SWAPPED FACE, SAME PERSON: MEASURING THE IDENTITY THAT FACE SWAPS LEAVE BEHIND

Anonymous authors
Paper under double-blind review

ABSTRACT

Face-swap methods report identity transfer as ArcFace similarity on a tight face-oval crop, and a low crop-similarity to the target is read as evidence that the target was de-identified. This practice is blind by construction: the region it discards—hair, ears, skull contour—is exactly the region a face-oval swapper never modifies. The general failure is that the metric is evaluated only on the *support of the transformation*, so any residual identity channel outside that support goes unmeasured; face swapping is the case where the gap is widest, but the construction is not specific to it. We localize that residual identity with *segmentation-defined* probes, replacing the arbitrary widened bounding box used in prior work with CelebAMask-HQ head regions, and in particular with a *pericranial ring* probe that sees only the pixels a face-oval swapper copies verbatim from the target. Because such regions have a much higher impostor-similarity floor than aligned face crops (mean 0.447 vs. 0.005 on our gallery), raw cosines are not comparable across probes, so we report calibrated rates instead: thresholds are set per probe at a stated false-accept rate on identity-disjoint impostor pairs, following the attacker-aware protocol of recent encoder-privacy work. At $\mathrm{FAR} = 10^{-3}$, the target is accepted from hair and ears alone for 57–97% of the outputs of the face-oval swappers shipped in a widely deployed open-source toolkit, against 0.3–61% under the standard crop; equal-area, equal-strength occlusion attribution confirms this is a property of *which* region is read, not of how many pixels it contains. The measurement is also discriminative: a full-head regeneration teacher that rewrites the pericranial region leaks 2.3% under the same probe, a 13–$42\times$ separation from every face-oval method that no face-crop metric reproduces. We release PHANTOM-150K's generation pipeline, per-sample metadata, and all evaluation code—but not generated images, since every one depicts real donor and target identities.

1 INTRODUCTION

Face-swap technology—transferring a source (donor) identity onto a target image while preserving pose and expression—underpins applications from privacy-preserving video redaction to entertainment. Evaluating whether a swap actually *protects* the target's identity, however, is not the same problem as evaluating whether it successfully *transfers* the donor's, and existing practice conflates the two.

Open swappers such as FaceFusion/inswapper, SimSwap (Chen et al., 2020), and HifiFace (Wang et al., 2021) operate on a tight face-oval crop. Hair, skull contour, and ears are not touched: they are the target's original pixels, composited back into the target's original frame. Standard ArcFace evaluation crops exactly that region away before embedding, so the metric cannot see what the method did not change. A swap can therefore score a comfortably "safe" target-similarity of ≈ 0.06 while the same recognizer still matches the person it was supposed to protect from the rest of the same image. Every claim in this paper is of that form—about what a recognizer accepts at a stated false-accept rate—and we make no claim about human identifiability, which we did not measure.

Our earlier attempt to measure this used a widened bounding box—a $2.2\times$ expansion of the detected face box—as a proxy for "the region the standard crop discards." That probe is easy to attack on two grounds, and both attacks are correct. First, the scale factor is arbitrary: nothing justifies 2.2

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

over 2.0 or 2.6. Second, and more seriously, a wider box contains *more pixels*, so a higher similarity under it may simply reflect more evidence rather than evidence of a different kind. Any claim built on comparing a small crop to a big box is confounded by area.

This paper removes both objections by replacing the box with regions that are *semantically* rather than geometrically defined, and by adopting measurement machinery that controls for area explicitly:

- **Segmentation-defined probes** (Sec. 4): ϕ_{head}, the whole head including hair, following the full-head framing of Jeong et al. (2025); and ϕ_{ring}, the *pericranial ring* alone — hair, ears and skull contour, face interior removed — which reads identity from *only* the pixels a face-oval swapper leaves untouched.
- **Area-controlled attribution** (Sec. 4.4): matched pixel budgets and matched strength, after George et al. (2026), separating *what* was occluded from *how much*.
- **Calibrated operating points** (Sec. 4.3): per-probe thresholds set on identity-disjoint impostor pairs, so "leaked" is defined at a stated false-accept rate, not a convention.
- **A benchmark that makes the measurement falsifiable** (Sec. 3): 156,111 verified triplets from a full-head teacher — the positive control no public swap corpus provides — plus a task-success criterion requiring transfer *and* non-leakage jointly.
- **A cross-representation test** (Sec. 5.5): the same regions read by four independently trained encoders, one never trained on faces, preserving the method ordering exactly (Spearman $\rho = +1.000$) and the teacher/face-oval separation at 5–42×.
- **An independent full-head system** (Sec. 5.6): a second, architecturally unrelated head-swap model scored under the same protocol, which separates the protocol's verdict from the teacher that built the benchmark; all rates carry bootstrap intervals over donor identities (Sec. 5.7).

What is and is not claimed. For a face-oval method the pericranial pixels are, by construction, copied from the target, so a high ϕ_{ring} score is closer to copy detection than to face recognition. We do not claim hair and ears are strong biometrics in general, and the contribution is not the observation that unchanged pixels remain unchanged. The claim is about *measurement*: a privacy metric evaluated only on the transformation's support cannot distinguish an image whose identity was removed from one where it was merely relocated outside the crop, and the field's metric is of that kind. What the protocol adds is a calibrated, region-conditioned test that makes the residual channel measurable, decides it at a stated false-accept rate rather than a convention, controls for area, and—as Sec. 5.5 shows—does not depend on the particular recognizer used to read it.

Calibration is load-bearing rather than cosmetic: the ring's impostor floor is an order of magnitude above an aligned face crop's, so its raw cosines look alarming for reasons unrelated to leakage (Table 5). The finding survives the correction. At $\alpha = 10^{-3}$ the ring probe accepts the target for 29–97% of face-oval outputs against 2.3% for the full-head teacher, and the separation is invisible to the standard protocol—under ϕ_{std} that teacher (9.7%) scores *worse* than a degenerate method transferring no identity at all (1.3%).

2 Related Work

Face-swap datasets. FaceForensics++ (Rössler et al., 2019), Celeb-DF (Li et al., 2020) and DFFD (Dang et al., 2020) supply manipulated media for forgery detection. None targets full-head replacement, stores per-sample quality metrics for threshold ablation, or defines a privacy-aware transfer/leakage protocol.

Face-swap methods. SimSwap (Chen et al., 2020), HifiFace (Wang et al., 2021), Ghost (Groshev et al., 2022) and the diffusion swappers DiffSwap (Zhao et al., 2023) and DiffFace (Kim et al., 2022) all operate on face-oval crops, as does the most widely deployed open checkpoint, **inswapper**, which we run through **FaceFusion** (FaceFusion Contributors, 2026) and call FaceFusion throughout. Moving beyond the oval, Jeong et al. (2025) perform zero-shot *head* swapping with a context-aware mask and a hair-injection module; their full-head formulation motivates our segmentation-defined probe regions.

Re-identification and privacy evaluation. FaceTracer (Zhang et al., 2024) traces a swap's *source* identity by ranked gallery retrieval; we adopt its methodology and extend it to *target* retrieval—can the person meant to be protected still be found? De-identification for surveillance (Newton et al., 2005) motivates the threat model. Closest to our stance, George et al. (2026) audit identity leakage in frozen general-purpose encoders under an attacker-aware protocol: low-FAR open-set verification, template inversion, and equal-area face–context attribution. We adapt two of their three instruments to face swapping, where the question is not whether an encoder stores identity but *which region of a manipulated image* still carries it.

3 THE PHANTOM-150K BENCHMARK

Why we needed head swaps. PHANTOM-150K exists because the measurement in Sec. 4 cannot be validated without it. A probe that reads high on every face-oval method is not by itself evidence that it detects anything: an instrument that merely saturates on hair-and-ears crops produces the same table. Falsifying that requires a *negative* case—a method that genuinely rewrites the pericranial region, which the probe should read as low.

No public corpus supplies one. FaceForensics++, Celeb-DF and DFFD are libraries of *face-oval* manipulations: they replace the inner face and composite it back, leaving hair, ears and skull contour untouched by construction, so nothing in the comparison set ever edits the region under test. We therefore generated full-head swaps ourselves. PHANTOM-150K's role here is instrumental rather than headline: a pool of (S, T, R) triplets large enough for threshold ablation, with per-sample provenance for *every* attempted generation rather than only the accepted ones (Fig. 2, Appendix B).

The control does its job. BFS-FLUX reads 2.3% under ϕ_{ring} where the five face-oval methods read 29–97% (Table 1), so the probe is discriminative rather than saturated; and ϕ_{ring}'s ret sits at the impostor floor for every method and every dataset (Table 10), so it is not merely reporting high similarity for everything it is shown.

Teacher. Generation uses FLUX.2-Klein-4B with **BFS Head V1**, a publicly released MIT-licensed LoRA adapter for full-head replacement trained by its author (Alissonerdx, 2026); we use it as a teacher and do *not* claim to have trained it, PHANTOM-150K's contribution being the quality-verified triplet pool and the protocol. A structured prompt preserves the target's lighting, background, pose and micro-expressions while replacing the head entirely. Sampling settings and artifact hashes are in Appendix B.1; we call this teacher **BFS-FLUX** throughout.

Identity and expression pools. Donors come from VGGFace2 (Cao et al., 2018) (540 identities from a training subset, 40 images each), targets from CelebA-HQ (Karras et al., 2018) (29,997 of 30K passing detection); drawing them from different corpora keeps the two identity sets distinct. Targets are sampled uniformly over k-means clusters (k=64) of 68-point landmarks (Bulat & Tzimiropoulos, 2017; Lloyd, 1982), so the ten largest clusters hold only 31% of triplets.

Notation. ρ_{don} denotes ArcFace cosine similarity between a result and its *source* (donor) image and ρ_{tgt} the same similarity to its *target*; these are released as CSV fields `arc_don`/`arc_tgt` and formalized as ID-Ret/ID-Leak in Eqs. 7–8. $\varepsilon_{\text{expr}}$ denotes 68-landmark expression RMSE between a result and its target (field `expr_rmse`).

Pipeline and quality gate. Pairs are enumerated exhaustively in a fixed shuffled order, with no pre-generation filter on source–target similarity (Appendix A). A generated result is accepted only if $\rho_{\text{don}} \geq 0.30$ (donor identity actually transferred), $\varepsilon_{\text{expr}} \leq 0.08$ (target pose/expression preserved), and `det_score` ≥ 0.60 (a face is confidently detected) all hold. Appendix F measures how loosening or tightening each threshold trades dataset yield against measured leakage; $\rho_{\text{don}} \geq 0.30$ sits near the knee of that trade-off.

Reconciled generation funnel. Of 1,461,318 pair keys logged across resumed runs, 994,293 (68.0%) could not be loaded and never reached generation; 467,025 were generated and scored. The gate rejected 308,536 (66.1%): 203,588 low ρ_{don}, 68,977 high $\varepsilon_{\text{expr}}$, 35,946 low `det_score`, 25 no face in the source. The first 10,000 accepted triplets were produced under an earlier, looser gate

($\rho_{\text{don}} \geq 0.28$, $\varepsilon_{\text{expr}} \leq 0.09$); we re-apply the final gate uniformly, excluding the 2,376 that fail it and 2 whose expression could not be measured, leaving **156,111 verified triplets**. Diffusion sampling ran for all 467,025 generated pairs over a 45-day wall-clock window on a single NVIDIA RTX PRO 6000 Blackwell (Fig. 7, Appendix G).

Release: protocol, not pixels. We do *not* redistribute generated images: every one depicts a real donor and a real target drawn from VGGFace2/CelebA-HQ, whose licences do not permit redistributing derivatives of those likenesses. We release instead the generation pipeline, the full protocol (Appendix A), a manifest of all 158,489 accepted triplets with identity and image indices, seeds, gate scores, cluster assignments, release flag and exclusion reason, canonical identity-disjoint splits, rejection counts by reason, and all evaluation code. A user with access to the source datasets can reconstruct the identical pair list and seeds; bit-exact images additionally depend on hardware (Appendix B.1).

4 METHODOLOGY

4.1 SETUP AND NOTATION

Let $\mathcal{X}$ be the space of RGB images. We fix a pretrained face recognizer $\phi : \mathcal{X} \to \mathbb{S}^{d-1}$ (InsightFace `buffalo_l`, d=512), ℓ_2-normalized and frozen: no part of this protocol trains or fine-tunes it, so every difference we report is attributable to what ϕ is shown rather than to its weights. We also fix a face parser $\Pi : \mathcal{X} \to \mathcal{L}^{H \times W}$ assigning each pixel one of the $|\mathcal{L}|$=19 CelebAMask-HQ classes $\mathcal{L} = \{0, \ldots, 18\}$, enumerated in Appendix B.

For a class subset $R \subseteq \mathcal{L}$, the induced region mask is

$$\big[M_R(x)\big]_{uv} = \mathbf{1}[\Pi(x)_{uv} \in R], \qquad u \in [H],\ v \in [W]. \tag{1}$$

Writing $\mathrm{bb}(\cdot)$ for the tight bounding box of a binary mask and ρ_k for a resize to $k \times k$, the *region operator* suppresses everything outside R, crops to what remains, and rescales to the recognizer's input resolution:

$$\Gamma_R(x) = \rho_{112}\Big(\big(x \odot M_R(x)\big)\big|_{\mathrm{bb}(M_R(x))}\Big). \tag{2}$$

A *probe* is the composition $\phi_R = \phi \circ \Gamma_R$, inducing the similarity

$$s_R(x, x') = \big\langle \phi_R(x),\ \phi_R(x') \big\rangle \in [-1, 1]. \tag{3}$$

Every probe below shares the same ϕ and differs only in R (or, for the two geometric baselines, only in Γ). Consequently s_R isolates the effect of *region choice*, which is the single quantity this paper varies.

4.2 THE PROBE FAMILY

We instantiate three segmentation-defined probes and retain two geometric ones for comparison. In terms of CelebAMask-HQ indices,

$$R_{\text{face}} = \{1, 2, 3, 4, 5, 6, 10, 11, 12, 13\} \quad \text{skin, brows, eyes, glasses, nose, mouth, lips,} \tag{4}$$

$$R_{\text{ring}} = \{7, 8, 9, 17, 18\} \quad \text{ears, earring, hair, hat,} \tag{5}$$

$$R_{\text{head}} = R_{\text{face}} \cup R_{\text{ring}}, \tag{6}$$

with $\mathcal{L} \setminus R_{\text{head}} = \{0, 14, 15, 16\}$ (background, neck, necklace, clothing) excluded from all three. The two geometric probes are not expressible as class sets: ϕ_{std} replaces Γ_R with the recognizer's own 5-point affine alignment $A(\cdot)$, giving $\phi_{\text{std}} = \phi \circ A$, and ϕ_{box} replaces it with a $2.2\times$ dilation of the detected face box. ϕ_{std} is what the field reports; ϕ_{box} is the proxy used in the earlier version of this work.

Given a triplet $(x_s, x_t, \hat{x})$ of donor, target and output, each probe yields two quantities:

$$\text{ID-Ret}_R(\hat{x}) = s_R(\hat{x}, x_s) \quad (\uparrow), \tag{7}$$

$$\text{ID-Leak}_R(\hat{x}) = s_R(\hat{x}, x_t) \quad (\downarrow). \tag{8}$$

ϕ_{ring} is the decisive instrument: for a face-oval method, $\Gamma_{R_{\text{ring}}}(\hat{x})$ consists of the target's own pixels, composited back into the target's own frame, so if identity survives there it cannot be argued away by a crop convention. Figure 3 in Appendix B shows all four views for one pair.

4.3 CALIBRATED OPERATING POINTS

Raw similarities are not comparable across probes. Γ_R changes the input distribution of ϕ, and a hair-and-ears crop is far outside the aligned-crop distribution ϕ was trained on, so each probe induces its own score scale. A single hand-set cutoff — the 0.35 convention — means a different false-accept rate under each. We therefore calibrate per probe, following the low-FAR protocol of George et al. (2026).

Let $G \subset \mathcal{X}$ be the benchmark's gallery and $\iota : G \to \mathcal{I}$ its identity map. The impostor set is the cross-identity pair set

$$\mathcal{P}_{\text{imp}} = \big\{(a, b) \in G \times G \;:\; a < b,\;\; \iota(a) \neq \iota(b)\big\}, \tag{9}$$

and for a target false-accept rate α the threshold is the corresponding upper quantile of the probe's own impostor scores,

$$\tau_R(\alpha) = Q_{1-\alpha}\Big(\big\{\, s_R(a, b) \;:\; (a, b) \in \mathcal{P}_{\text{imp}} \big\}\Big). \tag{10}$$

Reported quantities are then rates at a stated α rather than similarities:

$$\text{Leak@}\alpha = \Pr\big[\, s_R(\hat{x}, x_t) \;\geq\; \tau_R(\alpha) \,\big], \tag{11}$$

$$\text{Ret@}\alpha = \Pr\big[\, s_R(\hat{x}, x_s) \;\geq\; \tau_R(\alpha) \,\big]. \tag{12}$$

Because each probe is calibrated against its own $\mathcal{P}_{\text{imp}}$, these rates are commensurable in a way s_R is not.

Impostor hygiene and FAR precision. ι is not directly observable on our gallery, since donor identities can recur across pairs. We therefore approximate $\{\iota(a) \neq \iota(b)\}$ by excluding any pair with $s_{\text{std}}(a, b) > 0.50$ as presumed same-identity, and report the number excluded; contaminating $\mathcal{P}_{\text{imp}}$ with mated pairs would inflate τ_R and understate leakage. The smallest resolvable rate is the FAR precision $1/|\mathcal{P}_{\text{imp}}|$, which bounds the α we may report. Table 5 and Figure 6 (both Appendix C) give the thresholds, pair counts and the distributions they are drawn from: the ring's impostor mean is 0.447 against 0.005 for the aligned face crop, and $\tau_{\text{ring}}(10^{-3}) = 0.853$ against $\tau_{\text{std}}(10^{-3}) = 0.212$.

4.4 AREA-CONTROLLED ATTRIBUTION

Calibration removes the threshold objection but not the area objection: a region with more pixels may carry more identity merely by carrying more evidence. We adopt the equal-area, equal-strength construction of George et al. (2026)—report the area budget, perturb both regions with matched budget and matched strength, difference the effects—transposed from their face-versus-background setting to face-versus-pericranium.

The area budget is summarized by the face-coverage ratio

$$\text{FCR}(x) = \frac{\big|M_{R_{\text{face}}}(x)\big|}{H \cdot W}, \tag{13}$$

where $|\cdot|$ counts set pixels. Let $N(x) = \min\big(|M_{R_{\text{face}}}(x)|,\, |M_{R_{\text{ring}}}(x)|\big)$ be the common budget, and let $\mathcal{O}^N_R(x)$ denote x with the N pixels of R nearest that region's centroid set to a fixed fill — a deterministic, approximately contiguous occlusion of exactly N pixels, identical in size and strength for R_{face} and R_{ring}. Embedding all variants under ϕ_{head} and writing $s = s_{R_{\text{head}}}(\hat{x}, x_t)$,

$$\Delta_{\text{face}} = s - s_{R_{\text{head}}}\big(\mathcal{O}^N_{R_{\text{face}}}(\hat{x}),\, \mathcal{O}^N_{R_{\text{face}}}(x_t)\big), \tag{14}$$

$$\Delta_{\text{ring}} = s - s_{R_{\text{head}}}\big(\mathcal{O}^N_{R_{\text{ring}}}(\hat{x}),\, \mathcal{O}^N_{R_{\text{ring}}}(x_t)\big), \tag{15}$$

$$\text{FII} = \Delta_{\text{face}} - \Delta_{\text{ring}}. \tag{16}$$

$\text{FII} < 0$ indicates that the pericranial ring sustains more of the residual output-to-target match than the face interior does, at equal area. One departure deserves stating: George et al. (2026) evaluate these deltas over *mated* pairs, asking which region sustains a genuine match. Our pairs are an output and the person it was meant to protect, so we measure which region sustains the *residual* match—the construction is theirs, the population is not.

4.5 TASK-SUCCESS CRITERION

Leakage alone is not a privacy criterion: an output resembling neither party scores as low-leakage while protecting no one. We require transfer and non-leakage jointly,

$$\underbrace{\mathrm{Ret@}\alpha_{\,\phi_{\mathrm{std}}} \geq \tau_{\mathrm{transfer}}}_{\text{the donor was actually placed}} \;\wedge\; \underbrace{\mathrm{Leak@}\alpha_{\,\phi_{\mathrm{ring}}} \leq \tau_{\mathrm{leak}}}_{\text{the target does not survive}}, \tag{17}$$

with $\tau_{\mathrm{transfer}} = 50\%$, $\tau_{\mathrm{leak}} = 10\%$ and $\alpha = 10^{-3}$. Transfer is read from ϕ_{std} because that is where a swap is supposed to put the donor; leakage is read from ϕ_{ring} because that is where the target survives. Sec. 5.2 shows the two readings are near-independent, which is what makes the conjunction informative.

Definition (privacy-preserving face swap). An output is *privacy-preserving* for a donor/target pair iff it satisfies Eq. 17. Gallery retrieval (Sec. 5.3) is deliberately excluded and kept as an independent check: a method can satisfy Eq. 17 and still leave its target retrievable at a practical rank budget, so a pointwise criterion, however well calibrated, is necessary but not sufficient.

5 EXPERIMENTS

Setup. 300 PHANTOM-150K pairs were transferred to a second machine and run through four independently-developed swap methods, scored with our released protocol. The evaluation set is fixed at those 300 pairs: the comparison methods' outputs were produced on that machine and cannot be regenerated here, so n is a property of the transfer rather than a choice, and Sec. 5.7 reports what that n does and does not support. The 300 were drawn uniformly at random (seed 42) before the release filter; 7 fall in the excluded set, and removing them moves no headline rate by more than 0.7 points (Appendix A). **SimSwap** (Chen et al., 2020) and **HiFiFace** (Wang et al., 2021) are run from the ONNX exports distributed with FaceFusion (FaceFusion Contributors, 2026) (`simswap_256` and `hififace_unofficial_256`, each with its matching `crossface` identity adapter). We flag the provenance because it qualifies what the HiFiFace row means: the original authors released no weights, so that row measures a trained community reimplementation of the method, not a checkpoint from Wang et al. (2021). Both are trained models and are reported as such—unlike DiffSwap++ below, which is not. **DiffSwap++** follows Zhao et al. (2023) but with randomly initialised identity/landmark projection weights, no trained checkpoint being available; we therefore report it as an *untrained identity-conditioning ablation*, not a DiffSwap result—it establishes what zero identity conditioning looks like under this protocol. **Ghost** (Groshev et al., 2022) runs zero-shot. We also score **BFS-FLUX**, this paper's own generation teacher, on the identical pairs. BFS-FLUX is included as a *reference point*, not a like-for-like comparison: it is a full-head regeneration model in a different paradigm from the other four's face-oval warping, and its outputs are what PHANTOM-150K's quality gate was tuned to select. It shows what passing Eq. 17 looks like in practice, not that full-head regeneration beats face-oval warping in general; Sec. 5.6 meets the circularity objection directly, with a full-head system that had no part in building the benchmark.

5.1 THE BLIND SPOT, LOCALIZED

Table 1 reports all four probes at $\alpha = 10^{-3}$. Moving from ϕ_{std} to ϕ_{ring} raises the target-acceptance rate for every face-oval method — SimSwap $29.3 \rightarrow 69.3\%$, HiFiFace $61.0 \rightarrow 97.3\%$, Ghost $1.3 \rightarrow 29.3\%$ — so HiFiFace's outputs are accepted as their own targets, from hair and ears alone, in 97.3% of cases at a one-in-a-thousand false-accept rate. The ordering also inverts: under ϕ_{std} Ghost is the best entry in the table (1.3%) and BFS-FLUX is mediocre (9.7%); under ϕ_{ring} BFS-FLUX is best by an order of magnitude (2.3%) and Ghost is 29.3%. On the systems evaluated here, the standard crop does not merely understate leakage: it can reverse the relative ordering of methods with respect to residual target identity, ranking a method that leaves the target's whole head intact above one that regenerates it. Ghost's apparent excellence is in any case an artifact of failure: its $\mathrm{Ret@}10^{-3} = 2.3\%$ under ϕ_{std} means it transfers the donor almost never, so Eq. 17 rejects it, which a leakage-only criterion would not.

The clearest case is the swapper most people run. FaceFusion's inswapper, on the same pairs and thresholds, is the *best* entry under the field's own metric—0.3% leakage at $\alpha = 10^{-3}$ ($[0.0, 1.0]$)—while transferring the donor in every output ($\mathrm{Ret@}10^{-3} = 100\%$ under ϕ_{std}). It is not degenerate—it

Table 1: **Calibrated rates at** FAR $= 10^{-3}$, 300 pairs per method, thresholds from Table 5. Leak: fraction of outputs accepted as their own *target* (lower is better). Ret: fraction accepted as their own *source* (higher is better). † this paper's generation teacher, a reference point rather than a like-for-like comparison; ¶ FaceFusion/inswapper, run here on the same pairs (the other rows were produced on the second machine); ‡ untrained ablation; § zero-shot, fails transfer.

	Leak@10^{-3} ↓				Ret@10^{-3} ↑			
Method	ϕ_{std}	ϕ_{box}	ϕ_{head}	ϕ_{ring}	ϕ_{std}	ϕ_{box}	ϕ_{head}	ϕ_{ring}
BFS-FLUX†	9.7%	2.3%	2.7%	**2.3%**	99.7%	25.0%	31.7%	8.3%
SimSwap	29.3%	7.0%	15.7%	69.3%	100.0%	15.7%	32.7%	0.0%
HiFiFace	61.0%	31.3%	31.7%	**97.3%**	100.0%	8.3%	16.7%	0.0%
DiffSwap++‡	61.3%	34.3%	35.0%	84.3%	0.0%	0.0%	0.0%	0.0%
Ghost§	1.3%	0.0%	3.3%	29.3%	2.3%	0.0%	0.0%	0.0%
FaceFusion¶	**0.3%**	1.3%	5.7%	**57.2%**	**100.0%**	39.5%	54.2%	0.0%

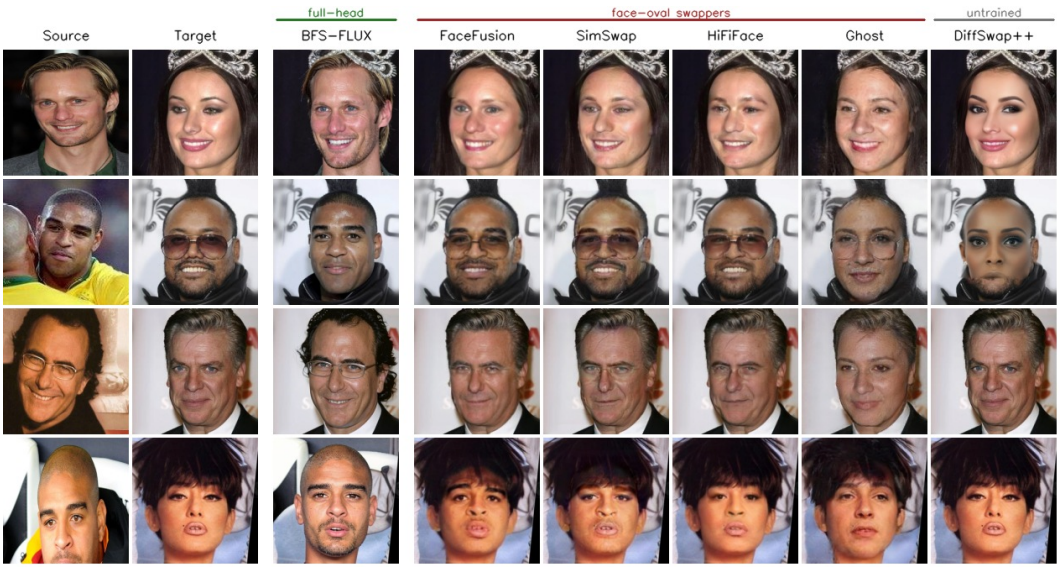


Figure 1: **Same pairs, six methods.** Every face-oval swapper (FaceFusion, SimSwap, HiFiFace, Ghost) places the donor's face inside the target's own hair, headwear and eyewear—exactly the pixels the standard crop never reads—while the full-head teacher replaces the whole head. DiffSwap++, an untrained ablation, largely returns the target. Rows were ranked by the teacher's transfer quality and curated for diversity, so they are illustrative rather than a random sample.

does exactly what a face swapper is asked to do— yet the ring probe accepts the target for 57.2% of those same outputs ($[51.2, 63.0]$). A method can satisfy the transfer half of Eq. 17 perfectly, beat every alternative on the standard crop, and still fail the criterion outright. Figure 1 shows why: on the same pairs, every face-oval method returns the donor's face inside the target's own hair, headwear and eyewear.

Under ϕ_{ring}, Ret is 0.0% for every face-oval method and 8.3% for BFS-FLUX: the ring carries essentially none of the *donor's* identity, making it empirically a near-pure target-identity channel. That asymmetry is why Eq. 17 reads transfer from ϕ_{std} and leakage from ϕ_{ring}.

5.2 Why reading the standard window alone does not work

The marginal rates above already show the standard window understates leakage. The sharper question is conditional: *of the outputs the standard window certifies as de-identified, how many does the ring probe identify?* Writing $C_{\text{std}} = \{\hat{x} : s_{\text{std}}(\hat{x}, x_t) < \tau_{\text{std}}(\alpha)\}$ for the certified-safe set

and $I_{\text{ring}} = \{\hat{x} : s_{\text{ring}}(\hat{x}, x_t) \geq \tau_{\text{ring}}(\alpha)\}$ for the ring-identified set, Table 4 (Appendix C) reports $\Pr[I_{\text{ring}}]$, $\Pr[C_{\text{std}}]$ and $\Pr[I_{\text{ring}} \mid C_{\text{std}}]$ at $\alpha = 10^{-3}$.

The conditional and marginal rates coincide to within a few points for every method (97.4 vs. 97.3 for HiFiFace, 70.3 vs. 69.3 for SimSwap, 29.4 vs. 29.3 for Ghost, 2.2 vs. 2.3 for BFS-FLUX), i.e.

$$\Pr\left[I_{\text{ring}} \mid C_{\text{std}}\right] \;\approx\; \Pr\left[I_{\text{ring}}\right], \tag{18}$$

so the two readings are very nearly independent. This is a stronger statement than "the standard window is optimistic". A conservative reader might hope a passing standard score at least shifts the odds — that certified outputs leak less often than average. Eq. 18 says it does not: passing the standard test carries almost no information about whether the target survives in the pericranial region. For HiFiFace, 39.0% of outputs are certified de-identified by the standard window and 97.4% of *those* are identified from hair and ears at the same false-accept rate. Figure 5 (Appendix C) shows three such outputs; in the third the standard probe reports a *negative* similarity to the target, its strongest possible verdict of non-identity, while the ring probe accepts at 0.924. Sec. 5.7 gives bootstrap intervals for the conditional–marginal gap; all five cover zero.

5.3 Gallery retrieval: can the protected person be found?

A pointwise threshold asks whether one pair is above τ. A realistic adversary ranks a gallery. Following FaceTracer (Zhang et al., 2024), we query each swap output against the benchmark's own 600-identity gallery and ask where the *target*—the person the swap was meant to protect—lands in the ranking (Table 9, Appendix C).

The gap is the largest in the paper. Under ϕ_{std} the protected person is found at rank 1 essentially never (0.0% SimSwap, 5.3% HiFiFace, 3.0% Ghost); under ϕ_{ring}—same gallery, same recognizer—for 85.7%, 98.0% and 61.7% of outputs. For the evaluated face-oval methods, target retrieval from the pericranial region substantially exceeds retrieval from the standard face crop—on this gallery, searching by hair and ears is the stronger query.

BFS-FLUX again separates: 4.3% target Rank-1 under ϕ_{ring}, the only entry that does not become substantially more retrievable when the adversary switches regions.

5.4 Is it area, or is it region?

With equal pixel budgets (≈14,000 px) removed from both regions, occluding the face *raises* result-to-target similarity while occluding the ring *lowers* it for every face-oval method: at matched area the two regions carry opposite-signed evidence, so pixel count cannot explain the effect (FII from -0.41 to -0.53, against -0.15 for the teacher; Table 8). The ring probe is also stable across photo composition where the 2.2× box is not (1.04× vs. 3.75× variation across four source datasets; Appendix E).

5.5 Does the conclusion depend on the recognizer?

Every number so far comes from one frozen ArcFace; if the pericranial reading were an artifact of that network the protocol would measure a quirk rather than a property of the image. We hold the regions fixed and vary the representation: ArcFace R50 (`buffalo_l`, the reference), ArcFace R100 (`antelopev2`, different depth and training corpus), FaceNet512 (different loss, different framework), and DINOv2 ViT-B/14, a general vision encoder never trained on faces. Each is calibrated separately on the same identity-disjoint gallery, because thresholds are not transferable between representations.

The method ordering under ϕ_{ring} is preserved exactly: Spearman $\rho = +1.000$ ($p < 10^{-4}$) against the reference for all three alternatives. The separation that matters is preserved too—the full-head teacher reads 0.3–2.3% while every face-oval method reads 11.7–97.3%, a factor of 5 to 42, with non-overlapping cluster-bootstrap intervals under every encoder (Table 3, Appendix C). Changing the recognizer, its training data, its objective, or removing face supervision altogether does not change which methods leave the target in the pericranial region.

Where the encoders disagree, and why it is informative. The *within-method* comparison does not transfer as cleanly. Under the three face recognizers every face-oval method reads higher on

ϕ_{ring} than on ϕ_{std}; under DINOv2 that ordering reverses for SimSwap (82.7% std vs. 55.3% ring), HiFiFace (96.3 vs. 94.3) and Ghost (38.3 vs. 11.7), while holding for DiffSwap++ (57.3 vs. 82.0). The calibration table explains it: DINOv2's impostor floor on the aligned face crop is 0.313, against ≤ 0.016 for the face recognizers. A general encoder is not identity-selective on a face crop—it reads lighting, skin tone and composition, which a face-oval swap inherits from the target—so its ϕ_{std} channel reports appearance, not identity, and stays high exactly when a method leaves that appearance intact. That is an argument for calibrating each probe against its own impostor distribution, not an objection to the finding.

5.6 An independently built full-head system

The teacher is the benchmark's own generator, so its showing invites the charge of circularity. The test that settles it is whether an *independently built* full-head system also comes out low. BFS-Qwen-V4 is one: a different base model (Qwen-Image-Edit-2511 rather than FLUX.2-Klein-4B), different adapter weights, and no role in tuning the gate that selected the 300-pair evaluation set. It shares no pairs with that set, so we score 300 of its own verified triplets (216 donor identities) under thresholds calibrated on its own gallery, with impostor pairs made disjoint by identity label (179,589 pairs, FAR precision 5.6×10^{-6}).

It leaks 1.7% at the ring (cluster-bootstrap CI $[0.3, 3.2]$), beside 2.3% for the teacher and 29–97% for the face-oval methods, while placing the donor in 98.3% of outputs under ϕ_{std} ($[96.8, 99.7]$) — so it satisfies Eq. 17 on its own evidence, and its low leakage is not the degenerate kind that Ghost exhibits. What the protocol rewards is therefore full-head regeneration as such, not the particular model that built PHANTOM-150K. Per-probe values are in Table 6 (Appendix C).

Its calibration is an unplanned replication: thresholds derived from a disjoint gallery land within 0.005 of the originals ($\tau_{\text{ring}} = 0.849$ vs. 0.853; ring impostor means 0.445 and 0.447), so the pericranial floor is a property of the probe, not of one gallery. Provenance is not fully independent—both share a line of work—but architecture, weights and gate-tuning are, which is what the objection concerns.

5.7 Uncertainty, dependence, and identity-disjoint re-analysis

The 300 pairs draw on 220 donor identities, 60 recurring. Under a cluster bootstrap over identity and on a one-pair-per-donor subset every rate holds (HiFiFace ring 97.3% overall, 98.2% identity-disjoint); exact McNemar rejects std/ring interchangeability for every method (HiFiFace $p = 5.7 \times 10^{-28}$); and the conditional gap of Sec. 5.2 has an interval covering zero for all five methods (Table 7, Appendix C).

6 Limitations

ArcFace was trained on aligned face crops, so pericranial inputs are out-of-distribution; the high ring impostor floor (Table 5) is the symptom, and calibration is what keeps the rates interpretable. For face-oval methods the ring probe is close to copy detection—it measures *what was left behind*, not the biometric power of hair. It does not, however, hinge on the parser's exact boundary: eroding or dilating the parsed region by up to 9 px and recalibrating at each radius moves ring leakage by a few points at most (HiFiFace 96.2–98.0%, the teacher 1.7–3.3%; Table 2). Split and modality caveats are in Appendix D.

7 Conclusion

Face-swap evaluation reads identity from the one region a face-oval swapper is guaranteed to change. Reading segmentation-defined head regions at calibrated operating points instead shows the target surviving in the pericranial ring for 57–97% of outputs from widely used swappers—one of which scores 0.3% under the standard crop—while that standard test says almost nothing about whether it does.

ETHICS STATEMENT

All identity images originate from VGGFace2 and CelebA-HQ, collected under those datasets' institutional approvals. Because every generated result depicts a real donor and a real target identity, we do not redistribute generated images (Sec. 3); we release only the generation pipeline, per-sample metadata, and evaluation code, under a non-commercial research-only licence prohibiting non-consensual face-swap creation, identity fraud, and surveillance evasion. This licence binds anyone who uses the code to regenerate the pool.

This work has a dual-use character we want to state plainly. A sharper instrument for finding residual identity in a swapped image is, by construction, also a sharper instrument for re-identifying someone who believed they had been anonymized. We judge publication to be net-positive because the failure it documents is already deployed: the face-oval swappers we evaluate ship in tools practitioners run today, and those practitioners use face-crop metrics to certify their outputs as de-identified. An attacker willing to look at hair does not need this paper; a practitioner relying on a face-crop metric does. We report no new attack capability beyond applying a standard recognizer to a standard segmentation of the image.

REPRODUCIBILITY STATEMENT

The generation pipeline, quality gate, seeds, per-sample metadata for every attempted generation (not only accepted ones), and all evaluation code are released. Sec. 3 describes the dataset construction and the reconciled funnel; Sec. 4 defines all four probes including the exact CelebAMask-HQ class lists; Sec. 4.3 specifies the impostor construction, the same-identity exclusion rule, and the FAR precision; Sec. 4.4 specifies the occlusion construction and fill. Appendix B gives model versions and preprocessing. Because source images derive from VGGFace2 and CelebA-HQ, reproducing the triplet pool requires the reader's own access to those datasets under their respective terms; the evaluation protocol itself can be run against any method's outputs without our data.

AI USE STATEMENT

Generative AI tools were used for code assistance (experiment scripts and plotting) and for editing and tightening prose. They were not used to generate research ideas, experimental designs, or results. All AI-assisted code was reviewed and executed by the authors, and every number reported in this paper was produced by running that code on the stated data; no results are AI-generated or estimated. The authors take responsibility for the final content, including all claims and artifacts.

REFERENCES

Alissonerdx. Bfs-best-face-swap: Head-swap LoRA adapters for FLUX.2-klein and Qwen-Image-Edit. https://huggingface.co/Alissonerdx/BFS-Best-Face-Swap, 2026. MIT license.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? In *ICCV*, 2017.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *FG*, 2018.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. SimSwap: An efficient framework for high fidelity face swapping. In *ACM MM*, 2020.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, 2020.

FaceFusion Contributors. FaceFusion: Industry leading face manipulation platform. https://github.com/facefusion/facefusion, 2026.

Daniel George, Charles Yeh, Daniel Lee, and Yifei Zhang. From measurement to mitigation: Quantifying and reducing identity leakage in image representation encoders with linear subspace removal. *arXiv:2604.05296*, 2026.

Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022.

Sohyun Jeong, Taewoong Kang, Hyojin Jang, and Jaegul Choo. Zero-shot head swapping in real-world scenarios. In *CVPR*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

Kihong Kim, Yunho Lee, Seokju Park, Ji-won Ha, and Seonghyeon Kim. DiffFace: Diffusion-based face swapping with facial guidance. *arXiv:2212.13344*, 2022.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *CVPR*, 2020.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

Elaine M. Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

Yuhan Wang, Xu Qi, Jing Chen, Junqi He, and Zongming Li. HifiFace: 3d shape and semantic prior guided high fidelity face swapping. In *IJCAI*, 2021.

Zhongyi Zhang, Jie Zhang, Wenbo Zhou, Xinghui Zhou, Qing Guo, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. Facetracer: Unveiling source identities from swapped face images and videos for fraud prevention. *arXiv:2412.08082*, 2024.

Wenliang Zhao, Yongming Rao, Weicheng Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. DiffSwap: High-fidelity and controllable face swapping via 3D-aware masked diffusion. In *CVPR*, 2023.

A Dataset generation protocol

Every step below is taken from the generation code as run; the release manifest records each triplet's inputs, seed and scores.

1. **Sources.** Donors: a 540-identity subset of the VGGFace2 training split (each identity has ≥ 102 images; the identity list is in the manifest). Targets: the 256×256 release of CelebA-HQ (30,000 images; 29,997 yield landmarks).
2. **Donor pool.** Per identity, images are sorted by filename and 40 are taken at stride $\lfloor n/40 \rfloor$.
3. **Expression clusters.** 68-point 2D landmarks (FAN) on every target; feature $(\ell_{17:68} - \ell_0)/\|\ell_{16} - \ell_0\|$ (102-D); MiniBatch k-means, $k = 64$, seed 42, 10 initialisations, 300 iterations, batch 2048.
4. **Pair enumeration.** All $540 \times 40 \times 64 = 1{,}382{,}400$ (identity, donor image, cluster) combinations, shuffled with NumPy `default_rng(42)`. The combination at position i takes the target at index $\mathrm{md5}(\text{id_img_cluster_}i)[:8] \bmod |\text{cluster}|$ within its cluster. The pair key is `identity_image_cluster_target`; resumed runs skip logged keys. Because the pair order changed between resumed runs, 1,461,318 distinct keys were logged in total.
5. **Prompt.** Picture 1 is the target (the base image), Picture 2 the donor (the head):

```
head_swap: start with Picture 1 as the base image, keeping its
lighting, environment, and background. Remove the head from
Picture 1 completely and replace it with the head from Picture
2, strictly preserving the hair, eye color, and nose structure
of Picture 2. Copy the direction of the eye, head rotation,
and micro-expressions from Picture 1. High quality, sharp
details. The expression in Picture 1 is: {desc}. Copy it
exactly.
```

`desc` is built from the target's landmarks, with distances divided by face width: "mouth open" if the inner-lip gap exceeds 0.06, otherwise "mouth closed"; "laughing"/"smiling" if mouth width exceeds 0.35; "eyes nearly closed" below 0.022 eye opening, "slightly squinting" below 0.032, otherwise "eyes open"; "eyebrows raised" above 0.25 brow–eye gap, "furrowed" below 0.15.

6. **Generation.** FLUX.2-Klein-4B (bfloat16) with the BFS Head V1 adapter; both images resized so the long side is 256 px, each dimension rounded to a multiple of 16 (Lanczos); output at the target's size; 20 steps; guidance 3.5 passed but ignored by the step-distilled model; seed $\mathrm{int}(\mathrm{md5}(\text{pair key})[:8], 16)$ on a CPU generator.
7. **Gate** (sequential; the first failure is recorded). Faces are detected with InsightFace `buffalo_l` at 640×640, keeping the highest-scoring face and retrying once with 30% replicate padding. (i) Reject if the result has no face or `det` < 0.60; (ii) reject if the donor has no face; (iii) ρ_{don} is the ArcFace cosine between result and donor, rejected if < 0.30; ρ_{tgt} is recorded but not gated; (iv) $\varepsilon_{\mathrm{expr}}$ is the RMSE over landmarks 17–67 between the result and the target (resized to the result's size), divided by the target's landmark x-extent, rejected if > 0.08.
8. **Release filter.** The first 10,000 accepted triplets were gated at $\rho_{\mathrm{don}} \geq 0.28$, $\varepsilon_{\mathrm{expr}} \leq 0.09$. Re-applying the gate above excludes 1,212 on ρ_{don} and 1,164 on $\varepsilon_{\mathrm{expr}}$, plus 2 whose expression landmarks failed: **156,111** released of 158,489 accepted.
9. **Splits.** Each donor identity gets $h = \mathrm{sha256}(s \,\|\, \text{identity}) \bmod 100$ with salt $s =$ `phantom150k-split-v1:`, and goes to train if $h < 80$, val if $h < 90$, test otherwise, giving 125,876 / 17,707 / 12,528 triplets from 438 / 58 / 44 identities. The strict split assigns each target image the same way and keeps a triplet only when both agree (100,243 / 1,701 / 1,250).
10. **Evaluation subset.** The 300 cross-method pairs were drawn with Python `random.sample` (seed 42) from all accepted triplets before the release filter; 7 fall in the excluded set. Excluding them changes $\mathrm{Leak@}10^{-3}$ by at most 0.7 points (SimSwap under ϕ_{ring}, $69.3 \rightarrow 68.6\%$), and by at most 0.4 points under ϕ_{std}.
11. **Manifest columns.** `triplet_id`, `pair_key`, `donor_identity`, `donor_image`, `target_image`, `expr_cluster`, `seed`, `arc_don`, `arc_tgt`, `expr_rmse`, `det_score`, `generated_at`, `in_release`, `exclusion_reason`, `split_donor`, `split_strict`. Paths are reduced to dataset-relative names.



Figure 2: **PHANTOM-150K is a source × target matrix of full-head swaps.** Four donors (rows) against four targets (columns); every interior cell is generated for that exact pair. The teacher replaces hair and skull contour along with the face, which is precisely the property that makes the corpus usable as a control for a pericranial probe—the face-oval methods we benchmark against cannot produce these.

B Implementation details

Recognizer. InsightFace `buffalo_l` (`w600k_r50`, ResNet-50, WebFace600K) for all probes, with SCRFD detection at 320×320. All probes feed this same network; only the pixels differ. ϕ_{std} uses the library's own 5-point aligned crop; ϕ_{box}, ϕ_{head} and ϕ_{ring} feed a 112×112 resize of the region to the recognition head directly.

Parser. BiSeNet ResNet-34 trained on CelebAMask-HQ, 512×512 input, ImageNet normalization, nearest-neighbour resize of the argmax label map back to image resolution. Region definitions: face interior $=$ {skin, brows, eyes, glasses, nose, mouth, lips}; pericranial ring $=$ {ears, earring, hair, hat}; head $=$ their union. Background, neck, necklace and clothing are always excluded. Masked pixels are zeroed, the region is cropped to its bounding box with 5% padding, then resized.

Calibration. Impostor pairs are all cross-identity pairs of the 600-image gallery (179,592 after excluding 108 pairs whose ϕ_{std} similarity exceeds 0.50 and are therefore presumed same-identity), giving a FAR precision of 5.6×10^{-6}. Thresholds are the $(1 - \text{FAR})$ quantiles of the per-probe impostor distribution.

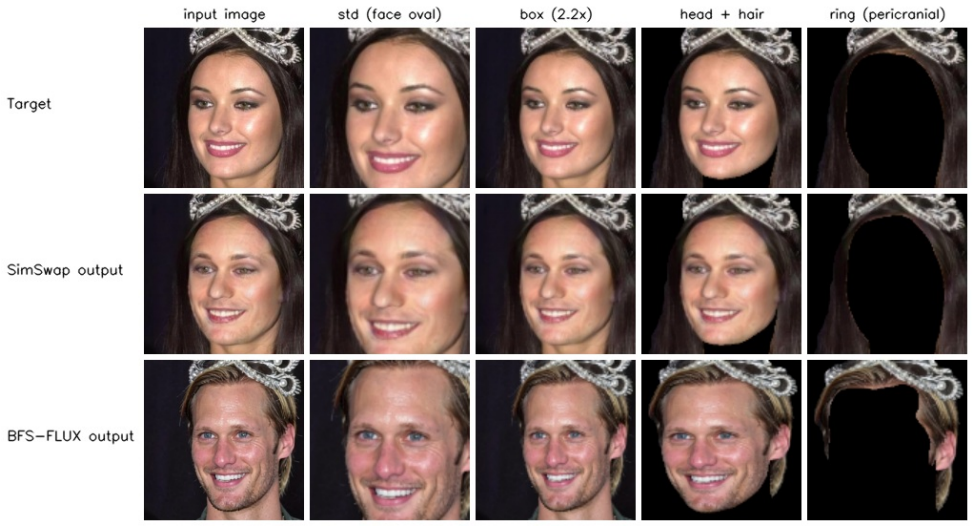


Figure 3: **What each probe sees**, for one target and two methods' outputs on the same pair. The ϕ_{ring} column is the argument in one picture: for the face-oval method (middle row) it is pixel-for-pixel the target's own hair and headwear, because that method never touched those pixels; for the full-head teacher (bottom row) it has been regenerated along with the rest of the head. The ϕ_{std} column—all that standard evaluation reads—looks equally "de-identified" in both rows.

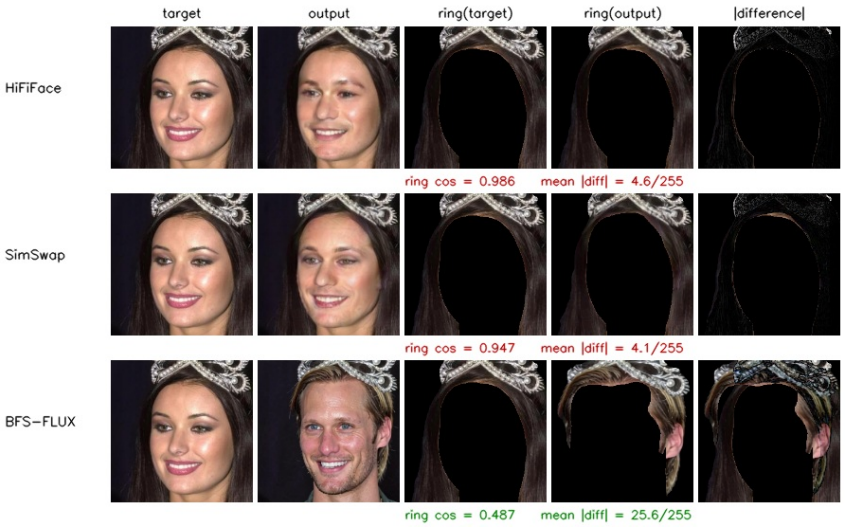


Figure 4: **The leakage, by inspection.** For one pair: the target, each method's output, the pericranial ring extracted from each, and the absolute difference between the two rings. For the face-oval methods the difference is near-black—mean $|\Delta|$ of 4.1–4.6 out of 255, the magnitude of JPEG recompression noise—because those are the target's own pixels, unmodified. That is the mechanism behind ring similarities of 0.95–0.99. BFS-FLUX regenerates the region and differs by $25.6/255$, with ring similarity 0.487, essentially the impostor floor.

B.1 Exact environment and artifact hashes

Software. Python 3.12.3 on Linux 6.17; PyTorch 2.7.1+cu128 (CUDA 12.8, cuDNN 9.7.1); diffusers 0.38.0; transformers 5.9.0; InsightFace 0.7.3; ONNX Runtime 1.26.0; DeepFace 0.0.100 (TensorFlow 2.21.0 backend, used only for the FaceNet512 cross-check); timm 1.0.28; OpenCV 4.13.0.92; NumPy 2.4.6; SciPy 1.17.1. Generation ran on a single NVIDIA RTX PRO 6000 Blackwell (96 GB, driver 580.173.02); all evaluation in this paper runs on CPU.

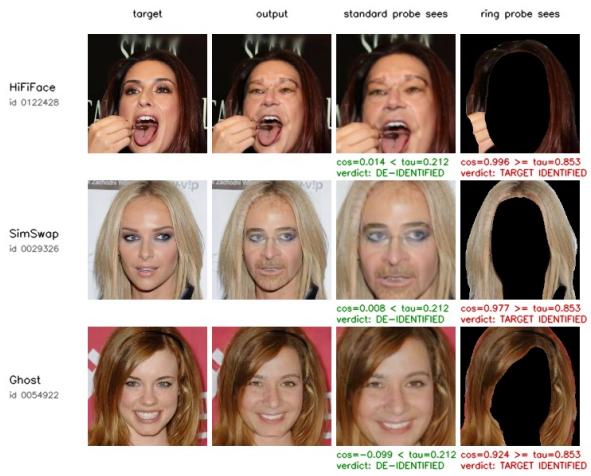


Figure 5: **Outputs the standard window certifies as de-identified and the ring probe identifies.** The last two columns show what each probe receives, scored against its own $\tau_R(10^{-3})$. The replaced face reads as non-identity; the unseen hair, ears and skull contour are the target's own pixels. Bottom row: standard similarity is negative, yet the ring accepts at 0.924.

Model artifacts (SHA-256, first 16 hex digits): parser `bisenet_resnet_34.onnx` `4a0b8c958a3c9389`; recognizer `w600k_r50.onnx` `4c06341c33c2ca1f` and detector `det_10g.onnx` `5838f7fe053675b1` (buffalo_l); `glintr100.onnx` `4ab1d6435d639628` (antelopev2); `facenet512_weights.h5` `3f76b5117a9ca574`; teacher adapter `bfs_head_v1_flux-klein_4b.safetensors` `77aabcd6144612dd`; independent full-head adapter `bfs_head_v4_qwen_image_edit_2509.safetensors` `cd3278c33c7df36b`; FaceFusion's `inswapper_128_fp16.onnx` `c4eccca86ad17758`. The SimSwap and HiFiFace exports (`simswap_256`, `hififace_unofficial_256` and their `crossface` adapters) were fetched on the second machine and are not present here, so we identify them by their FaceFusion asset releases (`models-3.0.0`, `models-3.1.0` and `models-3.4.0`) rather than by a hash we cannot compute.

Generation. FLUX.2-Klein-4B in bfloat16 with the teacher adapter above, 20 sampling steps, guidance scale 3.5 (passed to the pipeline but ignored: the model is step-distilled), 256 px long side, one prompt template with a description of each target's own expression, derived from its landmarks, appended (Appendix A). The seed is derived per pair, $\text{seed} = \text{int}(\text{md5}(\text{pair key})[:8], 16)$ on a CPU generator, so a pair's seed is a pure function of its identifier. The gate keeps a sample iff $\rho_{\text{don}} \geq 0.30$, $\varepsilon_{\text{expr}} \leq 0.08$ and $\det \geq 0.60$. The independent system (Sec. 5.6) is Qwen-Image-Edit-2511 in bfloat16 with its own adapter and the same gate.

Evaluation determinism. Bootstrap resampling uses seed 20260915 with $B = 10^4$; the Qwen pair sample uses the same seed. Every threshold in the paper is a quantile of a stored impostor array, so recalibration from the released score files reproduces the tables exactly without re-embedding.

What reproducibility means here. We do not claim bit-exact image regeneration: that would additionally require the same GPU, driver and dtype-reduction order. We claim reproducibility at two levels that do not depend on our hardware. *Pair level*: a pair key names a VGGFace2 identity and image and a CelebA-HQ index, so a user holding those datasets reconstructs the identical pair list and seeds. *Metric level*: every per-sample score, area, threshold and rate is released as JSON, so all tables, figures and statistics in this paper are recomputable end-to-end from the released files with no model inference at all.

C Additional result tables

Table 2: **The result does not hinge on the parser's boundary.** $\mathrm{Leak@}10^{-3}$ under ϕ_{ring} with the parsed region eroded ($-$) or dilated ($+$) by the stated pixel radius, recalibrated at every radius on identity-disjoint gallery pairs (bottom row). Erosion discards pixels near the face/hair boundary; dilation reaches into background and neck. Only one parser is available offline, so this perturbs its decision boundary rather than replacing the segmenter.

Radius (px)	-9	-6	-3	0	$+3$	$+6$	$+9$
BFS-FLUX†	1.7	1.7	1.7	3.3	2.3	2.0	2.7
SimSwap	63.2	64.5	67.9	72.7	73.0	70.3	64.0
HiFiFace	96.2	98.0	96.3	97.7	97.0	97.3	97.0
$\tau_{\mathrm{ring}}(10^{-3})$	0.915	0.901	0.870	0.838	0.824	0.807	0.783

Table 3: **The regional conclusion under four independently trained representations**, $\mathrm{Leak@}10^{-3}$ with cluster-bootstrap intervals over donor identity. Each encoder is calibrated on its own identity-disjoint impostor distribution. The ordering under ϕ_{ring} is identical across all four (Spearman $\rho = +1.000$); the teacher/face-oval separation is 5–42× everywhere. DINOv2's face-crop impostor floor (0.313 vs. ≤ 0.016) is why its ϕ_{std} column reads appearance rather than identity.

Method	ArcFace R50	ArcFace R100	FaceNet512	DINOv2 B/14
	$\mathrm{Leak@}10^{-3}$ under ϕ_{ring} [95% CI]			
BFS-FLUX†	**2.3** [0.7, 4.2]	**0.3** [0.0, 1.0]	**1.0** [0.0, 2.3]	**2.3** [0.7, 4.1]
SimSwap	69.3 [63.7, 74.6]	46.0 [40.4, 51.4]	54.3 [48.8, 59.9]	55.3 [50.0, 60.6]
HiFiFace	97.3 [95.4, 99.0]	95.3 [92.8, 97.6]	89.3 [85.9, 92.8]	94.3 [91.6, 96.9]
DiffSwap++‡	84.3 [80.3, 88.3]	71.7 [66.4, 76.4]	63.0 [57.6, 68.3]	82.0 [78.0, 86.1]
Ghost§	29.3 [24.5, 34.5]	14.0 [10.3, 17.8]	20.0 [15.6, 24.7]	11.7 [8.2, 15.4]
	for reference, $\mathrm{Leak@}10^{-3}$ under ϕ_{std}			
BFS-FLUX†	9.7	8.7	3.0	15.0
SimSwap	29.3	23.3	8.3	82.7
HiFiFace	61.0	52.7	18.3	96.3
DiffSwap++‡	61.3	58.0	34.0	57.3
Ghost§	1.3	2.0	1.3	38.3
impostor floor (std / ring)	0.005 / 0.447	0.003 / 0.373	0.016 / 0.690	0.313 / 0.466

Table 4: **The standard window is nearly uninformative about pericranial leakage**, at $\alpha = 10^{-3}$, $n = 300$ per method. C_{std}: standard window certifies de-identified. I_{ring}: ring probe identifies the target. Columns 3 and 4 agreeing is Eq. 18.

Method	$\Pr[C_{\mathrm{std}}]$	$\Pr[I_{\mathrm{ring}}]$	$\Pr[I_{\mathrm{ring}} \mid C_{\mathrm{std}}]$	$\lvert C_{\mathrm{std}} \cap I_{\mathrm{ring}} \rvert$
BFS-FLUX†	90.3%	2.3%	**2.2%**	6
SimSwap	70.7%	69.3%	**70.3%**	149
HiFiFace	39.0%	97.3%	**97.4%**	114
DiffSwap++‡	38.7%	84.3%	**81.0%**	94
Ghost§	98.7%	29.3%	**29.4%**	87

Table 5: **Impostor distributions differ sharply across probes**, measured over 179,592 cross-identity gallery pairs (108 pairs excluded as suspected same-identity; FAR precision 5.6×10^{-6}). The pericranial ring has a similarity floor an order of magnitude above the aligned face crop, so a single hand-set cutoff cannot serve both. All later results are reported at thresholds calibrated per probe.

Probe	Impostor similarity			Calibrated threshold τ_p		
	mean	p99	max	FAR 10^{-2}	FAR 10^{-3}	FAR 10^{-4}
ϕ_{std} (face oval)	0.005	0.147	0.497	0.147	0.212	0.284
ϕ_{box} (2.2× box)	0.030	0.242	0.667	0.242	0.356	0.477
ϕ_{head} (head + hair)	0.034	0.251	0.603	0.251	0.403	0.507
ϕ_{ring} (pericranial)	**0.447**	**0.771**	0.984	0.771	0.853	0.934

Table 6: **An independently built full-head system** (BFS-Qwen-V4) under the same protocol, $n = 300$ of its own verified triplets, thresholds calibrated on its own label-disjoint gallery. Intervals are cluster bootstraps over its 216 donor identities.

Probe	$\tau(10^{-3})$	Leak@10^{-3} ↓	Ret@10^{-3} ↑
ϕ_{std}	0.214	16.0% [12.2, 19.9]	**98.3%** [96.8, 99.7]
ϕ_{box}	0.374	1.7% [0.3, 3.2]	16.3% [12.3, 20.7]
ϕ_{head}	0.470	0.7% [0.0, 1.7]	21.0% [16.3, 26.0]
ϕ_{ring}	0.849	**1.7%** [0.3, 3.2]	5.3% [3.0, 7.9]

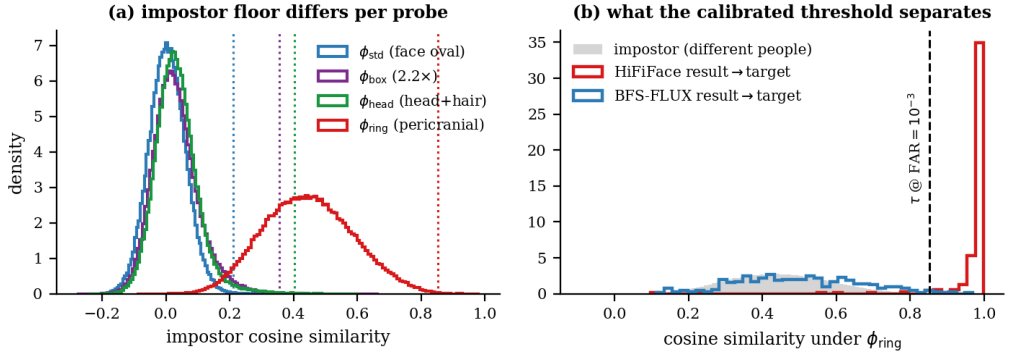


Figure 6: **Calibration is load-bearing.** (a) Impostor similarity over cross-identity gallery pairs. ϕ_{ring} sits an order of magnitude above the face crop, so one hand-set cutoff cannot serve both probes; dotted lines are the per-probe τ at FAR $= 10^{-3}$. (b) Under ϕ_{ring}, what that threshold separates: BFS-FLUX's result-to-target scores lie on top of the impostor distribution—indistinguishable from strangers—while HiFiFace's pile up near 1.0.

Table 7: **Uncertainty and identity-disjoint re-analysis** at $\alpha = 10^{-3}$, $n = 300$ per method, $B = 10^4$ bootstrap resamples. CI_{iid} resamples pairs; CI_{clu} resamples the 220 donor identities; *id-disj.* recomputes the rate on one pair per donor. p is the exact McNemar test on the paired ϕ_{std}/ϕ_{ring} accept decisions. Δ_{cond} is $\Pr[I_{\text{ring}} \mid C_{\text{std}}] - \Pr[I_{\text{ring}}]$ in points, with its bootstrap interval.

	Leak@10^{-3} ϕ_{std}		Leak@10^{-3} ϕ_{ring}			
Method	rate	CI_{clu}	rate	CI_{clu}	id-disj.	McNemar p
BFS-FLUX†	9.7%	[6.5, 13.2]	**2.3%**	[0.7, 4.1]	2.3%	1.9×10^{-4}
SimSwap	29.3%	[24.1, 34.7]	69.3%	[63.8, 74.6]	69.1%	1.2×10^{-20}
HiFiFace	61.0%	[55.4, 66.5]	**97.3%**	[95.4, 99.0]	98.2%	5.7×10^{-28}
DiffSwap++‡	61.3%	[55.9, 66.8]	84.3%	[80.3, 88.2]	83.6%	1.3×10^{-10}
Ghost§	1.3%	[0.3, 2.8]	29.3%	[24.4, 34.3]	28.2%	2.0×10^{-22}

Δ_{cond} (points, 95% CI): BFS-FLUX -0.1 $[-0.9, +0.4]$; SimSwap $+0.9$ $[-2.3, +4.4]$; HiFiFace $+0.1$ $[-2.3, +2.3]$; DiffSwap++ -3.3 $[-8.8, +2.0]$; Ghost $+0.1$ $[-0.6, +0.6]$ — all cover zero.

Table 8: **Equal-area, equal-strength occlusion attribution** (Eq. 16), 150 pairs per method, embedded under ϕ_{head}. Both occlusions remove the same number of pixels with the same fill. $\Delta_{\text{face}} < 0$ means occluding the face *raises* similarity to the target.

Method	base sim.	Δ_{face}	Δ_{ring}	FII	$P(\text{FII} < 0)$	occluded px
BFS-FLUX†	0.124	-0.164	-0.013	$\mathbf{-0.151}$	75%	13,262
SimSwap	0.276	-0.433	$+0.029$	-0.462	99%	14,372
HiFiFace	0.379	-0.484	$+0.049$	$\mathbf{-0.533}$	99%	14,040
DiffSwap++‡	0.384	-0.380	$+0.032$	-0.413	94%	14,228
Ghost§	0.125	-0.412	$+0.009$	-0.421	92%	15,055

Table 9: **Ranked retrieval against the 600-identity gallery.** Tgt: the person the swap should protect. Src: the donor. Under the standard crop the protected person is effectively unfindable; under the pericranial ring probe they are found at rank 1 in the large majority of cases.

Method	Tgt R@1 ↓		Tgt R@5 ↓		Src R@1 ↑	
	ϕ_{std}	ϕ_{ring}	ϕ_{std}	ϕ_{ring}	ϕ_{std}	ϕ_{ring}
BFS-FLUX[†]	2.3%	**4.3%**	17.7%	9.7%	95.0%	17.7%
SimSwap	0.0%	**85.7%**	48.7%	93.0%	100.0%	0.0%
HiFiFace	5.3%	**98.0%**	76.0%	98.3%	94.7%	0.0%
DiffSwap++[‡]	61.3%	92.7%	78.3%	96.0%	0.0%	0.0%
Ghost[§]	3.0%	**61.7%**	7.3%	75.0%	2.7%	0.0%

D EXTENDED LIMITATIONS

The recognizer was not trained for these regions. ArcFace is trained on tight aligned crops. Feeding it a hair-and-ears region is out-of-distribution, and the very high impostor floor we measure (Table 5) is a direct symptom: the embedding is not well-spread over such inputs. Calibration is what makes the resulting rates interpretable, and it is why we report rates rather than similarities throughout. A recognizer trained for pericranial regions would very likely give sharper separation; we do not have one, and building one is future work.

The ring probe is near-tautological for face-oval methods—by design. For SimSwap or HiFiFace the ring pixels are the target's own pixels, so a high match is close to a copy-detection result rather than a recognition result. We consider this the point rather than an objection: the field's metric reports "de-identified" for an image that still literally contains the target's hair, ears, and skull contour. But it does mean the ring probe should be read as measuring *what was left behind*, not as a claim about the discriminative power of hair as a biometric in general.

Segmentation is a dependency. The probe inherits the parser's failure modes—hats and heavy occlusion are the obvious ones, and the parser merges some accessories into the hair class. We use a single parser (BiSeNet/CelebAMask-HQ) and have not compared parsers, though the result is stable under ± 9 px perturbation of its boundary (Table 2).

Split, modality, and human judgment. PHANTOM-150K ships two canonical splits (Appendix A): one disjoint by donor identity, and a stricter one that is also disjoint by target image. Targets cannot be made disjoint by *person* without CelebA identity labels, which we did not use, so the same CelebA-HQ celebrity may still appear on both sides of the strict split. All metrics are single-frame, so temporal consistency is unmeasured, and every threshold here is grounded in measured impostor distributions rather than in human judgments of "does this look like person X"—gallery retrieval is only a partial proxy for that question.

E CROSS-DATASET PROBE STABILITY

A geometric probe inherits the photographer's framing. A $2.2\times$ box around a face in an in-the-wild snapshot reaches hair and shoulders; the same box on a tightly-cropped studio portrait runs out of image and captures almost nothing new. A segmentation-defined region does not have this problem: it finds the hair wherever the hair is.

Table 10 shows this is not hypothetical. Running the same FaceFusion face-oval baseline on fresh pairs from four sources, the box probe's mean target similarity varies by a factor of 3.7 across datasets (0.130 on FFHQ to 0.487 on VGGFace2) purely as a function of framing. The ring probe varies by 4% (0.855–0.888). The earlier version of this work had to treat that instability as a caveat—"the probe under-detects leakage on already-tightly-framed portraits"—and lean on retrieval instead. With a segmentation-defined region the caveat disappears, because the instability was a property of the box, not of the phenomenon.

Two further details support reading ϕ_{ring} as a target-identity channel. Its ret values (0.454–0.487) sit essentially on the impostor floor of 0.447 from Table 5: the ring carries no donor identity, on any dataset. And the standard crop is flat at 0.063–0.068 everywhere, confirming the blind spot is a property of the crop rather than of any one corpus.

Table 10: **Cross-dataset consistency**, FaceFusion (inswapper) face-oval baseline on fresh pairs from four sources, mean raw similarity. The box probe tracks photo composition; the segmentation probes do not. The ring column's ret values sit at the impostor floor (0.447), i.e. chance.

Dataset	n	mean ρ_{tgt} (leak) ↓ ϕ_{std}	ϕ_{box}	ϕ_{head}	ϕ_{ring}	ring ρ_{don} (chance = 0.447)
VGGFace2	268	0.068	0.487	0.423	0.862	0.487
CelebA-HQ	299	0.067	0.133	0.184	0.888	0.469
FFHQ	295	0.065	0.130	0.177	0.858	0.454
PHANTOM-150K val	198	0.063	0.137	0.187	0.855	0.479
max/min across datasets		1.08×	**3.75**×	2.39×	**1.04**×	1.07×

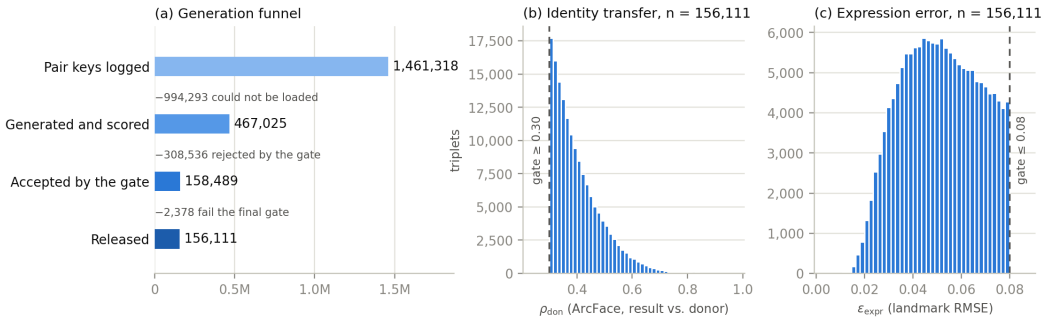


Figure 7: **Generation protocol and its yield.** (a) The reconciled funnel: diffusion sampling ran for all 467,025 pairs that reached generation, and the 308,536 gate rejections cost the same compute as the accepted ones. (b)/(c) Distributions of the two gate quantities on the accepted pool, with the gate thresholds dashed. Per-sample values are released for every accepted triplet, so the gate can be *tightened* without regenerating; loosening it would need the rejected samples, which were not retained.

F QUALITY-GATE THRESHOLD SENSITIVITY

Rather than assert that $0.30/0.08$ are the right gate thresholds, we measure their consequence on the released triplets' recorded scores: raising ρ_{don} from 0.30 to 0.50 retains 13.6% of the dataset at a 41% lower mean target leakage, while `det_score` barely affects leakage at any setting, being a detection-confidence filter rather than an identity signal. ρ_{don} is therefore the dominant privacy-relevant lever in the gate, which is why τ_{transfer} is anchored to it.

G COMPUTE AND ENERGY

Diffusion sampling ran for all $467{,}025$ attempted pairs, including the $308{,}536$ later rejected—a rejected sample costs exactly as much to generate as an accepted one, since rejection happens after generation and scoring. Generation spanned a 45-day wall-clock window (earliest and latest triplet file timestamps; intermittently paused, not continuous) on a single NVIDIA RTX PRO 6000 Blackwell (96 GB, 600 W rated TDP). The run's checkpoint counter records $329{,}464$ s (≈91.5 GPU-hours) of active compute, ≈0.71 s/attempt. We report this as a *lower bound*: the checkpoint's recorded start date postdates the earliest triplet on disk, so some early compute is not captured. At the rated TDP this is ≈54.9 kWh as an upper bound, or ≈41.2 kWh at 75% sustained utilization; both inherit the same lower-bound caveat. The separate cost of the $994{,}293$ candidate pairs that failed before generation (a single detection pass each) was not instrumented separately and is not included.