A. Few-Shot Verification: Full Protocol

This section provides complete implementation-ready details for the open-set verification experiments in Sec. 4.1, including dataset curation, pair construction, threshold calibration, hyperparameter selection, and resampling.

A.1. Dataset Curation

We curate two balanced identity subsets with strict identity separation:

CelebA-20. 6,348 identities with exactly 20 images each (126,960 total images). Open-set split: 320 train / 80 val / 80 test identities (1,600 test images).

VGGFace2-20. 480 identities with exactly 20 images each (9,600 total images). Open-set split: 320 train / 80 val / 80 test identities (1,600 test images).

All splits are identity-disjoint: $\mathcal{A}_{\text{train}} \cap \mathcal{A}_{\text{val}} \cap \mathcal{A}_{\text{test}} = \varnothing$, ensuring no identity appears in multiple splits.

The full CelebA and VGGFace2 datasets contain variable numbers of images per identity and heterogeneous crops (tight face crops, wider scenes, varying resolutions). Our balanced subsets standardize these nuisance factors so that FAR calibration and per-identity pair sampling are comparable across probes, models, and projection methods. This is important because unbalanced identity counts would distort the impostor distribution and make FAR estimates unreliable across experimental conditions.

A.2. Pair Construction

For each identity i in a given split, we form mated (same-identity) and impostor (different-identity) pairs exclusively from the query set Q_i; the support set S_i is used only to train the probe and never appears in evaluation pairs:

$$\mathcal{P}_{\text{mated}} = \{(z_a, z_b) : a, b \in Q_i,\ a \neq b\}, \tag{12}$$

$$\mathcal{P}_{\text{impostor}} = \{(z_a, z_b) : a \in Q_i,\ b \in Q_j,\ i \neq j\}. \tag{13}$$

Mated pairs are exhaustive within an identity: for an identity with n embeddings split into $|S_i| = k$ support and $|Q_i| = n - k$ query samples, all $\binom{n-k}{2}$ within-identity query pairs are included.

Impostor pairs are subsampled with *identity-balanced quotas*: for each identity i, we sample a fixed number of cross-identity comparisons per other identity j, ensuring that no single identity pair dominates the impostor distribution. This stabilizes the FAR denominator across different values of k and across resampling seeds.

A.3. Pair Counts and FAR Precision

FAR precision is $1/|\mathcal{P}_{\text{impostor}}|$, representing the smallest measurable FAR increment. Table 3 reports exact counts for each dataset.

Table 3. Test set pair counts. Min. FAR shows the minimum measurable FAR ($1/n_{\text{impostor}}$).

Dataset	Test IDs	Mated	Impostor	Min. FAR
CelebA-20	80	15,200	$\sim$6,400	1.6×10^{-4}
VGGFace2-20	80	15,200	$\sim$16,000	6.3×10^{-5}
LFW-20	13	78	1,326	7.5×10^{-4}

CelebA-20 and VGGFace2-20 evaluated at TAR@FAR$= 10^{-4}$. LFW-20 uses partial AUC at FAR$\in [0, 10^{-3}]$ due to insufficient impostor pairs.

A.4. Threshold Calibration

We calibrate a verification threshold τ on validation identities $\mathcal{A}_{\text{val}}$ only. The procedure is as follows:

1. Compute all mated and impostor similarity scores on $\mathcal{A}_{\text{val}}$.
2. When the impostor count permits ($|\mathcal{P}^{\text{val}}_{\text{impostor}}| \geq 10{,}000$), select τ such that $\text{FAR}(\tau) \approx 10^{-4}$.
3. When the impostor count is insufficient for stable 10^{-4} estimation (e.g., LFW-20), fall back to a fixed-head partial AUC criterion at FAR$\in [0, 10^{-3}]$ and still return a single τ.
4. A *separate* τ is calibrated for each (dataset, encoder, projection method, k) combination.
5. Once set on validation identities, τ is *held fixed* for all test resamplings—the five seeds that reshuffle S_i/Q_i per identity do not each get their own threshold.

This ensures that no information from test identities influences the operating point. We report TAR on test along with the achieved FAR (as both a rate and raw count of false accepts over total impostor comparisons) so that readers can verify the operating point is meaningful.

A.5. Hyperparameter Selection

We train a ridge regression verifier on support embeddings from $\mathcal{A}_{\text{train}}$ and evaluate TAR@FAR on $\mathcal{A}_{\text{test}}$. We sweep L_2 regularization strength $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ on $\mathcal{A}_{\text{val}}$ and select the setting that maximizes verification performance under the target operating point (TAR@FAR $\approx 10^{-4}$). The chosen α and threshold τ are frozen before any evaluation on $\mathcal{A}_{\text{test}}$.

A.6. k-Shot Sampling and Resampling

For k-shot evaluation where $k \in \{1, 4, 16\}$:

1. For each identity i, randomly split B_i into query set Q_i and support set S_i with $|S_i| = k$.
2. Train the probe (Ridge or MLP) on support embeddings from training identities only.
3. Calibrate τ on validation identities (as above).
4. Evaluate on test identities using the fixed τ.
5. Repeat with 5 random seeds for S_i/Q_i assignment; report mean TAR and 95% identity-aware confidence intervals.

"Identity-aware" refers to the fact that variance is computed across seeds (which reshuffle per-identity splits) rather than across individual pairs, avoiding pseudo-replication from correlated pairs within the same identity.

A.7. LFW-20 Sanity Check

LFW-20 has insufficient impostor pairs (1,326) for reliable TAR@FAR$= 10^{-4}$ calibration (minimum measurable FAR$= 7.5 \times 10^{-4}$). We include it as a sanity check using partial AUC at FAR$\in [0, 10^{-3}]$ and do not draw conclusions from it in the main paper.

Table 4. LFW-20 Ridge Open-Set: partial AUC (%) at FAR$\in [0, 10^{-3}]$ (13 test IDs). ISP-W is not applied to FR models.

	RAW			**ISP-W**		
Model	k=1	k=4	k=16	k=1	k=4	k=16
DINOv2	19.5	6.4	9.5	0.0	0.0	0.0
DINOv3	11.5	0.8	6.7	0.0	10.3	2.1
CLIP	55.6	40.8	9.0	10.8	15.4	0.0
SSCD	2.3	14.1	0.0	2.1	0.0	0.5
ArcFace	87.4	82.1	73.9	—	—	—
AdaFace	85.4	94.4	67.7	—	—	—

B. ISP Sensitivity and Cross-Dataset Transfer

This section provides the full sensitivity analyses and transfer experiments justifying our ISP rank selection and fitting procedure. All experiments use the within-dataset (ISP-W) protocol, Ridge attacker, TAR@FAR$= 10^{-4}$, and 3 evaluation seeds unless otherwise noted.

B.1. Cross-Dataset Transfer and Principal Angles

Transfer Protocol. To evaluate whether ISP generalizes across datasets:

1. Fit ISP on dataset A (e.g., CelebA-20, 320 train identities).
2. Apply P_A to embeddings from dataset B (e.g., VGGFace2-20).
3. Calibrate threshold on dataset B validation set.
4. Evaluate on dataset B test set.

We evaluate a full 2×2 transfer matrix:

- Within-dataset: $P_A \to A$, $P_B \to B$
- Cross-dataset: $P_A \to B$, $P_B \to A$

Principal Angle Analysis. To quantify subspace alignment between projections fitted on different datasets, we compute principal angles between the identity subspaces U_A and U_B:

$$\cos(\theta_i) = \sigma_i(U_A^\top U_B) \tag{14}$$

where σ_i are the singular values of $U_A^\top U_B$, giving the cosines of principal angles.

Cross-dataset alignment (same model, different datasets):

- DINOv2: $\max \cos(\theta) = 0.9998$
- DINOv3: $\max \cos(\theta) = 1.0000$
- CLIP: $\max \cos(\theta) = 0.9998$
- SSCD: $\max \cos(\theta) = 0.9977$

Near-perfect alignment ($\cos\theta \approx 1.0$) indicates identity subspaces are *universal* across datasets, not dataset-specific artifacts.

B.2. Rank Sweep

Table 5 shows TAR as r increases from 0 (raw, no projection) through 64, 96, 128, and 192 on both CelebA-20 and VGGFace2-20 with m=320 training identities. Protection generally improves with rank; CLIP requires higher r due to its larger embedding dimension and broader identity subspace. For VGGFace2-20, DINOv2/v3/SSCD reach 0.0% TAR already at moderate ranks, indicating near-complete suppression of linear identity access.

B.3. Identity-Count Sweep

Table 6 shows how TAR changes as the number of identities m used to fit the ISP projector scales from 320 to 2000, using the ISP-Large protocol (fit on a held-out pool disjoint from test identities, evaluated on CelebA-20 at r=192). Protection does not degrade with more identities; CLIP shows the largest improvement at scale (m=2000: 1.5% vs. m=320: 3.5%), suggesting its identity subspace benefits from a larger fit pool. DINOv3 and SSCD already reach sub-1% TAR at m=320 and remain stable.

Table 5. **Rank sweep.** TAR@FAR$= 10^{-4}$ (%) under ISP-W (Ridge, m=320, k=16). r=0 is the raw (unprojected) baseline.

Dataset	Model	r=0	r=64	r=96	r=128	r=192
CelebA-20	DINOv2	4.1	3.7	2.9	**1.3**	1.5
	DINOv3	6.2	2.3	1.4	**1.0**	1.5
	SSCD	2.9	1.4	1.5	**0.6**	1.0
	CLIP	7.3	6.6	6.7	5.7	**3.5**
VGGFace2-20	DINOv2	3.1	3.5	3.1	2.7	**2.2**
	DINOv3	2.9	2.5	**1.0**	2.2	1.5
	SSCD	1.3	2.0	2.1	1.8	**1.0**
	CLIP	7.1	4.7	**2.9**	4.2	2.3

Table 6. **Identity-count sweep.** TAR@FAR$= 10^{-4}$ (%) as the ISP fit pool m scales (ISP-Large, CelebA-20, Ridge, r=192, k=16). Fit identities are disjoint from test identities.

Model	m=320	m=640	m=1280	m=2000
DINOv2	1.5	1.9	2.9	2.7
DINOv3	1.5	5.4	4.0	3.9
SSCD	1.0	1.3	1.7	1.7
CLIP	**3.5**	6.7	6.7	**1.5**

C. Template Inversion: Protocols and Extended Results

This section provides per-attack protocol details, the shared cross-model verification criterion, and extended quantitative results.

C.1. Shared Evaluation Protocol

All template inversion attacks are evaluated under the same cross-model verification criterion. Given a reconstructed face $\hat{x}$ and true target x_{tgt}, inversion succeeds if

$$\frac{f_{\text{FR}}(\hat{x})^\top f_{\text{FR}}(x_{\text{tgt}})}{\|f_{\text{FR}}(\hat{x})\|_2 \, \|f_{\text{FR}}(x_{\text{tgt}})\|_2} \;\geq\; \tau_F, \tag{15}$$

where f_{FR} is a *disjoint* FR encoder (different from the one whose embedding is being inverted). We evaluate with two FR verifiers:

- **ArcFace**: $\tau_F = 0.1051$ (minimum EER on validation).
- **AdaFace**: $\tau_F = 0.1111$ (minimum EER on validation).

Verification rate is # accepted/# targets $\times\, 100\%$; all values in the main paper and tables below are averaged over both verifiers. 95% binomial confidence intervals are computed via the Wilson score method.

C.2. DiffMI (Wang et al., 2025)

DiffMI is a training-free white-box attack that uses an unconditional face diffusion prior to reconstruct faces from target embeddings alone.

Candidate pool generation.

- Pre-generate $V = 1{,}000$ latent codes $\ell \sim \mathcal{N}(0, I)$ with shape $[3, 256, 256]$.
- Decode via DDIM (20-step denoising) to produce 256×256 face images.
- Validate each latent: K^2 normality test ($p \geq 0.999$) and MTCNN face detection (confidence ≥ 0.999). Acceptance rate $\approx 10\%$; validated pool cached for reuse.

Top-N selection. For each target embedding z_{tgt}, compute cosine similarity with all pool images and select the $N = 3$ candidates with highest similarity as warm-start initialization.

Adversarial refinement (APGD).

- Objective: maximize $\cos\big(f(G(\ell)),\, z_{\text{tgt}}\big)$ where G is the DDPM decoder and f is the target encoder.
- L_2-bounded perturbations: $\|\Delta\ell\|_2 \leq \varepsilon = 35$.
- Budget: 100 iterations per target.
- Early stopping: halt if cosine $\geq \tau_C = 0.99$.

Diffusion prior. Unconditional DDPM trained on CelebA-HQ; DDIM decoder with 20 denoising steps; output 256×256 RGB.

Models and targets.

- **ArcFace** (512-dim), **CLIP ViT-B/32** (512-dim), **DINOv2-base** (768-dim), **DINOv3-small** (384-dim), **SSCD-ResNet50** (512-dim).
- $N = 200$ targets from CelebA-20 test set.

C.3. Bob (Kim et al., 2024)

Bob [37] algebraically inverts face recognition embeddings without iterative optimization, using only cosine similarity scores against a small set of probe images.

Method overview.

1. Build a local ArcFace embedding space model (iresnet50, 512-dim).
2. Pre-generate 99 δ-orthogonal face images (orthogonal face set; OFS).
3. For each target embedding z_{tgt}: compute 99 cosine similarity scores against OFS images, solve $z \approx A^\dagger s$ via pseudo-inverse where s is the score vector and A encodes the OFS geometry, then decode $\hat{x} = \text{NbNet}(z)$ via the NbNet inverse model.

Key parameters.

- **Local model**: ArcFace iresnet50 (512-dim).
- **Inverse model**: NbNet (512-dim embedding $\rightarrow 128\times128$ face).
- **OFS size**: 99 probe images.
- **Budget**: 99 embedding queries per target (closed-form; no iterations).

Non-FR encoder adaptations. The same OFS and NbNet decoder are used across all encoders. Encoder-specific preprocessing wrappers (resize, center-crop, or MTCNN alignment) ensure inputs are compatible.

Targets. $N = 200$ targets from CelebA-20 test set.

C.4. Vec2Face (Wu et al., 2025)

Vec2Face [39] uses a ViT-based conditional face generator trained to produce faces whose embeddings match a target vector.

Protocol for FR encoders. Use the pretrained Vec2Face model (trained on ArcFace embedding space) directly for ArcFace targets.

Protocol for non-FR encoders. Train a separate Vec2Face model per embedding type:

- Embedding dimensions: CLIP (512), DINOv2 (768), DINOv3 (384), SSCD (512).
- Training: 19 epochs, lr$= 0.001$ with MultiStepLR decay, batch size 32, cosine similarity loss.

Inference budget. 100 optimization steps per target with temporal latent averaging over the last 70 steps; top-K=10 candidate selection.

Targets. $N = 200$ targets (FR); $N = 200$ targets per non-FR encoder (reduced due to per-encoder training cost).

C.5. ALSUV (Jung et al., 2024)

ALSUV [38] searches a StyleGAN2 W+ latent space via Adam with multi-latent parallel search and unsupervised validation via latent averaging.

Generator and latent space. StyleGAN2 FFHQ; W+ space (14 layers $\times$ 512 dims).

Optimization.
- 10 latents per batch, optimized in parallel.
- Adam: lr $= 0.01$, weight decay 10^{-4}, LR drop at step 50 (factor 0.1).
- Budget: 100 steps per target.
- Latent averaging: average over the last 70 steps of the optimization trajectory (unsupervised validation).

Non-FR encoder adaptations. Encoder adapters for CLIP, DINOv2, DINOv3, and SSCD with optional ISP projection support are applied before the cosine similarity objective.

Targets. $N = 100$ targets from CelebA-20 test set.

Table 7. **Template inversion verification rates (%), RAW vs. +ISP.** Non-FR encoders only (ISP not applicable to FR encoders). N=200 per encoder; ALSUV N=100. Values averaged over ArcFace (τ=0.1051) and AdaFace (τ=0.1111) verifiers.

	Bob		**Vec2Face**		**ALSUV**	
Encoder	RAW	+ISP	RAW	+ISP	RAW	+ISP
CLIP	6.0	6.5	5.0	5.0	7.0	2.0
DINOv2	7.0	8.5	4.0	5.0	1.0	2.0
DINOv3	8.5	7.5	5.0	5.0	1.0	0.0
SSCD	9.0	5.5	3.0	3.0	1.0	0.0

C.6. Extended Results and ISP Comparison

Table 7 reports raw and post-ISP verification rates for non-FR encoders across all attack families. ISP has no applicable variant for FR encoders (ArcFace/AdaFace). Rates are averaged over ArcFace and AdaFace verifiers.

D. Face–Context Attribution: Complete Methods

This section provides extended related work, formal definitions, and full implementation details for the three face–context attribution diagnostics (FII, CPI, B^*) described in Sec. 3.2.

D.1. Extended Related Work

Modern encoders can entangle identity evidence with non-facial context (backgrounds, clothing, hair, scene), a phenomenon widely studied in object recognition as "context bias" or "shortcut" reliance. Classic analyses show CNNs tend toward local texture and background cues [16, 17], while transformer backbones propagate long-range information via self-attention, amplifying non-local context unless constrained [15]. In biometrics, explainable AI work has primarily focused on *where* a face recognition (FR) model looks—via occlusion/perturbation maps and saliency—rather than quantifying the *relative importance* of face vs. context. Representative perturbation methods include meaningful masking [13] and randomized occlusion (RISE) [14], and recent black-box explanations tailored to faces (e.g., MinPlus [19]) produce stable, interpretable saliency for FR match decisions.

Beyond visualization, several studies probe context attribution quantitatively. In generic recognition, Adhikari et al. report "volume attribution" that partitions saliency mass between object and background, finding non-trivial background reliance even for correct predictions [20]. For vision–language encoders, CLIP has been shown to produce attention and saliency that sometimes concentrates on background rather than foreground, motivating architectural or inference-time adjustments such as attention surgery and segmentation-aware debiasing [21, 22]. For face embeddings specifically, open-set attribution is more challenging: similarity scores are defined over image *pairs*, so explanation must reason about identity consistency across images rather than single-image class logits.

Our attribution-as-measurement framework addresses this gap with pairwise diagnostics designed for frozen embeddings and calibrated to biometric operating points.

D.2. Face Coverage Ratio (FCR) Normalization

Faces occupy different fractions of an image; naively "masking the face" in a tight crop removes more pixels (and thus more evidence) than in a wide shot. We therefore standardize the face budget before any perturbation by targeting a common face coverage ratio:

$$\mathrm{FCR}(x) = \frac{\mathrm{area}(\text{face mask in } x)}{\mathrm{area}(x)}. \tag{16}$$

After resizing/cropping to reach a target FCR, we only compare perturbations that are equal-area and equal-strength; this lets us attribute any similarity change to *what* we touched (face vs. background), not *how much* we touched. Implementation details including mask construction and area tolerance are in Sec. D.6 below.

Algorithm. **Input**: Original image, face bounding box (x_1, y_1, x_2, y_2) from RetinaFace detector.

Target: FCR $= 0.33$ (33% face, 67% background).

1. **Compute face area**: $A_{\text{face}} = (x_2 - x_1) \times (y_2 - y_1)$.
2. **Determine crop size**: $s = \sqrt{A_{\text{face}}/0.33}$.
3. **Center crop**: Face center $= ((x_1{+}x_2)/2,\ (y_1{+}y_2)/2)$. Crop window: $[c_x - s/2,\ c_y - s/2,\ c_x + s/2,\ c_y + s/2]$.
4. **Handle boundaries**: If crop extends beyond the image, pad with mean color. Track padding ratio: $r_{\text{pad}} = \sum_{\text{sides}} \text{pad}/(4s)$.
5. **Quality control**: Reject if $r_{\text{pad}} > 0.15$ (excessive artificial background).
6. **Resize**: Crop region to 224×224 (or 288×288 for SSCD).

Results: From 24,000 VGGFace2 images, 23,728 retained (272 rejected for padding). Achieved FCR: mean $= 0.330$, std $= 0.015$.

Face Mask Generation. We fit an ellipse to the 5-point landmarks (left eye, right eye, nose, left mouth, right mouth):

- **Center**: Midpoint of (eye center, mouth center).
- **Width axis**: $1.2\times$ inter-ocular distance.
- **Height axis**: $1.3\times$ eye-to-mouth distance.
- **Rotation**: Aligned to eye-line angle.
- **Feathering**: 3-pixel Gaussian blur on mask boundary.

Binary mask stored as uint8 (255 = face, 0 = background).

D.3. Face Importance Index (FII)

If equal-area occlusions to face and background have different effects on pairwise similarity, the difference is informative about where identity evidence lives. Let $\mathbf{z}^{\text{bg-occ}}$ and $\mathbf{z}^{\text{face-occ}}$ denote the ℓ_2-normalized embeddings of image x with an equal-area occlusion mask applied to the background region and face, respectively. For a query–reference pair $(\mathbf{z}_q, \mathbf{z}_r)$ we measure:

$$\Delta_{\text{face}} = \cos(\mathbf{z}_q, \mathbf{z}_r) - \cos\big(\mathbf{z}_q^{\text{face-occ}},\ \mathbf{z}_r^{\text{face-occ}}\big), \tag{17}$$

$$\Delta_{\text{bg}} = \cos(\mathbf{z}_q, \mathbf{z}_r) - \cos\big(\mathbf{z}_q^{\text{bg-occ}},\ \mathbf{z}_r^{\text{bg-occ}}\big), \tag{18}$$

and define $\mathrm{FII} = \Delta_{\text{face}} - \Delta_{\text{bg}}$, averaged over pairs. Positive FII indicates face-dominant similarity; negative FII indicates context dominance.

Implementation: equal-area occlusion. **Face occlusion**: Apply Gaussian blur ($\sigma = 8$, kernel 49×49) to face region, blended using the face mask with 3-pixel feathering.

Background occlusion (equal-area): Create an annulus mask surrounding the face by expanding the face mask outward by w pixels via morphological dilation, then subtracting the original face mask: $M_{\text{annulus}} = \text{dilate}(M, w) - M$. Use binary search to find w such that $A_{\text{annulus}} = A_{\text{face}} \pm 2\%$. Apply the same Gaussian blur ($\sigma = 8$) to the annulus region.

Aggregate: $\text{FII}_{\text{mean}} = \text{mean}(\text{FII})$ over all pairs; 95% CI via percentile bootstrap (2,000 samples).

D.4. Context Preference Index (CPI)

For each image i, we construct two reference images: an identity-matched image $\mathbf{z}^i_{\text{id}}$ (same person, different context) and a context-matched image $\mathbf{z}^i_{\text{ctx}}$ (different person via face-swap, same context). We apply Gaussian face blur with strength σ to all three images and compute their embeddings. CPI measures how often the blurred query prefers context over identity:

$$\text{CPI}(\sigma) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big[\cos\big(\mathbf{z}^i_{q,\sigma},\ \mathbf{z}^i_{\text{ctx},\sigma}\big)\ \geq\ \cos\big(\mathbf{z}^i_{q,\sigma},\ \mathbf{z}^i_{\text{id},\sigma}\big)\Big], \tag{19}$$

and we report $\Delta\text{CPI} = \text{CPI}(\sigma_{\max}) - \text{CPI}(\sigma_{\min})$ and the crossover σ^* when defined.

Implementation: face blur series. **Triplet construction**: For each query image, form:

- **SP-DC** (Same Person, Different Context): same identity, different scene/clothing.
- **DP-SC** (Different Person, Same Context): face-swapped into query's exact background.

Blur series: Apply Gaussian blur to face region at $\sigma \in \{0, 1, 2, 4, 6, 8\}$, blended symmetrically across query, SP-DC, and DP-SC.

Derived metrics:

- $\Delta\text{CPI} = \text{CPI}(\sigma = 8) - \text{CPI}(\sigma = 0)$: Change in context preference.
- σ^*: Crossover blur level where CPI $= 0.5$ (if exists).

D.5. Background Revelation Threshold (B^*)

As a complementary stress test, we monotonically reveal more background while keeping the face crop constant and ask when context overtakes identity. Let $B \in [0, 1]$ be the revealed background fraction in a monotone series that preserves the face crop. For each triplet (query, identity-matched, context-matched), define:

$$B^*_{\text{query}} = \min\Big\{B\ :\ \cos\big(\mathbf{z}_{q,B},\ \mathbf{z}_{\text{ctx}}\big)\ \geq\ \cos\big(\mathbf{z}_{q,B},\ \mathbf{z}_{\text{id}}\big)\Big\}, \tag{20}$$

and report the median across triplets (with censoring at 0 and 1). $B^* \to 0$ indicates context dominance (minimal background suffices to overtake identity); $B^* \to 1$ indicates face dominance (identity resists background revelation). Implementation parameters are in Sec. D.6 below.

D.6. Implementation Parameters and Quality Controls

Blur parameters. All Gaussian blurs use kernel size $= 2 \times \lceil 3\sigma \rceil + 1$. For $\sigma = 8$ (heavy occlusion), kernel $= 49 \times 49$.

Face-swap quality. DP-SC pairs generated via INSwapper (InsightFace):

- Success rate: 98.0% (23,254 / 23,728 attempted).
- Poisson blending with 3-pixel feather.
- Histogram matching for color consistency.

Visual inspection: $< 2\%$ show minor artifacts (acceptable for stress-test purposes).

FCR matching. For SP-DC pairs, FCR matched within $\pm 2\%$ tolerance. Actual FCR difference: mean $= 0.70\%$, max $= 2.0\%$.

Equal-area tolerance. FII background annulus matched to face area within $\pm 2\%$. Actual area difference: mean $= 0.8\%$, max $= 2.0\%$.

Monotonicity checks. For B^*, we verify:

- $\text{Sim}_{\text{DPSC}}(B)$ increases with B (94% of triplets).
- $\text{Sim}_{\text{SPDC}}(B)$ approximately flat (98% of triplets).

Non-monotonic cases ($< 6\%$) attributed to noise in the 5-point B-level grid.

E. Non-Linear Robustness: MLP Verifier Results

To test whether identity information persists for moderate non-linear attackers after ISP, we train a projection-only MLP verifier under the same open-set, identity-disjoint protocol as the Ridge probe. The MLP (two hidden layers, ReLU) operates on cosine similarities between query and support embeddings; all hyperparameters and thresholds are frozen on A_{val} before evaluation on A_{test}. FR models (ArcFace, AdaFace) are included as pre-ISP baselines only; ISP is not applied to FR encoders (marked "—").

Table 8 reports TAR@FAR$= 10^{-4}$ (%) for $k \in \{1, 4, 16\}$ shots on CelebA-20. Non-FR encoders already exhibit low non-linear leakage in raw embeddings. After ISP, TAR drops further, with DINOv2/v3/SSCD reaching near zero across all k. CLIP, with its larger embedding dimension and broader identity subspace, retains modest residual leakage after ISP, consistent with the rank-sweep results in Table 5.

Table 8. **Projection-only MLP verifier (CelebA-20).** TAR@FAR$= 10^{-4}$ (%). Pre-ISP uses raw embeddings; Post-ISP applies within-dataset ISP-W. FR models (ArcFace, AdaFace) are shown pre-ISP only (ISP not evaluated on FR encoders).

Type	Model	CelebA-20					
		Pre-ISP			Post-ISP		
		k=1	k=4	k=16	k=1	k=4	k=16
Non-FR	DINOv2	1.4	2.2	3.8	0.0	0.8	0.4
	DINOv3	2.2	3.4	5.9	0.5	0.3	0.9
	SSCD	2.4	3.9	5.8	0.8	1.1	0.7
	CLIP	18.3	22.9	26.9	1.6	1.8	0.5
FR	ArcFace	91.9	90.9	92.8		—	
	AdaFace	92.2	91.2	93.0		—	

Table 9. **DISC2021 copy-detection (SSCD).** Percentage of Recall@k retained after projection, relative to unprojected embeddings.

Metric	ISP	LEACE
Recall@1	95.0%	95.1%
Recall@5	96.5%	96.4%
Recall@10	96.0%	96.1%

F. DISC2021 Copy-Detection Utility (SSCD)

Table 9 reports Recall@k retained after projection on the DISC2021 copy-detection benchmark for SSCD. Both ISP and LEACE preserve over 95% of baseline recall across all k, confirming that task-specific retrieval utility is largely unaffected.

G. Subspace Removal: Extended Background

There is a mature literature on removing sensitive information from fixed representations via linear subspace editing. Two broad families dominate.

Classifier-driven nulling. Iterative or minimax procedures train a linear adversary for a protected attribute and project onto (the intersection of) its nullspaces. INLP [7] repeatedly fits a linear classifier, projects the data onto its nullspace, and iterates until the attribute is no longer linearly predictable. RLACE [23] casts the same goal as a minimax program and solves it in closed form, yielding a rank-constrained projection that minimizes worst-case linear leakage. Both families aim for strong linear unpredictability guarantees but require repeated classifier fitting (or solving a minimax program) and can rotate or compress the feature geometry in ways that are difficult to audit post hoc.

Moment- and covariance-driven nulling. One-shot transforms estimated from class means and scatter matrices offer a complementary approach. SAL [24] removes guarded attribute information via spectral decomposition of class-conditioned representations. LEACE [25] derives a closed-form least-squares erasure that removes all linearly accessible information about a concept while minimizing distortion to the embedding, and comes with a formal optimality certificate under squared loss. Related fair-PCA methods [26] optimize Pareto trade-offs between utility and fairness without enforcing zero leakage, offering a softer alternative when complete erasure is unnecessary.

These moment-based methods explicitly target between-class mean structure (often in a whitened Fisher geometry) and are attractive for deployment because they require no adversary training and produce a fixed, exportable projection matrix.

Positioning of ISP. Prior concept-erasure work largely targets binary or low-cardinality attributes (e.g., gender, sentiment) and reports accuracy drops on supervised classification tasks. In contrast, facial identity in our setting is high-cardinality and open-set, and risk must be quantified at low FAR under disjoint identities. It is therefore unclear a priori whether the identity signal in non-FR embeddings: (a) concentrates in a compact, transferable subspace, (b) can be linearly "certified away" without harming non-biometric utility, and (c) remains suppressed under stronger, projection-only non-linear probes. Our study directly answers these questions with attacker-aware metrics (open-set TAR@low-FAR, projection-only MLP, template inversion) and with cross-dataset transfer and robustness checks, which are typically absent from concept-erasure evaluations.

ISP adopts a moment-based, one-shot design: compute per-identity mean differences, take the SVD, and project onto the orthogonal complement of the top-r "identity" directions; optionally whiten to obtain a Fisher-space certificate [25]. Under homoscedasticity, this removes between-class mean structure and yields a clear linear leakage certificate while preserving the complementary subspace. Compared to iterative/minimax methods (INLP/RLACE), ISP is:

1. *Auditable*: rank r directly controls the privacy-utility trade-off and enables energy diagnostics.
2. *Lightweight*: a single SVD with no adversary training or hyperparameter grids.
3. *Model-agnostic and exportable*: a fixed P that plugs into any retrieval pipeline.
4. *Deployment-friendly*: stable, deterministic, no retraining, sub-millisecond latency.

In our evaluations, this simple construction is sufficient to drive linear identity accessibility near chance while retaining most utility, and its fixed projector generalizes across datasets - properties that alternatives must match to be practical at scale.

We emphasize that linear subspace removal certifies against linear attackers; non-linear leakage may persist. In exchange, linear methods typically preserve utility far better than end-to-end adversarial training and serve as auditable, deployable baselines that stronger mitigations must beat. We provide empirical evidence on non-linear robustness via a projection-only MLP verifier in Appendix E and Sec. 4.2.

H. Attribute-Swap Identity Inference

Prior work has shown that identity can remain stable under semantic attribute changes (e.g., gender) in FR embeddings. A natural question is whether this holds for non-FR encoders after ISP. We evaluate PCA-based attribute-swap attacks that attempt to modify the gender attribute while preserving identity post-ISP, following the manipulation strategy of Kim et al. (NeurIPS '25).

H.1. Protocol

Setup. We work with CelebA-20 embeddings ($N = 500$ victims per direction). A linear gender classifier is trained on training-split embeddings; its decision boundary normal defines the gender direction $\mathbf{v}_{\text{gender}}$. PCA is used to find the top-k directions in the subspace orthogonal to within-identity variation that maximally covary with the gender label. A victim embedding z is projected along these k directions to produce a manipulated embedding $\tilde{z}$.

Metrics. We report three quantities at $k = 128$ components:

- **Accept**: fraction of victims where $\cos(z, \tilde{z}) \geq \tau$ (identity similarity threshold at FAR $= 10^{-3}$), i.e. the manipulated embedding is still identity-proximate.
- **Flip**: fraction of victims where the gender classifier predicts the *target* gender for $\tilde{z}$.
- **Joint**: fraction of victims satisfying *both* Accept and Flip simultaneously—the attacker's true success rate.

H.2. Results

Table 10. Attribute-swap identity inference at $k = 128$. **Joint =** fraction of victims where identity is preserved (sim $\geq \tau_{\text{FAR}=10^{-3}}$) *and* gender is flipped. FR models allow highly reliable joint manipulation; non-FR encoders do not. 500 victims per direction on CelebA-20.

System	Direction	Accept	Flip	Joint
DINOv2	F→M	96.8%	14.6%	14.0%
+ ISP ($r = 192$)	F→M	96.2%	30.2%	28.8%
DINOv2	M→F	98.4%	14.6%	14.4%
+ ISP ($r = 192$)	M→F	98.2%	26.6%	26.2%
DINOv3	F→M	99.6%	13.0%	12.8%
+ ISP ($r = 192$)	F→M	100.0%	17.8%	17.8%
DINOv3	M→F	100.0%	10.2%	10.2%
+ ISP ($r = 192$)	M→F	100.0%	17.4%	17.4%
CLIP	F→M	100.0%	9.2%	9.2%
+ ISP ($r = 192$)	F→M	98.0%	13.8%	13.4%
CLIP	M→F	100.0%	6.6%	6.6%
+ ISP ($r = 192$)	M→F	98.6%	12.0%	11.8%
SSCD	F→M	100.0%	37.0%	37.0%
+ ISP ($r = 192$)	F→M	100.0%	35.0%	35.0%
SSCD	M→F	100.0%	23.6%	23.6%
+ ISP ($r = 192$)	M→F	100.0%	29.0%	29.0%
Face Recognition Models (no ISP variant)				
ArcFace	F→M	99.8%	81.2%	**81.2%**
ArcFace	M→F	99.8%	73.6%	**73.6%**
AdaFace	F→M	99.8%	70.4%	**70.4%**
AdaFace	M→F	99.6%	52.2%	**52.2%**