuracy? A critical methodological consideration is whether to include architecture as a categorical variable or control for face verification (FV) accuracy. We observe strong collinearity between architecture and FV performance , as different architectures inherently achieve different base recognition capabilities.

This creates two valid but distinct research questions:

- **Controlling for architecture**: "Which training configurations work best for each architecture?"
- **Controlling for FV accuracy**: "Given a certain base recognition quality, how do training choices affect DAI accuracy?"

We essentially consider Model 1 for concluding:

1. Architecture and FV accuracy measure fundamentally the same construct (model capability)
2. Including both would introduce multicollinearity, inflating standard errors
3. FV accuracy provides a continuous, more granular measure of model quality
4. The coefficient on FV accuracy has a natural interpretation: DAI accuracy as a function of base recognition capability

For Model 2, we allow flexibility to examine either control strategy depending on the research question.

H.5.2 Results

Model 1: Optimal Configuration Analysis Figure 21 and Figure 22 present the marginal effects from Model 1. The regression achieves $R^2 = 0.856$ ($R^2_{\text{adj}} = 0.833$), indicating excellent model fit.

Training Set Size Doubling the number of training identities yields between **+1.21 and 2.25 percentage points** increase in DAI accuracy. This substantial effect demonstrates that DAI performance scales predictably with dataset diversity, consistent with prior work showing that representation matters for bias mitigation (Buolamwini & Gebru, 2018).

Samples per Identity Doubling samples per identity increases accuracy by **+0.64-1 percentage points**. While significant, this effect is roughly half that of adding identities, suggesting that *diversity* (more identities) is more valuable than *depth* (more samples per identity) for DAI accuracy.

Base Model Quality A 10-point increase in base FV accuracy corresponds to more than **+7.5 percentage point** improvement in DAI performance ($p < 0.001$). This is the strongest predictor in the model, indicating that DAI performance is highly sensitive to base recognition quality.

This finding has important implications: improving the underlying face verification model (through better architectures, pretraining, or optimization) yields larger DAI accuracy gains than dataset manipulation alone.

Balancing Strategy Unbalanced training shows a significant **negative** effect (between -0.5and -1 percentage points) compared to balanced training.

Model 2: Per-Epoch Analysis With the Model 2 Analysis, we confirm the performance results for the different backbones reported in Table 1. We also observe the same behavior for the impact of the number of identities and the number of samples per identity as with Model 1. Interestingly, though, the analysis on the balancing of the pretraining face recognition training set shows a non-significant effect, which contrasts with Model 1.

H.5.3 Marginal effects Conclusion

Marginal effects analysis reveals that the backbone choice (FV accuracy/architecture) is the strongest predictor of annotator performance, followed by training set diversity (number of identities) and depth (samples per identity).