QUERY-FREE ADVERSARIAL FACE GENERATION VIA REFERENCE-CONDITIONED COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Face recognition is widely deployed for identity verification in security-sensitive applications. Verification accepts faces whose representations lie sufficiently close to a reference, defining an acceptance region in a space learned to discriminate identities. Attribute information remains encoded alongside identity even though this objective neither specifies it nor explicitly constrains its variation within the accepted region. Motivated by this coexistence, we study query-free adversarial face generation using a public gallery and accessible recognition models as surrogates for an unseen target model. The goal is to find a representation that yields a face with a chosen race or gender different from the reference, while the target model accepts it as the reference identity. A global approximation of the chosen attribute class produces a candidate representation by projecting the reference representation onto it, but fitting the approximation without the reference obscures the class representations nearest to that reference. We instead form a new representation through reference-conditioned composition, ranking the chosen-class gallery by similarity to the reference and composing the nearest representations. Across independently trained models, gallery representations ranked near the reference tend to remain near it, as do their compositions; aggregating surrogate rankings further improves transfer. Our attack reaches success rates up to 94.3% on AWS Rekognition and 95.0% on Tencent Cloud at default thresholds; at the stricter settings recommended by AWS for law-enforcement use and documented by Tencent at a false match rate of 10^{-5}, it retains rates up to 55.0% and 80.0%, respectively.

1 INTRODUCTION

Face recognition (FR) has been widely used for identity verification, and it has recently been deployed in various security-sensitive applications, e.g., access control, border management systems, or eKYC. This has motivated extensive research on understanding security threats against these systems, and various attacks and countermeasures have been actively studied (Sharif et al., 2016; Dong et al., 2019; Zhang et al., 2021; Nguyen et al., 2022). Notably, recent studies have shown that such threats can be realized in practice, targeting commercial FR APIs that serve as a backbone in several real-world applications (An et al., 2023; Li et al., 2023; Kim et al., 2025; 2026; Paik et al., 2026b).

Meanwhile, a growing body of work investigates the intrinsic structure of the representations learned by FR models. Although these representations are trained to encode identity-related information by default, they are known to contain by-products as well. In particular, several studies have discovered that representations of faces sharing a demographic attribute, e.g., race or gender, form a cluster in the representation space (Hill et al., 2019; Leroy et al., 2025; Kim et al., 2025). Remarkably, these findings go beyond theory: Kim et al. (2025) showed that these structures can be exploited to craft *adversarial faces*, i.e., faces that possess different demographic attributes but are recognized as the same person. They showed that projecting a reference face onto an attribute cluster's PCA subspace yields such faces, which can pose a realistic threat to commercial FR deployments.

Motivated by these studies, we further investigate structural properties of manifolds formed by representations of faces sharing the same attribute, which we call *attribute manifolds*, beyond the PCA-based analyses adopted in prior works (Kim et al., 2025). We summarize our findings as follows:

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.

- PCA-based approximation may oversimplify attribute manifolds. We identify an inevitable trade-off between the manifold coverage and attribute consistency of the PCA subspace, constituting the intrinsic limit of Kim et al.'s approach.
- Projection onto the global PCA subspace may harm the local structure of the representation of the reference face—a source face image from which the adversary wishes to craft adversarial faces.

These observations lead us to devise a new black-box adversarial face generation method. Instead of projecting onto the PCA subspace, we aggregate representations in an attribute manifold near that of the reference face image. In addition, we propose a consensus-based ensemble of multiple FR models to determine "nearby" representations, which enhances the transferability of the resulting adversarial faces to unseen FR models. All these procedures can be performed over surrogate FR models held by the adversary, without requiring any interaction with the target FR model or its internal knowledge, such as its architecture, parameters, or training datasets.

We evaluate the effectiveness of the crafted adversarial faces on multiple open-source FR models and commercial FR APIs, including AWS CompareFaces (AWS, 2026) and Tencent Cloud (Tencent, 2026). Across the five attribute classes, the crafted faces are matched to the victim by AWS in 78.0%–94.3% of attempts and by Tencent in 63.7%–95.0%, at each provider's default operating point. Even with tighter threshold settings, e.g., recommended parameters for law-enforcement scenarios (AWS, 2026), the rates are 29.3%–55.0% and 42.7%–80.0%, respectively.

2 Background and Problem Formulation

Face Recognition and Verification. For a set $\mathcal{X}$ of face images, an FR model $E : \mathcal{X} \to \mathbb{S}^{d-1}$ maps each image to its representation, an unit vector. Such models are trained with metric-learning objectives under which the cosine similarity of two representations is large for the same identity and small for different identities. A model accepts a pair of images as the same identity when the cosine similarity of their representations exceeds a preset decision threshold τ.

Threat Model and Attack Objective. The attacker holds a public face gallery $\mathcal{I} \subset \mathcal{X}$. A fixed attribute classifier $a : \mathcal{X} \to \mathcal{C}$ supplies the label set $\mathcal{C}$ and partitions the gallery into class galleries $\mathcal{I}_c \subset \mathcal{I}$. The attacker also holds publicly available recognition models and, paired with one of them, a fixed pretrained inverse model $D : \mathbb{S}^{d-1} \to \mathcal{X}$, trained to generate a face image of the identity that a unit representation encodes. The existence of such an inverse model has been reported in various types of biometrics (Mai et al., 2018; Wijewardena et al., 2022; Hwang et al., 2026). The deployed verifier uses a target model E_T, which the attacker neither holds nor can query.

Starting from a reference image x, the attacker chooses a class $c \in \mathcal{C}$ with $c \neq a(x)$. Using only the gallery and the models it holds, the attacker constructs a unit representation z and generates $\widehat{x} = D(z)$ without optimization in image space or in the generator latent space.

The chosen class is a property of the generated image and cannot be determined from z, so a evaluates it on $\widehat{x}$ while the verifier evaluates acceptance as the identity score. The joint attack succeeds when $a(\widehat{x}) = c$ and $\langle E_T(\widehat{x}), E_T(x) \rangle > \tau$.

3 Global Approximation of Attribute Representations

The structure of attribute manifolds has been studied extensively, and the resulting geometric observations have been used to construct adversarial faces (Kim et al., 2025). In this section, we examine PCA-based attribute manifold approximation into a subsphere and identify where its global approximation can be improved. To this end, we analyze how subsphere dimension affects manifold coverage and attribute preservation on the subsphere, and how projection changes the pattern of similarities between a reference and individual images in the observed class gallery.

3.1 Revisiting Subsphere Approximations of Attribute Manifolds

For an attribute label c, let us denote the gallery of corresponding faces as $\mathcal{I}_c$. For attribute c, we call the representation-space region sampled by $E(\mathcal{I}_c)$ its *attribute manifold*. Prior studies have shown

that the attribute manifold forms a clustered structure over $\mathbb{S}^{d-1}$ (Leroy et al., 2025). Furthermore, Kim et al. (2025) established a hypothesis that there is a k-subsphere that well approximates $E(\mathcal{I}_c)$, providing its candidate $\mathbb{S}_c^k$ obtained from the subspace spanned by top-k principal components.

Connection to Adversarial Faces. Notably, projecting a reference face onto such an attribute manifold enables injecting the target attribute into that face. For a reference image x whose attribute differs from c, consider the metric projection z' of $z = E(x)$ onto $E(\mathcal{I}_c)$. Kim et al. (2025) discovered that the corresponding image $x' := D(z')$ possesses the attribute c. They further showed that x and x' are recognized as the same person in various FR models, providing its justification by analyzing the average projection distance onto the k-subsphere. We call their method NAFG.

3.2 How Faithfully Does the Subsphere Represent the Attribute Class?

We first study how faithfully the attribute subsphere approximates the attribute manifold. To this end, we evaluate the following two quantities for the attribute subsphere $\mathbb{S}_c^k$ with the face recognition model E, its paired inverse model D, and the attribute classification model a:

- *Coverage*: For a distribution $\mathcal{I}_c$ of facial images with an attribute c and the basis matrix $B_{c,k}$ of $\mathbb{S}_c^k$, we define $C_{c,k} := \mathbb{E}_{x\sim\mathcal{I}_c}[\|B_{c,k}^\top E(x)\|_2]$ as the coverage of $\mathbb{S}_c^k$.
- *Validity*: We define $V_{c,k} := \Pr[a(D(z)) = c \mid z \xleftarrow{\$} \mathbb{S}_c^k]$ as the validity of $\mathbb{S}_c^k$.

where, $\xleftarrow{\$}$ denotes a uniform random sampling. Following NAFG, we assume $\mathbb{E}_{x\sim\mathcal{I}_c}[E(x)] = \vec{0}$ for analysis; Appendix B.1 derives the exact affine alternative and finds no material change in result.

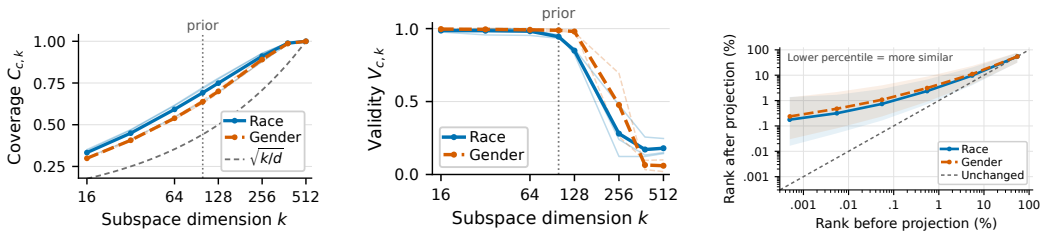
Note that $\|B_{c,k}^\top E(x)\|_2^2 = \langle B_{c,k}^\top E(x), B_{c,k}^\top E(x)\rangle = \langle B_{c,k}B_{c,k}^\top E(x), E(x)\rangle$, implying $C_{c,k}$ measures the average cosine similarity between $E(x)$ and its metric projection onto $\mathbb{S}_c^k$. We also remark that D was used for measuring $V_{c,k}$ to bridge the sampled embedding to the corresponding facial image, which can then be fed to the attribute classification model.

Coverage of the Attribute Subsphere. We estimate $C_{c,k}$ using facial images from the FairFace dataset (Karkkainen & Joo, 2021), whose attribute labeling coincides with our selected a. Note that, by definition, $C_{c,k}$ tends to 1 as a larger portion of the manifold overlaps with $\mathbb{S}_c^k$. As a baseline, we also provide the projection of a random vector z uniformly sampled from $\mathbb{S}^{d-1}$. Here, one can easily show that $\mathbb{E}[\|B_{c,k}^\top z\|_2^2] = \frac{k}{d}$ holds (Kim et al., 2025).

Fig. 1a provides the estimation results of $C_{c,k}$ with sweeping k. We can observe that measured $C_{c,k}$ exceeds the random baseline $\sqrt{k/d}$ in all swept k, meaning that the k-subsphere derived from PCA can capture non-uniformity of the attribute manifold to some extent. Nevertheless, a non-negligible gap from the correct projection still exists, and this gap is unlikely to be filled under k-subsphere approximation approaches; note that $B_{c,k}$ serves as the optimal linear fit of the manifold from the perspective of minimizing the projection distance, i.e., $1 - \|B_{c,k}^\top E(x)\|_2^2$ in our setting.

Validity: Demystifying the Attribute Subsphere Conjecture. The attribute subsphere conjecture by Kim et al. (2025) posits that the representation set associated with an attribute contains a k-subsphere with high probability. They attempted to justify this conjecture by constructing a k-subsphere via PCA principal components, with $k = 100$ fixed. However, the effective choice of k and the extent to which the subsphere's elements attain the chosen attribute remain less explored. We analyze these points via decoded validity $V_{c,k}$.

To this end, we design the following experiment: we uniformly sample N points $z_1, \ldots, z_N$ from $\mathbb{S}_c^k$ and find the corresponding image $\hat{x}_i := D(z_i)$. This can be realized by (i) sampling $\gamma_i \sim \mathcal{N}(0, I_k)$ and (ii) setting $z_i \leftarrow B_{c,k} \cdot \frac{\gamma_i}{\|\gamma_i\|_2}$. We then check $a(\hat{x}_i) \stackrel{?}{=} c$ with an attribute classification model a. Note that this procedure corresponds to estimating $V_{c,k}$ through Monte-Carlo. Figure 1b reports $V_{c,k}$ for the five evaluated attribute classes as k ranges from 16 to 512. Under FairFace, validity drops sharply for every class as k increases from 128 to 256.



(a) Coverage (b) Validity (c) Reference–gallery ranks

Figure 1: Coverage $C_{c,k}$ (a) and validity $V_{c,k}$ (b) evaluated from the PCA k-subsphere over attribute classes: White, Black, and Asian for races, and Male and Female for gender. The vertical line marks the $k = 100$ setting used by NAFG. Figure (c) shows the post-projection positions of groups defined by pre-projection rank percentile; lines are attribute-family means, and shading summarizes classwise interquartile ranges. Throughout, E is ArcFace (Deng et al., 2019), D is Arc2Face (Papantoniou et al., 2024), and a is FairFace (Karkkainen & Joo, 2021).

Coverage-Validity Trade-Off. From the results in Fig. 1a and 1b, we can observe a clear trade-off between $C_{c,k}$ and $V_{c,k}$: as k increases, $C_{c,k}$ increases while $V_{c,k}$ sharply decays. While such a trade-off is not surprising in itself, we emphasize that reducing this trade-off seems challenging via subsphere-based approximation: PCA already finds the optimal k-subsphere that maximizes $C_{c,k}$ in principle, so there is no room for improvement. Thus, the trade-off cannot be resolved by finding another k-subsphere with greater linear coverage, motivating a construction beyond global subsphere approximation. Appendix B.2 confirms these trends under alternative model choices and measures validity under the projection-induced distribution rather than the uniform one.

3.3 Reference–Gallery Similarities under Global Projection

For the joint objective, the class gallery provides observed images assigned to c, while their cosine similarities to the reference under E quantify how strongly each image matches the reference identity. The attributed subsphere is fitted from this gallery without the reference. Only after this reference-independent reduction does the prior attack introduce the reference through projection. Consequently, distinctions among gallery images that are not represented by the fitted span cannot contribute to the projected representation. We therefore ask whether this reduction preserves the original reference–gallery similarities, particularly among gallery images most similar to the reference. We answer this question by decomposing each similarity into the component represented by the fitted span and the component excluded from it, and then measuring how projection changes the corresponding similarity ranks.

For a reference image x and a gallery image $g \in \mathcal{I}_c$, let $u = E(x)$, $v = E(g)$, $P_{c,k} = B_{c,k}B_{c,k}^\top$, and $P_{c,k}^\perp = I - P_{c,k}$. For $P_{c,k}u \neq 0$, NAFG produces $z_{c,k}^{\mathrm{proj}} = P_{c,k}u/\|P_{c,k}u\|_2$. Orthogonality gives

$$\langle u, v\rangle = \underbrace{\langle P_{c,k}u, P_{c,k}v\rangle}_{\text{retained}} + \underbrace{\langle P_{c,k}^\perp u, P_{c,k}^\perp v\rangle}_{\text{excluded}}. \tag{1}$$

The retained term satisfies $\langle P_{c,k}u, P_{c,k}v\rangle = \|P_{c,k}u\|_2\langle z_{c,k}^{\mathrm{proj}}, v\rangle$. Because $\|P_{c,k}u\|_2$ is fixed for a given reference, this term induces the same gallery ranking as the projected representation. The excluded term contributes to the original reference–gallery score but not to the projected score, so its variation across gallery images can change their relative ranks.

For a fixed reference, we define the rank percentile of a gallery image as the percentage of $\mathcal{I}_c$ with greater cosine similarity to that reference; a smaller percentile therefore indicates greater similarity. At the $k = 100$ setting of the prior attack, we partition the class gallery by its pre-projection rank percentiles and report the median post-projection percentile of the same images. Figure 1c shows that, relative to their original positions, these images are placed progressively deeper in the projected ranking as their pre-projection similarity increases. The disagreement between the two rankings therefore increases among gallery images that are more similar to the reference. Equation (1) identifies the component omitted from the projected ranking: when the reference is used only after

the gallery has been reduced to a fixed class span, the projection can use only the in-span component of each reference–gallery similarity. The disagreement is concentrated among the class images most similar to the reference, where the identity side of the joint objective is strongest. This observation motivates using the original reference–gallery similarities rather than reducing the gallery to an approximation. Section 4 develops and evaluates such a reference-conditioned construction.

4 Local Construction and Cross-Model Transfer

Motivated by the analysis in Section 3, we introduce a new adversarial face generation that does not rely on a PCA-based approximation. Rather, we devise a reference-conditioned gallery construction that selects chosen-class images by their original similarity to the reference and aggregates their representations. Given a neighborhood size ℓ, the construction uses the ℓ class images most similar to the reference under the source model E_S that the inverse model D was trained against, and composes their representations. We further analyze whether the selection advantage persists across different FR models, from the perspective of their shared nearest neighbor structure.

4.1 Reference-Conditioned Selection and Uniform Averaging

For a gallery image $g \in \mathcal{I}_c$, define $s_E(g, x) := \langle E(g), E(x) \rangle$. At set size ℓ, we select the ℓ-nearest class images under E_S and uniformly average their original representations:

$$\begin{aligned} \mathcal{N}_\ell^S(x, c) &:= \mathrm{kNN}_\ell(x; \mathcal{I}_c, s_{E_S}), && \text{selected images,} \\ z_\ell^S(x, c) &:= \frac{\sum_{g \in \mathcal{N}_\ell^S(x,c)} E_S(g)}{\left\| \sum_{g \in \mathcal{N}_\ell^S(x,c)} E_S(g) \right\|_2}, && \text{uniform mean.} \end{aligned} \tag{2}$$

We regard $D(z_\ell^S(x, c))$ as the adversarial face of the anchor x. In the following paragraphs, we justify our design choices of employing kNN over the representation space and uniform averaging.

Reference-Conditioned Selection. The class gallery contains images with the chosen attribute but does not specify which images are compatible with the reference identity. We therefore compare the selected neighborhood $\mathcal{N}_\ell^S(x, c)$ with a reference-independent random subset of the same class gallery. With uniform coefficients, reference-conditioned selection increases generated-image identity similarity in every class, whereas random selection produces similarity near zero in Figure 2. Appendix B.3 changes only the ranking score and finds higher generated-image source similarity in every class when using the full reference-gallery similarity rather than after global approximation.

Uniform Averaging vs. Convex Combination. Neighbor selection does not specify how to compose the representations. A seemingly natural choice is to find the closest convex combination of them. Precisely, if we denote $g_1, \ldots, g_\ell \in \mathcal{N}_\ell^S(x, c)$, one can try to select $z_\ell^* := \frac{\sum_{i=1}^{\ell} \lambda_i E_S(g_i)}{\| \sum_{i=1}^{\ell} \lambda_i E_S(g_i) \|_2}$ that maximize $\langle E_S(x), z_\ell^* \rangle$ under the constraints $\lambda_i \geq 0$. Since z_ℓ^* is closer to x than our uniform averaging choice $z_\ell^S(x, c)$, $D(z_\ell^*)$ is more likely to be recognized as the same identity of x than $D(z_\ell^S(x, c))$.

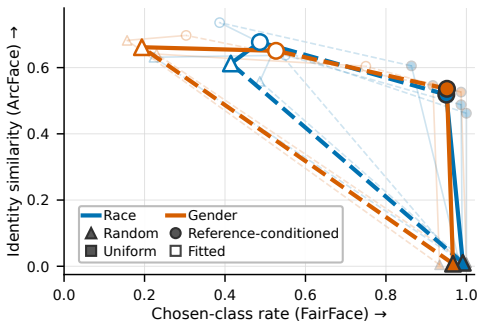
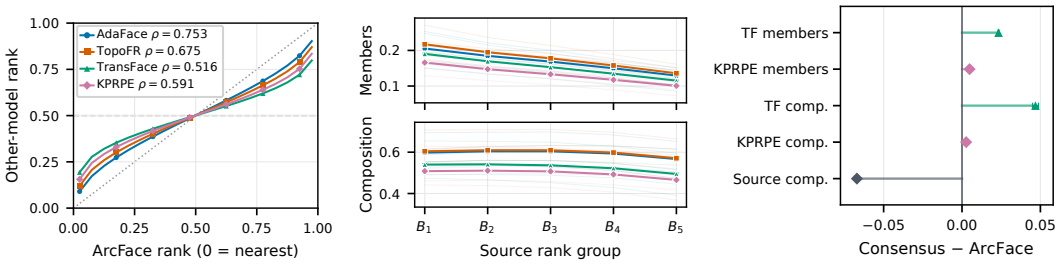


Figure 2: Selection and coefficient ablation at $\ell = 2^{10}$. Points report chosen-class assignment on FairFace and similarity on ArcFace.

Strikingly, we observe that the fitted z_ℓ^* from convex combination may not help the attack. This is because it significantly loses the chosen attribute, as the fitted objective contains no attribute term. To demonstrate this, we compare $D(z_\ell^*)$ and $D(z_\ell^S(x, c))$ in terms of the similarity with the anchor x and the chosen class assignment c. With the selected images fixed, Figure 2 shows that fitting raises source-model composition similarity but substantially reduces chosen-class assignment under FairFace in every race class. The same collapse occurs for a reference-independent random subset, indicating that it arises from the coefficients rather than reference conditioning.



(a) Broad relative ranks. (b) Selection and composition. (c) Effect of rank consensus.

Figure 3: Cross-model transfer and rank consensus. (a) Mean ranks along the ArcFace ordering; gray lines indicate identical ranks and shuffled pairs. (b) Member and composition similarities for lower-ranked groups; colors follow (a), faint curves show classes, bold curves their equal-class mean. (c) Changes from ArcFace-only selection to three-surrogate rank consensus at $\ell = 2^{10}$; rows separate model and metric, and Source comp. is the pre-decoding ArcFace composition similarity.

4.2 Cross-Model Transfer

As observed in prior works (McNeely-White et al., 2022; Kim et al., 2024b; Rubab et al., 2026), well-trained FR models tend to behave like almost-isometries, i.e., faces recognized as similar in one FR model remain so in others, and vice versa. This explains why adversarial faces from NAFG transfer to some extent. Prior studies represent this in terms of the magnitude of similarity scores, but they face a subtle issue: *what if the scales of models' similarity scores do not match?*

Instead, we take a different view in terms of the shared nearest-neighbor structure: faces ranked near the reference by the source FR model E_S should remain relatively close under the target FR model E_T. We state this relationship using gallery ranks, without relying on the absolute scale of similarity scores. For a gallery $\mathcal{G}$, let $\mathbf{r}_E(x;\mathcal{G}) := (r_E(g \mid x,\mathcal{G}))_{g\in\mathcal{G}}$ collect the normalized descending-similarity ranks defined in Section 3.3, and let $\widetilde{\mathbf{r}}_E$ denote its mean-centered version. Let P_{ref} denote the distribution of reference images.

Conjecture 1 (Cross-model transfer of relative ordering)**.** *For independently trained source and target models, define their reference-conditioned rank correlation and conjecture that*

$$\rho_{S,T}(x;\mathcal{G}) := \frac{\langle \widetilde{\mathbf{r}}_{E_S}, \widetilde{\mathbf{r}}_{E_T} \rangle}{\|\widetilde{\mathbf{r}}_{E_S}\|_2 \|\widetilde{\mathbf{r}}_{E_T}\|_2}, \quad \mathbb{E}_{x\sim P_{\text{ref}}}[\rho_{S,T}(x;\mathcal{G})] > 0. \tag{3}$$

Figure 3(a) first supports this relation across four comparison models on an attribute-unfiltered gallery, while permuting the reference–image pairing removes it. This broad correlation does not establish that the relation persists after restricting the gallery to a chosen class and to the upper source ordering on which the attack relies.

Chosen-Class Selection and Composition. Within each $\mathcal{I}_c$, let $B_1^S, \ldots, B_5^S$ be equal-size groups at progressively later positions in the source ordering. For a comparison model E_M, we measure the mean reference similarity $\bar{s}_M(B)$ of the individual members and the similarity $s_M^{\text{comp}}(B)$ of their normalized mean. The comparison set includes the other two available surrogates, AdaFace and TopoFR, and the two models excluded from selection, TransFace and KPRPE. The first quantity tests whether source-model selection remains informative under E_M. The second tests composition and does not follow from the first because $s_M^{\text{comp}}(B) = |B|\bar{s}_M(B)/\|\sum_{g\in B} E_M(g)\|_2$ also depends on alignment among the members. Figure 3(b) shows that member similarity decreases steadily under all four comparison models. Composition similarity remains nearly unchanged across the first three groups and decreases for the two later groups under all four models. Thus, the source ordering remains informative after composition, although member ordering alone does not determine the composed similarity. Appendix C.1 rules out exact neighbourhood overlap: despite low overlap, high source-ranked intervals remain near the top of both unseen-target orderings across surrogates.

Multi-Surrogate Selection. Under Conjecture 1, each surrogate provides a transferable ordering, while the limited overlap shows that its highest-ranked set remains model-specific. Each surrogate ordering therefore carries the part it shares with the target together with a deviation of its own, and an

attacker who cannot query the target has no way to tell whose deviation is smallest for it. Averaging the orderings leaves the shared part in place while deviations that differ between surrogates partly cancel. Write $E_1, \ldots, E_K$ for the recognition models the attacker holds, with $E_1 = E_S$. We average their normalized ranks rather than their scores, favoring images ranked near the reference across models without assuming that the models' score scales agree. For each E_j we compute the within-class rank $r_{E_j}(g \mid x, \mathcal{I}_c)$ and average these ranks:

$$\bar{r}_K(g \mid x, c) := \frac{1}{K} \sum_{j=1}^{K} r_{E_j}(g \mid x, \mathcal{I}_c), \qquad \mathcal{N}_\ell^{(K)}(x, c) \in \operatorname*{arg\,min}_{\substack{\mathcal{A} \subseteq \mathcal{I}_c \\ |\mathcal{A}| = \ell}} \sum_{g \in \mathcal{A}} \bar{r}_K(g \mid x, c). \tag{4}$$

Thus, rank consensus selects the ℓ images with the smallest mean rank across the available surrogate models. Figure 3(c) compares ArcFace-only selection with three-surrogate rank consensus at $\ell = 2^{10}$. The source row is not another target-model result: it reports the cosine similarity, before decoding, between the ArcFace representation of the reference and the normalized mean of the selected gallery images encoded by ArcFace. Consensus increases both member and composition similarity under the two unseen models despite reducing this ArcFace similarity. The transfer gain therefore does not arise from moving the construction closer under the source model. Appendices C.2–C.3 show that rank agreement orders transfer where raw cosine does not and that consensus attains the highest mean agreement with both unseen targets.

5 Evaluation

5.1 Experimental Setup

Datasets, Attribute Labels, and Models. The gallery is the cleaned WebFace42M (Zhu et al., 2021), partitioned into Male, Female, White, Black, and Asian galleries by using FaceXFormer (Narayan et al., 2025) as an attribute classifier. Reference images are taken from LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016), and AgeDB (Moschoglou et al., 2017). FaceXFormer is used only to form each class gallery $\mathcal{I}_c$, whereas FairFace (Karkkainen & Joo, 2021) determines the attribute of every reference and generated image. A reference is evaluated against every chosen class different from its FairFace label, yielding 115,390 reference-class pairs.

To instantiate the proposed method, we use a fixed surrogate pool comprising ArcFace (Deng et al., 2019), AdaFace (Kim et al., 2022), and TopoFR-R100 (Dan et al., 2024). We denote Ours as the attack that uses ArcFace only, whereas Ours-C as that uses all three surrogate models. We fix $\ell = 1024$ by default. The generated faces are then evaluated by ArcFace, AdaFace, TopoFR-R100, TransFace-L (Dan et al., 2023), and ViT-KPRPE (Kim et al., 2024a). For the inverse model, we employ Arc2Face (Papantoniou et al., 2024), which is aligned with ArcFace. Appendices A.1-A.2 detail model roles and evaluation pairs; Appendix D.3 varies the gallery data and size.

Evaluation Metric. For open-source models, we report results at several FMR levels α, denoting the corresponding threshold as τ_α. We calibrate τ_α from the empirical upper tail of cross-identity similarities on CASIA-WebFace (Yi et al., 2014). We estimate strict-FMR thresholds by repeatedly sampling one image per identity and averaging cross-identity quantiles. For the evaluated pair set $\mathcal{P}$, we define the attack success rate (ASR) as $\mathrm{ASR}_{T,\alpha} = |\mathcal{P}|^{-1} \sum_{(x,c)\in\mathcal{P}} \mathbf{1}[a(\hat{x}_{x,c}) = c]\mathbf{1}[\langle E_T(\hat{x}_{x,c}), E_T(x)\rangle \geq \tau_\alpha]$. On the other hand, the commercial services report similarity on their own scale rather than as a cosine, so we use the scores and thresholds each provider defines: AWS (Default: 80, Strict: 99) and Tencent (Default: 50, Strict: 60). AWS recommends its strict score for law-enforcement use, and Tencent documents its two scores as FMRs of 10^{-4} and 10^{-5}. Both services are queried on the same 300 reference-class pairs per chosen class under every construction. Threshold calibration details are in Appendix A.3.

5.2 Experimental Analysis

Open-Source and Commercial Evaluation. Table 1 compares NAFG (Kim et al., 2025) and Ours(-C) using the same reference images, inverse model, and operating points. Ours consistently improves ASR on both open-source targets and commercial services, with larger gains at stricter

Table 1: ASR against two open-source target models and two commercial services. Open-source columns are headed by the FMR, commercial ones by each provider's own scores (default, strict)

		Open source						Commercial			
		TransFace-L			ViT-KPRPE			AWS		Tencent	
Class	Method	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}	80	99	50	60
Male	NAFG	78.3%	51.0%	1.0%	77.1%	42.3%	0.3%	45.3%	9.0%	31.0%	18.0%
	Ours	86.4%	78.6%	15.2%	90.1%	87.1%	7.8%	74.7%	38.3%	83.7%	68.0%
	Ours-C	**95.0%**	**91.0%**	**24.6%**	**96.2%**	**94.6%**	**7.9%**	**94.0%**	**48.0%**	**95.0%**	**78.7%**
Female	NAFG	62.8%	32.0%	0.2%	71.1%	35.6%	0.1%	38.0%	8.0%	29.0%	17.0%
	Ours	86.0%	70.2%	5.6%	96.3%	87.6%	2.4%	70.7%	30.7%	78.3%	62.7%
	Ours-C	**95.6%**	**87.0%**	**11.9%**	**98.4%**	**93.6%**	**2.7%**	**90.7%**	**40.7%**	**94.0%**	**80.0%**
White	NAFG	73.4%	47.7%	1.5%	73.1%	45.5%	0.4%	50.7%	10.3%	29.0%	14.0%
	Ours	83.9%	76.9%	23.7%	89.0%	85.2%	19.4%	84.3%	47.0%	78.3%	66.7%
	Ours-C	**89.2%**	**85.4%**	**33.6%**	**90.2%**	**88.4%**	**22.1%**	**94.3%**	**55.0%**	**92.7%**	**77.0%**
Black	NAFG	53.6%	24.7%	0.1%	53.1%	21.3%	0.1%	27.7%	6.0%	26.3%	17.7%
	Ours	74.2%	50.8%	1.4%	87.9%	67.6%	0.9%	68.7%	29.3%	64.3%	46.7%
	Ours-C	**85.6%**	**65.7%**	**2.4%**	**91.2%**	**73.0%**	**1.0%**	**83.3%**	**35.3%**	**76.7%**	**53.3%**
Asian	NAFG	50.3%	21.2%	0.0%	58.4%	24.6%	0.0%	25.0%	4.0%	14.7%	6.7%
	Ours	69.8%	46.5%	1.2%	86.5%	64.0%	0.2%	59.3%	22.3%	51.0%	32.0%
	Ours-C	**86.4%**	**66.8%**	**2.6%**	**92.7%**	**74.2%**	**0.3%**	**78.0%**	**29.3%**	**63.7%**	**42.7%**

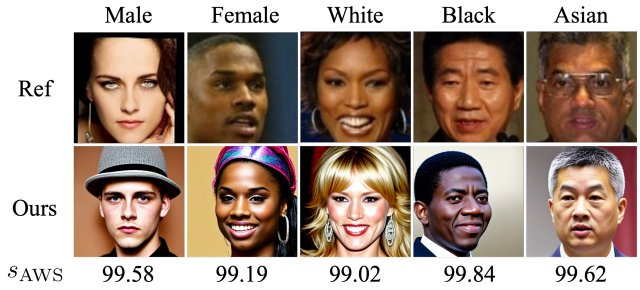


Figure 4: Reference images (top) and generated images (bottom). All outputs receive the chosen FairFace label and exceed the AWS threshold (default: 80, strict: 99).

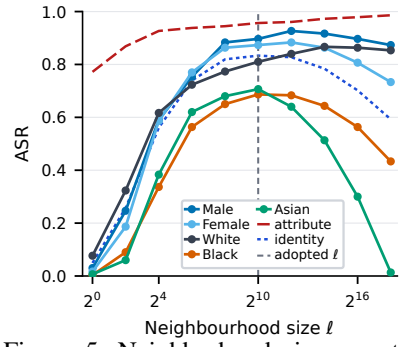


Figure 5: Neighborhood size swept with all other settings fixed, evaluated on TransFace-L at FMR 10^{-5}.

open-source operating points. The classwise analysis in Appendix D.1 separates the two ASR factors. The surrogate-subset analysis in Appendix D.6 shows that the full pool performs best on both unseen targets despite different best single models. Figure 4 presents one successful reference-output pair for each chosen class, with additional examples in Appendix D.8.

Effect of Neighborhood Size ℓ. Figure 5 sweeps ℓ, which determines how far the construction extends down the reference-conditioned ranking. Identity acceptance initially increases, but broader neighborhoods reduce ASR for some classes. The five-model sweep in Appendix D.5 finds different maximizing depths across classes and models, yet supports $\ell = 2^{10}$ as a stable common setting. We therefore fix it throughout the evaluation without any class- or target-specific tuning.

Effect of the Choice of Inverse Models D. To evaluate sensitivity to the inverse model, we replace Arc2Face with modified NbNet (Paik et al., 2026a) under the same setting as Table 1. Absolute ASR changes, yet the ordering NAFG $<$ Ours $<$ Ours-C is preserved in almost every cell of Table 2 across open-source and commercial targets. The comparison holds across both of the inverse models; Appendix D.7 evaluates a different inverse model aligned with another recognition model.

Additional Analyses. To test whether identity impersonation extends beyond the particular reference image used for construction, Appendix D.1 re-verifies the outputs against another image of the same identity; Ours-C remains highest. Appendix D.4 further compares Ours(-C) against NAFG's

Table 2: ASR with NbNet as the inverse model across open-source and commercial targets.

		Open source						Commercial			
		TransFace-L			ViT-KPRPE			AWS		Tencent	
Group	Method	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}	80	99	50	60
Gender	NAFG	88.1%	70.1%	2.5%	89.0%	64.2%	0.7%	51.0%	10.8%	36.7%	22.3%
	Ours	88.7%	85.6%	39.2%	90.0%	88.5%	**18.1%**	85.0%	41.2%	89.8%	80.2%
	Ours-C	**95.1%**	**94.1%**	**54.8%**	**95.4%**	**94.7%**	17.8%	**93.0%**	**43.5%**	**96.2%**	**86.8%**
Race	NAFG	81.8%	58.3%	2.0%	78.8%	49.1%	0.8%	46.3%	5.9%	35.8%	19.9%
	Ours	94.1%	85.9%	25.3%	96.1%	88.9%	14.3%	77.9%	36.3%	76.6%	60.6%
	Ours-C	**97.6%**	**93.5%**	**33.1%**	**97.5%**	**92.1%**	**15.1%**	**88.3%**	**42.6%**	**86.1%**	**68.7%**

target-query variant using the same reference images and AWS thresholds; queried NAFG leads at the default threshold, whereas query-free Ours-C leads at the strict threshold.

Possible Mitigation in the Representation Space. The proposed attack's success raises a question prior to post-hoc defense: which property of the representation space makes such images findable? Our method relies on cross-model agreement in reference-conditioned gallery orderings, formalized by Conjecture 1. The positive relation between agreement strength and ASR observed where both are measured supports this transfer account and points to a representation-level defense, as detailed in Appendix C.2. We therefore test a prototype target whose agreement with every other model is less than half the weakest among them. At FMR 10^{-4}, Ours-C reaches at most 28.7%; Table 1 reports at least 85.6% on both open-source targets. However, the analysis in Appendix C.4 shows that weaker agreement is not free: the prototype's true match rate falls below both open-source targets at 10^{-4} and 10^{-5}, and it still shares part of the ordering at every reference. Separating what the attack consumes from what the FR model needs is therefore a hard problem, and how far the ordering can be misaligned without that loss remains open.

6 Related Works

Adversarial Attacks on Face Recognition. A broad range of adversarial attacks and threat models has been studied for face recognition (Sharif et al., 2016; Dong et al., 2019; An et al., 2023; Jia et al., 2022). Among these, attacks designed for impersonation can satisfy our output criterion when an image from the chosen attribute class is used as the source and the reference identity as the target. Although this comparison introduces a designated source image absent from our protocol, we evaluate representative attacks (Li et al., 2023; Salar et al., 2025) in Appendix D.2. Despite this functional overlap, the problems differ in formulation: these attacks optimize a designated source for impersonation, whereas we exploit the geometry of FR representations to construct a new representation satisfying the attribute and identity objectives jointly. NAFG first studied this joint objective; we introduce a new reference-conditioned construction and analyze its transferability to unseen models.

Attribute Information in Face Representations. Although FR models are trained for identity discrimination, their representations retain demographic attributes and other facial properties (Terhörst et al., 2020). Prior work has studied gender–identity organization (Hill et al., 2019), attribute geometry (Leroy et al., 2025), and interpretable directions related to facial appearance (Plesh et al., 2024). These studies establish that attribute information is measurable and geometrically structured within FR representations. We evaluate the security consequences of this structure by asking whether it can support a chosen attribute while maintaining the reference identity.

7 Conclusion

In this paper, we study how the coexistence of identity and attribute information in learned face representations can be exploited to craft adversarial faces. Based on our analysis results of attribute manifolds, we propose a new, query-free adversarial face generation method with the reference-conditioned composition technique. We demonstrate the effectiveness of our method across multiple open-source and commercial FR models, achieving non-trivial ASR on both usual and strict threshold settings. More broadly, our findings motivate examining how unintended information in learned representations can affect the trustworthiness of vision systems beyond face recognition.

AI USE STATEMENT

In this work, we used generative AI tools, including Anthropic Claude Opus 5 and OpenAI ChatGPT 5.6 Sol, to assist with literature retrieval and to edit and polish author-provided manuscript text. We also used these tools to assist with the implementation of author-designed experiments, including drafting and debugging data-processing, evaluation, and visualization code. The authors checked the retrieved literature against primary sources, reviewed and tested AI-assisted code, and verified the reported results against the corresponding experimental artifacts. The authors made the final decisions regarding the experimental design, interpretation, claims, and manuscript content, and take responsibility for the final content of this work.

ETHICS STATEMENT

Research Purpose and Societal Relevance. This work studies a security failure mode in face recognition: a generated face may receive a chosen race or gender label while being accepted as a reference identity. Our purpose is to characterize how identity and attribute information jointly encoded in learned representations can produce this failure, and to support the evaluation and mitigation of such risks. The experiments are not used for operational identification, access decisions, or decisions about any depicted individual. Characterizing this attack surface under a query-free threat model can inform representation auditing, verification policies, and future defenses.

Interpretation of Demographic Attributes. Race and gender labels are used as operational categories for constructing and evaluating the attack, not as biological definitions or universal demographic ground truth. They derive from dataset annotations and automated classifiers rather than self-identification, and their definitions, coverage, and errors may differ across groups. Differences in attack outcomes across the reported classes may therefore reflect the data and attribute evaluators and should not be interpreted as evidence that any demographic group is inherently more vulnerable. Our conclusions are limited to the stated labeling and evaluation protocols and make no claims about the identities or attributes of depicted individuals.

Data, Models, and Commercial Services. This study did not recruit participants, interact with individuals, or collect new human-subject data. We used existing face datasets for academic research, and controlled-access datasets were obtained through the providers' prescribed institutional channels. Third-party models were obtained from public research releases or authorized contributors and were used subject to the applicable terms. Commercial API experiments were limited to pairwise face comparisons, did not create persistent enrollments or identity databases, and were conducted using the available provider privacy settings.

Potential Misuse and Responsible Disclosure. Although this work is intended to support security analysis and defense, the method could be misused to facilitate impersonation or deceptive demographic presentation. No deployed system was attacked: the commercial evaluation consists of pairwise comparisons of benchmark images, which create no enrollment and involve no person outside those datasets. We disclosed the findings to the evaluated providers before submission, and the review artifact withholds the datasets, the gallery, and the model weights that a run would require.

REPRODUCIBILITY STATEMENT

Section 2 defines the threat model and joint success criterion, Section 4 specifies the proposed constructions, and Section 5 describes the datasets, model roles, evaluation metrics, and operating points. The appendix provides the preprocessing and gallery-construction procedures, threshold calibration, hyperparameters, commercial evaluation protocols, and complete class-wise results and ablations. An anonymous review repository[1] provides the construction and the evaluation code, the fixed configurations, the calibrated thresholds, the environment specification, and instructions for obtaining each dataset and model. It stops where a representation would be handed to an inverse model, which is third-party work that the artifact neither carries nor drives. Third-party datasets and model weights, the attribute-partitioned gallery, and the intermediate representation and generated-image collections are not redistributed, because of their access conditions, licenses, and storage

[1] https://anonymous.4open.science/r/reference-conditioned-composition-EEE6

requirements. The repository is an anonymized copy made for review and is not permanent; whether a public release follows, and what it would contain, we leave to the recommendation of the review.

REFERENCES

Shengwei An, Yuan Yao, Qiuling Xu, Shiqing Ma, Guanhong Tao, Siyuan Cheng, Kaiyuan Zhang, Yingqi Liu, Guangyu Shen, Ian Kelk, and Xiangyu Zhang. Imu: Physical impersonating attack for face recognition system with natural style changes. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 899–916. IEEE, 2023.

AWS. Use cases that involve public safety. Amazon Rekognition Developer Guide, 2026. URL `https://docs.aws.amazon.com/rekognition/latest/dg/considerations-public-safety-use-cases.html`. Accessed: 2026-09-24.

AWS. Amazon rekognition, 2026. URL `https://aws.amazon.com/rekognition/`. Accessed: 2026-09-23.

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

Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20585–20596. IEEE, 2023.

Jun Dan, Yang Liu, Jiankang Deng, Haoyu Xie, Siyuan Li, Baigui Sun, and Shan Luo. Topofr: A closer look at topology alignment on face recognition. *Advances in Neural Information Processing Systems*, 37:37213–37240, 2024.

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.

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7706–7714. IEEE, 2019.

Matthew Q Hill, Connor J Parde, Carlos D Castillo, Y Ivette Colon, Rajeev Ranjan, Jun-Cheng Chen, Volker Blanz, and Alice J O'Toole. Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, 1(11):522–529, 2019.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

Chanwoo Hwang, Sunpill Kim, Yong Kiam Tan, Tianchi Liu, Seunghun Paik, Dongsoo Kim, Mondal Soumik, Khin Mi Mi Aung, and Jae Hong Seo. Scores know bobs voice: Speaker impersonation attack. *arXiv preprint arXiv:2603.02781*, 2026.

Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34136–34147, 2022.

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

Changjin Kim, Seunghun Paik, Dongsoo Kim, and Jae Hong Seo. Breaking high confidence: Practical face impersonation under high-security thresholds. In *European Conference on Computer Vision*, pp. 551–570. Springer, 2026.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18729–18738. IEEE, 2022.

Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–255, 2024a.

Sunpill Kim, Yong Kiam Tan, Bora Jeong, Soumik Mondal, Khin Mi Mi Aung, and Jae Hong Seo. Scores tell everything about bob: Non-adaptive face reconstruction on face recognition systems. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1684–1702. IEEE, 2024b.

Sunpill Kim, Seunghun Paik, Chanwoo Hwang, Minsu Kim, and Jae Hong Seo. Non-adaptive adversarial face generation. *Advances in Neural Information Processing Systems*, 38:37315–37354, 2025.

Pierrick Leroy, Antonio Mastropietro, Marco Nurisso, and Francesco Vaccarino. Attributes shape the embedding space of face recognition models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33960–33983. PMLR, 2025.

Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24626–24637. IEEE, 2023.

Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5): 1188–1202, 2018.

David McNeely-White, Ben Sattelberg, Nathaniel Blanchard, and Ross Beveridge. Canonical face embeddings. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(2):197–209, 2022.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 51–59, 2017.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11369–11382, 2025.

Huy H Nguyen, Sébastien Marcel, Junichi Yamagishi, and Isao Echizen. Master face attacks on face recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3): 398–411, 2022.

Seunghun Paik, Chanwoo Hwang, Sunpill Kim, and Jae Hong Seo. On the reversibility of locality-sensitive hashing-based biometric template protections. *IEEE Transactions on Dependable and Secure Computing*, 23(2):3477–3494, 2026a.

Seunghun Paik, Sunpill Kim, Chanwoo Hwang, and Jae Hong Seo. Casting the net! revisiting masterface impersonation attacks. *arXiv preprint arXiv:2608.06952*, 2026b.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, pp. 241–261. Springer, 2024.

Richard Plesh, Janez Križaj, Keivan Bahmani, Mahesh Banavar, Vitomir Štruc, and Stephanie Schuckers. Discovering interpretable feature directions in the embedding space of face recognition models. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2024.

Fizza Rubab, Yiying Tong, and Arun Ross. Are face embeddings compatible across deep neural network models? *arXiv preprint arXiv:2604.07282*, 2026.

Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. Enhancing facial privacy protection via weakening diffusion purification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8235–8244. IEEE, 2025.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9. IEEE, 2016.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540, 2016.

Tencent. Tencent cloud, 2026. URL `https://www.tencentcloud.com/`. Accessed: 2026-09-23.

Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Beyond identity: What information is stored in biometric face templates? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2020.

Kanishka P Wijewardena, Steven A Grosz, Kai Cao, and Anil K Jain. Fingerprint template invertibility: Minutiae vs. deep templates. *IEEE Transactions on Information Forensics and Security*, 18:744–757, 2022.

Haiyu Wu, Jaskirat Singh, Sicong Tian, Liang Zheng, and Kevin Bowyer. Vec2face: Scaling face dataset generation with loosely constrained vectors. In *International Conference on Learning Representations*, volume 2025, pp. 23490–23503, 2025.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Mipgan—generating strong and high-quality morphing attacks using identity prior-driven gan. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):365–383, 2021.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10487–10497. IEEE, 2021.

A Experimental Protocol

This section specifies the model roles and construction settings, the data and eligible evaluation pairs, and the operating points used to interpret verification scores. Each experiment was run on a single NVIDIA L40S GPU with 48 GB of memory.

A.1 Models and Construction Settings

Face Recognition Models and Construction Roles. We evaluate generated images with five open-source recognition models, all of which return a 512-dimensional unit representation. ArcFace, AdaFace, and TopoFR-R100 also form the surrogate pool available to the attacker, whereas TransFace-L and ViT-KPRPE are used only as held-out targets. Ours ranks the class gallery with ArcFace, while Ours-C averages percentile ranks from all three surrogates. Both constructions uniformly average the selected images' ArcFace representations, so the representation supplied to the inverse model remains in the ArcFace space. NAFG uses $k = 100$, and both proposed constructions use $\ell = 2^{10}$ throughout the main evaluation. Table 3 records these roles together with the model specifications.

Table 3: Face recognition, inverse models, and commercial APIs and their experimental roles.

Model	Role	Architecture / objective	Training set
ArcFace (Deng et al., 2019)	surrogate, composer, target	ResNet-100 / ArcFace	WebFace42M
AdaFace (Kim et al., 2022)	surrogate, target	IResNet-101 / AdaFace	WebFace12M
TopoFR-R100 (Dan et al., 2024)	surrogate, target	ResNet-100 / TopoFR	MS1MV2
TransFace-L (Dan et al., 2023)	held-out target	ViT-L / TransFace	MS1MV2
ViT-KPRPE (Kim et al., 2024a)	held-out target	ViT-B / KPRPE	WebFace12M
Arc2Face (Papantoniou et al., 2024)	default inverse	Latent diffusion / ID conditioning	WebFace42M
NbNet (Paik et al., 2026a)	inverse ablation	Convolutional / perceptual	MS1MV3
AWS Rekognition	commercial target	`CompareFaces`	not disclosed
Tencent Cloud	commercial target	`CompareFace`	not disclosed

Inverse Models. Arc2Face and NbNet receive a unit ArcFace representation and generate an image at 512×512 and 128×128, respectively. Arc2Face is the default and uses four consistency-sampler steps, guidance 1.0, and seed 0 for every image. NbNet is deterministic and is evaluated in Appendix D.7.

A.2 Datasets and Evaluation Pairs

Table 4 lists every dataset and the role it plays. Reference inputs are the image occurrences in the standard verification pairs of LFW, CFP-FP, and AgeDB-30.

Table 4: Datasets and their experimental roles. TP and FP denote the genuine and impostor pairs in each verification protocol.

Dataset	Role	Images	Identities	TP / FP
LFW (Huang et al., 2008)	reference pool	13,233	5,749	3,000 / 3,000
CFP-FP (Sengupta et al., 2016)	reference pool	7,000	500	3,500 / 3,500
AgeDB-30 (Moschoglou et al., 2017)	reference pool	16,488	568	3,000 / 3,000
WebFace42M (Zhu et al., 2021)	attacker gallery	41,857,480	2,029,906	–
CASIA-WebFace (Yi et al., 2014)	threshold calibration	490,623	10,572	–

Attribute Labels. FaceXFormer (Narayan et al., 2025) partitions WebFace42M into the class galleries $\mathcal{I}_c$, and a gallery image enters $\mathcal{I}_c$ only when its posterior for c reaches 0.9999. FairFace (Karkkainen & Joo, 2021) independently assigns the attribute of every reference and generated

image using its four-race checkpoint. Using separate models for gallery construction and output evaluation avoids judging generated images with the classifier that admitted their source material.

A reference is eligible for intended class c when FairFace does not already assign c to it, so no pair asks a construction to produce an attribute the reference already has. Table 5 gives the resulting counts against each class gallery after the purity gate. Pair count and gallery size move in opposite directions across classes, so class-level rates rest on very different sample sizes.

Table 5: Eligible reference–class pairs by reference dataset and intended class.

Class	LFW	CFP-FP	AgeDB-30	Pairs	Class gallery
Male	3,132	4,780	5,409	13,321	21,651,320
Female	8,868	9,220	6,591	24,679	10,640,016
White	2,335	3,400	485	6,220	8,084,681
Black	11,129	12,069	11,905	35,103	1,592,275
Asian	10,952	13,262	11,853	36,067	553,099
Total	36,416	42,731	36,243	115,390	

Open-source models are evaluated on all 115,390 eligible pairs. For each intended class, the commercial evaluation uses a fixed sample of 300 eligible pairs shared by every method.

A.3 Operating Points and Score Scales

Open-Source Operating Points. The verification benchmarks cannot supply a threshold at the rates this paper reports, since LFW carries 3,000 negative pairs and a threshold at 10^{-5} would rest on a fraction of one pair. We therefore follow the calibration protocol of Kim et al. (2026): draw one image per identity, score every cross-identity pair, and read the threshold from that distribution. Since two of the models are trained on MS1MV2, we calibrate on CASIA-WebFace, which is not MS1M-derived. We use all 10,572 identities, giving 55,878,306 impostor pairs in each of 1,000 repetitions and placing the 10^{-6} threshold on the 56th largest score per repetition. Table 6 reports the mean and standard deviation of each threshold across repetitions. The spread stays below 0.0015 down to 10^{-5} and widens by an order of magnitude at 10^{-6}, where the estimate rests on the extreme tail. Thresholds also differ across models at the same rate, which is why a raw similarity is not comparable between them.

Table 6: Impostor thresholds calibrated on CASIA-WebFace, mean and standard deviation over repetitions.

	false match rate			
Model	10^{-3}	10^{-4}	10^{-5}	10^{-6}
ArcFace	0.1989 ± 0.0002	0.2539 ± 0.0004	0.3108 ± 0.0011	0.5483 ± 0.0166
AdaFace	0.1973 ± 0.0002	0.2528 ± 0.0004	0.3103 ± 0.0012	0.5401 ± 0.0167
TopoFR-R100	0.2281 ± 0.0002	0.2909 ± 0.0004	0.3534 ± 0.0012	0.5557 ± 0.0191
TransFace-L	0.2443 ± 0.0002	0.3043 ± 0.0004	0.3635 ± 0.0011	0.5463 ± 0.0184
ViT-KPRPE	0.1933 ± 0.0002	0.2466 ± 0.0003	0.3025 ± 0.0012	0.5357 ± 0.0165

True Match Rate (TMR) at Calibrated Thresholds. Table 7 reports the true match rate on each reference pool. Every model stays above 87% down to 10^{-5}, but at 10^{-6} the rate falls to 44–52% on CFP-FP and 26–46% on AgeDB-30. The 10^{-6} column of every attack table should be read as a regime a deployment would adopt only where a false match is far more costly than a rejection.

Table 7: True match rate (%) of each model on the three reference pools at the calibrated operating points.

	LFW				CFP-FP				AgeDB-30			
Model	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
ArcFace	99.70	99.70	99.63	93.93	98.74	98.34	97.29	52.00	96.97	95.67	93.27	31.73
AdaFace	99.73	99.73	99.63	94.30	98.71	98.29	96.34	50.69	96.90	95.67	92.53	31.97
TopoFR-R100	99.73	99.70	99.57	95.97	97.60	95.74	91.49	48.83	96.77	95.20	91.97	45.77
TransFace-L	99.70	99.67	99.50	96.03	96.66	93.57	87.83	44.31	96.30	94.47	90.57	44.43
ViT-KPRPE	99.67	99.67	99.60	94.13	98.69	98.00	96.31	46.83	96.43	95.03	91.73	25.97

Commercial Operating Points. We query AWS Rekognition at face model version 7 and Tencent Cloud with `FaceModelVersion` 3.0, as deployed in September 2026. AWS names score 80 as its default and recommends 99 where a false match is most costly, but publishes no false match rate for either score. Tencent names 50 as its decision point and documents scores 50 and 60 at false match rates of 10^{-4} and 10^{-5}, respectively. Each generated image is compared with its reference once, and the returned score is stored without conversion. A detection failure or returned non-match counts as an attack failure.

B The Global Subsphere

This section supplements Section 3 by testing its zero-mean approximation, repeating its measurements across models, checking validity under the distribution induced by projection, and isolating the effect of the reference–gallery similarity component excluded by projection.

B.1 The Zero-Mean Assumption

The subsphere $\mathbb{S}_c^k$ of Section 3 passes through the origin, whereas the class representations it approximates have a non-zero mean μ_c. We measure this mean, derive the closest point on the corresponding affine set, and evaluate its effect on attack success.

Magnitude of the Class Mean. Table 8 gives $\|\mu_c\|_2$ and the part of it the basis does not span, $m_{c,k}^{\perp} := (I - P_{c,k})\mu_c$. The mean is short on the unit sphere, between 0.029 and 0.058 across the five classes, and shorter still outside the span.

The Exact Closest Point. Accounting for the mean replaces $\mathrm{span}(B_{c,k})$ with the affine set $\mathcal{A}_{c,k} = \mu_c + \mathrm{span}(B_{c,k})$. We seek the point of $\mathcal{A}_{c,k} \cap \mathbb{S}^{d-1}$ closest to the reference representation $u = E(x)$. Since $P_{c,k}\mu_c$ already lies in the span, $\mathcal{A}_{c,k} = m_{c,k}^{\perp} + \mathrm{span}(B_{c,k})$, and $m_{c,k}^{\perp}$ is orthogonal to that span. A point $m_{c,k}^{\perp} + B_{c,k}w$ is therefore a unit vector exactly when $\|w\|_2 = \eta_{c,k} := \sqrt{1 - \|m_{c,k}^{\perp}\|_2^2}$, so the intersection is a sphere of radius $\eta_{c,k}$ inside $\mathcal{A}_{c,k}$. Maximizing $\langle u, m_{c,k}^{\perp} + B_{c,k}w \rangle$ over that sphere leaves the first term fixed and aligns w with $B_{c,k}^{\top}u$, giving

$$z_{c,k}^{\mathrm{aff}} = m_{c,k}^{\perp} + \eta_{c,k} \frac{P_{c,k}u}{\|P_{c,k}u\|_2} = m_{c,k}^{\perp} + \eta_{c,k}\, z_{c,k}^{\mathrm{proj}}. \tag{5}$$

The affine point is the projected representation scaled by $\eta_{c,k}$ and offset by $m_{c,k}^{\perp}$. Because the offset is orthogonal to the span, $\langle z_{c,k}^{\mathrm{proj}}, z_{c,k}^{\mathrm{aff}} \rangle = \eta_{c,k}$. Table 8 places $\eta_{c,k}$ above 0.999 in every class, corresponding to an angular difference of at most 2.3 degrees. We next test whether this difference changes the attack outcome.

Table 8: Class mean, affine-sphere radius $\eta_{c,k}$, and ASR after mean correction. Classes are ordered by gallery size.

					$\alpha = 10^{-4}$		$\alpha = 10^{-5}$	
Class	Gallery	$\|\mu_c\|_2$	$\|m_{c,k}^{\perp}\|_2$	$\eta_{c,k}$	NAFG	Affine	NAFG	Affine
Male	21,651,320	0.0288	0.0213	0.99977	73.7%	73.7%	42.7%	43.0%
Female	10,640,016	0.0327	0.0226	0.99975	58.7%	59.0%	30.7%	30.7%
White	8,084,681	0.0399	0.0284	0.99960	83.0%	83.0%	59.7%	59.7%
Black	1,592,275	0.0424	0.0256	0.99967	55.3%	57.0%	28.7%	27.7%
Asian	553,099	0.0580	0.0396	0.99922	58.3%	57.3%	29.3%	29.0%

Effect on Attack Success Rate. The right-hand columns of Table 8 compare both methods on the same 300 references per class under the same basis, gate, and inverse model. The absolute ASR difference is at most 1.7 percentage points, the attribute-assignment difference is at most one point, and neither construction leads consistently. Under the evaluated configuration, accounting for the class mean therefore produces no consistent or material change in attack success, supporting the zero-mean approximation used in the main analysis.

B.2 Generality Across Models and Projection Inputs

Figure 1 measures coverage, validity, and reference–gallery ranks with one recognition and inverse-model configuration. We repeat each measurement with alternative models while keeping the gate, class galleries, classes, and k grid fixed. We also compare uniform validity with the validity of points induced by projecting out-of-class inputs.

Coverage. Coverage depends only on the recognition model used to represent the class gallery. Figure 6 places coverage above the random baseline at every k under all five recognition models. Class representations are therefore more concentrated in the fitted span than uniformly sampled representations across each evaluated recognition space. The absolute values vary by model, but their increase with k matches the trend in Figure 1.

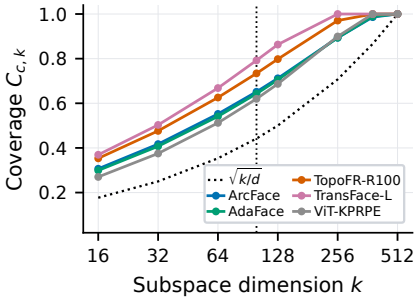


Figure 6: Coverage across recognition models; vertical rule: $k = 100$, unmarked curve: $\sqrt{k/d}$.

Validity under Projection. The uniform probe measures validity across the fitted subsphere, whereas the attack reaches only points induced by projecting out-of-class references. We therefore evaluate the same chosen-class condition over the same k grid at these projected points. As a control, we repeat the projection with images from the other classes in the same attribute family. The projection-induced and uniform family means agree within 0.02 through $k = 128$; at $k = 100$, the projected references attain 0.98 validity for race and 1.00 for gender. The two input sources also agree within 0.02 through $k = 256$. Thus, the high validity at the prior operating dimension is not an artifact of uniform sampling. At larger k, projection approaches the out-of-class input itself, so its subsequent decline does not characterize uniform validity over the subsphere.

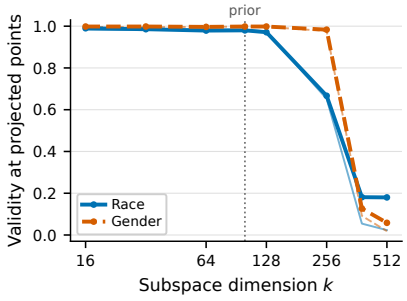


Figure 7: Validity at projected inputs. Bold: evaluation references; faint: other-class gallery images; vertical rule: $k = 100$.

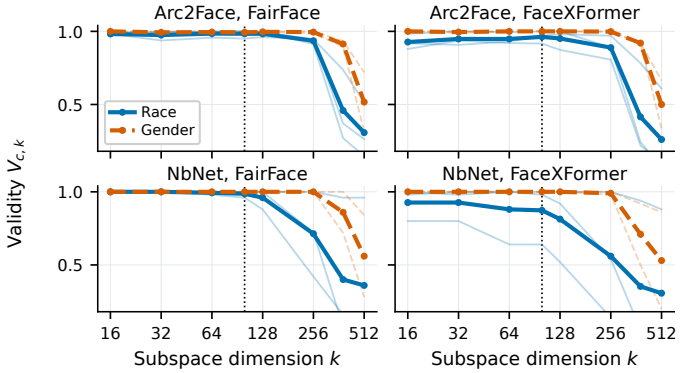


Figure 8: Validity $V_{c,k}$ across the evaluated inverse models and attribute classifiers. Faint curves are classes, bold curves are attribute-family means.

Validity across Measurement Models. Validity additionally depends on the inverse model and attribute classifier used to interpret a sampled representation. Figure 8 shows the same decrease with k under every evaluated pair. The absolute level varies across pairs, but the race classes begin their sharp decline one grid step before the gender classes, and validity at $k = 100$ remains above 0.85 in every panel. The coverage–validity trend is therefore not specific to the measurement pipeline.

Reference–Gallery Ranks. The rank measurement depends only on the recognition model and requires neither an inverse model nor an attribute classifier. Figure 9 reproduces the body observation under the other four recognition models: the post-projection displacement grows among gallery images with greater original similarity to the reference. For the nearest band, the median post-projection percentile ranges from 0.012% to 0.130%, compared with 0.0005% under an unchanged ranking.

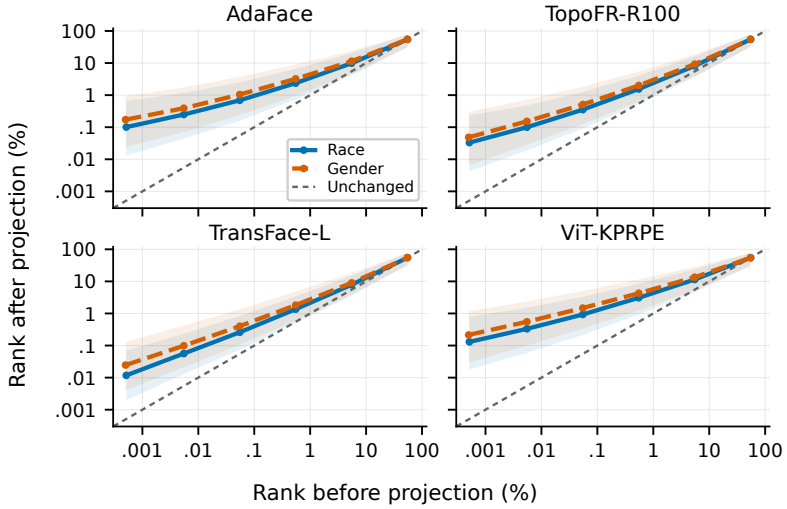


Figure 9: Reference–gallery ranks at $k = 100$ under the four additional recognition models. Curves are attribute-family means, shading gives class-wise interquartile ranges, and the diagonal marks unchanged ranks.

B.3 Effect of the Similarity Component Excluded by Projection

Section 3.3 shows that the k-dimensional class span cannot represent the component of reference–gallery similarity outside that span. The decomposition alone does not establish whether this component changes a representation constructed from the gallery. We isolate its effect on 300 references per class by changing only the score used to choose gallery images.

For each reference, we choose $\ell{=}2^{10}$ class images using either the retained score $\langle P_{c,k}E_S(x), P_{c,k}E_S(g)\rangle$ or the full score $s_S(g,x)$. Both conditions uniformly average the chosen images' original representations and use the same inverse model.

Figure 10 shows class-level transitions in faint lines, and attribute-family means in bold. Using the full score increases generated-image ArcFace similarity in every race and gender class, by 0.111 to 0.174. FairFace chosen-class assignment changes by at most 0.013 in three classes, but falls by 0.134 in White and 0.080 in Male. The excluded similarity component therefore affects which observed representations enter the composition and consistently improves source-model identity similarity, but it does not improve both objectives uniformly.

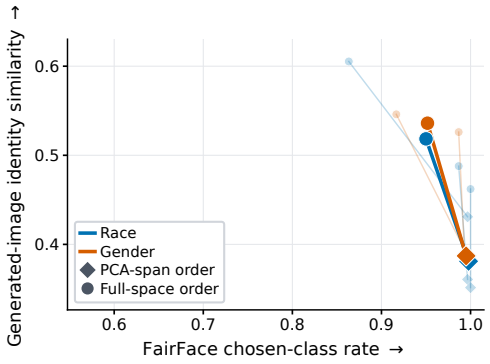


Figure 10: Gallery choice using the full or retained score. Points report chosen-class assignment (x) and generated-image ArcFace similarity to the reference (y).

C Locality and Transfer

This section tests whether transfer requires shared neighborhoods and whether ordering agreement tracks transfer strength. It then evaluates surrogate consensus and a strict-threshold target.

C.1 Transfer Without Shared Neighborhoods

The construction uses a surrogate ordering but is evaluated by an unseen target model. Transfer would be immediate if both models retrieved the same gallery images, so we first test whether exact set overlap explains the result.

Overlap of the Retrieved Sets. For a reference x and class gallery $\mathcal{I}_c$, we compare $\mathcal{N}_\ell^S(x,c)$ with the set each target model would retrieve. Across the three race classes and $\ell \in \{64, 2^{10}\}$, their Jaccard index stays below 0.08 and the surrogate set recovers under 15% of the target set. TransFace-L and ViT-KPRPE overlap still less with each other, reaching a Jaccard index of 0.045 at $\ell{=}64$. The transferred construction hence cannot be explained by both models retrieving the same images.

Placement in the Target Ordering. Exact overlap is unnecessary if the surrogate-ranked images remain concentrated near the top of the target ordering. Figure 11 groups images by their surrogate rank and reports the fraction that falls within the target's top h. The fraction decreases monotonically with surrogate rank, while the first 16 images enter the target's top 2^{10} more than 450 times as often as random images. Both permutation controls remain near the random rate.

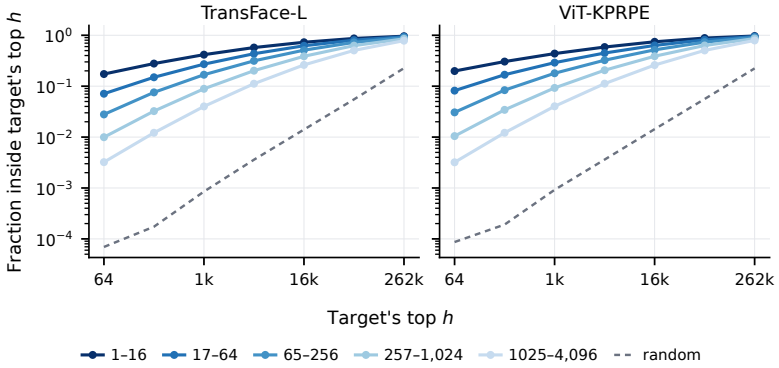


Figure 11: Target-order placement by surrogate-rank interval. Both axes are logarithmic, and the dashed line is the random baseline.

C.2 Ordering Agreement and Transfer Strength

Conjecture 1 requires positive ordering agreement but does not state whether its strength is informative about transfer. We measure $\mathbb{E}_x[\rho_{S,T}(x;\mathcal{G})]$ on an attribute-unfiltered gallery $\mathcal{G}$ of 500,000 images and 300 references. The value is positive for all four comparison models and zero after permuting the reference–gallery pairing. Transfer strength is read on a common axis because raw similarities are not comparable across models. We use the empirical impostor tail $\mathrm{FMR}_T(s) = \Pr[\langle E_T(x_i), E_T(x_j)\rangle \geq s \mid i \neq j]$, estimated from 50 independent repetitions of the CASIA-WebFace protocol of Appendix A.3 and totaling 2.79×10^9 impostor pairs per model. Reading a similarity as $\log_{10} \mathrm{FMR}_T(s)$ places every model on one axis without assuming a parametric tail. Across these four models, stronger ordering agreement coincides with better transfer, giving a Spearman correlation of -1 with mean implied false match rate. The quantity is symmetric on a fixed gallery, and it stays between 0.499 and 0.602 for every surrogate and held-out target pair, so no surrogate is a special choice of source.

Figure 12 contrasts this relation with raw target-model cosine similarity. The raw cosine orders the models differently and is not comparable across models because their operating thresholds differ. Within the evaluated models, ordering agreement therefore tracks transfer more directly than the uncalibrated cosine score. Among the two held-out targets, ViT-KPRPE has stronger ordering agreement with ArcFace and generally higher ASR for Ours and Ours-C than TransFace-L (Table 1).

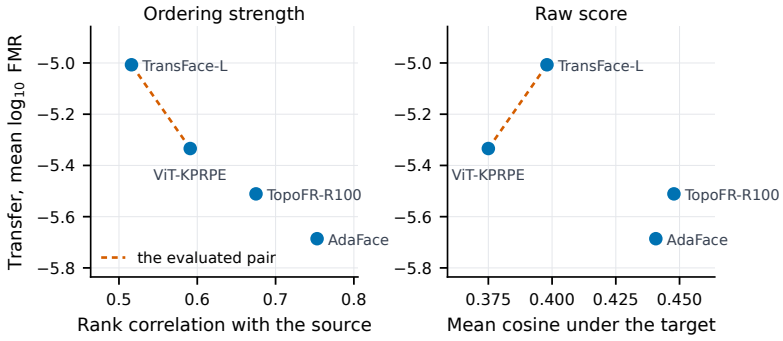


Figure 12: Transfer strength against ordering agreement and raw cosine. Lower on the vertical axis indicates better transfer; the dashed segment joins the two targets evaluated in the body.

C.3 Consensus Across Surrogate Orderings

The attacker cannot know which surrogate ordering agrees best with an unseen target. Equation 4 therefore averages the surrogate ranks rather than committing to one ordering. Figure 13 tests whether this consensus agrees more consistently with both target models.

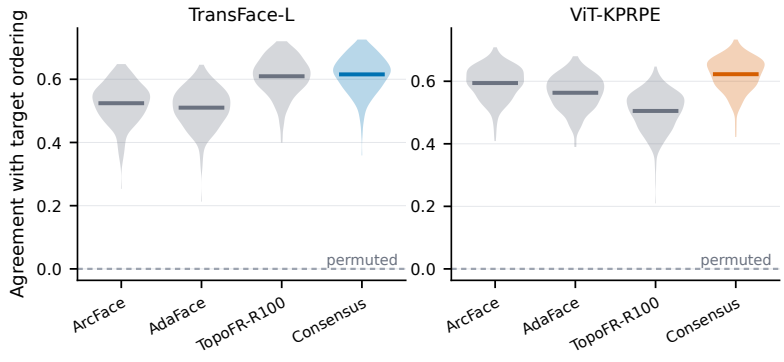


Figure 13: Agreement between each surrogate ordering and the target ordering. Bars are means, and the shapes show reference-level distributions.

The best individual surrogate differs between the two targets, whereas consensus attains the highest mean agreement with both. It also exceeds the best individual surrogate on 195 of 300 references for TransFace-L and 299 of 300 for ViT-KPRPE.

C.4 A Target Trained for a Strict Threshold

This section evaluates a prototype trained in the mitigation direction proposed by Kim et al. (2025) and measures its attack success, ordering agreement, and recognition utility.

The Model. The mitigation direction of Kim et al. (2025) is to train a recognition model for operation at a high acceptance threshold and evaluate the attack against that model. We evaluate an IResNet-100 prototype trained in this direction with a bilateral ArcFace margin of 0.5 and a logit scale annealed from 32 to 12. The prior work reports an Inception ResNet-101, so this prototype tests the training direction rather than reproducing its model. The prototype attains verification accuracies of 99.53% on LFW, 97.18% on AgeDB-30, and 98.51% on CFP-FP. We re-embed the images evaluated in Table 1 under the prototype without reconstructing representations or generating new ones, except for the selection probe of Table 10, which requires its own renders.

What the Strict Threshold Costs. Calibrated as in Appendix A.3, the prototype requires a similarity of 0.5766 at FMR 10^{-4} against 0.3043 for TransFace-L and 0.2466 for ViT-KPRPE. A larger numerical threshold is not by itself a security gain because recognition models have different score scales, so we compare them at matched FMRs. Table 9 shows that, at FMR 10^{-4} and 10^{-5}, the prototype accepts fewer genuine pairs than both main target models on all three reference datasets.

Table 9: True match rate (%) at each model's threshold on the three reference datasets.

	LFW			AgeDB-30			CFP-FP		
Model	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}
TransFace-L	99.7	99.5	96.0	94.5	90.6	44.4	93.6	87.8	44.3
ViT-KPRPE	99.7	99.6	94.1	95.0	91.7	26.0	98.0	96.3	46.8
Prototype	98.1	93.8	75.3	87.4	74.7	42.6	83.1	64.5	32.1

ASR against the Prototype. Table 10 applies the ASR criterion from Table 1 against the prototype, by chosen class, on the same $1{,}500$ reference faces used in evaluation against the commercial APIs. At FMR 10^{-4}, Ours-C falls from 90.0–100.0% on TransFace-L to 5.3–28.7% on the prototype, while Ours and Ours-C remain within 1.3 points in every class. As a diagnostic unavailable to the attacker, the final row ranks the gallery with the prototype but retains ArcFace composition; it raises mean ASR from 19.9% to 26.0% at FMR 10^{-4} yet falls below Ours-C at 10^{-6}. These rates include the accompanying attribute loss, including a White assignment rate of 66.0% compared with at least 90% under the other constructions.

Table 10: ASR against the prototype on 1,500 frozen references, resolved by chosen class. The prototype-selection row ranks with the prototype and composes in the ArcFace space.

Operating point	Method	Male	Female	White	Black	Asian
10^{-4}	NAFG	0.7%	0.7%	0.3%	1.0%	0.7%
	Ours	26.0%	23.0%	27.3%	20.3%	6.0%
	Ours-C	25.0%	21.7%	28.7%	19.0%	5.3%
	Ours, prototype selection	32.3%	32.3%	36.7%	22.0%	6.7%
10^{-5}	NAFG	0.0%	0.0%	0.3%	0.7%	0.0%
	Ours	23.7%	19.0%	21.3%	14.7%	2.0%
	Ours-C	22.3%	17.7%	25.7%	14.3%	3.0%
	Ours, prototype selection	23.3%	25.3%	27.0%	17.0%	4.0%
10^{-6}	NAFG	0.0%	0.0%	0.0%	0.3%	0.0%
	Ours	14.3%	12.0%	10.3%	2.3%	1.0%
	Ours-C	14.0%	13.3%	14.0%	4.3%	1.0%
	Ours, prototype selection	10.0%	15.7%	11.7%	7.0%	1.0%

Ordering Agreement. Measured as in Appendix C.2, the prototype's ordering agreement with each of the five other models is 0.134 to 0.218, against 0.497 to 0.753 for every pair among those five. The agreement remains positive at all 300 evaluated references, with a minimum of 0.044. The prototype therefore lies at the low end of both ordering agreement and attack success, although one target model cannot establish a causal relation between the two.

D Extended Evaluation

This section expands the main comparison between ours and prior works across operating points and evaluates attack baselines, galleries, query access, neighborhood depth, surrogate pools, inverse models, and qualitative outputs.

D.1 The Main Comparison in Full

Complete Operating-Point Grid. Table 11 extends the pooled comparison in the body across the complete grid of operating points.

Table 11: ASRs across the complete operating-point grid on all 115,390 eligible reference–class pairs. Shading marks models used by Ours-C; ArcFace is omitted because every construction uses it. Each α is a false match rate, applied as the threshold τ_α calibrated on that target model.

Method	AdaFace	TopoFR-R100	TransFace-L	ViT-KPRPE
$\alpha = 10^{-3}$				
NAFG	98.7%	96.2%	82.8%	87.2%
Ours	97.7%	96.7%	89.7%	95.9%
Ours-C	98.7%	98.6%	96.1%	97.7%
$\alpha = 10^{-4}$				
NAFG	92.8%	82.8%	58.5%	62.5%
Ours	96.6%	92.4%	77.3%	89.6%
Ours-C	98.6%	98.2%	89.3%	93.7%
$\alpha = 10^{-5}$				
NAFG	69.5%	51.8%	29.4%	29.1%
Ours	91.4%	79.8%	58.2%	74.0%
Ours-C	97.5%	95.2%	74.6%	81.1%
$\alpha = 10^{-6}$				
NAFG	0.2%	0.4%	0.3%	0.1%
Ours	8.0%	9.2%	5.0%	2.8%
Ours-C	14.0%	21.0%	8.7%	3.1%

Attribute and Identity as Separate Factors. ASR is the conjunction of two indicators, and Table 12 reports them apart so that a gain in one is not read as a gain in the other. All methods retain high attribute assignment, whereas Ours and Ours-C substantially increase identity acceptance. The ASR improvement therefore does not result from relaxing the attribute condition.

Table 12: Attribute assignment and identity acceptance underlying ASR. Each identity cell is followed by the corresponding ASR; shading marks models used by Ours-C.

Method	Attribute	AdaFace	TopoFR-R100	TransFace-L	ViT-KPRPE
NAFG	99.5%	69.8 → 69.5%	52.1 → 51.8%	29.6 → 29.4%	29.3 → 29.1%
Ours	98.0%	93.3 → 91.4%	81.6 → 79.8%	59.8 → 58.2%	75.8 → 74.0%
Ours-C	98.7%	98.8 → 97.5%	96.5 → 95.2%	75.7 → 74.6%	82.3 → 81.1%

Exceptions. The grid is the five chosen classes and the two held-out targets of Table 1, evaluated at the four operating points of Table 11. At $\ell = 2^{10}$, Ours-C exceeds NAFG in all forty cells under Arc2Face and in 37 of 40 under NbNet, the inverse model of Table 2. The three exceptions occur for Male at the two loosest operating points, with a maximum deficit of 6.3 points.

Transfer to Another Image of the Same Identity. The main evaluation verifies each generated image against the same reference image used for construction. We examine whether the generated image is also accepted when verification uses a different image of the same identity, using the 773 references whose benchmark pair provides such an image. Without regenerating the outputs, we replace the verification reference with this paired same-identity image and recompute ASR in Table 13. ASR decreases for every method, but their ordering is unchanged and Ours-C remains highest at all four operating points. The nonzero rates show that acceptance extends beyond the construction reference, while the decrease shows substantial dependence on that image. As a negative control, replacing each of the remaining 727 references with its impostor-pair image yields 0.0% ASR in all twelve method–model–threshold cells.

Table 13: ASR against the construction reference and another image of the same identity for the 773 references in genuine benchmark pairs.

		TransFace-L		ViT-KPRPE	
Method	Verification image	10^{-4}	10^{-5}	10^{-4}	10^{-5}
NAFG	construction reference	55.2%	29.1%	62.6%	32.9%
	same-identity image	15.8%	6.3%	15.7%	5.6%
Ours	construction reference	86.5%	74.9%	95.2%	86.2%
	same-identity image	45.3%	30.8%	50.6%	34.7%
Ours-C	construction reference	96.1%	88.2%	98.1%	92.2%
	same-identity image	**54.1%**	**37.4%**	**57.1%**	**38.9%**

D.2 Comparison with Adversarial Face Attack Baselines

Face-recognition adversarial attacks provide functional baselines because an impersonation attack can satisfy the same success criterion when its source image belongs to the chosen attribute class. This functional overlap does not make the research questions equivalent. These baselines optimize a source image or its generator variables for impersonation, whereas we study whether the identity and attribute structure of the face-recognition representation space supports constructing a representation that satisfies both conditions. We compare against one pixel-space and one diffusion-based attack under a common evaluation protocol while accounting for their additional requirement of a source image.

Setting. Both attacks start from an image of a person assigned to the chosen class and modify it until the target model accepts it as the reference identity. For each reference, we draw this source uniformly from normalized ranks 0.50 to 0.80 under E_S and keep it fixed across methods, avoiding an image already close to the reference. Sibling-Attack (Li et al., 2023) is the pixel-space one,

run from the authors' released implementation at their schedule of 200 outer and 5 inner steps. Its attribute branch retains the released recognizer, while its face-recognition branch uses the same surrogate pool as our method; we report $\epsilon \in \{8/255, 16/255\}$ in L_∞. The diffusion baseline of Salar et al. (2025) inverts and optimizes the source image in its generator space using the released configuration. We report ASR as defined in Section 5 on the same 1,500 references for every method. The evaluation models and operating points follow Table 1. A service that finds no face in an image counts that image as a failure rather than dropping it.

Single and Ensemble Baselines. Commercial queries were run only for the three-surrogate versions of the baselines. Table 14 verifies this choice on the open-source targets: the ensemble increases ASR for both baselines at every operating point. The remaining comparisons therefore use the ensemble version of each baseline.

Table 14: ASR of the image-modifying baselines with one or three surrogates. Results are macro-averaged over the five chosen classes.

	TransFace-L			ViT-KPRPE		
Method	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}
Single surrogate						
Sibling, ϵ=8/255	6.6%	2.1%	0.1%	9.9%	3.3%	0.1%
Sibling, ϵ=16/255	20.4%	10.1%	0.7%	25.4%	12.4%	0.2%
Diffusion	32.7%	16.4%	0.9%	53.7%	26.9%	0.7%
Ensemble of three						
Sibling, ϵ=8/255	29.3%	14.7%	1.1%	38.6%	19.5%	0.7%
Sibling, ϵ=16/255	50.7%	35.2%	5.5%	59.0%	39.7%	3.5%
Diffusion	**77.4%**	**57.8%**	**9.3%**	**84.6%**	**65.4%**	**5.5%**

Comparison under the Common Success Criterion. Table 15 compares the methods on their shared functionality: producing the chosen attribute while inducing reference-identity acceptance. The adversarial baselines modify a supplied source image, whereas the representation-space constructions do not require one. Both representation-space methods exceed all image-modifying baselines in every column. Ours-C exceeds Ours in nine columns; Ours is higher by 0.2 points under ViT-KPRPE at FMR= 10^{-6}.

Table 15: ASR on the common 1,500-reference sample, macro-averaged over the five chosen classes. Sibling and Diffusion require a source image; NAFG, Ours, and Ours-C do not.

	Open source						Commercial			
	TransFace-L			ViT-KPRPE			AWS		Tencent	
Method	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}	80	99	50	60
Sibling, ϵ=8/255	29.3%	14.7%	1.1%	38.6%	19.5%	0.7%	11.9%	2.1%	16.5%	9.6%
Sibling, ϵ=16/255	50.7%	35.2%	5.5%	59.0%	39.7%	3.5%	25.5%	7.5%	34.7%	25.3%
NAFG	56.0%	32.6%	2.7%	62.3%	33.7%	1.1%	37.3%	7.5%	26.0%	14.7%
Diffusion	77.4%	57.8%	9.3%	84.6%	65.4%	5.5%	31.0%	7.3%	27.8%	17.3%
Ours	86.0%	72.9%	27.7%	95.4%	85.9%	**25.5%**	71.5%	33.5%	71.1%	55.2%
Ours-C	**95.8%**	**87.5%**	**31.9%**	**98.0%**	**92.5%**	25.3%	**88.1%**	**41.7%**	**84.4%**	**66.3%**

Qualitative Comparison. ASR places all methods under a common success criterion but does not show the different inputs and outputs behind it. Sibling and Diffusion transform a supplied source image from the chosen class, whereas Ours-C forms its output without such an image. Table 16 shows two cases in which every method satisfies the chosen-attribute condition, clears both open-source targets at FMR 10^{-4}, and clears the default thresholds of both commercial services. In both cases, the reference is assigned to the opposite class, and the supplied source is assigned to the chosen class.

Table 16: Outputs on two common successful cases. Source is required only by Sibling and Diffusion; the final row of each block reports AWS and Tencent similarity scores.

	Reference	Source	Ours-C	Diffusion	Sibling $\epsilon = \frac{8}{255}$	Sibling $\epsilon = \frac{16}{255}$
Chosen class: Female						
						
AWS / Tencent	—	2.3 / 19.1	93.0 / 81.3	92.7 / 72.9	99.3 / 71.1	99.9 / 100.0
Chosen class: Black						
						
AWS / Tencent	—	0.5 / 9.3	97.6 / 75.4	84.3 / 53.4	93.6 / 58.6	95.5 / 92.9

D.3 Effect of Gallery Database and Size

The main evaluation draws each class gallery from WebFace42M. We vary the source database and the number of available gallery images separately to test whether the observed advantage depends on either choice.

Gallery Database. We replace the WebFace42M gallery with VGGFace2 (Cao et al., 2018), align its class representations to the fixed inverse model, and retain all other evaluation settings. The attributed subsphere is also refitted on the replacement gallery.

Table 17: ASR with a VGGFace2 gallery on the same 1,500 references, macro-averaged within gender and race.

		TransFace-L			ViT-KPRPE		
Group	Method	10^{-4}	10^{-5}	10^{-6}	10^{-4}	10^{-5}	10^{-6}
Gender	NAFG	58.8%	35.5%	3.8%	69.5%	38.8%	0.8%
	Ours	82.0%	64.8%	19.8%	92.2%	81.2%	**17.7%**
	Ours-C	**92.0%**	**79.3%**	**20.3%**	**96.2%**	**84.5%**	17.5%
Race	NAFG	56.8%	30.8%	2.2%	58.4%	31.9%	1.1%
	Ours	69.9%	47.4%	12.9%	82.8%	63.2%	**12.9%**
	Ours-C	**79.4%**	**58.8%**	**13.7%**	**84.4%**	**64.3%**	12.1%

Both Ours and Ours-C exceed NAFG in every reported cell, and Ours-C is generally strongest except at the strictest ViT-KPRPE operating point. Thus, the advantage over NAFG persists with a gallery drawn from a different database. Because only the replacement gallery is aligned and a fixed ℓ represents different gallery fractions, this experiment compares method ordering rather than absolute rates with Table 1.

Gallery size. We next sweep the size of the gated WebFace42M gallery for Ours-C while fixing the same 1,500 references, $\ell = 2^{10}$, Arc2Face, and every other evaluation setting. For each class, one seeded permutation defines nested subsets, so only the number of available images changes across the sweep. The largest common size is 2^{19}, set by the smallest class gallery.

Table 18: ASR of Ours-C as the size of the gated WebFace42M class gallery is swept, on the same 1,500 references, macro-averaged within gender and within race. The draws are nested, $\ell = 2^{10}$ throughout, and 2^{19} is the largest size every class can supply.

		TransFace-L		ViT-KPRPE	
Group	Gallery rows	10^{-4}	10^{-5}	10^{-4}	10^{-5}
Gender	4,096	85.0%	62.7%	90.8%	69.2%
	16,384	91.7%	78.0%	96.0%	84.5%
	65,536	95.7%	85.5%	97.8%	92.5%
	262,144	96.0%	89.8%	97.8%	94.8%
	524,288	**97.3%**	**92.5%**	**98.5%**	**96.8%**
Race	4,096	52.1%	29.6%	60.2%	34.2%
	16,384	81.6%	56.6%	85.9%	62.4%
	65,536	88.8%	70.8%	93.7%	76.7%
	262,144	91.1%	76.2%	**95.2%**	81.8%
	524,288	**92.2%**	**77.4%**	94.9%	**85.1%**

ASR rises most rapidly through 2^{16} gallery images. Increasing the gallery from 2^{16} to 2^{19} adds at most 3.4 points at $\alpha = 10^{-4}$, while gains of 4.3–8.4 points remain at $\alpha = 10^{-5}$. The construction therefore obtains most of its loose-threshold performance without the full gallery, whereas stricter verification continues to benefit from additional images.

D.4 Comparison with Target-Query Access

The target-query variant of NAFG registers 100 class probes, obtains their target-model scores, and solves the score-based linear system with its correction matrix R. The class-level setup uses 100 registration calls and a further 100 calls to construct R; each reference then requires one target search. Table 19 evaluates this variant and our query-free constructions using the same reference images and AWS thresholds. The corrected query variant attains higher ASR at the default threshold, whereas Ours-C attains higher ASR at the strict threshold.

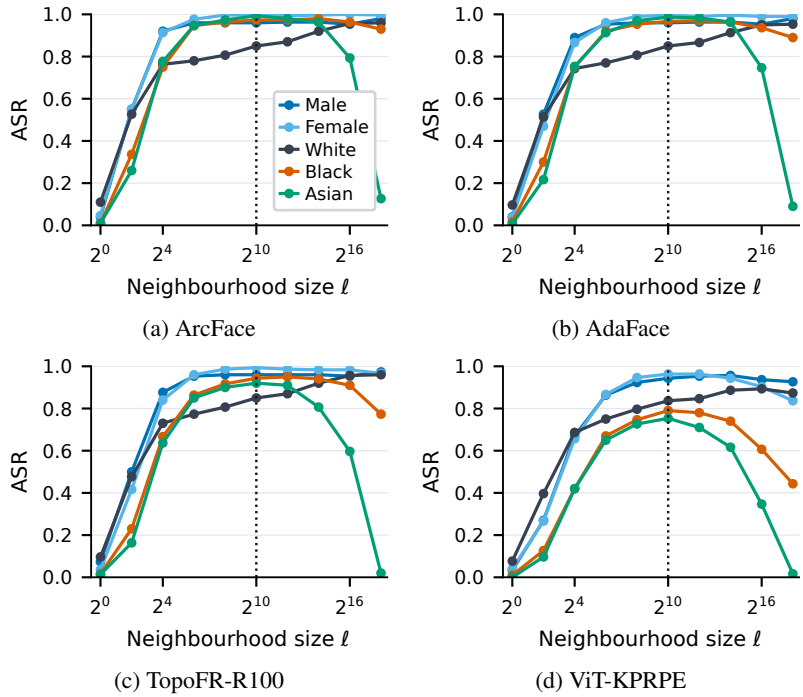
Table 19: ASR against AWS using the same reference images as the commercial evaluation. The queried rows are the prior work's black-box construction, with and without the correction its paper specifies.

Method	Target queries	80	99
NAFG, transferred	none	37.3%	7.5%
NAFG, queried, no R	per reference	49.8%	2.1%
NAFG, queried, with R	per reference + R setup	**92.2%**	18.5%
Ours	none	71.5%	33.5%
Ours-C	none	88.1%	**41.7%**

D.5 Neighborhood Depth

Full Sweep. Figure 14 sweeps the depth for every class under the four recognition models the body does not draw, at FMR$= 10^{-5}$, and Figure 5 is the fifth. The maximizing depth varies across both classes and evaluation models. Under TransFace-L and ViT-KPRPE, ASR flattens by 2^{12} and then decreases for Black and Asian, whereas several curves evaluated by models involved in construction continue to increase.

Where the Peak Sits. Across the twenty-five model–class cells, the best measured depth ranges from 2^{10} to 2^{18}. The fixed choice $\ell = 2^{10}$ is optimal in ten cells and remains within 3.0 ASR points of the best measured depth in every non-White cell. A target-dependent choice is unavailable without querying the target, so all reported attacks use the same fixed depth.



(a) ArcFace (b) AdaFace

(c) TopoFR-R100 (d) ViT-KPRPE

Figure 14: ASR across neighborhood depths under four additional recognition models. The vertical rule marks $\ell = 2^{10}$; ArcFace is the composing model rather than a transfer target.

D.6 Effect of the Surrogate Pool

Table 20 varies only the surrogate subset used for ordering the gallery. Rows specify the surrogate subset and columns the evaluation model; unshaded cells are transfer results.

Table 20: ASR by surrogate subset and evaluation model. Shading marks models involved in construction, and bold marks the best transfer result in each column.

Selected with	$\lvert\mathcal{S}\rvert$	ArcFace	AdaFace	TopoFR-R100	TransFace-L	ViT-KPRPE
NAFG	0	93.3%	69.5%	51.8%	29.4%	29.1%
ArcFace	1	97.2%	91.4%	79.8%	58.2%	74.0%
AdaFace	1	92.0%	97.8%	71.4%	53.9%	65.9%
TopoFR-R100	1	85.1%	77.0%	96.6%	62.7%	52.6%
ArcFace, AdaFace	2	97.7%	97.4%	**84.5%**	65.7%	79.8%
ArcFace, TopoFR-R100	2	97.9%	**93.6%**	95.7%	72.4%	77.1%
AdaFace, TopoFR-R100	2	96.0%	97.4%	95.4%	72.6%	75.0%
all three	3	97.9%	97.5%	95.2%	**74.6%**	**81.1%**

The full surrogate pool gives the highest ASR under both unseen target models. The best individual surrogate differs between TransFace-L and ViT-KPRPE, so one surrogate cannot be chosen reliably without target access. This result complements the ordering analysis in Appendix C.3: consensus reduces dependence on which surrogate best matches an unseen target.

D.7 Other Inverse Models

An Alternative ArcFace-Aligned Inverse Model. Table 21 replaces Arc2Face with NbNet (Paik et al., 2026a) while retaining the same composing representation space. Absolute ASR changes, but NAFG, Ours, and Ours-C retain their ordering under every evaluation model.

Table 21: Attribute assignment, identity acceptance, and ASR with NbNet. Each identity cell is followed by the corresponding ASR.

Method	Attribute	AdaFace	TopoFR-R100	TransFace-L	ViT-KPRPE
NAFG	98.3%	89.3 → 87.8%	78.5 → 77.3%	59.2 → 58.2%	51.2 → 50.2%
Ours	97.0%	98.9 → 96.0%	96.2 → 93.3%	87.5 → 84.6%	90.6 → 87.7%
Ours-C	98.4%	99.8 → 98.2%	99.7 → 98.1%	94.6 → 93.1%	93.3 → 91.7%

A Different Recognition–Inverse Pair. Vec2Face (Wu et al., 2025) uses a different ArcFace representation, so the gallery ordering, composition, and NAFG subspace are all recomputed in its aligned space. Table 22 therefore evaluates a changed recognition–inverse pair rather than an inverse-model replacement alone. Ours-C retains the highest ASR across both attribute families, all four evaluation models, and both operating points.

Table 22: ASR under the Vec2Face-aligned recognition–inverse pair, pooled by attribute family. Shading marks models used by Ours-C.

		AdaFace		TopoFR-R100		TransFace-L		ViT-KPRPE	
Family	Method	10^{-5}	10^{-6}	10^{-5}	10^{-6}	10^{-5}	10^{-6}	10^{-5}	10^{-6}
Gender	NAFG	86.2%	13.5%	84.2%	17.0%	69.2%	17.3%	44.2%	7.5%
	Ours	90.2%	58.8%	89.8%	64.0%	87.2%	56.3%	85.7%	32.3%
	Ours-C	98.2%	92.0%	98.2%	95.7%	97.8%	79.2%	97.3%	41.0%
Race	NAFG	82.9%	8.8%	85.1%	12.3%	65.8%	12.4%	43.3%	5.9%
	Ours	84.2%	33.9%	82.8%	39.2%	75.2%	32.8%	71.2%	19.9%
	Ours-C	93.2%	59.6%	93.2%	74.2%	91.3%	50.2%	86.8%	26.8%

Table 23 shows one reference–output pair per class after moving composition and inversion to the Vec2Face-aligned space.

Table 23: Reference–output pairs under Vec2Face; rows report target-model cosine similarities.

	Chosen class				
	Male	*Female*	*White*	*Black*	*Asian*
Reference					
Ours-C					
TransFace-L	0.697	0.681	0.680	0.638	0.480
ViT-KPRPE	0.571	0.568	0.573	0.430	0.411

D.8 Additional Qualitative Results

Tables 24 and 25 show six additional reference–output pairs per chosen class for Arc2Face and NbNet on disjoint references. Each reference differs from the chosen class, every pair clears both commercial defaults, and the tables expose inverse-model effects absent from ASR.

Table 24: Six reference–output pairs per chosen class. The final row of each block reports AWS and Tencent similarity scores.

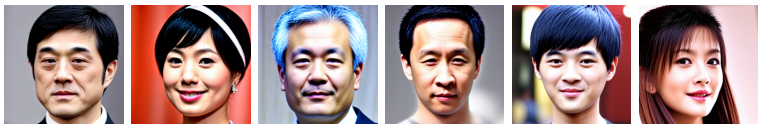
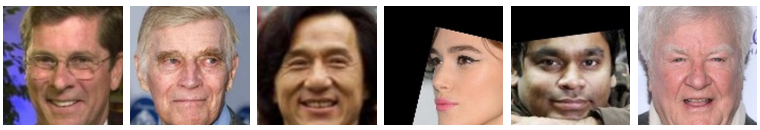
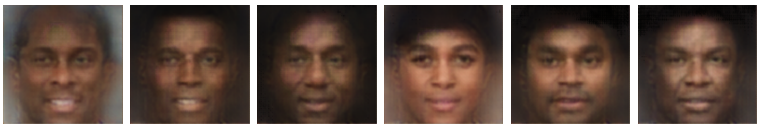
Chosen class: Male						
Reference						
Ours						
AWS / Tencent	95.5 / 85.2	98.9 / 64.5	97.4 / 90.8	97.7 / 84.5	99.3 / 85.0	96.9 / 68.3
Chosen class: Female						
Reference						
Ours						
AWS / Tencent	86.1 / 68.0	98.4 / 72.1	96.7 / 51.0	96.4 / 69.0	89.7 / 80.2	96.8 / 62.9
Chosen class: White						
Reference						
Ours						
AWS / Tencent	80.5 / 76.7	99.9 / 75.8	93.5 / 56.5	99.3 / 72.9	98.9 / 89.3	100.0 / 100.0
Chosen class: Black						
Reference						
Ours						
AWS / Tencent	96.0 / 62.9	99.3 / 56.8	96.8 / 50.8	99.0 / 74.4	99.8 / 100.0	99.0 / 81.2
Chosen class: Asian						
Reference						
Ours						
AWS / Tencent	97.0 / 59.1	97.4 / 54.1	99.7 / 64.0	99.0 / 67.7	91.0 / 53.5	96.1 / 61.8

Table 25: Six additional reference–output pairs per chosen class under NbNet. The final row of each block reports AWS and Tencent similarity scores.

Chosen class: Male						
Reference						
Ours						
AWS / Tencent	99.6 / 100.0	99.9 / 85.8	98.4 / 100.0	99.7 / 80.6	96.8 / 90.9	98.8 / 63.1
Chosen class: Female						
Reference						
Ours						
AWS / Tencent	91.5 / 80.7	89.3 / 66.8	96.5 / 79.1	92.3 / 65.0	96.9 / 60.7	85.9 / 66.1
Chosen class: White						
Reference						
Ours						
AWS / Tencent	93.6 / 57.4	99.3 / 94.9	99.0 / 64.4	84.4 / 54.0	99.4 / 96.2	95.7 / 60.8
Chosen class: Black						
Reference						
Ours						
AWS / Tencent	83.3 / 60.9	85.3 / 62.6	94.1 / 73.3	88.4 / 57.0	99.9 / 100.0	90.2 / 57.5
Chosen class: Asian						
Reference						
Ours						
AWS / Tencent	93.5 / 74.3	97.5 / 56.9	99.0 / 54.2	82.7 / 56.1	99.6 / 63.3	92.3 / 59.3