Rawafed Almarefa Journal مجلة روافد المعرفة
ISSN: 2709-0345 Volume: 7, Issue: 13, (2026), Pages: 133-154

Facial Attribute Mismatch in Face Verification: A Protocol-Controlled Study of Locally Trained ArcFace and CosFace on CelebA

Maamar Abdulsalam Alhaaj Meelad [1*] . Fathi Hajjaj [2] . Asmail Muftah [3*]

[1,2,3] Department of Computer Science, Faculty of Science, Azzytuna University, Libya.
fathihajjaj88@gmail.com; asmail.muftah@azu.edu.ly

Corresponding Author:(*)

Received: 03 – 11 -2026; Accepted: 07 - 11-2026; Published: 28- 11-2026

Abstract

Overall, face-verification metrics can conceal differences in genuine-pair acceptance associated with facial appearance. We examined within-identity attribute mismatch at fixed operating thresholds, distinguishing endpoint disagreement from attribute presence and assessing sensitivity to pair construction. Using CelebA's official identity-disjoint partitions, we trained two local iResNet18-style models with ArcFace and CosFace losses. Model-specific thresholds calibrated on Validation at a target false accept rate (FAR) of 1% were fixed before Test evaluation. Genuine acceptance was compared across six annotation-defined attributes using an approximate identity-node/dyadic bootstrap and a predefined 30-contrast Bonferroni adjustment for primary comparisons. Under the primary profile-matched impostor protocol, true accept rates (TARs) were 87.19% for ArcFace and 85.71% for CosFace; empirical FARs were 1.084% and 1.065%, respectively. Relative to pairs agreeing on all six attributes, any-mismatch pairs had TAR reductions of 9.05 and 9.83 percentage points, with approximate adjusted confidence intervals [5.77, 12.50] and [6.45, 13.50]. Adjusted intervals supported lower acceptance for eyeglasses, wearing a hat, heavy makeup, and bangs; mustache and sideburns remained inconclusive. Alternative impostor sampling changed the FAR without changing genuine acceptance, and duplicate-screened sensitivity retained the any-mismatch association. These observational comparisons show why attribute-specific acceptance and explicit evaluation protocols should accompany aggregate metrics. One training run per loss, annotation uncertainty, dependent pairs, and known cross-partition duplicate content limit causal interpretation and generalization to loss families.

Keywords: Face verification, facial attribute mismatch, CelebA, ArcFace, CosFace, and evaluation protocols.

عدم تطابق سمات الوجه في التحقق من الهوية: دراسة ببروتوكول تقييم مضبوط لنموذجيArcFace و CosFace المدرَّبين محليًا على CelebA

الملخص

قد تخفي المقاييس الإجمالية للتحقق من هوية الوجه تفاوتًا في قبول الأزواج الحقيقية يرتبط بمظهر الوجه. دُرس عدم تطابق السمات بين صور الشخص نفسه عند عتبات ثابتة، مع التمييز بين اختلاف السمات ووجودها، وفُحصت الحساسية لطريقة بناء الأزواج. استخدمنا مجموعة بيانات CelebA بأقسامها الرسمية المنفصلة من حيث الهويات، ودُرِّب نموذجان محليًا ببنية مستندة إلى iResNet18، باستخدام دالتَي الخسارة ArcFace و CosFace. عُيِّرت عتبة كل نموذج على مجموعة التحقق عند معدل قبول خاطئ مستهدف قدره 1%، وثُبِّتت قبل تقييم الاختبار. قورن قبول الأزواج الحقيقية عبر ست سمات موسومة باستخدام بوتستراب تقريبي للهويات وعلاقاتها الزوجية، مع تعديل بونفيروني لعائلة محددة مسبقًا من 30 مقارنة أساسية. في البروتوكول الأساسي، الذي يطابق ملف سمات صورة الاستعلام عند اختيار أزواج الهويات المختلفة، بلغ معدل القبول الحقيقي 87.19% لنموذج

This is an open access article under the CC BY-SA license.

ArcFaceو85.71% لنموذجCosFace ، وبلغ معدل القبول الخاطئ التجريبي 1.084% و1.065% على الترتيب. مقارنةً بالأزواج المتطابقة في السمات الست، انخفض قبول الأزواج المختلفة في سمة واحدة على الأقل بمقدار 9.05 و9.83 نقطة مئوية، بفترتَي ثقة تقريبيتين معدَّلتين [5.77, 12.50] و [6.45, 13.50]دعمت الفترات المعدَّلة انخفاض القبول لسمات ارتداء النظارات وارتداء القبعة والمكياج الكثيف والغُرّة، وظلت نتائج الشارب والسوالف غير حاسمة. غيّر الاختيار البديل لأزواج الهويات المختلفة معدل القبول الخاطئ دون أن يغيّر قبول الأزواج الحقيقية، واستمر ارتباط عدم التطابق بانخفاض القبول بعد استبعاد الصور المكررة. تبيّن هذه المقارنات الرصدية ضرورة إرفاق المقاييس الإجمالية بالقبول حسب السمات وببروتوكولات تقييم صريحة. ومن القيود التي تحدّ من التفسير السببي ومن التعميم على عائلات دوال الخسارة: إجراء تجربة تدريب واحدة لكل دالة خسارة، وعدم اليقين في توصيف السمات، وعدم استقلالية الأزواج، ووجود محتوى مكرر بين الأقسام.

الكلمات المفتاحية: *التحقق من هوية الوجه؛ عدم تطابق سمات الوجه CelebA؛ ArcFace؛ CosFace؛ بروتوكولات التقييم*.

1. Introduction

Face verification requires a system to accept different photographs of the same person while rejecting comparisons between people. Changes in visible appearance matter, even when identity remains unchanged. An aggregate verification score summarizes the sampled comparisons but can conceal differences in genuine acceptance between appearance conditions. Attribute-based representations and analyses beyond demographic categories have established the relevance of facial attributes to recognition (Kumar et al., 2009; Terhörst et al., 2022). The remaining question is how to characterize this dependence for a specified set of image pairs and operating thresholds.

Natural eyeglass status and facial hair differences have already been examined in verification experiments (Guo et al., 2018; Wu, Bezold, Bhatta, et al., 2023). Here, we study six CelebA annotations: eyeglasses, heavy _ makeup, wearing _ hat, bangs, mustache, and sideburns (Liu et al., 2015). A genuine pair is classified as mismatched when its two photographs have unequal annotations for a selected attribute. This comparison differs from asking whether an attribute is present: both the agreement between endpoints and the images entering each group determine what is measured. Unequal labels alone establish neither a chronological appearance change nor an isolated physical intervention.

The operating rule and comparison population are equally important for interpretation. Non-match population composition can influence recognition estimates and calibration, while challenging natural genuine and impostor comparisons expose weaknesses that easier test sets may conceal (O'Toole et al., 2012; Wu et al., 2026). Consequently, a mismatch analysis should report the threshold used to accept a genuine pair and the impostor population against which false acceptance is assessed. Keeping genuine pairs and thresholds fixed when changing impostors allows these two sources of variation to be distinguished.

We addressed this measurement problem using two locally trained iResNet18-style baselines with ArcFace and CosFace objectives, a shared training budget, and common evaluation pairs (Deng et al., 2019; Wang et al., 2018). Model-specific thresholds calibrated on the validation are transported unchanged to the test. Within a common CelebA experiment, we compared attribute-specific and mismatch-count acceptance at fixed operating points and assessed sensitivity to impostor construction and predefined duplicate screening.

Four questions organize the frozen comparisons: whether any mismatch is associated with lower genuine acceptance (RQ1); how associations differ by attribute and mismatch count (RQ2); whether the fitted ArcFace and CosFace instances differ in overall acceptance or mismatch reductions (RQ3); and how impostor sampling and duplicate screening change the evaluation (RQ4). Exploratory Validation informed calibration and the local plan, which was frozen before Test scoring. These questions concern observational, pair-weighted associations. Annotation uncertainty, uncontrolled acquisition differences, dependent pairs, one

This is an open access article under the CC BY-SA license.

training realization per objective, and known cross-partition image-content overlap limit generalization beyond these comparisons (Wu, Bezold, Günther, et al., 2023).

2. Related Work

Earlier attribute and simile classifiers used facial properties as representations for verification (Kumar et al., 2009), while CelebA made identity-labeled, attribute-annotated imagery available for appearance research (Liu et al., 2015). More recent studies have examined attributes in recognition systems. Terhörst et al. (2022) analyzed 47 attributes using attribute-presence strata and controls; Leroy et al. (2025) examined attribute structure within and between identities and invariance in synthetic experiments. Together, these studies show that attribute dependence can be studied through both verification outcomes and embedding geometry. Our outcome is genuine acceptance for endpoint-agreement groups at a fixed threshold; therefore, neither presence-stratum errors nor embedding measures can be interchangeable with the contrasts reported here. ArcFace and CosFace supply established training objectives for this evaluation (Deng et al., 2019; Wang et al., 2018); their original benchmark scores are not reference values for the local models and protocol used here.

Natural appearance studies provide closer comparisons. Guo et al. (2018) distinguished matched and differing eyeglass-status configurations using real MeGlass test images, with synthetic eyeglass augmentation used in training. Before-and-after cosmetic databases have also been used to examine recognition across makeup application (Dantcheva et al., 2012). Facial-hair studies distinguish richer similarity and difference categories in genuine and impostor pairs and examine training-data balance (Ozturk et al., 2024, 2025; Wu et al., 2024; Wu, Bezold, Bhatta, et al., 2023). These designs support the study of natural endpoint differences, but their attribute definitions and acquisition protocols constrain comparison. CelebA's Heavy_Makeup label does not establish a before-and-after acquisition, and Mustache and Sideburns do not reproduce detailed beard-area categories.

Synthetic manipulation offers a different way to study attribute sensitivity. Banerjee et al. (2023) included reconstruction controls when evaluating GAN-induced edits because changes in a recognition representation can accompany reconstruction as well as the intended edit. Comparing distinct natural photographs avoids the editing pipeline but allows pose, illumination, image quality, and other acquisition differences to vary. Therefore, the two designs address different sources of uncertainty; the effect magnitude measured in edited images cannot be assumed to describe naturally occurring mismatch pairs.

Evaluation-population studies also shaped the present design. O'Toole et al. (2012) showed why the non-match population matters for recognition estimates and operating thresholds. Goldilocks test sets emphasize natural facial-hair or exposure differences in genuine pairs together with challenging impostors (Wu et al., 2026). Their curated comparisons and fold-specific evaluation differ from our score-blind manifests and thresholds transported from a separate Validation partition. P1 conditions an impostor probe on the reference genuine probe's six-attribute profile; P2 changes the impostors while retaining the genuine comparisons. This construction permits protocol sensitivity to be examined without conflating it with a change in fixed-threshold genuine acceptance.

Annotation and content provenance further limit these comparisons. Inconsistent CelebA attribute values, including contradictions among duplicate faces, weaken the literal interpretation of unequal labels as visible changes (Wu, Bezold, Günther, et al., 2023). Work on duplicate detection likewise motivates the checking of image content rather than relying on identity labels alone (Schlett et al., 2024). Neither publication supplies an error or leakage estimate for our models and annotations. Our duplicate-screened view tests the persistence

This is an open access article under the CC BY-SA license.

of an association under a defined exclusion policy, with changed image availability and pair composition; it cannot identify the effect of contamination alone.

3. Materials and Methods

3.1. Study Design, Data, and Preprocessing

We compared genuine-pair acceptance between annotation-defined appearance groups using two fixed trained models and a predefined set of image pairs. Each pair contributed equally to the group rates; the comparisons were observational and did not isolate causal effects of individual attributes. Exploratory Validation informed threshold calibration and the study protocol, including an amendment finalized before Test scoring. The protocol was specified locally rather than publicly preregistered. The unfiltered Official Test view was primary, and the duplicate-screened Test was secondary. The predefined 30-contrast family covered the primary P1/Official Test comparisons addressing RQ1–RQ3. The protocol and duplicate-screening analyses under RQ4 were secondary and not included in this family.

We used the official aligned-PNG CelebA with its supplied identity, attribute, and evaluation-partition annotations (Liu et al., 2015). Metadata filenames ending in.jpg were mapped to.png files using their six-digit stems. Parsing checked the agreement between filename sets, unique filenames, valid partition labels, and −1/+1 attribute values. The experiments used the official partitions. Identity labels and filenames were pairwise disjoint across partitions, although the image content was not fully independent. Classification labels were assigned only to the Train identities. Official Train, Validation, and Test contained 162,770, 19,867, and 19,962 images from 8,192, 985, and 1,000 identities, respectively. The screened view retained 19,934 Test images and all 1,000 identities.

The attributes were eyeglasses, heavy _ makeup, wearing _ hat, bangs, mustache, and sideburns. Presence was +1 and absence was −1. The six annotations span eyewear, headwear, cosmetics, scalp hair, and facial hair. They constitute a focused subset of CelebA's binary attributes rather than an exhaustive representation of facial appearance variation. Annotation disagreement does not necessarily identify physical change (Wu, Bezold, Günther, et al., 2023).

The images were 178×218 RGB PNGs. Images were converted to RGB and resized directly to 112×112 using Pillow bilinear interpolation, changing their aspect ratio. No additional cropping, landmark alignment, EXIF transposition, or ICC transformation was applied. Channel-first tensors used Equation 1 for unsigned 8-bit channel value p. Training added horizontal reflection with a probability of 0.5. Evaluation used no augmentation, cohort normalization, or ensembling. Aspect-ratio distortion and the absence of additional landmark alignment limit comparability with standard face-recognition pipelines.

$$x = \frac{\frac{p}{255} - 0.5}{0.5} \quad (1)$$

3.2. Models and Training

Both local iResNet18-style backbones were trained from scratch and used a 3×3, 64-channel convolutional stem with batch normalization/PReLU, followed by stages of widths 64,128,256,512 with two blocks each. First blocks downsampled by two. The residual branches used BN–3×3 convolution–BN/PReLU–3×3 convolution–BN; required shortcuts used stride-matched 1×1 convolution/BN. The final 512×7×7 activation passed through BN, flattening, a 512-dimensional linear layer, and BN. There was no dropout. Convolutional weights used Kaiming-normal initialization, linear weights used normal initialization with SD 0.01 and zero

This is an open access article under the CC BY-SA license.

biases, and classifier weights used Xavier-uniform initialization. These were local implementations; equivalence to official pretrained models was not established.

An 8,192-class margin head was jointly trained with identity cross-entropy. The normalized class cosine c_j followed Equation 2, with $\varepsilon=10^{-12}$ and FP32 arithmetic. CosFace used target logit $64(c_\gamma-0.35)$ and other logits $64c_j$, without post-margin clamping/fallback (Wang et al., 2018). ArcFace used $64\varphi(c_\gamma)$, m=0.5 radians, and Equation 3, including the implemented numerical floor and monotonic fallback (Deng et al., 2019). Other logits were $64c_j$. The cross-entropy was $-z_\gamma+\log \Sigma_j \exp z_j$. No attribute loss or attribute-aware sampler was used. Verification discarded the head.

$$c_j = \operatorname{clip}\left(\frac{W_j^{\top} f}{\max\left(\|W_j\|_2, \varepsilon\right)\max(\|f\|_2, \varepsilon)}, -1, 1\right) \quad (2)$$

$$\varphi(c) = \begin{cases} c\cos m - \sqrt{\max(1 - c^2, 10^{-7})}\sin m, & c > \cos(\pi - m) \\ c - m\sin(\pi - m), & \text{otherwise} \end{cases} \quad (3)$$

All training images were retained, including singleton identities and known duplicate content. Sorted identity labels were mapped to contiguous classification indices. Each epoch visited every image once in a shuffled order, without replacement, omission, or balancing. A dedicated sampler generator used seed 42 plus the zero-based epoch, and the Python, NumPy, and PyTorch random generators used seed 42. Only one training run was completed for each objective.

SGD used a learning rate of 0.025, momentum of 0.9, and weight decay of 0.0005. Gradients were accumulated over two FP32 microbatches of 32; the summed cross-entropy gradients were divided by the actual window size before the combined backbone/head gradient norm was clipped to 5. Ordinary windows had 64 images; final windows had 18. The BN operated on microbatches. Each epoch contained 5,087 microbatches and 2,544 updates. The schedule multiplier in Equation 4 used W=2,500 and T=61,056; learning rate $0.025a(k)$ was set before update k and the scheduler advanced after success. Each run completed 24 epochs, 61,056 updates, and 3,906,480 image presentations. Final epoch-24 checkpoints were used without verification-best selection.

$$a(k) = \begin{cases} \dfrac{k+1}{W}, & k < W \\ \dfrac{1}{2}\left[1 + \cos\left(\pi \operatorname{clip}\left(\dfrac{k-W}{T-W}, 0, 1\right)\right)\right], & k \geq W \end{cases} \quad (4)$$

Deterministic algorithms were requested, cuDNN benchmarking and TF32 were disabled, the CUDA workspace was set to:4096:8, and loaders used zero workers and no pinned memory. Training logs identified CUDA device 0 but did not independently record the physical GPU model; verification logs identified an RTX 2080 Ti. Recorded evaluation versions were Python 3.9.21, PyTorch 2.2.2+cu121, CUDA build 12.1, NumPy 1.26.4, and Pillow 10.2.0. Driver/cuDNN records were incomplete. The historical ArcFace entrypoint hash differed from the current script, although shared architecture/loss/transform/loader/trainer hashes matched. Exact historical entrypoint replay and cross-machine bitwise reproducibility were not established.

3.3. Verification Pairs, Mismatch, and Calibration

Pair construction used metadata and seeds without model scores. Within each partition/view, identities were sorted numerically, and images were sorted lexicographically. Each identity with at least two images used its

This is an open access article under the CC BY-SA license.

first image as the sole enrollment; up to 20 remaining images were sampled without replacement and sorted as genuine probes. Singletons remained eligible only as impostor probes. Enrollment was reproducible and not randomized or quality-balanced. P1's Python generator used string seeds 42:validation or 42:test.

P1 created an impostor slot for each genuine row, retaining the enrollment, and conditioning the replacement probe on the reference genuine probe's joint six-attribute profile. This did not require equal attributes between the impostor endpoints. Pools excluded non-singleton enrollment images and were ordered by identity/filename, followed by lexicographically sorted singletons. A random starting index followed by circular scanning selected an eligible different-identity probe with an unused, unordered pair key. If no exact-profile candidate existed, the same procedure was used for the full eligible pool and marked as a fallback. The algorithm was not uniform across valid pairs/identities. Insufficient eligible test candidates caused the pair construction to fail.

P2 retained P1's genuine rows, their order, and the enrollment/reference slots. For each of the seeds 101,202,303,404,505, the generator selected uniformly from sorted, currently eligible different-identity probe identities, then uniformly selected and removed an unused image from that identity's sorted pool. An exhausted identity was removed for that enrollment, and exhaustion of the full pool caused failure without fallback. Probe use was tracked separately for each enrollment; therefore, images could recur with other enrollments. The resulting procedure was sequential identity-uniform sampling among the remaining candidates, rather than a globally balanced allocation.

Each protocol contained 15,718 genuine and 15,718 impostor pairs in Validation, 15,769 per class in Official Test, and 15,767 per class in the screened Test. P1 used 5, 4, and 3 fallback rows in those populations, respectively.

Within each manifest, self-pairs, same-identity impostors, and duplicate unordered filename pairs were prohibited. However, images and identities could recur in distinct pairs and across protocols. Both models used identical frozen manifests within a view. Because P2 preserved the genuine pairs and thresholds, its genuine acceptance was identical to that of P1 by construction. The separate exact-profile-only sensitivity instead removed fallback impostors and their referenced genuine rows, changing the genuine set.

For selected attribute a, b_{ia} was 1 for +1 and 0 otherwise. Equation 5 defines mismatch and its six-attribute count. Groups were K=0,1,2,≥3; any mismatch meant K≥1. Attribute-matched groups pooled 0→0/1→1 and mismatched groups pooled 0→1/1→0. Arrows indicate enrollment/probe order, not time. Groups overlapped across attributes/identities. No single-change restriction, covariate balancing, or independent manipulation was performed.

$$d_{ij,a} = \mathbb{1}(b_{ia} \neq b_{ja}); \quad K_{ij} = \sum_{a \in \mathcal{A}} d_{ij,a}; \quad |\mathcal{A}| = 6 \quad (5)$$

Embedding extraction used FP32 inference in the evaluation and inference modes, batches of 32, and deterministic image order. Each finite, nonzero 512-dimensional embedding was L2-normalized and cached. Embeddings were extracted once for the Official Test, and the corresponding subset was reused for the screened view. Pair scoring normalized the cached vectors again, calculated their dot product, and clipped it to [−1,1]. Scores were stored as float64; feature extraction remained FP32. Encoded-image hashes were checked before the corresponding bytes were decoded in memory.

For each model and target α, P1 Validation scores were sorted in descending order with exact ties grouped. The lowest empirical ROC threshold satisfying FAR≤α was selected, including a reject-all threshold above

This is an open access article under the CC BY-SA license.

the maximum score if needed. Acceptance was scored ≥threshold; no interpolation, subgroup fitting, or test recalibration was used. At the primary target α=0.01, the thresholds were 0.25424203276634216 for ArcFace and 0.20998413860797882 for CosFace. At exploratory α=0.001, they were 0.33777374029159546 and 0.27491146326065063, respectively. These values were unchanged across groups, protocols, and Test views.

3.4. Statistical Analysis, Duplicate Screening, and Integrity

Equation 6 defines the true accept rate (TAR) for genuine pairs and the false accept rate (FAR) for impostor pairs. The rate differences were expressed in percentage points after multiplication by 100. The receiver operating characteristic area under the curve (ROC-AUC) used trapezoidal integration with exact-tie accumulation and the reject-all origin. The equal error rate (EER) was linearly interpolated at the crossing FAR=1−TAR; it did not determine the fixed operating threshold. Attribute reductions were matched-minus-mismatched TAR, and any-mismatch or count-group reductions used zero mismatch as the reference. Cross-model differences were CosFace minus ArcFace, and protocol differences were P2 minus P1. P2 means, ranges, and population standard deviations (ddof=0) describe sampling sensitivity, not repeated training.

$$\mathrm{TAR}(t) = \frac{\sum_{i \in G} \mathbb{1}(s_i \geq t)}{|G|}; \quad \mathrm{FAR}(t) = \frac{\sum_{i \in I} \mathbb{1}(s_i \geq t)}{|I|} \quad (6)$$

Shared images induce dependence between similarity-score pairs, complicating uncertainty estimation (Conti & Clémençon, 2024). We used an approximate identity-node/dyadic bootstrap conditional on the observed pair network. Within each view, the sorted union of endpoint identities across six protocols contained 1,000 identities. NumPy default_rng seed 2707 generated 20,000 multinomial count vectors c with N trials and probabilities 1/N. The same draws were used across models, protocols, thresholds, and groups. Genuine rows from u had weight c_u; impostor rows connecting u≠v had $c_u c_v$ (Equation 7). Contrasts used these shared draws, and weighted ROC calculations retained exact ties. Thresholds, images, and pair manifests were not refitted/resampled.

$$\hat{R}_C^{(b)}(t) = \frac{\sum_{i \in C} w_i^{(b)} \mathbb{1}(s_i \geq t)}{\sum_{i \in C} w_i^{(b)}}$$
$$w_i^{(b)} = \begin{cases} c_u^{(b)}, & i \in G \\ c_u^{(b)} c_v^{(b)}, & i \in I, \quad u \neq v \end{cases} \quad (7)$$

Pointwise intervals used finite-draw percentiles 2.5/97.5. The primary Official Test/P1 family at α=0.01 comprised 30 contrasts: two any-mismatch, 12 attribute, six count-versus-zero, one overall model, six model attribute-reduction, and three model count-reduction differences. Bonferroni percentiles were 0.05/(2×30) and its complement (nominal marginal coverage 99.8333%). Approximately 17 ordered draws occupied each adjusted tail, limiting precision. Adjustment does not prove simultaneous coverage for approximate bootstrap marginals. Comparisons involving P2, the screened view, directed states, α=0.001, or cross-model AUC/EER were secondary or exploratory. They were not promoted through gatekeeping, and no exact p-values were reported.

Each component of a genuine contrast required ≥30 pairs and ≥10 identities; overall intervals required ≥10 relevant-class identities and nonzero resampled weight. At least 95% of the requested draws had to be finite; otherwise, the intervals were null with a reason and no regrouping. The family size remained 30. These minimum-support requirements do not guarantee statistical power or interval coverage. The low-FAR heuristic was |I|α≥100, with count resolution 1/|I|; dependent pairs are not independent count evidence, and

This is an open access article under the CC BY-SA license.

zero false accepts would not prove zero population FAR. Intervals omitted training variation, calibration uncertainty, new within-identity acquisitions, and nuisance adjustment.

Before Test scoring, all 202,599 PNGs were decoded and screened for duplicates using the hashes of encoded bytes and decoded dimensions, mode, and RGB bytes. No decoding failures or identical encoded-image groups were found. Decoded content identified 131 duplicate groups containing 263 images; 38 groups crossed partitions (24 Train–Validation, 12 Train–Test, 2 Validation–Test), 89 had differing identity labels, and 99 had differing attributes. The labels were not reassigned. This is consistent with the need to distinguish label separation from content separation (Schlett et al., 2024).

Fourteen test images were cross-partitioned members. The locally frozen amendment retained unfiltered Test as primary and excluded all 28 known duplicate-member Test images for sensitivity: those 14 plus 14 members of seven within-Test, single-label groups. Exclusion before enrollment/pair selection retained 19,934 images and 1,000 labels. Screening could change enrollments/pairs, did not undo training/calibration exposure, and did not exclude unknown near duplicates or identify a causal contamination effect.

SHA-256 hashes linked configurations, training counters, checkpoints, calibration records, pair manifests, source snapshots and Test outputs. Software, preprocessing, inputs, checkpoints, and thresholds were checked against these records before and after scoring. Detailed configurations and historical validation records were retained locally. These records support traceability; however, incomplete environment and entrypoint records limit the exact reproduction.

4. Results

4.1. Overall Verification Performance and Attribute Mismatch

The primary Official Test view contained 19,962 images with 1,000 supplied identity labels. P1 comprised 15,769 genuine and 15,769 impostor pairs, shared by both models. Unless stated otherwise, operating-point results use the model-specific thresholds calibrated to P1 Validation FAR 1% and then frozen. AUC is dimensionless, rates are percentages, and rate differences are percentage points (pp). Pointwise intervals describe individual estimates; adjusted intervals apply only to the frozen 30-comparison family. An interval excluding zero indicates direction under that approximate conditional analysis. Meeting a reportability floor alone does not establish a nonzero contrast.

Both models combined high aggregate discrimination with genuine rejections at the primary operating point (Table 1). The empirical test FAR point estimates were above the validation target in both models; calibration did not enforce that target on the test. The adjusted comparison between the models is reported in Section 4.3.

Table 1. Overall face verification performance

This is an open access article under the CC BY-SA license.

Model	ROC-AUC [95% CI]	EER % [95% CI]	TAR % [95% CI]	Test FAR % [95% CI]
ArcFace	0.9793 [0.9751, 0.9830]	6.24 [5.54, 6.94]	87.19 [85.79, 88.53]	1.084 [0.815, 1.382]
CosFace	0.9767 [0.9722, 0.9806]	6.83 [6.15, 7.57]	85.71 [84.28, 87.09]	1.065 [0.800, 1.357]

Note. Official Test Primary/P1, with 15,769 pairs per class. Intervals are pointwise approximate bootstrap intervals. Genuine accepted counts were 13,749 (ArcFace) and 13,515 (CosFace); false accepts were 171 and 168. Model-specific validation thresholds were frozen before the test; the validation target FAR was 1%.

Zero-mismatch pairs agreed on all six annotations, whereas any-mismatch pairs differed on at least one. The groups contained 8,652 and 7,117 pairs from 920 and 883 identities, respectively. Identity membership overlapped, so they were not independent cohorts. ArcFace TAR was 91.27% for zero mismatch and 82.23% for any mismatch; corresponding CosFace rates were 90.14% and 80.31%.

The zero-minus-any TAR reductions were 9.05 pp for ArcFace and 9.83 pp for CosFace, with adjusted intervals of [5.77, 12.50] and [6.45, 13.50] pp. Both intervals were positive, supporting an association between annotation mismatch and lower genuine acceptance for RQ1. The contrast compares observed groups and does not estimate what would happen if appearance alone were changed.

4.2. Attribute-Specific and Mismatch-Count Analysis

The any-mismatch association varied across attributes (Table 2; Figure 1). Adjusted intervals were positive for Eyeglasses, Wearing Hat, Heavy Makeup, and Bangs in both models. Eyeglasses and Wearing Hat had among the largest point reductions, but their numerical ordering was not tested as an attribute-ranking hypothesis.

Mustache and Sideburns were inconclusive under the same adjusted policy. Mustache point reductions were positive, whereas Sideburns estimates were slightly negative; all four intervals included zero. These results support neither an absence of association nor improved robustness for mismatched facial hair. Mismatched support was smaller for Mustache and Sideburns than for Heavy Makeup, as reported alongside the intervals in Table 2.

This is an open access article under the CC BY-SA license.

Table 2. Attribute-Specific Mismatch Contrasts

Attribute	Mismatched pairs / identities	ArcFace reduction pp [adjusted CI]	CosFace reduction pp [adjusted CI]
Eyeglasses	1,173 / 306	12.50 [5.21, 20.43]	12.00 [4.24, 20.27]
Wearing Hat	1,017 / 286	12.16 [4.33, 21.23]	15.20 [5.79, 25.50]
Heavy Makeup	2,973 / 489	9.37 [4.90, 14.19]	10.32 [5.71, 15.34]
Bangs	2,307 / 476	6.57 [1.62, 12.22]	7.83 [2.65, 13.80]
Mustache	559 / 132	3.60 [-4.75, 13.41]	4.28 [-4.86, 15.65]
Sideburns	845 / 155	-0.16 [-6.87, 8.52]	-0.10 [-7.92, 9.69]

Note. Official Test Primary/P1 at primary fixed thresholds. Reduction is matched-minus-mismatched genuine TAR in percentage points (pp). Intervals follow the frozen 30-contrast Bonferroni policy; coverage is approximate and conditional. Each attribute partitions the same 15,769 genuine pairs, and identities may occur in both groups or across attributes. The first four attributes have positive adjusted intervals in both models; Mustache and Sideburns include zero.

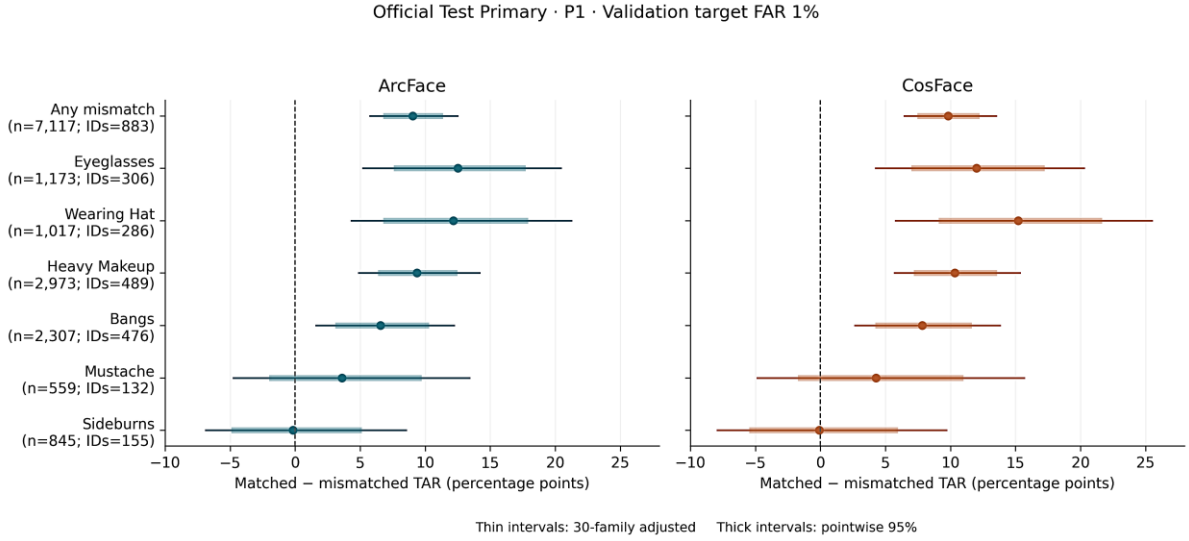


Figure 1. Genuine TAR contrasts for any mismatch and all six attributes in Official Test Primary/P1, at each model's fixed threshold calibrated to Validation FAR 1%. Thin intervals use the frozen 30-comparison Bonferroni policy; thick intervals are pointwise 95%. Labels report mismatched pairs and identity support. Positive values denote lower mismatched acceptance. Intervals are approximate and conditional; overlapping attributes are not independent interventions.

Acceptance was declined across the zero, one, two, and three-or-more mismatch groups in both models (Figure 2). The four groups contained 8,652, 5,523, 1,439, and 155 pairs from 920, 866, 437, and 89 identities, respectively. ArcFace TAR was 91.27%, 84.74%, 75.47% and 55.48%; CosFace TAR was 90.14%, 83.43%, 71.58% and 50.32%, respectively.

All six nonzero-count-versus-zero reductions had positive adjusted intervals. For one mismatch, reductions were 6.54 pp [3.31, 9.97] for ArcFace and 6.71 pp [3.50, 10.05] for CosFace; for two, they were 15.80 pp [10.01, 22.35] and 18.56 pp [11.90, 25.69]. The three-or-more group had the largest observed reductions and

This is an open access article under the CC BY-SA license.

substantially wider intervals: 35.79 pp ([21.17, 49.70]) and 39.82 pp ([25.09, 54.34]). The ordering was descriptive. No adjusted pairwise comparison between adjacent nonzero groups, monotonic trend test, or causal dose-response analysis was performed.

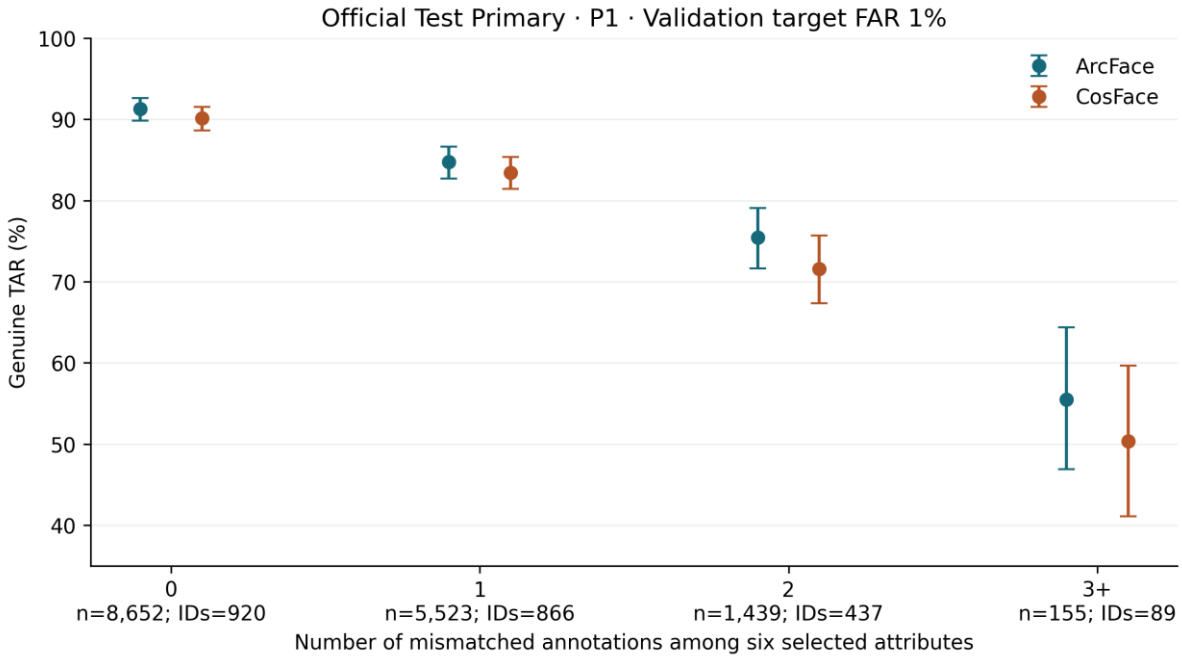


Figure 2. Genuine TAR across mismatch-count groups, Official Test Primary/P1 at unchanged model-specific thresholds calibrated to Validation FAR 1%. Bars are pointwise 95% approximate identity-node intervals; group supports are shared across models, and identity overlap occurs across groups. The plotted categories do not represent a causal dose or an interpolated trend. Adjusted contrasts versus zero are reported in Section 4.2.

4.3. Comparison of ArcFace and CosFace

The primary overall CosFace-minus-ArcFace TAR difference was -1.48 pp, with an adjusted interval of [-2.40, -0.62] pp. This supports the lower overall TAR for this CosFace instance under the specified transport operating rules. It did not compare the loss of families at a common empirical Test FAR or imply that their numerical score thresholds shared a calibration scale.

All six cross-model attribute-reduction-adjusted intervals included zero, as did all three cross-model count-reduction intervals (Supplementary Table S1). In particular, the two-mismatch reduction difference was 2.76 pp: its pointwise interval [0.44, 5.18] excluded zero, whereas its adjusted interval [-0.83, 6.79] did not. Therefore, the adjusted interval did not establish a difference in mismatch robustness despite the overall TAR difference. Exploratory CosFace-minus-ArcFace AUC and EER differences were -0.002649 (pointwise interval [-0.004467, -0.000818]) and 0.59 pp ([0.20, 1.02]), respectively; these summaries were outside the primary family.

4.4. Evaluation Protocol and Duplicate-Screening Sensitivity

P1's profile matching and the five P2 random-impostor realizations produced different estimates from the same embeddings (Figure 3). Mean P2 AUC was 0.981451 for ArcFace and 0.978853 for CosFace. The mean empirical FAR was 0.602% and 0.571%, compared with P1's 1.084% and 1.065%, respectively. Across P2

This is an open access article under the CC BY-SA license.

seeds, FAR ranged from 0.476% to 0.729% for ArcFace and 0.463% to 0.723% for CosFace. The paired mean-P2-minus-P1 FAR differences were -0.482 pp [−0.790, −0.198] for ArcFace and -0.495 pp [−0.794, −0.220] for CosFace, using secondary pointwise 95% intervals.

Genuine acceptance was identical across P1 and all five P2 samples because neither the genuine pairs nor the thresholds changed. Consequently, the lower P2 FAR and EER and higher AUC describe the changed impostor population for fixed representations rather than an improvement in genuine recognition. Figure 3 shows every realization, and Supplementary Table S2 reports the point estimates and pointwise intervals for both test views.

P1 included four fallback rows in the Official Test. The exact-profile-only sensitivity analysis retained 15,765 pairs per class and gave AUC 0.979322/0.976674 and TAR 87.193%/85.715% for ArcFace/CosFace (Supplementary Table S5). These descriptive estimates were close to the fallback-including estimates. Separate inferential intervals were not available for this restriction, and removing referenced genuine rows meant that its TAR was not subject to P1/P2 invariance.

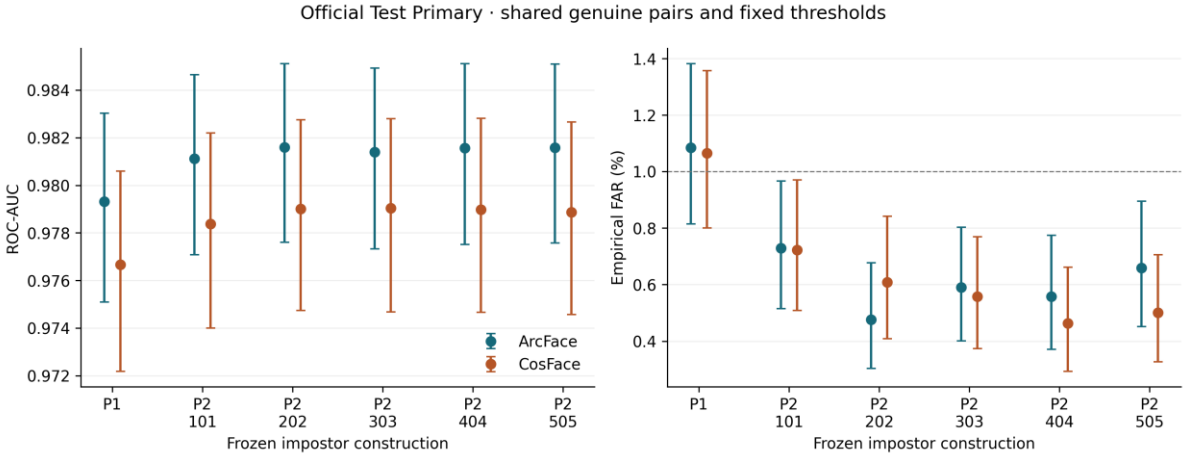


Figure 3. ROC-AUC and empirical Test FAR for P1 and five P2 realizations in Official Test Primary. Bars are pointwise 95% approximate dyadic-bootstrap intervals; comparisons share identity draws and are dependent. FAR uses a fixed model-specific threshold calibrated to P1 Validation FAR 1% (dashed reference). There are 15,769 genuine and 15,769 impostor pairs per realization; genuine TAR is invariant within each model.

The predefined screened view retained 19,934 images and 1,000 identities after the exclusions described in Section 3.4. Genuine acceptance and empirical FAR at the unchanged thresholds remained close to the Official Test estimates (Table 3). The screened AUC was 0.979355/0.977148 and the EER was 6.279%/6.780% for ArcFace/CosFace, respectively; the corresponding pointwise intervals are in Supplementary Table S2.

The view-specific pointwise intervals for screened any-mismatch reductions remained positive (Table 3), supporting persistence under this exclusion policy. The screened group support and rates are in Supplementary Table S3. This secondary comparison does not inherit the primary adjusted-family conclusion. No inferential test of the between-view difference was available, and changed enrollment and pair composition prevented interpreting screening as an isolated contamination effect.

This is an open access article under the CC BY-SA license.

Table 3. Duplicate-Screened Sensitivity at Unchanged Thresholds

Model / Test view	TAR % [95% CI]	FAR % [95% CI]	Any-mismatch reduction pp [95% CI]
ArcFace / Official	87.19 [85.79, 88.53]	1.084 [0.815, 1.382]	9.05 [6.99, 11.15]
ArcFace / Screened	87.20 [85.80, 88.54]	1.097 [0.811, 1.421]	8.89 [6.84, 10.99]
CosFace / Official	85.71 [84.28, 87.09]	1.065 [0.800, 1.357]	9.83 [7.66, 12.00]
CosFace / Screened	85.65 [84.20, 87.04]	1.008 [0.751, 1.299]	9.62 [7.44, 11.80]

Note. P1; Official/screened views contain 15,769/15,767 pairs per class and 19,962/19,934 images, respectively, with 1,000 identities retained. All intervals here are pointwise and view-specific, not intervals for between-view differences. The primary adjusted any-mismatch intervals are reported in Section 4.1. Screening changes available images, enrollments, and pairs; it does not isolate a contamination effect.

At the exploratory thresholds calibrated to P1 Validation FAR 0.1%, Official Test P1 genuine acceptance was 12,135/15,769 (76.95%) for ArcFace and 11,956/15,769 (75.82%) for CosFace. Only 14 and 18 impostor pairs were accepted, giving empirical FAR 0.089% (pointwise interval [0.020, 0.190]) and 0.114% ([0.039, 0.220]). Supplementary Table S4 lists the exact thresholds, TAR intervals, and corresponding screened results.

In this secondary exploratory analysis, both models retained lower acceptance for any mismatch and for eyeglasses, wearing hats, heavy makeup, and bangs, with descriptively descending TAR across mismatch-count groups. These directional patterns do not extend the primary family adjusted conclusions to the 0.1% target group.

At this target, approximately 16 false accepts were expected per test realization, below the predefined 100-count heuristic; dependent pairs further limit precision. These results were exploratory and were not used to modify thresholds, enlarge the primary family, or make operational low-FAR guarantees.

5. Discussion

The central result is that high aggregate discrimination did not ensure similar genuine acceptance across appearance groups. At the same model-specific operating rule, pairs with unequal annotations were less often accepted than pairs agreeing on all selected attributes. Terhörst et al. (2022) demonstrated attribute-associated variation through presence strata and controls; our endpoint-agreement analysis addresses a related but different comparison. Sharing an identity within each genuine pair does not make the contrast between groups identity-adjusted, because identities participate unequally and some contribute to both groups. The bootstrap accounts for dependence without balancing these populations.

The eyeglasses association is consistent with the relevance of natural eyeglasses-status differences studied by Guo et al. (2018). It is not a replication of their augmentation or gallery/probe design, and its magnitude should not be compared directly across protocols. Wearing a hat, heavy makeup, and bangs also had positive adjusted contrasts in the present experiment. Altered visible cues or occlusion may be plausible explanations, but pose, quality, and identity composition were not separated from the annotations. The evidence therefore identifies less successful comparison groups rather than a verified mechanism of recognition failure.

The inconclusive facial-hair results place an equally important boundary on that interpretation. Positive Mustache estimates and slightly negative Sideburns estimates both had adjusted intervals spanning zero. Their uncertainty leaves room for associations in either direction; the negative point estimates do not establish

This is an open access article under the CC BY-SA license.

a benefit from mismatch, and the zero-containing intervals do not demonstrate invariance. Detailed facial-hair studies examine categories or areas that these two binary labels cannot reproduce (Ozturk et al., 2024, 2025; Wu et al., 2024; Wu, Bezold, Bhatta, et al., 2023). Differences in category definition and comparison therefore limit direct agreement or disagreement with those studies. Larger support alone would also not resolve the confounding inherent in the current groups.

Mismatch count summarizes another aspect of pair composition. The descending rates suggest that pairs differing on more selected annotations are harder for these fitted models, but the count gives each annotation equal weight and does not hold other image properties constant. The sparse highest-count group has correspondingly wider uncertainty, and adjacent nonzero groups were not formally compared. Annotation inconsistency adds another source of ambiguity: unequal CelebA labels need not indicate a true appearance difference, while equal labels need not imply visual similarity (Wu, Bezold, Günther, et al., 2023). Natural-image comparisons inherit acquisition confounding, whereas GAN-based analyses require attention to reconstruction and editing artifacts (Banerjee et al., 2023). These designs offer complementary evidence without establishing the same causal contrast.

The model comparison separates overall acceptance from mismatch robustness. ArcFace's overall TAR advantage for the fitted instances did not extend to adjusted differences in attribute- or count-specific reductions. In particular, the pointwise interval for the two-mismatch contrast excluded zero, but its prespecified family-adjusted interval included zero. Selecting that pointwise result would overstate the evidence for an objective-associated robustness difference. A shared backbone, budget, and evaluation pairs make this comparison interpretable within the experiment; however, one run per objective cannot characterize training variability. Separate transported thresholds also compare operating policies with different achieved Test FARs. The results therefore support neither a general ranking of the loss families nor equivalence between them.

Replacing P1 impostors with P2 samples altered the AUC, EER, and FAR without changing genuine acceptance. The gain in those summaries consequently reflects a different non-match population rather than a change to the model or its genuine comparisons. This is consistent with O'Toole et al. (2012) and the emphasis on comparison difficulty in Goldilocks test sets (Wu et al., 2026), although their populations and evaluation rules differ from ours. P1's annotation-profile conditioning is an explicit sampling rule, not a guarantee of uniform semantic difficulty. Reporting that rule, its fallbacks, and the calibration population is necessary to understand the observed protocol differences; five P2 seeds characterize only the specified sampler.

Duplicate screening retained the broad any-mismatch association; thus, the finding was not confined to the complete unfiltered pair set. This persistence is evidence of one predefined exclusion policy. It does not show what performance would have been without duplicated content during training or calibration because exclusion also changed the available enrollments and pair composition. Unknown near duplicates and label inconsistencies can remain. Duplicate-detection research motivates this distinction between documented content overlap and an estimate of its effect (Schlett et al., 2024). The screened results qualify the primary analysis rather than restoring full content independence.

Selective fine-tuning of VGG-based models on CelebA has also been explored for efficient face recognition (Muftah et al., 2024). For future resource-constrained deployments, layer selection and pruning (Kaushal et al., 2025) and edge–cloud application placement (Almurshed et al., 2024) provide complementary efficiency directions. Neither approach was evaluated in the verification experiments.

This is an open access article under the CC BY-SA license.

6. Limitations

The study is limited to CelebA, whose celebrity imagery and imperfect binary annotations restrict the interpretation of appearance mismatch. Because comparisons used observational groups with equal weight per pair and no covariate adjustment, differences in pose, age, illumination, image quality, identity participation, and correlated attributes remain possible confounders. Neither causal effects nor demographic fairness were evaluated. One completed training run per objective does not establish seed robustness or loss-family superiority, and deterministic selection of the first eligible enrollment image makes acceptance conditional on particular references. Repeated images and identities create dependent pairs. The identity-node/dyadic bootstrap approximates that dependence conditional on fixed models, labels, thresholds, and pairs, while omitting training variation, calibration uncertainty, and new acquisitions; reportability floors and Bonferroni adjustment do not guarantee coverage, especially with limited extreme-tail precision. Known cross-partition image-content duplicates compromise content independence. Screening changes enrollments and pairs, cannot undo training or calibration exposure, and cannot exclude unknown near duplicates or isolate contamination effects. Incomplete historical ArcFace entrypoint provenance and environment records further limit exact replay despite matching core-source hashes. The selected operating thresholds, impostor protocols, and sparse exploratory low-FAR support constrain generalization to other datasets, architectures, populations, and deployment conditions.

7. Future Work

External validation should examine whether the associations persist across additional identity-disjoint datasets and acquisition conditions, with exact and near-duplicate audits completed before calibration. Multiple training seeds and additional architectures would help distinguish fitted-instance variability from objective-associated behavior. Annotation review and designs that balance pose, quality, illumination, identity participation, and correlated attributes could make mismatch comparisons more informative. Alternative enrollment images, multiple references, impostor populations, and operating targets should be assessed under prospectively fixed protocols.

8. Conclusion

For the two locally trained CelebA baselines, annotation-defined mismatch was associated with lower genuine acceptance at fixed validation-calibrated thresholds. The association differed by attribute: four had positive adjusted contrasts, while Mustache and Sideburns remained inconclusive. Acceptance also declined descriptively across mismatch-count groups. These observations characterize naturally sampled comparison groups; they do not establish independent attribute effects or a causal dose response.

The overall difference between the fitted models did not establish attribute-specific loss-family superiority. Impostor sampling changed discrimination and false acceptance, while genuine acceptance remained fixed, and duplicate screening retained the broad mismatch association within its secondary scope. Within this experiment, interpreting verification robustness therefore requires subgroup support, pair construction, calibration, and content-overlap limitations to accompany aggregate performance.

9. Disclosure of AI Use

OpenAI tools were used for manuscript preparation, including language editing, drafting, translation, and formatting. The authors are responsible for the scientific content.

This is an open access article under the CC BY-SA license.

10. References

Almurshed, O., Meshoul, S., Muftah, A., Kaushal, A. K., Almoghamis, O., Petri, I., Auluck, N., & Rana, O. (2024). A framework for performance optimization of internet of things applications. In *Euro-Par 2023: Parallel Processing Workshops* (Lecture Notes in Computer Science, Vol. 14351, pp. 165–176). Springer. https://doi.org/10.1007/978-3-031-50684-0_13

Banerjee, S., Aggarwal, A., & Ross, A. (2023). Can GAN-induced attribute manipulations impact face recognition? In *Pattern recognition, computer vision, and image processing: ICPR 2022 international workshops and challenges* (Lecture Notes in Computer Science, Vol. 13645, pp. 521–535). Springer. https://doi.org/10.1007/978-3-031-37731-0_38

Conti, J.-R., & Clémençon, S. (2024). Assessing uncertainty in similarity scoring: Performance & fairness in face recognition. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8d2c36836fb0e7d78fe68762ff8b5f1e-Abstract-Conference.html

Dantcheva, A., Chen, C., & Ross, A. (2012). Can facial cosmetics affect the matching accuracy of face recognition systems? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems* (pp. 391–398). https://doi.org/10.1109/BTAS.2012.6374605

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). 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).

https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html

Guo, J., Zhu, X., Lei, Z., & Li, S. Z. (2018). Face synthesis for eyeglass-robust face recognition. In *Biometric Recognition: CCBR 2018* (pp. 275–284). Springer. https://doi.org/10.1007/978-3-319-97909-0_30

Kaushal, A., Almurshed, O., Muftah, A., Auluck, N., & Rana, O. (2025). ToSiM-IoT: Toward a sustainable optimization of machine learning tasks in Internet of Things. *IEEE Internet of Things Journal, 12*(11), 16829–16840. https://doi.org/10.1109/JIOT.2025.3537169

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 365–372). https://www.cs.columbia.edu/CAVE/publications/pdfs/Kumar_ICCV09.pdf

Leroy, P., Mastropietro, A., Nurisso, M., & Vaccarino, F. (2025). Attributes shape the embedding space of face recognition models. In *Proceedings of the 42nd International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 267, pp. 33960–33983). https://proceedings.mlr.press/v267/leroy25a.html

This is an open access article under the CC BY-SA license.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3730–3738). https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html

Muftah, A., Almurshed, O., Bennasar, M., Price, B., Laurence, S., & Pike, G. (2024). Towards efficient AI solutions for facial recognition in the wild. *Procedia Computer Science*, *246*, 1752–1761. https://doi.org/10.1016/j.procs.2024.09.672

O'Toole, A. J., Phillips, P. J., An, X., & Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, *30*(3), 169–176. https://doi.org/10.1016/j.imavis.2011.12.007

Ozturk, K., Bezold, G., Wu, H., Bhatta, A., & Bowyer, K. W. (2025). Effects of facial hair on face recognition. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition* (pp. 1–10). https://doi.org/10.1109/FG61629.2025.11099142

Ozturk, K., Wu, H., & Bowyer, K. W. (2024). Can the accuracy bias by facial hairstyle be reduced through balancing the training data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1519–1528). https://openaccess.thecvf.com/content/CVPR2024W/BIOMET/html/Ozturk_Can_the_Accuracy_Bias_by_Facial_Hairstyle_be_Reduced_Through_CVPRW_2024_paper.html

Schlett, T., Rathgeb, C., Tapia, J., & Busch, C. (2024). Double trouble? Impact and detection of duplicates in face image datasets. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods* (pp. 801–808). https://doi.org/10.5220/0012422500003654

Terhörst, P., Kolf, J. N., Huber, M., Kirchbuchner, F., Damer, N., Morales, A., Fierrez, J., & Kuijper, A. (2022). A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, *3*(1), 16–30. https://doi.org/10.1109/TTS.2021.3111823

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5265–5274). https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_CosFace_Large_Margin_CVPR_2018_paper.html

Wu, H., Bezold, G., Bhatta, A., & Bowyer, K. W. (2023). Logical consistency and greater descriptive power for facial hair attribute learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8588–8597). https://openaccess.thecvf.com/content/CVPR2023/html/Wu_Logical_Consistency_and_Greater_Descriptive_Power_for_Facial_Hair_Attribute_CVPR_2023_paper.html

This is an open access article under the CC BY-SA license.

Wu, H., Bezold, G., Günther, M., Boult, T., King, M. C., & Bowyer, K. W. (2023). Consistency and accuracy of CelebA attribute values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 3258–3266). https://openaccess.thecvf.com/content/CVPR2023W/VDU/html/Wu_Consistency_and_Accuracy_of_CelebA_Attribute_Values_CVPRW_2023_paper.html

Wu, H., Tian, S., Bhatta, A., Gutierrez, J., Bezold, G., Argueta, G., Ricanek, K., King, M. C., & Bowyer, K. W. (2026). Goldilocks test sets for face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 35504–35513). https://openaccess.thecvf.com/content/CVPR2026/papers/Wu_Goldilocks_Test_Sets_for_Face_Verification_CVPR_2026_paper.pdf

Wu, H., Tian, S., Bhatta, A., Öztürk, K., Ricanek, K., & Bowyer, K. W. (2024). Facial hair area in face recognition across demographics: Small size, big effect. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops* (pp. 1131–1140). https://openaccess.thecvf.com/content/WACV2024W/DVPBA/html/Wu_Facial_Hair_Area_in_Face_Recognition_Across_Demographics_Small_Size_WACVW_2024_paper.html

Supplementary Tables

Supplementary Table S1. Cross-model comparisons

Comparison	Difference (pp)	Pointwise 95% CI	Bonferroni-adjusted CI	Adjusted direction
Overall TAR	-1.48	[-2.05, -0.93]	[-2.40, -0.62]	below zero
Eyeglasses reduction	-0.50	[-3.31, 2.21]	[-5.19, 3.75]	includes zero
Heavy makeup reduction	0.95	[-0.51, 2.39]	[-1.37, 3.30]	includes zero
Wearing hat reduction	3.04	[-0.01, 6.25]	[-1.55, 8.32]	includes zero
Bangs reduction	1.26	[-0.26, 2.76]	[-1.17, 3.70]	includes zero
Mustache reduction	0.69	[-1.92, 3.49]	[-3.21, 5.36]	includes zero
Sideburns reduction	0.06	[-2.18, 2.68]	[-3.25, 4.69]	includes zero
One versus zero mismatches	0.17	[-0.81, 1.15]	[-1.41, 1.72]	includes zero
Two versus zero mismatches	2.76	[0.44, 5.18]	[-0.83, 6.79]	includes zero
Three or more versus zero mismatches	4.03	[-2.46, 10.95]	[-6.09, 15.30]	includes zero

Note. Official Test, P1. Model-specific thresholds were calibrated on P1 Validation at FAR 1% and held fixed for Test. Direction is CosFace minus ArcFace. Attribute entries compare matched-minus-mismatched TAR reductions; count entries compare the reduction from zero mismatches. Estimates and intervals are in pp. Pointwise intervals have nominal 95% coverage; Bonferroni intervals belong to the primary 30-comparison family (nominal marginal coverage 99.8333%). Overall TAR is lower for this CosFace instance: its adjusted interval excludes zero. All nine adjusted mismatch-reduction intervals include zero.

Interval method. Saved approximate identity-node/dyadic bootstrap intervals are conditional on the fixed models, labels, thresholds and pairs. They omit training and calibration uncertainty; nominal coverage is approximate.

TAR, true accept rate; FAR, false accept rate; ROC-AUC, area under the receiver operating characteristic curve; EER, equal error rate; CI, confidence interval; pp, percentage points. Screened denotes the predefined duplicate-screened Test view.

This is an open access article under the CC BY-SA license.

Supplementary Table S2. Evaluation protocol results

Test view	Model	Protocol	ROC-AUC [95% CI]	EER (%) [95% CI]	FAR (%) [95% CI]
Official	ArcFace	P1	0.979324 [0.975102, 0.983025]	6.240 [5.544, 6.941]	1.084 [0.815, 1.382]
Official	ArcFace	P2 (101)	0.981125 [0.977085, 0.984650]	5.758 [5.147, 6.443]	0.729 [0.515, 0.966]
Official	ArcFace	P2 (202)	0.981602 [0.977601, 0.985104]	5.733 [5.094, 6.400]	0.476 [0.303, 0.677]
Official	ArcFace	P2 (303)	0.981392 [0.977335, 0.984918]	5.733 [5.068, 6.433]	0.590 [0.401, 0.803]
Official	ArcFace	P2 (404)	0.981558 [0.977512, 0.985108]	5.631 [4.992, 6.351]	0.558 [0.371, 0.775]
Official	ArcFace	P2 (505)	0.981580 [0.977581, 0.985094]	5.726 [5.107, 6.418]	0.660 [0.451, 0.895]
Official	CosFace	P1	0.976675 [0.972191, 0.980591]	6.830 [6.150, 7.569]	1.065 [0.800, 1.357]
Official	CosFace	P2 (101)	0.978371 [0.974008, 0.982199]	6.373 [5.730, 7.020]	0.723 [0.508, 0.970]
Official	CosFace	P2 (202)	0.979006 [0.974743, 0.982751]	6.221 [5.581, 6.854]	0.609 [0.409, 0.842]
Official	CosFace	P2 (303)	0.979045 [0.974683, 0.982804]	6.227 [5.582, 6.886]	0.558 [0.374, 0.769]
Official	CosFace	P2 (404)	0.978980 [0.974663, 0.982807]	6.259 [5.628, 6.910]	0.463 [0.294, 0.661]
Official	CosFace	P2 (505)	0.978863 [0.974578, 0.982657]	6.246 [5.595, 6.893]	0.501 [0.327, 0.706]
Screened	ArcFace	P1	0.979355 [0.975096, 0.983037]	6.279 [5.613, 7.032]	1.097 [0.811, 1.421]
Screened	ArcFace	P2 (101)	0.981216 [0.977152, 0.984732]	5.746 [5.128, 6.426]	0.717 [0.507, 0.955]
Screened	ArcFace	P2 (202)	0.981676 [0.977656, 0.985191]	5.721 [5.088, 6.404]	0.482 [0.309, 0.687]
Screened	ArcFace	P2 (303)	0.981346 [0.977282, 0.984883]	5.784 [5.098, 6.464]	0.571 [0.382, 0.786]
Screened	ArcFace	P2 (404)	0.981560 [0.977491, 0.985123]	5.664 [5.005, 6.370]	0.564 [0.376, 0.776]
Screened	ArcFace	P2 (505)	0.981734 [0.977693, 0.985235]	5.689 [5.066, 6.379]	0.647 [0.442, 0.888]
Screened	CosFace	P1	0.977148 [0.972683, 0.981053]	6.780 [6.080, 7.516]	1.008 [0.751, 1.299]
Screened	CosFace	P2 (101)	0.978559 [0.974211, 0.982367]	6.374 [5.726, 7.031]	0.736 [0.517, 0.987]
Screened	CosFace	P2 (202)	0.979206 [0.974938, 0.982948]	6.196 [5.550, 6.845]	0.596 [0.398, 0.827]
Screened	CosFace	P2 (303)	0.978861 [0.974517, 0.982671]	6.279 [5.639, 6.953]	0.622 [0.425, 0.849]
Screened	CosFace	P2 (404)	0.979196 [0.974882, 0.982982]	6.260 [5.625, 6.909]	0.488 [0.313, 0.690]
Screened	CosFace	P2 (505)	0.979049 [0.974728, 0.982816]	6.241 [5.593, 6.878]	0.609 [0.414, 0.835]

Note. P1 uses profile-matched impostors; P2 uses random impostors, with seeds shown in parentheses. Each realization has 15,769 pairs per class in Official Test and 15,767 in screened Test. Model-specific thresholds calibrated on P1 Validation at FAR 1% remain fixed. P2 changes neither genuine pairs nor genuine TAR within a model and view. ROC-AUC is unitless; EER and empirical FAR are percentages. Brackets contain saved pointwise 95% CIs. These protocol and view sensitivity results are secondary.

Interval method. Saved approximate identity-node/dyadic bootstrap intervals are conditional on the fixed models, labels, thresholds and pairs. They omit training and calibration uncertainty; nominal coverage is approximate.

TAR, true accept rate; FAR, false accept rate; ROC-AUC, area under the receiver operating characteristic curve; EER, equal error rate; CI, confidence interval; pp, percentage points. Screened denotes the predefined duplicate-screened Test view.

This is an open access article under the CC BY-SA license.

Supplementary Table S3. Screened genuine acceptance

Model	Genuine group	Pairs	Distinct identities	Accepted pairs	TAR (%) [95% CI]
ArcFace	Overall	15767	996	13749	87.20 [85.80, 88.54]
ArcFace	Zero mismatches	8650	919	7890	91.21 [89.76, 92.57]
ArcFace	Any mismatch	7117	882	5859	82.32 [80.26, 84.28]
ArcFace	One mismatch	5530	866	4688	84.77 [82.78, 86.66]
ArcFace	Two mismatches	1427	440	1082	75.82 [72.07, 79.38]
ArcFace	Three or more mismatches	160	89	89	55.62 [47.06, 64.67]
CosFace	Overall	15767	996	13504	85.65 [84.20, 87.04]
CosFace	Zero mismatches	8650	919	7784	89.99 [88.47, 91.38]
CosFace	Any mismatch	7117	882	5720	80.37 [78.23, 82.45]
CosFace	One mismatch	5530	866	4608	83.33 [81.33, 85.24]
CosFace	Two mismatches	1427	440	1028	72.04 [67.79, 76.19]
CosFace	Three or more mismatches	160	89	84	52.50 [43.38, 61.59]

Note. Screened Test, P1, at the unchanged thresholds calibrated on P1 Validation at FAR 1%. Mismatch counts refer to disagreement on the six evaluated attributes. Pairs and accepted counts refer to genuine pairs; identities are distinct identities within each group. Any mismatch combines the one-, two- and three-or-more-mismatch groups. Identities can occur across groups and must not be summed. TAR and bracketed pointwise 95% CIs are percentages. These results are secondary and outside the primary 30-comparison adjusted family.

Interval method. Saved approximate identity-node/dyadic bootstrap intervals are conditional on the fixed models, labels, thresholds and pairs. They omit training and calibration uncertainty; nominal coverage is approximate.

TAR, true accept rate; FAR, false accept rate; ROC-AUC, area under the receiver operating characteristic curve; EER, equal error rate; CI, confidence interval; pp, percentage points. Screened denotes the predefined duplicate-screened Test view.

This is an open access article under the CC BY-SA license.

Supplementary Table S4. Exploratory FAR 0.1%

Test view	Model	Threshold	Genuine accepted/total	TAR (%) [95% CI]	Impostor accepted/total	FAR (%) [95% CI]
Official	ArcFace	0.33777374029159546	12135/15769	76.95 [75.11, 78.77]	14/15769	0.089 [0.020, 0.190]
Official	CosFace	0.27491146326065063	11956/15769	75.82 [74.00, 77.63]	18/15769	0.114 [0.039, 0.220]
Screened	ArcFace	0.33777374029159546	12132/15767	76.95 [75.07, 78.76]	21/15767	0.133 [0.050, 0.247]
Screened	CosFace	0.27491146326065063	11949/15767	75.78 [73.96, 77.60]	20/15767	0.127 [0.046, 0.239]

Note. Exploratory P1 results at model-specific thresholds calibrated on P1 Validation at FAR 0.1%, held fixed for Test. Thresholds retain the exact saved decimal representation. TAR and empirical FAR, including bracketed pointwise 95% CIs, are percentages. This analysis is outside the primary 30-comparison family. At the target, approximately 16 false accepts are expected per Test realization, below the predefined 100-count heuristic; dependent pairs further limit precision. These results do not provide operational low-FAR guarantees.

Interval method. Saved approximate identity-node/dyadic bootstrap intervals are conditional on the fixed models, labels, thresholds and pairs. They omit training and calibration uncertainty; nominal coverage is approximate.

TAR, true accept rate; FAR, false accept rate; ROC-AUC, area under the receiver operating characteristic curve; EER, equal error rate; CI, confidence interval; pp, percentage points. Screened denotes the predefined duplicate-screened Test view.

This is an open access article under the CC BY-SA license.

Supplementary Table S5. Exact-profile-only sensitivity

Test view	Model	Pairs per class	Distinct identities	ROC-AUC	EER (%)	TAR (%)	FAR (%)
Official	ArcFace	15765	999	0.979322	6.242	87.193	1.085
Official	CosFace	15765	999	0.976674	6.832	85.715	1.066
Screened	ArcFace	15764	997	0.979355	6.274	87.205	1.097
Screened	CosFace	15764	997	0.977145	6.781	85.657	1.009

Note. P1 descriptive sensitivity at the unchanged primary thresholds calibrated on P1 Validation at FAR 1%. Excluding fallback impostor rows also removes their corresponding genuine rows; each class has the stated pair count. Distinct identities are the union across both pair classes. ROC-AUC is unitless; EER, TAR and FAR are percentages. Separate CIs were not calculated and saved for this restriction. Unlike P2, this restriction changes the genuine set, so genuine TAR need not remain invariant.

No inferential intervals are supplied for these descriptive estimates.

TAR, true accept rate; FAR, false accept rate; ROC-AUC, area under the receiver operating characteristic curve; EER, equal error rate; CI, confidence interval; pp, percentage points. Screened denotes the predefined duplicate-screened Test view.

This is an open access article under the CC BY-SA license.