# Representation Rigidity in Face Embeddings: Orthogonal Identifiability and Backfill-Free Compatibility

**Yongle Zhao  Zhichao Chen  Yin Xie  Jun Wang  Ziyong Feng***
DeepGlint
{yonglezhao, zhichaochen, yinxie, junwang, ziyongfeng}@deepglint.com

## Abstract

Upgrading a deployed face-recognition encoder often makes new query embeddings incompatible with the legacy gallery, requiring stored embeddings and the approximate-nearest-neighbor index to be rebuilt. We investigate when this incompatibility can instead be resolved by a query-side change of coordinates. Across many independently trained angular-margin face models, we observe that the legacy and upgraded embedding spaces are approximately related by a single global orthogonal transformation, a property we call *representation rigidity*. We formalize this empirical condition as *Procrustes rigidity*, derive a conditional finite-calibration recovery bound, and estimate the orthogonal adapter in closed form from a small set of unlabeled images processed by both fixed encoders. The resulting backfill-free protocol maps upgraded queries into legacy coordinates while leaving the gallery and index unchanged. Experiments on 45 models spanning three architectures, two losses, three data scales, and five seeds show strong within-family rigidity and non-monotonic variation with training-set scale. On a derived IFRT-MR $1:N$ split with a 1.6M-image gallery, orthogonal transport retains at least $99.8\%$ of legacy (old–old) Rank-1 accuracy for the tested same-family upgrades; in a more challenging setting, an optional residual correction substantially narrows the remaining gap.

## 1 Introduction

**A model upgrade should not require rebuilding the gallery.** Large-scale face-recognition systems store embeddings for millions or billions of enrolled images and perform identification through nearest-neighbor search in a normalized feature space [5, 17, 23, 26–28]. When the encoder is upgraded with a better backbone [8, 12], more training data [4, 10], or an improved angular-margin objective [3, 5, 6, 13, 16, 18, 28, 30], the resulting query embeddings are usually incompatible with the legacy gallery because their coordinates have changed. The standard remedy is *backfilling*: re-embedding every stored gallery item with the upgraded model. Although conceptually straightforward, backfilling has a cost that grows with the gallery and can dominate the deployment cost of a routine model upgrade.

**Compatibility may be a geometry problem, not only a training problem.** Backward-compatible representation learning (BCRL) avoids backfilling by training the upgraded encoder to remain compatible with the legacy system [20, 24, 29, 31]. We ask a complementary question: before enforcing compatibility during training, how much of the apparent incompatibility can be removed by choosing an appropriate coordinate system for two independently trained face models? We frame the question as:

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

*To what extent do independently trained face embedding spaces share a geometry that can be recovered by a single global adapter?*

For many same-family upgrades, we find that models $A$ and $B$ differ mainly by one global orthogonal map $R_\star \in \mathbb{O}(d)$, so that $z_B(x) \approx R_\star z_A(x)$ over the deployment distribution. We call this empirical behavior **representation rigidity**.

**Why the default adapter is orthogonal.** Angular-margin face losses depend on normalized features and classifier prototypes only through inner products. Rotating all features and prototypes together therefore leaves logits and losses unchanged. This symmetry identifies an orthogonal coordinate ambiguity, but it does not ensure that two independent runs preserve the same relational geometry. It therefore motivates, rather than guarantees, orthogonal equivalence: when the learned geometries agree, they can be related by a global rotation or reflection. We formalize this condition as *Procrustes rigidity* and estimate the default adapter by closed-form orthogonal Procrustes from shared images processed by both encoders. The resulting finite-sample estimate of $R_\star$ is denoted by $\hat{R}$.

**The deployment payoff is query-side feature transport.** Once $\hat{R}$ is estimated, each upgraded query embedding $z_B(x_q)$ is transported by $\hat{R}^\top$ into the legacy coordinate system and searched against the existing index. The stored gallery embeddings and ANN index remain unchanged. For geometry-preserving upgrades, this closed-form adapter is the complete method. For harder pairs whose relational geometry is not captured by a single rotation, we optionally apply a lightweight residual MLP after orthogonal transport. This extension remains query-side and backfill-free, but unlike the Procrustes stage it is learned and does not preserve inner products by construction.

**Contributions.**

- **A geometry-first view of compatibility.** We characterize representation rigidity in face embeddings, formalize it as Procrustes rigidity, and distinguish the loss symmetry that motivates orthogonal alignment from the empirical rigidity that makes alignment effective.
- **A backfill-free transport protocol with a bounded extension.** We estimate a closed-form orthogonal adapter from shared calibration images processed by both encoders and reuse the legacy gallery unchanged. For high-residual cross-architecture pairs, an optional residual MLP corrects part of the remaining non-orthogonal deformation without modifying the gallery.
- **Evidence across upgrade axes.** Across 45 models, we quantify rigidity under changes in seed, loss, depth, architecture family, and data scale, and use calibration-size and target-domain diagnostics to separate estimation error from non-orthogonal deformation.
- **Operational validation and limitations.** On the derived IFRT-MR identification split with a 1.6-million-image gallery, Procrustes retains at least $99.8\%$ of old–old Rank-1 for the tested seed, loss, and depth upgrades. CNN-to-ViT and low-quality identification remain harder regimes in which residual correction narrows, but does not eliminate, the gap.

## 2 Related work

**Positioning.** Our work connects four lines of research that are often studied separately: angular-margin face losses define hyperspherical embedding geometry; BCRL methods enforce compatibility during training; Procrustes alignment provides a post-hoc estimator; and representation-convergence studies ask why independently trained models may share geometry. Unlike training-based methods, we keep both encoders fixed and test whether they can be related by a simple change of coordinates.

### 2.1 Hyperspherical face embeddings and angular-margin losses

Face recognition has evolved from identification and verification supervision and metric learning toward hyperspherical embeddings trained with margin-based softmax objectives [5, 17, 23, 25–28]. ArcFace and CosFace popularized normalized features and classifier weights with explicit angular or cosine margins [5, 28]. Subsequent variants address label noise and sample difficulty [3, 6, 13, 16, 18, 30]. Because these objectives depend on normalized inner products, orthogonal transformations form a natural coordinate ambiguity of the learned embedding space.

### 2.2 Backward-compatible representation learning (BCRL)

BCRL methods aim to deploy upgraded models without backfilling. BCT introduced backward-compatible training with a legacy model and classifier [24]; $BT^2$ explores basis transformation and augmentation to alleviate compatibility constraints [31]. AdvBCT uses boundary-aware and adversarial objectives to improve retrieval compatibility [20], while broader formulations seek universal backward compatibility [29]. Our approach is complementary: training-based BCRL optimizes a new encoder with legacy-aware objectives, whereas Procrustes keeps both encoders fixed and changes only the query-side coordinate system. $\lambda$-Orthogonality regularizes a learned alignment map [21]; we instead ask whether a usable orthogonal map emerges without retraining. This distinction makes computational and deployment costs important alongside verification accuracy.

### 2.3 Procrustes alignment and stability

Orthogonal Procrustes has a classical closed-form SVD solution [9, 22]. Perturbation analyses relate its stability to residual energy and conditioning [1]. We use it as both an estimator and a diagnostic: a low residual indicates that the cross-model mismatch is largely orthogonal, whereas a high residual marks regimes in which query-side rotation may be insufficient.

### 2.4 Representation convergence and model stitching

Recent work suggests that independently trained networks may converge toward shared internal representations. The *Platonic Representation Hypothesis* [14] proposes that, as models scale, representations across architectures and modalities converge toward a common statistical model of reality. Model-stitching studies [2] show that intermediate representations of differently trained networks can often be linearly mapped with little loss in accuracy, suggesting shared geometry. Relatedly, $\lambda$-Orthogonality regularization encourages a learned alignment map to remain orthogonal in backward-compatible representation learning. Our results provide face-recognition-specific evidence relevant to this hypothesis and connect it to deployment: sufficiently orthogonal shared structure can support backward compatibility without encoder retraining. Multimodal canonicalization [11] also fits an orthogonal map from shared unlabeled anchors, enabling upgrades without re-embedding. Our face-specific contribution is a 45-model characterization, million-scale legacy-gallery evaluation, and diagnostics of calibration mismatch and representation deformation.

## 3 Representation rigidity and feature transport

This section presents the method pipeline. First, normalized angular-margin losses admit a global orthogonal coordinate ambiguity. Second, Procrustes rigidity states when two independently trained models are well approximated by that ambiguity, and a finite calibration set estimates the corresponding adapter in closed form. Third, for model pairs with substantial non-orthogonal deformation, an optional residual MLP acts after orthogonal transport. Finally, the query-side adapter maps upgraded queries into the legacy coordinate system while leaving the gallery unchanged.

### 3.1 Orthogonal symmetry of normalized margin losses

We consider a face-recognition model $f_\theta : \mathcal{X} \to \mathbb{R}^d$ that produces $\ell_2$-normalized embeddings $z(x) = f_\theta(x)/\|f_\theta(x)\|_2 \in \mathbb{S}^{d-1}$. The classification layer is a prototype matrix $W = [w_1, \dots, w_C] \in \mathbb{R}^{d\times C}$ with $\|w_c\|_2 = 1$. Angular-margin objectives such as ArcFace and CosFace operate on normalized inner products $w_c^\top z(x) = \cos\theta_c(x)$ and modify the positive-class logit using an angular or cosine margin [5, 28]. These objectives admit a global orthogonal symmetry.

**Lemma 3.1** (Orthogonal Equivariance of Normalized Margin Losses)**.** Let $R \in \mathbb{O}(d)$ be any orthogonal matrix. Define transformed features and prototypes as $z'(x) = Rz(x)$ and $W' = RW$. Then all logits are preserved: $(Rw_c)^\top(Rz) = w_c^\top z$. Consequently, any loss built solely from these normalized inner products (including ArcFace/CosFace) is invariant under $(z, W) \mapsto (Rz, RW)$.

Lemma 3.1 implies that the training objective cannot identify an absolute coordinate system on $\mathbb{S}^{d-1}$; it admits a global orthogonal coordinate ambiguity. This symmetry motivates alignment within the orthogonal equivalence class, but it does not imply that two independent training runs preserve the

same relational geometry. The rigidity defined below is therefore an empirical property, and the recoverability result is conditional on the assumed residual model.

### 3.2 Procrustes rigidity and adapter estimation

Given $N$ inputs evaluated by both models, define their feature matrices $Z_A = [z_A(x_i)^\top]_{i=1}^N \in \mathbb{R}^{N\times d}$ and $Z_B = [z_B(x_i)^\top]_{i=1}^N \in \mathbb{R}^{N\times d}$. Row $i$ in both matrices represents the same image $x_i$; no identity matching between different images is required.

**Definition 3.1** (Procrustes Rigidity on a Sample Set)**.** Define the optimal orthogonal alignment

$$\hat{R} = \arg\min_{R\in\mathbb{O}(d)} \|Z_B - Z_A R^\top\|_F^2, \tag{1}$$

and the normalized residual $\mathrm{Res}(Z_A, Z_B) \triangleq N^{-1/2}\|Z_B - Z_A\hat{R}^\top\|_F$. We say $(f_A, f_B)$ exhibit $(\varepsilon, N)$-*Procrustes rigidity* on $\{x_i\}$ if $\mathrm{Res}(Z_A, Z_B) \le \varepsilon$.

The adapter has a closed-form solution. Let $Z_A^\top Z_B = U\Sigma V^\top$ denote its singular value decomposition. Then the minimizer of Eq. (1) is

$$\hat{R} = VU^\top. \tag{2}$$

Computing this SVD is the only fitting step in the default orthogonal protocol; both encoders remain fixed.

### 3.3 Recoverability from finite calibration

Let $Z_A^{(cal)}, Z_B^{(cal)} \in \mathbb{R}^{K\times d}$ denote the feature matrices for a calibration set of size $K$. The estimate is useful at deployment only if it transfers beyond these calibration samples. Assume $X = Z_A^{(cal)}$ has full column rank and $Y = Z_B^{(cal)} = XR_\star^\top + E$ for $R_\star \in \mathbb{O}(d)$. Procrustes optimality and the triangle inequality then give

$$\|\hat{R} - R_\star\|_F \le \frac{2\|E\|_F}{\sigma_{\min}(X)}. \tag{3}$$

Specifically, $\|X(\hat{R} - R_\star)^\top\|_F \le \|Y - X\hat{R}^\top\|_F + \|E\|_F \le 2\|E\|_F$. This conditional bound motivates monitoring residuals and conditioning [1], but it does not guarantee transfer under a shift in the calibration distribution.

### 3.4 Alignment error metrics

The central empirical question is whether an orthogonal adapter fits these same-input correspondences closely enough to preserve nearest-neighbor decisions. We report the global residual (4), angular agreement (5), and pairwise geometric distortion. Given $(Z_A, Z_B)$ and Procrustes solution $\hat{R}$:

**(1) Global Procrustes residual.**

$$\mathrm{Res}(Z_A, Z_B) = \frac{1}{\sqrt{N}}\|Z_B - Z_A\hat{R}^\top\|_F. \tag{4}$$

**(2) Angular consistency score (ACS).**

$$\mathrm{ACS}(Z_A, Z_B) = \frac{1}{N}\sum_{i=1}^N z_{A,i}^\top \hat{R}^\top z_{B,i}. \tag{5}$$

**(3) Pairwise angular distortion.** For each sampled pair $(i, j)$, $\Delta_\theta(i, j)$ is the absolute difference between the legacy angle $\arccos(z_{A,i}^\top z_{A,j})$ and the transported upgraded angle $\arccos((\hat{R}^\top z_{B,i})^\top(\hat{R}^\top z_{B,j}))$.

**Why normalization alone is insufficient.** Unit normalization fixes norms, but models may learn different pairwise angles, inducing retrieval-margin shifts. Rotation aligns shared geometry yet preserves these differences; for same-input $X, Y$, an exact fit exists iff $YY^\top = XX^\top$ (App. F).

$$\min_{Q\in\mathbb{O}(d)} \|YQ - X\|_F \ge \frac{\|YY^\top - XX^\top\|_F}{\|Y\|_2 + \|X\|_2}. \tag{6}$$

### 3.5 Optional residual correction beyond a global rotation

Residual correction therefore follows rather than replaces Procrustes: the rotation recovers the dominant global transport, and a lightweight function models the remaining sample-dependent, non-orthogonal deviations that additional calibration cannot remove. For an upgraded embedding, let $u(x) = \hat{R}^\top z_B(x)$ and define

$$T_\phi(z_B(x)) = \frac{u(x) + g_\phi(u(x))}{\|u(x) + g_\phi(u(x))\|_2}, \tag{7}$$

where $g_\phi$ is a two-layer MLP with hidden width $d/2$ and a ReLU nonlinearity. Using the same-image calibration correspondences, we minimize

$$\frac{1}{K}\sum_{i=1}^{K}\left[1 - z_A(x_i)^\top T_\phi(z_B(x_i))\right]. \tag{8}$$

In (8), both encoders and $\hat{R}$ remain fixed; only $g_\phi$ is learned. The residual adapter is therefore a query-side extension rather than part of the closed-form Procrustes estimator. It need not preserve inner products, so the orthogonal equivalence and score-perturbation results below apply to the Procrustes path, not to the learned correction.

### 3.6 Backfill-free query-side transport

**Problem setting.** Let model $A$ be the legacy system with embeddings $z_A(x) \in \mathbb{S}^{d-1}$ and model $B$ be an upgraded system producing $z_B(x) \in \mathbb{S}^{d-1}$. The legacy gallery contains embeddings for a large set of enrolled identities. Recomputing all gallery features with model $B$ ("backfilling") is the operation we seek to avoid.

We assume that a single orthogonal map $R_\star \in \mathbb{O}(d)$ approximately relates the two embedding spaces, $z_B(x) \approx R_\star z_A(x)$. We estimate $\hat{R}$ from a small calibration set using Procrustes; under the residual model in Section 3.3, its stability depends on the residual and calibration conditioning.

The task is **cross-model retrieval**: given a query embedding from model $B$, retrieve its nearest neighbors from a legacy gallery built with model $A$ without re-embedding the gallery.

**Offline (legacy gallery).** The legacy system computes and stores $z_A(x_g)$ for each gallery item $x_g$. These embeddings define the coordinate system to which upgraded queries must be mapped.

**Calibration (one-time).** Pass each image in a small calibration set $\{x_i\}_{i=1}^K$ through both models, stack the normalized embeddings as rows of $Z_A, Z_B \in \mathbb{R}^{K\times d}$, and compute $\hat{R}$ using Eq. (2). Fitting uses these correspondences only; it does not require identity labels, classifier prototypes, training-set access, or gallery embeddings. Throughout the paper, $R$ denotes the legacy-to-upgraded direction ($z_B \approx R z_A$). The deployed query adapter is therefore the inverse rotation $\hat{R}^\top$, which maps upgraded queries into the legacy gallery coordinates. For the optional residual path, the same features are used to fit $g_\phi$ after $\hat{R}$ is fixed.

**Online (cross-model query).** For a query $x_q$ processed by model $B$, the default path computes $\tilde{z}_q = \hat{R}^\top z_B(x_q)$ in the legacy coordinate system. The optional residual path replaces $\tilde{z}_q$ with $T_\phi(z_B(x_q))$. In either case, the system queries the existing index without modifying the gallery-side data path.

### 3.7 Why orthogonal transport preserves similarity

**Exact rigidity implies exact compatibility.** If rigidity were exact, i.e. $z_B(x) = R_\star z_A(x)$ for all $x$, then the transported query satisfies $R_\star^\top z_B(x_q) = z_A(x_q)$. Consequently, for any legacy gallery item $x_g$, the transported cross-model score $(R_\star^\top z_B(x_q))^\top z_A(x_g)$ equals the legacy same-model score $z_A(x_q)^\top z_A(x_g)$. Cross-model retrieval therefore reproduces legacy same-model retrieval.

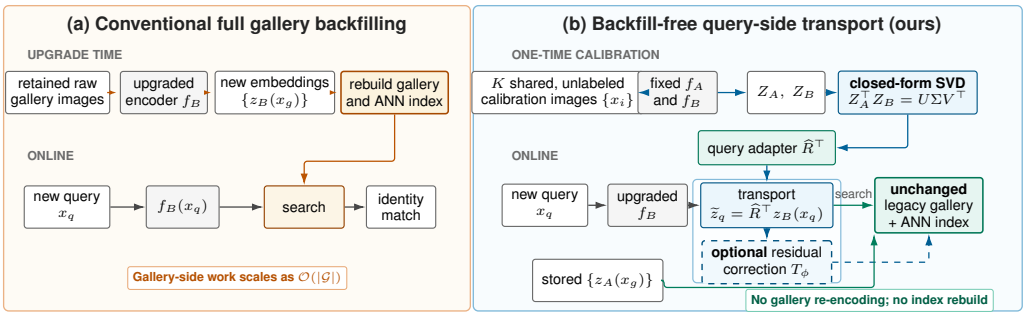


Figure 1: Full backfilling rebuilds the gallery features and index. Query-side transport processes the same unlabeled calibration images with both encoders and fits an orthogonal adapter while retaining the gallery and index. A residual correction may follow for high-residual model pairs.

**Approximate rigidity implies controlled similarity error.** In practice, the transported query differs from the legacy query by the residual vector $e_q = \tilde{z}_q - z_A(x_q)$. For any unit-norm gallery embedding $z_A(x_g)$, the similarity error satisfies $|\tilde{z}_q^\top z_A(x_g) - z_A(x_q)^\top z_A(x_g)| \le \|e_q\|_2$. Thus each gallery score changes by at most the per-query residual norm; under the model in Section 3.3, adapter recovery also depends on the calibration residual and conditioning. If the legacy best-positive minus best-negative score margin exceeds $2\|e_q\|_2$, correct Rank-1 is preserved (Appendix F). This sufficient condition preserves the legacy ranking; it does not guarantee recovery of full-backfill performance.

### 3.8 Complexity and deployment notes

The gallery remains unchanged (Figure 1); backfill-free does not mean calibration-free or equivalent to full backfilling. The default protocol estimates $\hat{R}$ once from $K$ shared calibration images and adds one $d \times d$ matrix–vector product per query. The optional path additionally fits and evaluates the residual MLP, and is therefore not training-free. With $K$ and $d$ fixed, both paths have $O(1)$ gallery-migration cost, in contrast to the $O(|\mathcal{G}|)$ cost of backfilling.

**Label-set mismatch.** The Procrustes protocol does *not* require identical training label sets. In realistic upgrades, the upgraded model may be trained on a much larger identity set ($C_B \gg C_A$), so the prototype matrices differ in width. The adapter remains well defined because $\hat{R}$ is estimated from same-image feature correspondences using Procrustes (Definition 3.1), not from classifier prototypes. Appendix D reports an existing-model scale-transfer proxy using WF4M/WF12M/WF42M models; Section 4.6 additionally evaluates cross-corpus transfer without assuming training-label inclusion.

## 4 Experiments

The experiments proceed from geometry to deployment: we diagnose when orthogonal transport fits, compare the closed-form method with training-based baselines, test residual correction for cross-architecture pairs, and evaluate large-gallery, cross-corpus, and low-quality conditions.

### 4.1 Experimental setup and protocols

**Training data and model suite.** The core study uses WebFace4M (4.2M images, 206K identities), WebFace12M (12.7M, 618K), and WebFace42M (42.5M, 2.06M) [32]. We consider nine configurations: R50-CosFace at all three data scales; R100-CosFace and ViT-B-CosFace at 12M and 42M; and R100-ArcFace and ViT-B-ArcFace at 12M [5, 8, 12, 28]. With $S$=5 training seeds per configuration and embedding dimension $d$=512, this suite contains 45 models. We evaluate rigidity on $N$=5,000 in-domain images. For cross-corpus transfer, we additionally evaluate MS1MV2-R100-ArcFace $\rightarrow$ WebFace12M-R100-ArcFace.

**Calibration.** Unless noted, both fixed encoders process the same $K$=5,000 in-domain calibration images, independently of the evaluation split. Procrustes uses same-input feature correspondences but no identity labels; labels only construct disjoint splits. The residual adapter reuses the calibration

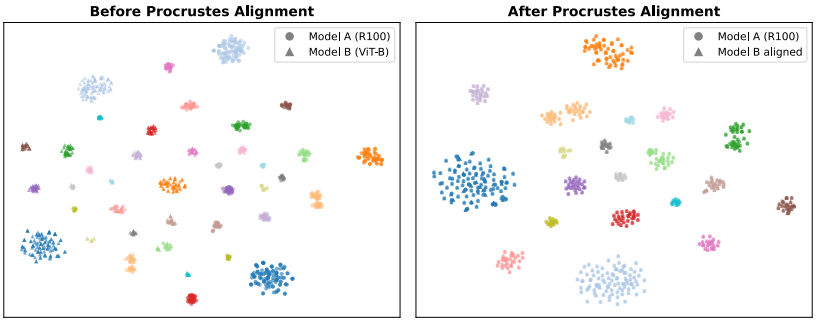


Figure 2: t-SNE visualization of R100 (circles) and ViT-B (triangles) embeddings, colored by identity. Procrustes improves the visual correspondence of same-identity points (right versus left), although the quantitative residual remains nonzero.

features after $\hat{R}$ is fixed. Cross-corpus, IJB-B/C, and TinyFace instead use $K$=3,000, with sources given below and in Appendix G. At this size, the same-architecture seed upgrade remains above 98% throughout the benchmark-level calibration-shift matrix, whereas $K$=100 is unstable (Appendix E).

**Evaluation protocols and references.** LFW, CFP-FP, and AgeDB-30 report accuracy for cross-model verification. IJB-B/C report TAR at FAR $10^{-4}$ from symmetric mixed-model template scores; these $1:1$ compatibility diagnostics differ from query-only old-gallery deployment. TinyFace evaluates new-model probes against the unchanged old-model gallery and reports Rank-1, Rank-5, and mAP. For derived IFRT-MR $1:N$ identification, Table 4 retains the original split for comparisons across upgrade axes, whereas Table 5a uses a controlled split to study cross-corpus transfer and calibration-domain mismatch, with the reserved IFRT calibration identities excluded from retrieval. Appendix G details both constructions and their interpretation boundaries. Both derived settings differ from the official IFRT/MFR-Ongoing all-to-all $1:1$ verification protocol over 1,624,305 images and 242,143 identities [7, 15]. Whenever available, we report old–old, direct, Procrustes, and new–new/full-backfill references.

### 4.2 When are face embeddings orthogonally rigid?

We first ask whether incompatibility along each upgrade axis is primarily a coordinate mismatch, using the residual $\mathrm{Res}$, angular consistency $\mathrm{ACS}$, and pairwise angular distortion (Table 1).

**Rigidity is strongest within architecture families.** Seed-only changes yield $\mathrm{Res} \approx 0.44$ for ResNets and $\approx 0.36$ for ViT-B, with $\mathrm{ACS} = 0.898$–$0.935$. CosFace–ArcFace changes remain close to the seed-only residuals for matched backbones. Increasing depth from R50 to R100 raises $\mathrm{Res}$ to $0.506$, whereas CNN$\rightarrow$ViT yields $\mathrm{Res} \approx 0.85$–$0.88$ and $\mathrm{ACS} \approx 0.61$–$0.65$. Figure 2 visualizes the coordinate correction; the nonzero residual shows that the fit remains approximate.

**Calibration size separates estimation error from deformation.** Figure 3 shows a block-diagonal rigidity map (left), with lower residuals within the CNN and ViT families than across them. In the calibration-size curves (right), $K$=1,000 images provide more than 85% of asymptotic alignment quality for same-family pairs; cross-architecture pairs require $K \geq$2,000 and still plateau at a higher residual.

**Target-domain diagnostics identify non-orthogonal deformation.** To separate calibration mismatch from model-pair deformation, we fit an oracle Procrustes map on 10,000 target probes for diagnosis only. Representative seed, loss, and depth changes yield $(D_G, \mathrm{Res}_{\mathrm{oracle}}) = (0.538, 0.450)$–$(0.608, 0.512)$, whereas CNN$\rightarrow$ViT yields $(1.169, 0.849)$ (Appendix G.3).

Overall, the studied model pairs are most rigid under within-family seed, loss, and depth changes; increasing $K$ reduces estimation error but not CNN$\rightarrow$ViT deformation, so calibration mismatch is not its sole cause. $D_G$ is comparative, not a universal retrieval threshold.

Table 1: Rigidity across seeds, losses, and architectures on WebFace12M ($K$=5,000 calibration samples). Values are reported as mean±s.d. over the 10 unordered model pairs formed by five training seeds. Cross-configuration rows use 3–5 model pairs, and rows without ± report one canonical pair. $\Delta\theta$ is the mean pairwise angular distortion.

| Model Pair | Change | Res ↓ | ACS ↑ | $\Delta\theta$ (°) |
|---|---|---|---|---|
| R50(Cos) vs R50(Cos) | seed only | $0.436 \pm 0.001$ | $0.905 \pm 0.000$ | 1.29 |
| R100(Cos) vs R100(Cos) | seed only | $0.440 \pm 0.002$ | $0.903 \pm 0.001$ | 1.24 |
| R100(Arc) vs R100(Arc) | seed only | $0.451 \pm 0.003$ | $0.898 \pm 0.001$ | 1.26 |
| ViT-B(Cos) vs ViT-B(Cos) | seed only | $0.361 \pm 0.002$ | $0.935 \pm 0.001$ | 1.37 |
| ViT-B(Arc) vs ViT-B(Arc) | seed only | $0.360 \pm 0.002$ | $0.935 \pm 0.001$ | 1.37 |
| R100(Cos) vs R100(Arc) | loss | $0.445 \pm 0.005$ | $0.901 \pm 0.002$ | 1.25 |
| ViT-B(Cos) vs ViT-B(Arc) | loss | $0.347 \pm 0.005$ | $0.940 \pm 0.002$ | 1.35 |
| R50(Cos) vs R100(Cos) | depth | $0.506 \pm 0.003$ | $0.872 \pm 0.002$ | 1.46 |
| R100(Cos) vs ViT-B(Cos) | arch | $0.879 \pm 0.004$ | $0.614 \pm 0.004$ | 2.81 |
| R50(Cos) vs ViT-B(Cos) | arch | 0.843 | 0.645 | 2.72 |
| R50(Cos) vs R100(Arc) | depth+loss | 0.518 | 0.866 | 1.48 |
| R100(Cos) vs ViT-B(Arc) | arch+loss | 0.881 | 0.612 | 2.82 |

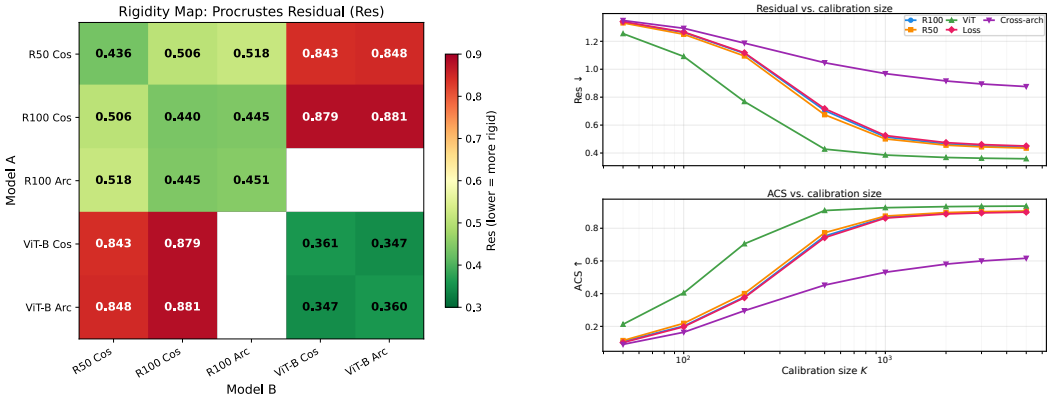


Figure 3: Rigidity across model configurations (left) and alignment quality versus calibration size $K$ (right). The block-diagonal map indicates stronger within-family rigidity; same-architecture pairs approach their plateau by $K \approx 1{,}000$, whereas cross-architecture pairs retain a higher residual.

### 4.3 Does rigidity restore standard verification efficiently?

We next test whether within-family rigidity restores standard cross-model verification without legacy-aware retraining. Table 2 compares closed-form Procrustes with training-based BCRL on the R100-CosFace seed upgrade across LFW, CFP-FP, and AgeDB-30. On this protocol, Procrustes matches the strongest reproduced baselines, using about 30 seconds for dual-encoder feature extraction and less than one second for SVD instead of approximately 23–34 GPU-hours of retraining. This comparison is limited to a same-architecture seed upgrade and excludes residual correction; it does not establish BCRL parity for other upgrades or deployment protocols.

### 4.4 Residual correction for cross-architecture upgrades

CNN→ViT departs most from the single-rotation model (Section 4.2), motivating the optional residual correction in Section 3.5. Table 3 compares direct matching, relative representations [19], a two-layer MLP, Procrustes, and the correction.

The correction lowers the AgeDB-30 residual from 0.636 to 0.450 and raises accuracy from 97.93% to 98.13%; LFW and CFP-FP change by at most 0.02 points. On IFRT (Section 4.5), it improves CNN→ViT compatibility but does not close the old–old retrieval gap.

Table 2: Cost and cross-model verification accuracy (%) for the R100-CosFace seed upgrade. $\Delta_{\mathrm{ref}}$ is the mean accuracy drop from each method's own same-model reference.

| Method | Type | Retrain? | GPU-h | $\Delta_{\mathrm{ref}} \downarrow$ | LFW | CFP-FP | AgeDB-30 |
|---|---|---|---|---|---|---|---|
| Same-model (oracle) | — | — | — | — | 99.85 | 99.31 | 98.23 |
| Direct cross-model | — | — | — | 48.61 | 49.37 | 51.69 | 50.50 |
| BCT [24] | train | yes | $\sim$23 | 0.04 | 99.73 | 98.29 | 97.37 |
| BT$^2$ [31] | train | yes | $\sim$23 | 1.44 | 99.58 | 97.69 | 95.77 |
| AdvBCT [20] | train | yes | $\sim$34 (est.) | 0.06 | 99.82 | 99.21 | 98.10 |
| UniBCT [29] | train | yes | $\sim$25 | 0.07 | 99.85 | 99.21 | 98.03 |
| **Ours (Procrustes)** | **post-hoc** | **no** | $\mathbf{< 0.01}$ | **0.07** | **99.85** | **99.20** | **98.12** |

Table 3: CNN→ViT verification accuracy (%) and AgeDB-30 residual. Source: R100-ArcFace; target: ViT-B-ArcFace; $K$=5,000 shared calibration images.

| Method | LFW | CFP-FP | AgeDB-30 | Res (AgeDB-30) |
|---|---|---|---|---|
| No alignment | 49.50 | 50.00 | 49.80 | — |
| Relative repr | 99.32 | 96.43 | 95.73 | — |
| MLP (2-layer) | 99.73 | 99.01 | 97.67 | 0.540 |
| Procrustes | 99.80 | 99.07 | 97.93 | 0.636 |
| **Proc + MLP-res** | **99.82** | **99.07** | **98.13** | **0.450** |
| Same-model (ViT-B) | 99.82 | 99.29 | 98.15 | — |

### 4.5 Does compatibility survive million-scale identification?

Pairwise verification cannot determine whether small score errors change rankings in a million-scale gallery. We therefore test old-gallery/new-query identification on the original derived IFRT-MR $1:N$ split (Appendix G).

Table 4 shows that direct matching nearly collapses after seed, depth, or architecture changes, whereas loss changes retain partial compatibility. For the tested seed, loss, and depth upgrades, Procrustes recovers at least $99.8\%$ of old–old Rank-1; residual correction provides no consistent additional gain. CNN→ViT remains harder: for R100→ViT-B, correction raises Procrustes Rank-1/Rank-5 from 96.29/98.29 to 98.01/99.23, still below 99.27/99.78 old–old. Thus, compatibility survives million-scale identification for the tested seed, loss, and depth upgrades, but cross-architecture recovery remains incomplete. Backfill-free denotes an unchanged gallery and index, not universal equivalence to full backfilling. The controlled experiment below uses a different split and calibration construction, so its absolute values are reported separately rather than pooled.

### 4.6 Does compatibility generalize across data and protocols?

**Cross-corpus transfer.** To test compatibility beyond a nested training-corpus upgrade, we evaluate MS1MV2-R100-ArcFace→WebFace12M-R100-ArcFace; the checkpoints share the R100 backbone and ArcFace loss but come from different training pipelines. Verification uses $K$=3,000 LFW calibration for CFP-FP/AgeDB-30, whereas identification uses the controlled IFRT setting; Appendix G specifies both protocols. Table 5a shows that Procrustes approaches old–old across all three evaluations but does not recover full new–new performance. Adapter fitting accesses neither training corpus, so neither an inclusion relation between the corpora nor training-set access is an estimator requirement. Together, these results extend compatibility beyond nested-corpus upgrades to independently trained cross-corpus checkpoints with matched backbone and loss.

**Standard and low-quality benchmarks.** Table 5b tests data- and backbone-scale upgrades with external $K$=3,000 LFW calibration. On IJB-B/C, direct matching nearly collapses and Procrustes approaches new–new at FAR $10^{-4}$. Bidirectional IJB-B/C scoring removes pair-order dependence; TinyFace and derived IFRT-MR test query-only deployment. On TinyFace, Procrustes improves over direct matching but does not consistently reach old–old or new–new Rank-1/Rank-5/mAP. Thus, IJB-B/C recovery does not imply complete recovery under low-quality identification.

Table 4: Derived IFRT-MR $1:N$ identification on the original split (Appendix G). Entries report Rank-1/Rank-5 (%); old–old is the legacy reference.

| Upgrade | Old–old | Direct | Procrustes | Proc.+MLP |
|---|---|---|---|---|
| R100(Cos) s1→s2 | 99.27/99.78 | 0.00/0.00 | 99.23/99.76 | 99.20/99.75 |
| ViT-B(Cos) s1→s2 | 98.94/99.67 | 0.00/0.00 | 98.95/99.59 | 98.90/99.61 |
| R100: Cos→Arc | 99.27/99.78 | 86.43/92.32 | 99.18/99.72 | 99.15/99.70 |
| ViT-B: Cos→Arc | 98.94/99.67 | 95.81/98.25 | 98.93/99.62 | 98.89/99.55 |
| R50(Cos)→R100(Cos) | 98.95/99.60 | 0.00/0.00 | 99.10/99.70 | 99.05/99.68 |
| R100(Cos)→ViT-B(Cos) | 99.27/99.78 | 0.00/0.01 | 96.29/98.29 | 98.01/99.23 |
| R50(Cos)→ViT-B(Cos) | 98.95/99.60 | 0.00/0.00 | 96.43/98.48 | 97.97/99.26 |

Table 5: Four-way compatibility evaluations (%). (a) Cross-corpus verification and controlled derived IFRT-MR Rank-1/Rank-5; protocols are in Appendix G. (b) IJB-B/C symmetric TAR at FAR $10^{-4}$ (B/C) and TinyFace Rank-1/Rank-5/mAP.

| (a) Cross-corpus evaluation | Old–old | Direct | Procrustes | New–new |
|---|---|---|---|---|
| CFP-FP accuracy | 98.51 | 49.33 | 98.70 | 99.30 |
| AgeDB-30 accuracy | 98.25 | 49.83 | 97.60 | 98.37 |
| Controlled IFRT R1/R5 | 97.18/98.63 | 0.00/0.00 | 97.29/98.80 | 99.12/99.66 |

| (b) Benchmark | Upgrade | Old–old | Direct | Procrustes | New–new |
|---|---|---|---|---|---|
| IJB-B/C | Data | 95.22/96.87 | 4.02/4.73 | 95.43/97.06 | 95.40/97.02 |
| IJB-B/C | Depth | 95.40/97.02 | 0.24/0.16 | 95.87/97.34 | 96.14/97.49 |
| TinyFace | Data | 26.23/48.77/25.11 | 0.03/0.11/0.04 | 23.95/45.57/23.35 | 26.29/49.25/25.59 |
| TinyFace | Depth | 26.29/49.25/25.59 | 0.00/0.03/0.01 | 26.98/48.52/26.19 | 28.70/52.25/28.78 |

## 5 Limitations and safeguards

The evidence is strongest for same-architecture or same-family upgrades with matched embedding dimensionality and representative calibration. Cross-architecture upgrades remain a stress case, and calibration-distribution mismatch or severe image degradation can leave measurable compatibility gaps. Because calibration residual alone does not reliably predict deployment performance, target-domain validation remains necessary. Appendix D provides an existing-model proxy for class-count/data-scale transfer. However, strict nested-label upgrades ($C_A \subset C_B$), demographic- or device-specific calibration shifts, and different embedding dimensions remain outside the present evidence and require additional validation before deployment.

Because the method lowers the cost of face-recognition upgrades, deployments should include access control, audit logging, consent and policy checks, and subgroup- and quality-stratified false-match audits. The method should not be used to expand non-consensual or high-risk surveillance.

## 6 Conclusion

This paper frames face-model compatibility as coordinate recovery. Across 45 models, many independently trained angular-margin encoders exhibit *representation rigidity*: a global orthogonal transformation approximately relates their embedding spaces, enabling a Procrustes adapter to transport upgraded queries into legacy coordinates. Rigidity is strong within architecture families, largely stable across ArcFace/CosFace changes, and non-monotonic with data scale.

The benefit is clearest for same-family upgrades: on the derived IFRT-MR $1:N$ split with a 1.6M-image gallery, Procrustes retains at least $99.8\%$ of old-old Rank-1 without gallery backfilling. Cross-architecture upgrades remain harder. For CNN↔ViT, Procrustes reaches 97.9–99.8% verification accuracy; the residual MLP substantially narrows the gap observed in large-gallery evaluation.

WF4M/WF12M/WF42M provide class-count/data-scale transfer proxies, while strict nested-label upgrades and demographic- or device-specific calibration shifts remain outside the current evidence. Cross-corpus and low-quality evaluations further show that compatibility recovery can fall short of the full new-model gain.

## Acknowledgments and disclosures

This research was supported by DeepGlint. All authors are employed by DeepGlint. Beyond this disclosed relationship, the authors declare no other competing interests.

## References

[1] Ery Arias-Castro, Adel Javanmard, and Bruno Pelletier. Perturbation bounds for Procrustes, classical scaling, and trilateration, with applications to manifold learning. *Journal of Machine Learning Research*, 21(15):1–37, 2020.

[2] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 225–236, 2021.

[3] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. ElasticFace: Elastic margin loss for deep face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1578–1587, 2022.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 67–74, 2018. doi: 10.1109/fg.2018.00020.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[6] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Computer Vision – ECCV 2020*, volume 12356 of *Lecture Notes in Computer Science*, pages 741–757. Springer, 2020. doi: 10.1007/978-3-030-58621-8_43.

[7] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The InsightFace track report. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1437–1444, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] John C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Computer Vision – ECCV 2016*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016. doi: 10.1007/978-3-319-46487-9_6.

[11] Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas Garg. Canonicalizing multimodal contrastive representation learning. *arXiv preprint arXiv:2602.17584*, 2026.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: Adaptive curriculum learning loss for deep face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020.

[14] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proc. International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024.

[15] InsightFace. MFR-Ongoing (IFRT: InsightFace Recognition Test). `https://github.com/deepinsight/insightface/blob/master/challenges/mfr/README.md`, 2021. Accessed October 2026.

[16] Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–220, 2017.

[18] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021.

[19] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023.

[20] Tan Pan, Furong Xu, Xudong Yang, Sifeng He, Chen Jiang, Qingpei Guo, Feng Qian, Xiaobo Zhang, Yuan Cheng, Lei Yang, and Wei Chu. Boundary-aware backward-compatible representation via adversarial learning in image retrieval. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15201–15210, 2023.

[21] Simone Ricci, Niccolò Biondi, Federico Pernici, Ioannis Patras, and Alberto Del Bimbo. $\lambda$-orthogonality regularization for compatible representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 29036–29063, 2025. doi: 10.52202/085713-0974.

[22] Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1): 1–10, 1966. doi: 10.1007/BF02289451.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[24] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6368–6377, 2020.

[25] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 1988–1996, 2014.

[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, 2014.

[27] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[28] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[29] Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Fanzi Wu, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Towards universal backward-compatible representation learning. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1615–1621, 2022. doi: 10.24963/ijcai.2022/225.

[30] Jiancan Zhou, Xi Jia, Qiufu Li, Linlin Shen, and Jinming Duan. UniFace: Unified cross-entropy loss for deep face recognition. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20730–20739, 2023.

[31] Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. $BT^2$: Backward-compatible training with basis transformation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11229–11238, 2023.

[32] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, 2021.

# A Reproduction of BCRL baselines

This appendix documents the training-based baselines used in the cost–accuracy comparison in Table 2. All reproductions follow the same R100-CosFace seed-upgrade setting as the main text: seed 1 encodes the legacy gallery, while seed 2 is trained as the upgraded query encoder under each compatibility objective. Evaluation follows the BCT old-gallery/new-query verification protocol.

## A.1 Reproduced baselines

**BCT [24].** We follow the original training recipe, keeping the legacy model and classifier fixed while training the new encoder with compatibility constraints that preserve comparability with legacy embeddings.

**BT$^2$ [31].** We reproduce the basis-transformation and augmented-representation components of the original method, matching the architecture and training budget of our main models where possible.

**Universal BCRL / UniBCT [29].** We apply the universal compatibility objective to the same-label R100-CosFace upgrade used in Table 2. The extended-class variant of UniBCT is outside the scope of this reproduction.

**AdvBCT [20].** We adapt boundary-aware adversarial compatibility to our embedding-training pipeline. When benchmark differences preclude a direct comparison, we additionally report retrieval results alongside face-specific metrics.

Table 6: Scope of the BCRL reproductions used in Table 2. All methods follow the same R100-CosFace seed-upgrade setting and are evaluated by old-gallery/new-query verification.

| Method | Objective family | Legacy frozen? | Scope | GPU-h |
|---|---|---|---|---|
| BCT | legacy-classifier compatibility | yes | reproduced | $\sim$23 |
| BT$^2$ | basis transformation | yes | reproduced | $\sim$23 |
| AdvBCT | boundary/adversarial compatibility | yes | reproduced | $\sim$34 |
| UniBCT | universal compatibility | yes | reproduced | $\sim$25 |

## A.2 Reporting conventions

The comparison accounts for same-model and cross-model performance, training compute, and any additional components required at deployment. Table 2 reports the resulting cost–accuracy trade-off.

# B Cross-architecture calibration curve

This appendix extends the cross-architecture stress test in Section 4.4 by varying the calibration size $K$. Figure 4 shows that verification accuracy rises rapidly and stabilizes near $K{=}3{,}000$ for all adapters. Procrustes alone performs strongly in pairwise verification, whereas the benefit of residual correction is more pronounced in large-gallery identification (Section 4.5).

# C Data-scale ablation of representation rigidity

This ablation examines whether increasing the training-data scale makes independently trained feature spaces more rigid. We compare R50-CosFace across all three WebFace scales and R100/ViT-B-CosFace at 12M and 42M, measuring within-configuration rigidity over $S{=}5$ seeds. Figure 5 shows a non-monotonic relationship. For R50, $\mathrm{Res}$ decreases from $0.464$ at 4M to $0.442$ at 12M, but remains similar at 42M ($0.446$). For ViT-B, it increases from $0.375$ at 12M to $0.428$ at 42M ($+14\%$ relative), indicating lower rigidity. Thus, improvements in recognition performance with more training data need not coincide with greater geometric similarity between independent runs. At both shared scales, ViT-B models remain more rigid than ResNets ($\mathrm{Res} \approx 0.36$–$0.43$ vs. $0.44$–$0.46$); we treat this difference as an empirical regularity rather than evidence of an architectural mechanism.

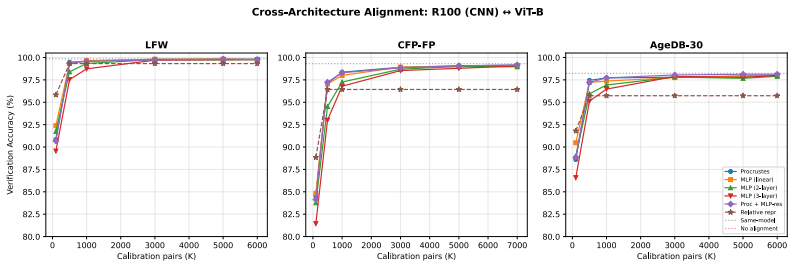


Figure 4: Cross-architecture verification accuracy versus the number of same-input calibration samples $K$ for R100→ViT-B on LFW, CFP-FP, and AgeDB-30. Adapter performance stabilizes near $K$=3,000; horizontal lines denote fixed references.

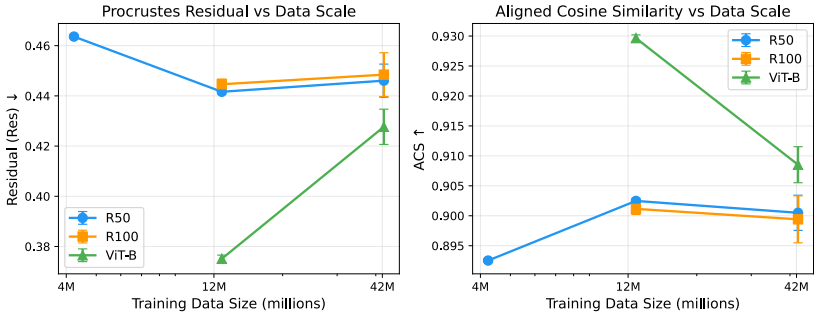


Figure 5: Representation rigidity versus training-data scale: Procrustes residual Res (left; lower is better) and aligned cosine similarity ACS (right; higher is better). Error bars show $\pm 1$ standard deviation across the 10 unordered model pairs formed by five seeds. Both metrics vary non-monotonically with data scale.

## D Existing-model class-count/data-scale transfer proxy

This appendix evaluates an existing-model transfer proxy using R50-CosFace checkpoints trained on WebFace4M, WebFace12M, and WebFace42M. The checkpoints differ in both training-set scale and classifier class count, allowing us to test whether feature-space alignment remains effective across these model upgrades.

The available checkpoints are WF4M-R50-CosFace (205,990 identities, 4.24M images), WF12M-R50-CosFace (617,970 identities, 12.72M images), and WF42M-R50-CosFace (2,059,906 identities, 42.47M images). Because their label sets are not documented as nested, this experiment is a class-count/data-scale proxy rather than a validation of strict $C_A \subset C_B$ upgrades.

Table 7: Existing-model class-count/data-scale transfer on LFW, CFP-FP, and AgeDB-30. Each row averages two seed-matched R50-CosFace upgrade pairs across the three benchmarks. "Direct" compares raw new-query and legacy-gallery features; "Proc." applies the LFW-calibrated orthogonal adapter.

| Upgrade | Class ratio | Res/ACS | Direct Acc. (%) | Proc. Acc. (%) |
|---|---|---|---|---|
| WF4M→WF12M | 3.0× | 0.443 / 0.901 | 66.88 | 98.85 |
| WF12M→WF42M | 3.3× | 0.395 / 0.921 | 52.00 | 98.98 |
| WF4M→WF42M | 10.0× | 0.523 / 0.861 | 51.81 | 98.78 |

Table 7 shows that direct cross-model accuracy falls to 51.81–66.88% as class count and data scale change, whereas Procrustes reaches 98.78–98.98%. In this proxy, feature-space alignment remains effective even when classifier widths differ by up to 10×. This result does not establish compatibility for strict nested-label upgrades.

Table 8: Calibration-domain proxy matrix. Each cell reports cross-model verification accuracy (%) when $\hat{R}$ is estimated on the calibration benchmark and evaluated on the target benchmark. Direct unaligned cross-model accuracy is near chance (49.37/51.69/50.50 on LFW/CFP-FP/AgeDB-30).

| Calibration | Cal. Res/ACS | Target: LFW | Target: CFP-FP | Target: AgeDB-30 |
|---|---|---|---|---|
| LFW | 0.360 / 0.935 | 99.82 | 99.31 | 98.18 |
| CFP-FP | 0.334 / 0.944 | 99.85 | 99.29 | 98.05 |
| AgeDB-30 | 0.309 / 0.952 | 99.82 | 99.27 | 98.13 |

# E Calibration domain shift: proxy matrix experiments

This appendix addresses a deployment question left open by the main experiments: does the calibration set need to match the target domain? Table 8 reports a proxy calibration-shift experiment where the adapter is calibrated on one standard verification benchmark and evaluated on another. These benchmarks differ in pose and age composition (LFW, CFP-FP, AgeDB-30), so the experiment tests distribution transfer at the benchmark level; it does not replace demographic, sensor, or deployment-site audits.

## E.1 Domain definitions

Let $\mathcal{D}_{\text{cal}}$ be the calibration benchmark and $\mathcal{D}_{\text{tgt}}$ the target benchmark. We use legacy R100-CosFace seed 1 as the gallery encoder and R100-CosFace seed 2 as the upgraded query encoder. For each calibration benchmark, we estimate $\hat{R}$ using up to $K{=}3{,}000$ first images from verification pairs and evaluate old-gallery/new-query cross-verification on the target benchmark.

## E.2 Calibration–target matrix

Table 8 shows that cross-benchmark calibration remains stable for this same-architecture seed upgrade once $K$ is sufficiently large: all nine calibration–target combinations stay above $98\%$ verification accuracy. This subsection is therefore useful as the evidence table for the calibration-shift claim, while Section E.3 records the failure mode when calibration is too small.

## E.3 Calibration-size sensitivity and limitations

The proxy matrix is stable once $K$ is moderately large: averaged over all 9 calibration–target pairs, accuracy is 99.02% at $K{=}500$, 99.13% at $K{=}1{,}000$, and 99.08% at $K{=}3{,}000$. With only $K{=}100$, however, the average drops to 89.41% and the worst cell is 84.92%, showing that under-specified calibration is a real failure mode. In deployment, we recommend target-domain recalibration when possible, or a multi-adapter strategy estimated per domain/quality cluster. We leave demographic, sensor, and site-specific calibration audits to future work because the public verification benchmarks used here do not provide the necessary operational metadata.

# F Geometry, calibration transfer, and retrieval margins

Orthogonal symmetry makes a global rotation a plausible source of cross-model incompatibility, but it does not guarantee that two independently trained encoders differ only by such a rotation. Even when an orthogonal map fits a calibration set well, deployment additionally depends on whether the calibration distribution represents the target queries and whether the remaining score perturbations are small relative to the gallery margins. This appendix makes these dependencies explicit.

The analysis follows the operational chain from geometry to retrieval. We first state the estimator assumptions and characterize exact and approximate orthogonal alignability through Gram geometry. We then separate finite-sample recovery from calibration-distribution transfer before translating a per-query residual into a sufficient Rank-$k$ preservation condition. These results distinguish complementary failure modes; none is intended as a universal success threshold.

**Conventions.** Embeddings $z_A, z_B$ are columns, and $R$ maps legacy to upgraded columns: $z_B \approx Rz_A$. For row matrices $X = Z_A$ and $Y = Z_B$ computed from the same inputs, $Y \approx XR^\top$. If $X^\top Y = U\Sigma V^\top$, then $\hat{R} = VU^\top$. At deployment, new columns use $\hat{R}^\top z_B$; equivalently, new rows use $Y\hat{R}$. The results below apply to the orthogonal adapter; a fitted residual MLP need not preserve inner products.

**Estimator assumptions.** The orthogonal estimator requires equal embedding dimensionality and same-input calibration features from the two fixed encoders. It does not require identity labels, an inclusion relation between the training sets, classifier prototypes, or gallery embeddings. The correspondence is between the two feature vectors computed from each calibration image; it does not pair different images by identity. Transfer beyond calibration remains conditional on approximate geometric compatibility and a representative calibration distribution.

**Exact Gram equivalence.** For $X, Y \in \mathbb{R}^{n\times d}$, there exists an orthogonal $Q$ with $YQ = X$ if and only if $YY^\top = XX^\top$. The forward implication follows immediately from $QQ^\top = I$. For the converse, take a compact eigendecomposition of the common Gram matrix, $P\Lambda P^\top$, with rank $r$. The corresponding factorizations have the form $X = P\Lambda^{1/2}V_X^\top$ and $Y = P\Lambda^{1/2}V_Y^\top$, where $V_X, V_Y$ have orthonormal columns. Extend $V_X$ and $V_Y$ to orthogonal bases $\bar{V}_X$ and $\bar{V}_Y$, and set $Q = \bar{V}_Y\bar{V}_X^\top$. Then $V_Y^\top Q = V_X^\top$, and hence $YQ = X$. The zero-rank case is immediate. Full column rank is not required for existence, although deficient rank can make the map non-unique outside the observed span.

**Approximate Gram obstruction.** For any $Q \in \mathbb{O}(d)$, let $E_Q = YQ - X$. Then

$$YY^\top - XX^\top = E_Q(YQ)^\top + XE_Q^\top, \qquad \|YY^\top - XX^\top\|_F \le (\|Y\|_2 + \|X\|_2)\|E_Q\|_F.$$

Minimizing over $Q$ proves Eq. (6). Here $\|\cdot\|_2$ is the spectral matrix norm. Unit rows imply only equality of the Gram diagonals. In contrast, the unaligned feature-coordinate covariance difference $\|X^\top X - Y^\top Y\|_F$ need not vanish under exact orthogonal equivalence: the covariance changes by orthogonal conjugation. We therefore use the sample Gram matrices for rotation-invariant mismatch diagnostics. A small empirical Gram discrepancy is not, by itself, a distribution-free guarantee of small deployment error.

**Finite-sample recovery and conditioning.** Under the calibration model $Y = XR_\star^\top + E$ with $\sigma_{\min}(X) > 0$, Procrustes optimality gives $\|X(\hat{R} - R_\star)^\top\|_F \le 2\|E\|_F$, which yields Eq. (3). The bound separates two finite-sample factors: the residual magnitude $\|E\|_F$ and the conditioning of the legacy calibration features through $\sigma_{\min}(X)$. It is conditional on a specified $R_\star$ and does not assert a universal shared coordinate system across arbitrary training runs.

**Transfer across calibration distributions.** Conditioning alone does not ensure that the calibration objective matches deployment. Let $D$ be a distribution over unit-normalized, same-input row features $(x, y)$, and define

$$L_D(Q) = \mathbb{E}_D\|yQ - x\|_2^2 = 2 - 2\,\mathrm{tr}(Q^\top M_D), \qquad M_D = \mathbb{E}_D[y^\top x]. \tag{9}$$

Let $Q_C$ minimize the calibration loss $L_C$ defined by Eq. (9), and let $Q_T$ minimize the corresponding target loss $L_T$. With $\Delta M = M_T - M_C$, their target-loss gap satisfies

$$L_T(Q_C) - L_T(Q_T) \le 4\|\Delta M\|_* \le 4\sqrt{d}\,\|\Delta M\|_F, \tag{10}$$

where $\|\cdot\|_*$ is the nuclear norm. Indeed,

$$\begin{aligned} L_T(Q_C) - L_T(Q_T) = {} & [L_T(Q_C) - L_C(Q_C)] + [L_C(Q_C) - L_C(Q_T)] \\ & + [L_C(Q_T) - L_T(Q_T)]. \end{aligned} \tag{11}$$

The middle term in Eq. (11) is non-positive by optimality of $Q_C$. The remaining terms are bounded using $|\,\mathrm{tr}(Q^\top \Delta M)| \le \|\Delta M\|_*$ for orthogonal $Q$, followed by $\|\Delta M\|_* \le \sqrt{d}\|\Delta M\|_F$.

The bound in Eq. (10) separates estimation from transfer. Calibration size and conditioning govern how accurately the empirical adapter estimates $Q_C$, whereas a change from $M_C$ to $M_T$ changes the population objective itself. More samples from a mismatched calibration distribution therefore

need not recover $Q_T$. The oracle loss $L_T(Q_T)$ measures the deformation left after the best target-domain orthogonal fit, while $L_T(Q_C) - L_T(Q_T)$ measures the additional calibration-transfer error. Consequently, a low in-sample fitting residual is not sufficient for adapter selection; representative calibration and held-out validation against the intended gallery remain necessary. Appendix G.3 reports the corresponding controlled diagnostics.

**Score perturbation and Rank-$k$ preservation.** For a query, write $e_q = \hat{R}^\top z_B(x_q) - z_A(x_q)$ and assume all legacy gallery vectors have unit norm. Cauchy–Schwarz gives $|\langle e_q, g\rangle| \le \|e_q\|_2$ for every gallery item. Thus every gallery score changes by at most $\|e_q\|_2$.

Let $s_+$ be the maximum legacy score among same-identity gallery items, and let $s_{-,k}$ be the $k$-th largest score among different-identity items. After transport, the best-positive score is at least $s_+ - \|e_q\|_2$, while the $k$-th largest negative score is at most $s_{-,k} + \|e_q\|_2$. Therefore,

$$s_+ - s_{-,k} > 2\|e_q\|_2 \tag{12}$$

Condition (12) is sufficient for at least one positive to remain within the top $k$, with strict inequality avoiding ties. The case $k = 1$ gives the Rank-1 condition used in the main text. For a fixed gallery, replacing $\|e_q\|_2$ by $\delta_q = \max_{g\in\mathcal{G}} |\langle e_q, g\rangle|$ gives a tighter certificate.

These conditions are sufficient, not necessary. They do not imply that every positive query is certified, that the full ranking is unchanged, or that transported queries attain new–new/full-backfill performance. Computing the certificate requires same-input old/new query features, and the positive-versus-negative margin additionally uses identity labels; neither requirement is part of adapter fitting.

# G Supplementary protocols and diagnostics

This appendix documents the evaluation protocols and diagnostic definitions used in Section 4.6. Table 5 reports the main identification and verification results.

## G.1 Cross-corpus protocols and additional results

For the MS1MV2-to-WebFace12M R100-ArcFace upgrade, we calibrate on $K$=3,000 LFW image instances sampled with seed 20260724 and evaluate on CFP-FP and AgeDB-30. For both calibration and evaluation, we sum the original-view and horizontal-flip features before normalization. Each encoder retains its native input normalization, and verification accuracy is measured by ten-fold cross-validation.

Table 9: Cross-corpus verification for MS1MV2-R100-ArcFace → WebFace12M-R100-ArcFace. Accuracy (%) uses ten-fold cross-validation; adapter calibration uses $K$=3,000 LFW image instances with original+flip features.

| Benchmark | Old–old | Direct | Procrustes | New–new |
|---|---|---|---|---|
| CFP-FP | 98.51 | 49.33 | 98.70 | 99.30 |
| AgeDB-30 | 98.25 | 49.83 | 97.60 | 98.37 |

**Complementary derived IFRT-MR settings.** Table 4 retains the original broad-comparison setting: 10,000 new-model queries, a 1,614,305-image old-model gallery, and $K$=5,000 calibration images. The query–gallery split is generated deterministically from the ordered IFRT label vector with `numpy.random.RandomState(42)`. Among rows with nonnegative labels, we sample 10,000 identities without replacement and then one query image per selected identity; the gallery is the complement of these query rows. The calibration images are sampled with an independently initialized `RandomState(42)` from a separate, ordered in-domain calibration pool that is disjoint from the query–gallery image set. Thus, adapter fitting uses no evaluation images, although this original setting does not require calibration and evaluation identities to be disjoint. Given the ordered IFRT label vector and calibration pool, the query, gallery, and calibration manifests are recoverable from the fixed seed. This protocol compares seed, loss, depth, and architecture upgrades. Table 5a instead uses a controlled setting for two analyses: cross-corpus compatibility without requiring

training-corpus inclusion, and calibration-source transfer under a fixed retrieval population. Beyond image-level separation, this split enforces identity-level separation: it reserves 3,000 calibration identities outside the 10,000 query identities and removes their 20,109 images from the gallery, leaving $1{,}624{,}305 - 20{,}109 - 10{,}000 = 1{,}594{,}196$ gallery images. All calibration-source conditions share these queries and gallery and use $K$=3,000 single-view features. Identity labels define the disjoint split and evaluation ground truth; adapter fitting uses only same-input feature correspondences across the two encoders. The cross-corpus comparison accesses neither training corpus, while the calibration-source comparison changes only the calibration data. Because the original and controlled settings use different split and calibration constructions, their absolute results are not pooled. We do not assume that the external corpora are identity-disjoint. For the additional CFP-FP calibration condition, cross-corpus Procrustes reaches 87.92/92.58% Rank-1/Rank-5; old–old, direct, and new–new are identical to Table 5a.

### G.2 Benchmark scoring and preprocessing

**IJB-B/C score direction.** For normalized legacy and upgraded row-template features $x_i$ and $y_i$, respectively, the row adapter is $\widehat{Q} = \widehat{R}$ (Appendix F). We define the forward and reverse scores as $s_{\mathrm{f}}(i,j) = \langle x_i, y_j \widehat{Q} \rangle$ and $s_{\mathrm{r}}(i,j) = \langle y_i \widehat{Q}, x_j \rangle$. Their mean gives the symmetric scores in Table 5b. Direct matching replaces $\widehat{Q}$ with the identity. Table 10 reports the complementary forward-only results in the official pair order.

Table 10: Forward-order IJB-B/C $1{:}1$ template verification. TAR (%) at nominal FAR $10^{-4}$ uses the nearest observed ROC operating point; Table 5b reports symmetric scoring.

| Dataset | Upgrade | Old–old | Direct | Procrustes | New–new |
|---|---|---|---|---|---|
| IJB-B | Data scale | 95.22 | 0.47 | 95.18 | 95.40 |
| IJB-B | Backbone | 95.40 | 0.00 | 95.55 | 96.14 |
| IJB-C | Data scale | 96.87 | 0.66 | 96.85 | 97.02 |
| IJB-C | Backbone | 97.02 | 0.00 | 97.25 | 97.49 |

**FAR convention and sensitivity.** The primary evaluation reports the observed ROC point nearest the requested FAR, so the realized FAR may lie on either side of the target. For a conservative check, we choose an order-statistic threshold from the $\lfloor \alpha N_- \rfloor$ highest negative scores and accept scores greater than or equal to that threshold. In these runs, the resulting FAR does not exceed $\alpha$, although this rule need not maximize TAR subject to that ceiling. At $\alpha = 10^{-4}$, the two conventions differ by at most 0.01 percentage points across both score directions and all four comparisons in Tables 5b and 10.

Table 11: Extreme-FAR sensitivity of symmetric IJB-B/C scoring. TAR (%) is reported at nominal FAR $10^{-6}$ using either the nearest observed ROC point or a conservative negative-score order statistic whose measured FAR is at most $10^{-6}$ in these runs.

| Dataset | Upgrade | Threshold | Old–old | Direct | Procrustes | New–new |
|---|---|---|---|---|---|---|
| IJB-B | Data scale | Nearest | 42.44 | 0.59 | 41.87 | 43.52 |
| IJB-B | Data scale | Conservative | 42.27 | 0.58 | 37.12 | 43.40 |
| IJB-B | Backbone | Nearest | 43.52 | 0.01 | 36.40 | 45.20 |
| IJB-B | Backbone | Conservative | 43.40 | 0.00 | 35.76 | 45.05 |
| IJB-C | Data scale | Nearest | 89.48 | 0.82 | 91.48 | 89.55 |
| IJB-C | Data scale | Conservative | 88.75 | 0.78 | 91.31 | 88.46 |
| IJB-C | Backbone | Nearest | 89.55 | 0.01 | 91.58 | 89.01 |
| IJB-C | Backbone | Conservative | 88.46 | 0.00 | 90.67 | 88.46 |

At FAR $10^{-6}$, the threshold convention has a larger effect, and IJB-B retains a material gap to new–new. Thus, the near-new–new results at FAR $10^{-4}$ do not establish uniform equivalence across operating points. Small differences are descriptive point estimates rather than evidence of statistically significant superiority.

**TinyFace preprocessing.** TinyFace contains 3,728 probes, 4,443 matched-gallery images, and 153,428 distractors, for a total gallery size of 157,871. We convert the released cropped faces to RGB and resize them to $112 \times 112$ by bilinear interpolation, without landmark alignment. The released probe and matched-gallery order is retained, while distractor filenames are sorted. Rank-1 and Rank-5 follow the released CMC convention, and mAP uses trapezoidal AP. Cosine and Euclidean distance induce the same ranking for normalized features. Table 5b reports the complete four-way comparison; mixed-model conditions use new-model queries against the unchanged legacy gallery.

### G.3 Geometry and calibration diagnostics

For unit-normalized legacy and upgraded target-feature matrices $X, Y \in \mathbb{R}^{n \times d}$ computed from the same inputs, define

$$D_G = \frac{\|YY^\top - XX^\top\|_F}{\|XX^\top\|_F}, \qquad e_* = \min_{Q^\top Q = I} \frac{\|YQ - X\|_F}{\sqrt{n}}. \tag{13}$$

In Eq. (13), the target-fitted oracle residual $e_*$ is the error after the best orthogonal fit on the target probes; this fit is used only for diagnosis. The calibrated residual $e_{\text{cal}} = \|Y\widehat{Q} - X\|_F/\sqrt{n}$ instead evaluates an independently calibrated map. Except for the cross-corpus pair, all models are trained on WebFace12M. Seed comparisons use CosFace seeds 1 and 2; loss comparisons use ArcFace and CosFace at seed 1. Depth and CNN–ViT comparisons use CosFace seed 1. All calibrated conditions use $K$=3,000 single-view images.

Table 12: Geometry diagnostics on the controlled 10,000-probe split. $D_G$ is the legacy-normalized sample-Gram discrepancy; $e_*$ and $e_{\text{cal}}$ are root-mean-square same-input residuals. The oracle is fitted on target probes for diagnosis only. "Cert. R1/R5" is the fraction (%) satisfying the $\ell_2$ sufficient condition, not retrieval accuracy.

| Pair | Calibration | $D_G$ | $e_*$ | $e_{\text{cal}}$ | Cert. R1/R5 |
|---|---|---|---|---|---|
| R100 seed | LFW | 0.538 | 0.450 | 0.514 | 0.00/0.00 |
| ViT seed | LFW | 0.410 | 0.353 | 0.393 | 0.13/0.25 |
| R100 Arc→Cos | LFW | 0.547 | 0.457 | 0.522 | 0.00/0.00 |
| ViT Arc→Cos | LFW | 0.399 | 0.342 | 0.377 | 0.26/0.55 |
| R50–R100 | LFW | 0.608 | 0.512 | 0.583 | 0.00/0.00 |
| R100–ViT | LFW | 1.169 | 0.849 | 1.028 | 0.00/0.00 |
| Cross-corpus | LFW | 0.850 | 0.861 | 1.029 | 0.00/0.00 |
| Cross-corpus | IFRT | 0.850 | 0.861 | 0.908 | 0.00/0.00 |

Table 13 separates in-sample calibration fit from target-probe error. In the evaluated conditions, external calibration yields a lower fitting residual but a larger held-out residual; the corresponding identification accuracy is also lower than with IFRT calibration.

Table 13: In-sample calibration and held-out same-input root-mean-square residuals under the controlled IFRT setting. All conditions use $K$=3,000 single-view images.

| Pair | Calibration | $e_{\text{fit}}$ | $e_{\text{heldout}}$ |
|---|---|---|---|
| Cross-corpus | LFW | 0.788 | 1.029 |
| Cross-corpus | CFP-FP | 0.787 | 1.046 |
| Cross-corpus | IFRT | 0.832 | 0.908 |
| CNN–ViT | LFW | 0.789 | 1.028 |
| CNN–ViT | IFRT | 0.818 | 0.900 |

**Interpreting the certificates.** For query residual $r_q = y_q\widehat{Q} - x_q$, Appendix F gives the sufficient Rank-$k$ condition $s_+ - s_{-,k} > 2\|r_q\|_2$. Rank-1 uses the highest negative score ($k = 1$), whereas Rank-5 uses the fifth-highest negative score ($k = 5$). Queries without a positive gallery match are not certified. The fractions in Table 12 report certificate coverage rather than accuracy. Coverage is near zero for most pairs even when identification accuracy is high, reflecting the conservativeness of the sufficient condition. The certificates are therefore not stand-alone deployment guarantees; retrieval must be validated against the intended gallery.

# H Training and adapter configurations

This appendix specifies the training configurations for the 45-model core suite and the optional residual adapter. Appendices E and G document evaluation-specific calibration sources, splits, and scoring rules.

## H.1 Core 45-model suite

Each of the nine configurations in Table 14 is trained with seeds $\{2048, 2049, 2050, 2051, 2052\}$, yielding 45 checkpoints. All encoders output 512-dimensional features and train for 20 epochs with mixed precision and a full classifier (sampling rate 1.0). The angular-softmax scale is 64; CosFace uses $(m_1, m_2, m_3) = (1, 0, 0.4)$ and ArcFace uses $(1, 0.5, 0)$.

Table 14: Training matrix for the 45-model core suite. Each row uses all five seeds; counts are those used for training.

| Training set | Images | Identities | Backbone | Objective |
|---|---|---|---|---|
| WebFace4M | 4,235,242 | 205,990 | IResNet-50 | CosFace |
| WebFace12M | 12,720,066 | 617,970 | IResNet-50 | CosFace |
| WebFace42M | 42,474,557 | 2,059,906 | IResNet-50 | CosFace |
| WebFace12M | 12,720,066 | 617,970 | IResNet-100 | CosFace |
| WebFace42M | 42,474,557 | 2,059,906 | IResNet-100 | CosFace |
| WebFace12M | 12,720,066 | 617,970 | ViT-B | CosFace |
| WebFace42M | 42,474,557 | 2,059,906 | ViT-B | CosFace |
| WebFace12M | 12,720,066 | 617,970 | IResNet-100 | ArcFace |
| WebFace12M | 12,720,066 | 617,970 | ViT-B | ArcFace |

**IResNet optimization.** IResNet-50 and IResNet-100 use SGD with momentum 0.9, weight decay $5 \times 10^{-4}$, learning rate 0.2, per-GPU batch size 256, and no gradient accumulation. The learning rate warms up linearly for two epochs and then follows first-order polynomial decay to zero; backbone gradients are clipped to norm 5.

**ViT-B optimization and architecture.** ViT-B uses $112 \times 112$ inputs, $9 \times 9$ non-overlapping patches, width 512, 24 transformer blocks, eight attention heads, stochastic-depth rate 0.05, and mask ratio 0.05. AdamW uses learning rate $10^{-3}$, weight decay 0.05, per-GPU micro-batch size 512, and eight-step gradient accumulation. The optimizer updates every eight micro-batches, whereas the historical implementation advances the scheduler after each micro-batch. The learning rate warms up for four epochs and then follows first-order polynomial decay to zero; backbone gradients are clipped to norm 5.

**Input pipeline and compute.** Runs use eight NVIDIA L20 GPUs (46 GB each). Images are decoded as RGB at $112 \times 112$ and normalized with channel-wise mean and standard deviation 127.5. DALI applies horizontal flipping ($p$=0.5), downsize–resize degradation ($p$=0.1; intermediate side 56–89 pixels), Gaussian blur ($p$=0.2; kernel size 3 or 5), HSV perturbation ($p$=0.2; hue $[0, 20]$, saturation $[1, 1.2]$), and grayscale conversion ($p$=0.1). Python, NumPy, and PyTorch use the configuration seed.

## H.2 Optional residual correction

The residual path in Eq. (7) takes the normalized Procrustes-transported feature $u \in \mathbb{R}^{512}$. A two-layer MLP, $512 \to 256 \to 512$, with an intermediate ReLU predicts a residual that is added to $u$ and L2-normalized. Both encoders and the fitted Procrustes map remain frozen.

The adapter reuses the $K$ same-input calibration features from the corresponding evaluation protocol. Python, NumPy, and PyTorch use seed 42 for initialization and minibatch ordering. Training runs for 200 epochs with Adam (learning rate $10^{-3}$, $(\beta_1, \beta_2)$=$(0.9, 0.999)$, no weight decay, batch size 512) and cosine decay to zero; samples are reshuffled each epoch. We minimize one minus the mean cosine similarity to legacy features and retain the state with the lowest mean training loss. Large-gallery experiments use $K$=5,000; training and inference each use one L20 GPU.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 describe the empirical scope of representation rigidity, the closed-form query-side transport protocol, and the optional residual correction. They do not claim universal orthogonal equivalence or parity with full backfilling. Section 5 and the appendices state the tested settings and remaining limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sections 4.2–4.6, Section 5, and Appendices D–G discuss cross-architecture and low-quality gaps, small or mismatched calibration sets, and the limits of the class-count/data-scale proxy. Strict nested-label upgrades, demographic- or device-specific shifts, and unequal embedding dimensions remain outside the present evidence.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Lemma 3.1 is proved directly, and the conditional finite-calibration recovery bound follows from Procrustes optimality and the triangle inequality. Appendix F derives the Gram obstruction, calibration-distribution transfer bound, and Rank-$k$ condition. The paper does not claim convergence of independent training runs or universal orthogonal equivalence.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix H specifies the complete architectures, training configurations, and seeds needed to regenerate the 45-model suite and residual adapter. Appendix G gives the deterministic seed-42 construction of the original derived IFRT-MR query, gallery, and calibration manifests, as well as the controlled and added evaluation protocols. Together with the estimators and metrics in the main text, these details disclose the information needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No public code or checkpoint package accompanies this paper. The evaluated face datasets and checkpoints remain subject to their original access and license conditions, and the algorithms and protocols in the manuscript are not an open package sufficient to reproduce every reported result.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies datasets, model families, losses, seeds, calibration sizes, metrics, and evaluation protocols; Appendix G details the added benchmarks; and Appendix H gives the optimizer, schedule, preprocessing, architecture, compute, and residual-adapter configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Table 1 and Appendix C report mean±standard deviation over the 10 unordered pairs formed by five seeds; these pairs share models and are not independent replicates. The deployment and added evaluations use fixed checkpoint pairs and splits without confidence intervals or repeated-run variance, so uncertainty is not characterized for every result.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Experiments used NVIDIA L20 GPUs (46 GB). Section 4.3 reports representative feature-extraction and SVD time, and Table 2 reports BCRL GPU-hours. Per-experiment wall-clock time, peak memory, and worker counts are not reported for all residual-adapter and added-benchmark runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and conform with the NeurIPS Code of Ethics. Section 5 discusses misuse risks specific to face recognition and recommends consent, policy, access-control, and auditing safeguards.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 notes that backfill-free upgrades can reduce deployment cost but may also lower the barrier to larger or more frequent face-recognition deployments. It recommends access control, consent and policy checks, and subgroup-, quality-, and calibration-shift audits.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No new dataset or model checkpoint is released with this paper. Section 5 nevertheless recommends access control, audit logging, consent and policy checks, prohibition of non-consensual or high-risk surveillance use, and target-domain, subgroup, and quality audits for deployments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the principal training datasets, architectures, objectives, and IFRT source. It does not explicitly document the license and access terms for every benchmark and pretrained checkpoint used, including LFW, CFP-FP, AgeDB-30, IJB-B/C, TinyFace, and MS1MV2.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new dataset, model checkpoint, or public software artifact is released. The contributions are the method, theoretical analysis, and empirical evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not recruit participants or conduct crowdsourcing; it evaluates existing face-recognition datasets and checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study involves no participant recruitment, intervention, or new collection of identifiable human-subject data; it uses existing benchmarks and checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used in the core method, data processing, model training, adapter fitting, or evaluation. Any writing or editing assistance does not affect the scientific methodology or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.