# Certified Dental Biometric Verification Under Partial Overlap

Krishi Attri
Seoul National University · Independent research
krishiattriwork@gmail.com github.com/Archerkattri/toothprint

## Abstract

Dental identification is usually evaluated as closed-set retrieval: given a dental scan, rank the enrolled gallery and report whether the correct subject appears first. This framing hides two operational requirements. First, forensic and clinical deployments need an *accept/abstain* decision with a controlled false-match rate (FMR), not only a Rank-1 score. Second, practical queries are often partial: field-of-view limits, trauma, extractions, and missing teeth break the rigid alignment assumption used by most 3D dental matching pipelines. We present ToothPrint, a dental biometric verification pipeline for 3D intraoral scans and 2D radiographs that combines rigid surface verification, learned partial-overlap correspondence, and split-conformal calibration. On 200 public 3D dental arches, rigid surface verification achieves Rank-1 0.995 and EER 0.005 under full coverage. Under realistic whole-tooth dropout, however, rigid GICP falls to Rank-1 0.23 at 50% tooth retention. A learned point-correspondence verifier lifts this regime to Rank-1 0.867 (AUC 0.984), while a crop-hardened embedding baseline reaches 0.635. The same score family supports a conformal open-set decision: full-coverage FNIR at FMR=1% is 0.00, while heavy tooth loss is correctly exposed as an abstention regime. We then move from synthetic scans to *real* intraoral arches: on 150 Teeth3DS+ upper arches acquired ungated and md5-verified, full-coverage identity reproduces Rank-1 1.000 (N=40), and an off-the-shelf zero-shot registrar (BUFFER-X) with no dental training recovers partial overlap to Rank-1 $1.00 \pm 0.00$ at 50% tooth retention and 0.95 at 30% over three crop-seed repetitions—evidence that the partial-overlap leg may be better served by a general-purpose zero-shot registrar than by a bespoke correspondence network, without touching the certified pipeline. A negative control is equally informative: a frozen indoor self-supervised encoder (Sonata/PTv3) with an ArcFace head does not transfer to dental identity (Rank-1 0.275 at full coverage). We report DET curves, full per-condition partial-overlap tables, a sensor-perturbation robustness sweep, hard-negative calibration, an untrained-correspondence control, and a cross-dataset transfer test, together with auxiliary results that carry the same split-conformal machinery to 2D radiographs, restoration patterns, and longitudinal bone-level and 3D surface change—each with a finite-sample false-alarm bound, and each on single-session synthetic data. The central limitation remains explicit: the new real-arch results are still single-session, so identity queries are synthetic reacquisitions and the real cross-session gate stays open. The method is therefore evidence for a research direction, not a clinical claim.

## 1 Introduction

Dental structures are among the most durable biometric signals. Forensic odontology has long used dental charting, restorations, and radiographs for post-mortem identification; modern intraoral scanners now provide dense 3D crown and gingival surfaces. Recent work has shown that 3D intraoral scans can identify individuals with high closed-set accuracy using registration pipelines based on hand-crafted descriptors and ICP refinement [18]. Yet a high Rank-1 number is not the same as a deployable biometric decision.

Two gaps motivate this paper. The first is *certification*. A forensic examiner or clinical reviewer needs to know when a system should accept a match, reject it, or decline to decide. Standard dental-identification papers typically report Rank-$N$ retrieval, not a finite-sample bound on the false-match rate of an individual accept decision. The second is *partial overlap*. A full dental arch is the favourable case: a rigid surface alignment can exploit the entire crown geometry. A real query can be incomplete because the scan did not cover the full arch, because teeth are missing, or because a fragment is available. Rigid methods then fail before scoring: the geometry used to initialize and refine the alignment is no longer the same geometry as the enrolled scan.

We study dental biometric verification under these two requirements. ToothPrint is a research pipeline with three layers. A rigid point-to-surface verifier provides a strong full-coverage baseline. A learned correspondence verifier emits a descriptor per point and verifies partial queries by mutual nearest-neighbor correspondences followed by weighted Procrustes alignment. A split-conformal calibration layer converts similarity scores into accept/abstain decisions with a target FMR.

The contribution is not a new clinical device. All 3D identity experiments here use synthetic reacquisition of single-session scans, whether those scans are the synthetic-benchmark Poseidon3D arches or the real Teeth3DS+ arches we add in this work. The contribution is a re-

producible benchmark and method study showing that (i) certified dental verification is feasible under full coverage, (ii) partial overlap is the dominant failure mode of rigid dental biometrics, and (iii) learned point correspondence—and, we now find, an off-the-shelf zero-shot registrar—substantially raises performance in that regime while still exposing when open-set rejection must abstain.

**Contributions.**

- We formulate dental identity as open-set verification with a conformal accept/abstain rule, rather than only closed-set retrieval.
- We introduce a point-correspondence verifier for partial dental arches and compare it against rigid GICP and crop-hardened embedding baselines under realistic whole-tooth dropout.
- We move from synthetic benchmark scans to *real* Teeth3DS+ intraoral arches, reproducing full-coverage identity and—crucially—measuring partial overlap on real dental geometry with error bars.
- We show that a general-purpose zero-shot registrar (BUFFER-X) recovers the real-arch partial-overlap regime, and position it as an optional drop-in for the correspondence leg while leaving the certified pipeline unchanged.
- We report a negative-transfer result: a frozen point-cloud foundation-model encoder (Sonata/PTv3) with a metric-learning head does not transfer to dental identity under a head-only recipe.
- We report full DET curves, complete per-condition partial-overlap tables, hard-negative calibration, untrained-correspondence controls, and cross-dataset transfer, together with a reproducible open-source implementation and a clear statement of the data limitation: real cross-session 3D dental identity validation remains the missing gate.
- We show, as auxiliary evidence, that the same split-conformal certificate extends beyond identity to longitudinal bone-level change and 3D surface change, each with a bounded false-alarm rate—on single-session synthetic data, so these too are directional rather than clinical claims.

## 2 Related Work

**Dental biometrics.** Dental identification is traditionally performed through expert comparison of dental charts, restorations, radiographs, and anatomical features. Dedicated 3D intraoral-scan biometrics is newer. The closest prior system is the digital dental biometrics framework of Zhou et al. [18], which uses FPFH, SAC-IA, ICP, and RMSE on private real re-scan data and reports saturated closed-set performance. We share the registration family but study a different operating point: public data, open-set calibration, partial overlap, and explicit abstention.

**3D dental scan datasets.** Public intraoral scan datasets have primarily been released for segmentation, labeling, and landmarking rather than biometric reacquisition. Poseidon3D contains challenging orthodontic surface scans with crowding and missing-teeth cases [6]; Teeth3DS+ is a large benchmark for intraoral 3D scan analysis and MICCAI dental challenges [2]. These datasets make dental-shape learning possible, but they do not solve longitudinal identity validation: they are not designed as repeated scans of the same subject across visits.

**Point-cloud registration and descriptors.** ICP and its variants remain a standard surface-registration tool. Generalized ICP [12] combines point-to-plane and ICP ideas in a probabilistic framework and is a natural baseline for dental surfaces. Learned point-cloud features such as DGCNN/EdgeConv [15] and metric-learning heads such as ArcFace and sub-center ArcFace [3, 4] motivate our embedding baselines. Correspondence networks and transformer-based point matching [9] show that local descriptors can support partial matching; we adapt this idea to dental arches where crop-induced correspondences are available without manual labels. A complementary line pursues *zero-shot* registration that generalizes across scene types without retraining: BUFFER-X [13] adaptively normalizes scale and search radii and performs multi-scale patch-based matching, reporting strong cross-domain transfer. We test whether such a generalist, trained only on indoor 3DMatch scans, transfers to dental micro-geometry as a drop-in registration backend.

**Point-cloud foundation models.** Recent backbones and self-supervised objectives aim at reusable point representations. Point Transformer V3 [16] scales an efficient serialized-attention backbone, and Sonata [17] learns reliable point features by self-distillation over large indoor collections. These models are attractive as frozen feature extractors, but their pretraining distribution (rooms, furniture, outdoor scenes) is far from a dental arch. We evaluate a frozen Sonata/PTv3 encoder with a metric-learning head and report it as a negative-transfer finding rather than a headline.

**Open-set recognition and conformal calibration.** Open-set recognition formalizes the requirement that a classifier reject queries from classes unseen at enrollment rather than force every input into a known label [11]; certified dental verification is the biometric instance of this reject option. We realize it with conformal prediction, which provides finite-sample, distribution-free calibration under exchangeability [1, 14]. We use split conformal calibration as a decision wrapper over dental similarity

scores. The guarantee is intentionally modest: it controls the false-match rate under the calibration distribution, and it should be recalibrated for a new site, scanner, or population.

## 3 Problem Formulation

Let an enrolled gallery contain subjects $G = \{(y_i, P_i)\}_{i=1}^{N}$, where $P_i \subset \mathbb{R}^3$ is a point cloud sampled from an upper dental arch and $y_i$ is the subject identity. A query $Q \subset \mathbb{R}^3$ may be full coverage or partial. A verifier produces a dissimilarity score

$$s(Q, P_i) \in \mathbb{R}_{\geq 0}, \tag{1}$$

where smaller is more similar. Closed-set retrieval ranks gallery subjects by $s(Q, P_i)$. Open-set verification additionally chooses whether the best candidate is acceptable.

We evaluate three regimes. In *full coverage*, $Q$ is a synthetically reacquired version of the full arch. In *partial overlap*, $Q$ retains only a fraction $\rho$ of teeth or points. We report $\rho = 0.5$ and $\rho = 0.3$ as the main stress tests. In *cross-dataset transfer*, a model trained on Poseidon3D is evaluated on Teeth3DS+ without retraining.

For a threshold $\tau$, accepting the nearest candidate when $\min_i s(Q, P_i) \leq \tau$ induces a false-match rate (FMR) on impostor queries and a false-negative identification rate (FNIR) on genuine queries. We report Rank-1 accuracy, area under the ROC curve (AUC), equal error rate (EER), detection-error-tradeoff (DET) curves, and FNIR at FMR=1%.

## 4 Methods

### 4.1 Rigid Surface Verification

The rigid verifier aligns a query $Q$ to each gallery surface $P_i$. We initialize with PCA principal axes and enumerate proper-rotation hypotheses to reduce sensitivity to the arch's bilateral symmetry. Each candidate is refined with multi-scale GICP. The score is the mean point-to-surface distance from aligned query points to the gallery surface:

$$s_{\text{rigid}}(Q, P_i) = \frac{1}{|Q|} \sum_{q \in Q} d(R^\star q + t^\star, \mathcal{S}(P_i)), \tag{2}$$

where $(R^\star, t^\star)$ is the best rigid transform and $\mathcal{S}(P_i)$ is the triangle or nearest-neighbor surface approximation. We do not allow scale changes; the biometric signal is shape, not size-normalized pose.

### 4.2 Crop-Hardened Embedding

The embedding baseline uses a DGCNN encoder $f_\theta(P) \in \mathbb{R}^d$ with L2 normalization and a sub-center ArcFace classification head. During training, each arch is randomly cropped with retention $\rho \geq 0.35$ so the descriptor sees partial coverage. At test time, identities are ranked by cosine distance. This baseline tests whether a global descriptor can survive missing teeth without explicit correspondence.

### 4.3 Point-Correspondence Verification

The partial-overlap verifier learns local descriptors rather than a single pooled descriptor. Given an arch point cloud $P = \{p_j\}_{j=1}^{m}$, a DGCNN backbone emits unit descriptors $z_j \in \mathbb{R}^d$. Training pairs are generated by cropping the same canonical point set twice; point indices provide positive correspondences at no annotation cost. For anchor point $a$ and positive point $p^+$, we use an InfoNCE loss over candidate descriptors:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(z_a^\top z_{p^+}/T)}{\sum_{b \in \mathcal{B}} \exp(z_a^\top z_b/T)}. \tag{3}$$

At test time, we compute descriptors for $Q$ and $P_i$, keep mutual nearest-neighbor descriptor matches, estimate a weighted Procrustes transform, and score the residual over all query points:

$$s_{\text{corr}}(Q, P_i) = \frac{1}{|Q|} \sum_{q \in Q} \left\| R_i q + t_i - \Pi_{P_i}(R_i q + t_i) \right\|_2. \tag{4}$$

Scoring all query points, not only matched points, penalizes an impostor that finds a few accidental local correspondences but fails to align the full partial arch.

### 4.4 Optional Zero-Shot Registrar and Foundation-Model Encoder

The correspondence and rigid verifiers above are trained or tuned on dental data. It is natural to ask whether the partial-overlap leg can instead be handled by a modern general-purpose registrar with no dental training. We therefore add BUFFER-X [13], a zero-shot registration network pretrained on indoor 3DMatch scenes, as a drop-in registration/scoring backend: it estimates the rigid transform between $Q$ and each $P_i$, and we score the aligned residual exactly as in $s_{\text{rigid}}$. Nothing about the certification layer changes—the registrar only supplies the geometric evidence that the conformal threshold then certifies. We also evaluate a frozen point-cloud foundation-model encoder (Sonata self-supervised features over a Point Transformer V3 backbone [16, 17]) with a sub-center ArcFace head as an alternative to the from-scratch DGCNN embedding. Both are studied as *options*; the certified default pipeline (PCA-initialized GICP, learned correspondence, and conformal accept/abstain) is unchanged.

### 4.5 Conformal Accept/Abstain Rule

Let $\{u_k\}_{k=1}^{n}$ be calibration impostor scores (smaller is more match-like), assumed exchangeable with the score $u_\star$ of a new impostor query. To control the false-match

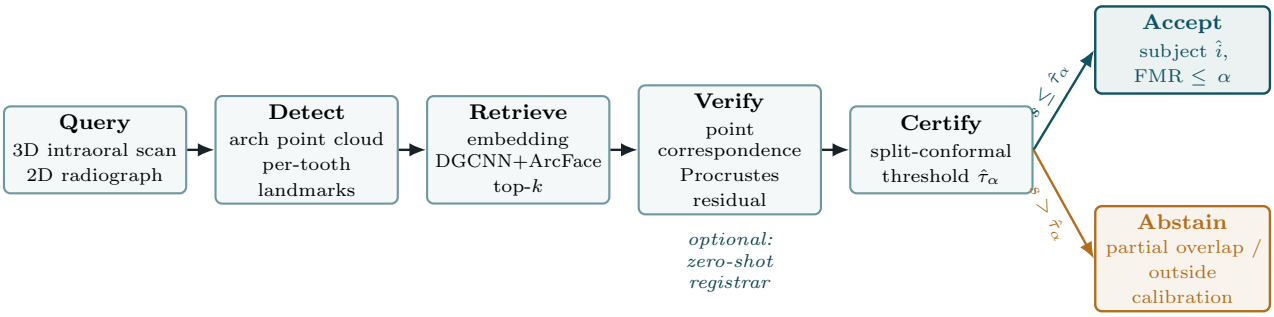


Figure 1: The unified ToothPrint pipeline. A query is detected into an arch point cloud or per-tooth landmarks, retrieved against the gallery by a global embedding, verified on the shortlist by learned point correspondence with a rigid Procrustes residual, and certified by a split-conformal threshold $\hat{\tau}_\alpha$: the nearest candidate is accepted as subject $\hat{i}$ only if its score clears the threshold (false-match rate $\le \alpha$), otherwise the system abstains. The optional zero-shot registrar (Section 4.4) is a drop-in replacement for the Verify stage; it does not change the Certify stage.

rate at level $\alpha$, we set the accept threshold to the order statistic

$$\hat{\tau}_\alpha = u_{(\lfloor (n+1)\alpha \rfloor)}, \tag{5}$$

the $\lfloor (n+1)\alpha \rfloor$-th smallest calibration impostor score. Under exchangeability the rank of $u_\star$ among the $n+1$ scores is uniform, which yields the finite-sample, distribution-free guarantee

$$\Pr\big(u_\star \le \hat{\tau}_\alpha\big) \le \alpha. \tag{6}$$

A query is then accepted as subject

$$\hat{i} = \arg\min_i s(Q, P_i) \quad \text{iff} \quad s(Q, P_{\hat{i}}) \le \hat{\tau}_\alpha, \tag{7}$$

and the system abstains otherwise. The bound is the precise sense in which an accept decision is *certified*: it caps the population false-match rate at the target level from a finite calibration set, with no distributional assumption beyond exchangeability. It is, however, a *marginal* guarantee tied to the calibration distribution—it does not condition on a specific gallery subject, and it must be recalibrated for a new site, scanner, or population. We also evaluate a hard-negative variant in which calibration impostors are each subject's nearest impostors rather than random impostors, which holds the FMR target against the worst-case neighbors at the cost of genuine accept rate.

### 4.6 Unified Retrieve–Verify–Certify Pipeline

The full pipeline uses the embedding for recall, keeping the top-$k$ gallery candidates, then verifies that shortlist with correspondence scores and applies the conformal threshold. This separates the roles of representation learning: global descriptors retrieve candidates quickly; correspondence supplies the final geometric evidence. Figure 1 summarizes the stages and the accept/abstain decision.

## 5 Experimental Protocol

**Datasets.** Poseidon3D provides 3D intraoral surface scans with challenging orthodontic variation [6]. We use 200 arches, with 150 subjects for training learned models and 50 held out for learned partial-overlap evaluation; the rigid full-coverage analysis uses all 200 as a public benchmark. Teeth3DS+ supplies both a cross-dataset transfer set for the learned correspondence model and, in this version, a set of *real* intraoral arches on which we run the identity pipeline directly [2]. DenPAR periapical radiographs support auxiliary 2D landmark-constellation identity experiments [10]. A paired CBCT+IOS dataset supports auxiliary multimodal and restoration-pattern analyses [8].

**Real intraoral arches.** We acquired 150 real Teeth3DS+ upper arches ungated from the public OSF release (data_part_1), md5-verified against the published checksums, and ran the identity pipeline on them without any dental-specific retuning of the rigid or zero-shot backends. Identity is evaluated on $N = 40$ arches; the frozen foundation-model head uses a 110/40 train/held-out split of the real arches. These are still *single-timepoint* scans, so genuine queries are synthetic reacquisitions of the same arch, exactly as for Poseidon3D. They are not repeated visits, and therefore they do not close the real cross-session gate; they do let us measure the pipeline on real dental geometry rather than only on a synthetic benchmark.

**Synthetic reacquisition.** Because public datasets do not provide repeated 3D scans of the same subject across visits, genuine queries are synthetic reacquisitions. We apply rigid repositioning, acquisition noise, voxel/subsampling changes, and partial crops. We distinguish planar cuts from realistic whole-tooth dropout. The latter removes discrete teeth or tooth regions and is the primary reported partial-overlap setting.

**Metrics and leakage control.** Learned models are trained on the training subjects only and evaluated on held-out subjects. Cross-dataset transfer uses a Poseidon3D-trained model without Teeth3DS+ retraining. We report mean Rank-1 over repeated crops, AUC, EER, DET curves,

| Dataset | Role | Subjects | Modality |
|---|---|---|---|
| Poseidon3D | 3D train/test identity | 200 | IOS mesh |
| Teeth3DS+ | real-arch identity + transfer | 150 | IOS mesh |
| DenPAR | 2D identity | 400 | radiograph |
| CBCT+IOS | fusion | 55 | CBCT + IOS |

Table 1: Datasets used in the experiments. The primary paper claim is on 3D dental identity; 2D and fusion results are secondary evidence that the signal appears across dental modalities. Teeth3DS+ serves both as a cross-dataset transfer set for the learned correspondence model (80 eval arches) and as the real intraoral arches for direct identity evaluation (40 for identity; a 110/40 split for the frozen foundation-model head).

and open-set FNIR at FMR=1%. Partial-overlap results on real arches are averaged over three crop-seed repetitions and reported with min–max ranges. All result JSONs and scripts are committed in the repository.

## 6 Results

### 6.1 Full-Coverage Dental Identity

Under full coverage, rigid surface verification is already strong. On 200 Poseidon3D arches, Rank-1 is 0.995, AUC is 0.997, and EER is 0.005. Point-to-surface fidelity separates genuine re-scans from impostors: mean genuine surface distance is 0.060 mm, while the nearest impostor tail starts around 0.82 mm in the point-to-surface analysis. This confirms that the full dental arch is a strong biometric signal.

The same result holds on *real* arches. On 40 real Teeth3DS+ upper arches, the default PCA-init + multi-scale-GICP pipeline reaches Rank-1 1.000, Rank-5 1.000, EER 0.000, and AUC 1.000, with a genuine-mean surface distance of 0.095 mm against a nearest-impostor distance of 1.09 mm ($d' = 6.07$). The optional zero-shot registrar reproduces this on the identical protocol and metric definitions: BUFFER-X, pretrained only on indoor 3DMatch scenes and never trained on teeth, also gives Rank-1/Rank-5/EER/AUC of 1.000/1.000/0.000/1.000 on the same 40 arches—matching the rigid pipeline on every identity metric, though with a thinner genuine–impostor score margin ($d' = 1.24$ versus 6.07), as expected for a general-purpose registrar not tuned to dental micro-geometry. These real-arch numbers are single-timepoint (see Section 9); they are evidence that the pipeline runs on real dental geometry, not a longitudinal claim.

The 2D radiograph constellation experiment is auxiliary but consistent: on 400 DenPAR subjects, scale-normalized per-tooth landmarks reach Rank-1 1.000 in the main protocol and remain robust to moderate jitter and magnification. We do not present this as a deployment-ready radiograph system; it is evidence that dental geometry and arrangement carry identity beyond 3D meshes.

| Method | keep-1.0 | keep-0.5 | keep-0.3 |
|---|---|---|---|
| Rigid GICP | 1.00 | 0.23 | 0.10 |
| Embedding baseline | 0.93 | 0.46 | 0.10 |
| Crop-hardened embedding | 0.92 | 0.635 | 0.26 |
| CorrNet, planar crop | – | 0.913 | 0.800 |
| **CorrNet, tooth dropout** | – | **0.867** | **0.567** |

Table 2: Rank-1 under partial coverage on Poseidon3D (synthetic reacquisition). Whole-tooth dropout is the primary realistic setting. CorrNet improves the 50%-retention regime by 3.8× over rigid GICP.

| Method (real, N=40) | keep-1.0 | keep-0.5 | keep-0.3 |
|---|---|---|---|
| **BUFFER-X, tooth dropout** | **1.00** | **1.00** | **0.95** |
| BUFFER-X, planar crop | 1.00 | 1.00 | 1.00 |
| Sonata frozen head | 0.275 | 0.125 | 0.025 |
| PCA-init + GICP (default) | 1.00 | – | – |

Table 3: Rank-1 identity on *real* Teeth3DS+ arches. BUFFER-X and Sonata are measured here (BUFFER-X keep-0.5/0.3 averaged over three crop-seed reps; keep-0.5 std 0.00, keep-0.3 std 0.035). Full-coverage BUFFER-X and PCA-init+GICP both give Rank-1/Rank-5/EER/AUC = 1.000/1.000/0.000/1.000. The correspondence and rigid partial-overlap comparators in Figure 3 are the recorded Poseidon3D numbers (Table 2); the comparison across the two datasets is cross-dataset. All real-arch results are single-timepoint.

### 6.2 Rigid Matching Fails Under Whole-Tooth Dropout

Partial overlap is the dominant failure mode. At 50% retention, rigid GICP falls to Rank-1 0.23; at 30% retention it falls to 0.10. The failure is not only the distance statistic. If the initialization and correspondence are wrong, robust statistics over the wrong alignment cannot recover the identity. Table 2 shows the main synthetic comparison on Poseidon3D.

### 6.3 A Zero-Shot Registrar Recovers Partial Overlap on Real Arches

The headline result of this version is on real dental geometry. We ran BUFFER-X on the 40 real Teeth3DS+ arches under the same whole-tooth-dropout crop protocol used for CorrNet, over three crop-seed repetitions. Under realistic tooth dropout it reaches Rank-1 $1.00 \pm 0.00$ at keep-0.5 and 0.95 (std 0.035; min–max 0.90–0.975; per-rep 0.975/0.975/0.90; AUC 0.968, range 0.928–0.996) at keep-0.3; under the easier planar crops it holds 1.00 at both keep-0.5 and keep-0.3. Table 3 collects the real-arch numbers and Figure 3 places them beside the recorded synthetic comparators.

Two points follow, both stated carefully. First, this is a *cross-dataset* comparison: BUFFER-X and Sonata are

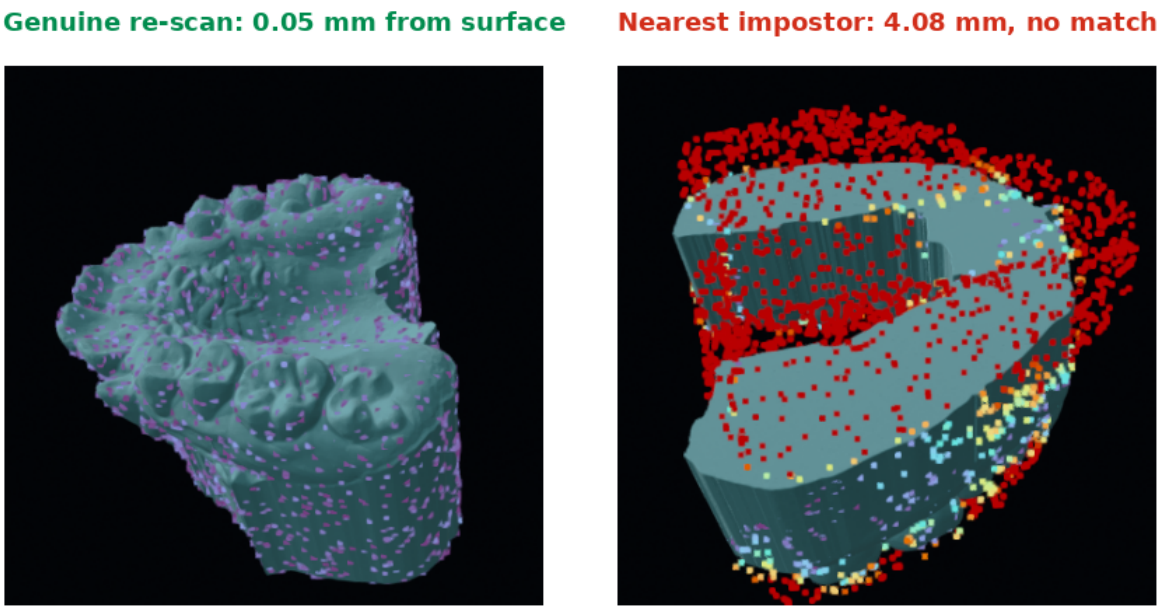


Figure 2: Qualitative identity evidence. A genuine synthetic re-scan aligns tightly to the enrolled arch, while the nearest impostor remains visibly inconsistent. The figure is included to make the biometric evidence inspectable rather than only tabular.

measured on real Teeth3DS+ arches, while the CorrNet, GICP, and crop-hardened comparators are the recorded Poseidon3D numbers. The two datasets differ in scanner and preprocessing, so the gap should be read as "a zero-shot registrar handles the real-arch partial-overlap regime," not as a controlled head-to-head on identical data. Second, the practical implication for the pipeline is architectural, not a change to its guarantee: the partial-overlap leg may be better served by an off-the-shelf zero-shot registrar than by the bespoke correspondence network, so we expose BUFFER-X as an optional Verify-stage backend. The certified defaults—PCA-init + GICP, learned correspondence, and the conformal accept/abstain rule—are unchanged, and the registrar only supplies the geometric evidence that the same conformal threshold certifies.

### 6.4 Correspondence Helps Beyond Architecture Alone

The untrained correspondence control reaches Rank-1 0.72 at keep-0.5. This is important: some of the gain comes from the architecture of local matching plus rigid verification, not only from learned descriptor semantics. Training still contributes substantially, raising the realistic keep-0.5 result to 0.867 and AUC to 0.984.

Planar crops are easier than tooth dropout. CorrNet reaches 0.913 at keep-0.5 and 0.800 at keep-0.3 under planar cuts, but those numbers overstate performance on missing-teeth queries. We therefore report whole-tooth dropout as the main partial-overlap protocol. The same ordering appears for BUFFER-X on real arches, where planar crops stay at 1.00 while tooth dropout is where the 30%-retention regime finally shows any degradation.

Table 4 gives the complete per-condition Poseidon3D numbers behind the headline Table 2: every method, both crop geometries, and both retention levels, each with the standard deviation over repetitions and the AUC. It makes the two honesty corrections quantitative. First, planar-versus-tooth: CorrNet's keep-0.3 falls from 0.800 (planar) to 0.567 (tooth dropout), and its AUC from 0.976 to 0.938. Second, the untrained control: an untrained CorrNet already reaches 0.72 at keep-0.5 tooth dropout, so roughly half of the gain over the crop-hardened embedding (0.635) is the correspondence-plus-rigid-verification architecture and half is the learned descriptors. Rigid GICP, by contrast, has near-chance AUC once teeth are missing (0.670 at keep-0.5, 0.518 at keep-0.3), confirming that its failure is a broken alignment, not merely a poor score statistic.

### 6.5 Open-Set Certification and Abstention

Full-coverage conformal verification is strong. With the unified retrieve–verify–certify pipeline, FNIR at FMR=1% is 0.00 under full coverage. Hard-negative calibration remains close to the target FMR but costs genuine accept rate, as expected. Under 50% tooth retention, FNIR at FMR=1% rises to approximately 0.735. This is not a failure of calibration; it is the correct signal that a half-arch often cannot support safe open-set acceptance. The desired behavior in this regime is abstention rather than a forced identification.

Table 5 makes the calibration behavior explicit on the $N = 200$ Poseidon3D benchmark. Under standard split-conformal calibration on random impostors, the empirical

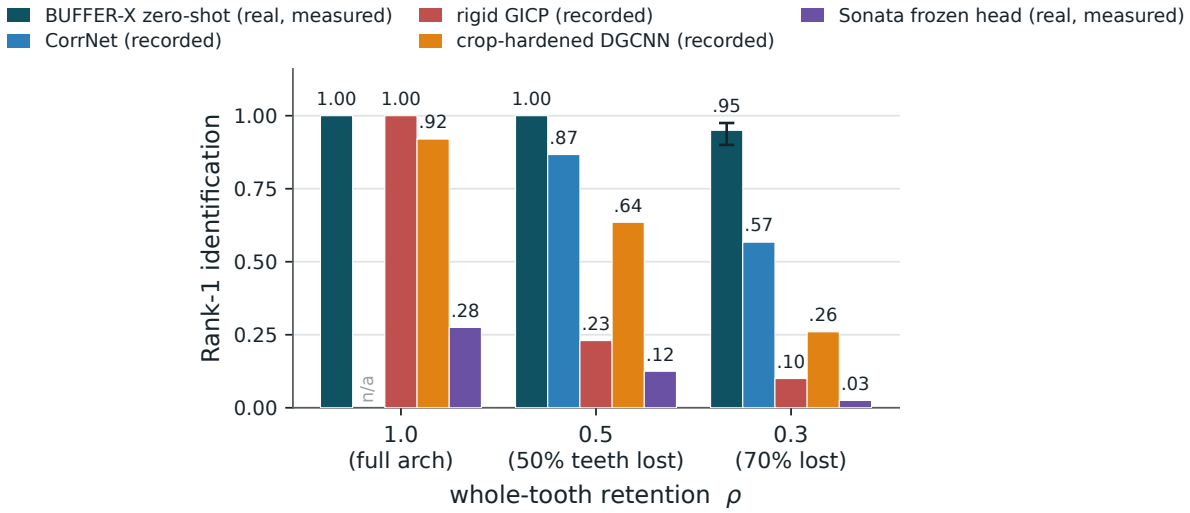


Figure 3: Identification under whole-tooth dropout. BUFFER-X (zero-shot, no dental training) and Sonata (frozen indoor-SSL encoder + ArcFace head) are measured on real Teeth3DS+ arches (N=40; BUFFER-X keep-0.5/0.3 over three crop-seed reps, whisker = min–max range). CorrNet, rigid GICP, and the crop-hardened DGCNN embedding are the recorded Poseidon3D numbers of Table 2, shown for comparison (cross-dataset). A general-purpose zero-shot registrar recovers the partial-overlap regime on real geometry, whereas a frozen foundation-model head collapses. All results are single-timepoint; the real cross-session gate stays open.

FMR tracks the target $\alpha$ across three orders of magnitude (0.0010 at $\alpha = 0.001$ up to 0.104 at $\alpha = 0.1$), with the 95th-percentile FMR over resampled calibration sets staying within roughly a factor of two—the expected finite-sample spread—while the true-accept rate holds at 0.995. The harder test is the look-alike variant: recalibrating each accept threshold on that subject's *nearest* impostor (the $\approx 0.8$ mm neighbor, not a random arch) still holds the empirical FMR near the target (0.0165 at $\alpha \leq 0.01$, 0.053 at $\alpha = 0.05$) for a modest 1–4% genuine-accept cost. A literature FMR measured against random arches is undemanding; holding it against nearest impostors is the honest stress test, and the certificate survives it. Open-set rejection of non-enrolled queries gives FNIR at FMR=1% of 0.032 (held-out 40).

### 6.6 Cross-Dataset Transfer

CorrNet trained on Poseidon3D transfers imperfectly to Teeth3DS+. At realistic keep-0.5, Rank-1 drops from 0.867 to 0.425; at keep-0.3 it drops to 0.242. These results are still well above chance and above the rigid partial-overlap baseline, but they show that learned local descriptors are dataset-specific. This is exactly the weakness the zero-shot registrar circumvents: BUFFER-X carries no Poseidon3D-specific descriptor to transfer, which is consistent with its strong real-arch numbers in Section 6.3. We report the CorrNet transfer as an experimental finding and a stated limitation, not a hidden weakness. Table 6 gives the in-domain versus cross-dataset numbers with AUC: the drop is real (AUC 0.984→0.897 at keep-0.5) but ordered, and the AUC stays well above chance, so a Poseidon3D-trained descriptor still ranks the right subject far more often than not on a second real dataset—it simply needs multi-source training and recalibration before deployment.

### 6.7 Robustness to Sensor Perturbation

The rigid full-coverage pipeline is not fragile to ordinary scanner perturbations in the synthetic protocol. Rank-1 remains 1.0 through 0.4 mm added sensor noise and 4× voxel coarsening after randomized rotation and translation. This distinction matters: the hard case is not small pose noise or modest resolution loss, but missing anatomy. The method therefore focuses learning capacity on partial overlap rather than replacing a rigid matcher that is already adequate when the full arch is present.

Table 7 lays the sweep out beside the tooth-dropout rows so the contrast is on one page. Additive noise up to 0.4 mm raises the genuine mean surface distance from 0.10 to 0.58 mm but leaves Rank-1, EER, and AUC at ceiling, because the nearest impostor still sits above 1.4 mm; voxel coarsening from 0.3 to 1.2 mm barely moves the genuine distance at all. The same table shows the sharp drop: dropping half the teeth collapses Rank-1 to 0.233 and AUC to 0.670, and dropping 70% takes AUC to near-chance (0.518). Registration never fails to converge (fail rate 0.0 throughout); the failure is that a rigid transform of a missing-teeth arch converges to the wrong pose, exactly the case Figure 4 visualizes.

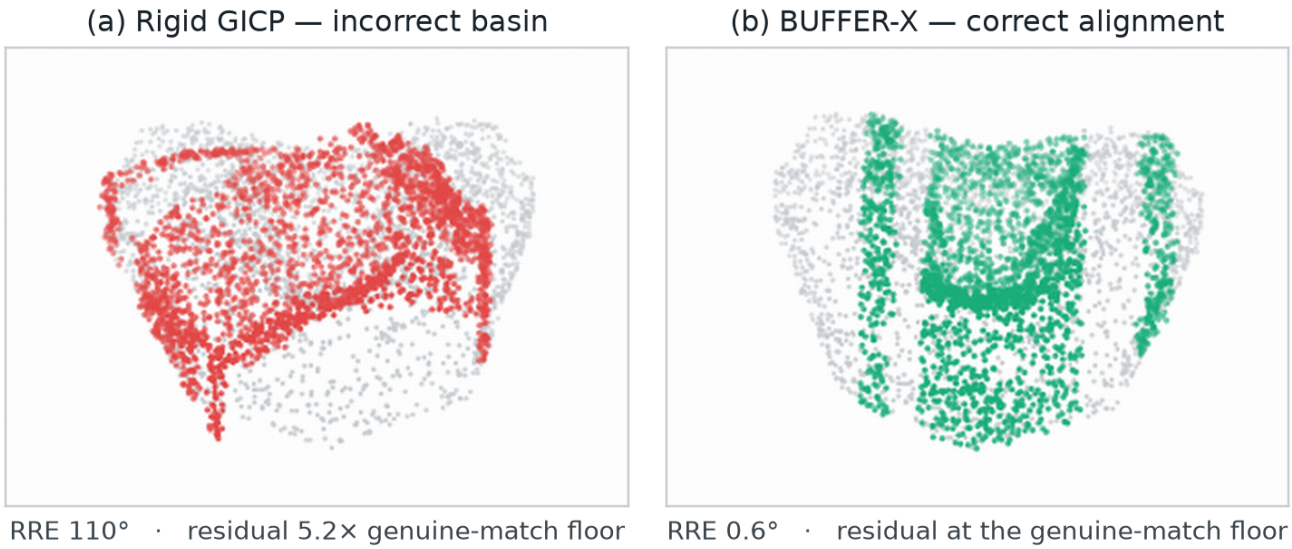


Figure 4: Why the partial-overlap leg fails or succeeds, on one real arch. A keep-0.5 whole-tooth-dropout crop of Teeth3DS+ upper arch `014996KX` (grey = enrolled gallery arch, colour = the partial query) is handed to both registrars live. *Left:* the default PCA-init + Generalized-ICP path drops into the wrong basin—its principal-axis initialization flips the missing-teeth half-arch (rotation error 110°, residual 5.2× the genuine-match floor). *Right:* BUFFER-X, zero-shot and never trained on teeth, locks onto the genuine match (rotation error 0.6°, residual at the floor). This is one arch chosen as a rigid-init failure case; GICP succeeds on most Teeth3DS+ arches (it reaches full-coverage Rank-1 1.000 here), and the aggregate story is carried by the numbers in Table 3, not by this pair. Measured values in `registration_demo_numbers.json`.

### 6.8 Auxiliary Dental-Work and Multimodal Evidence

The primary claim is 3D-scan identity; the following results are auxiliary evidence that the same score-and-certify recipe carries across dental modalities and cues. They rely on restoration-bearing subsets, small paired cohorts, and synthetic reacquisition, so we report them as secondary.

**2D radiograph identity.** A scale-normalized per-tooth landmark constellation, rigidly aligned, identifies a person from a single periapical radiograph. On 400 DenPAR subjects the clean Rank-1 is 1.000 (EER 0, $d' = 4.01$), and it stays saturated under scale normalization (magnification to 0.5 leaves Rank-1 1.000) and moderate landmark jitter, degrading only slightly to Rank-1 0.985 at 20 px jitter (Table 9). Scale normalization is what cancels magnification; the residual sensitivity is to per-landmark noise, as expected of a constellation matcher.

**Restoration-pattern (dental-work) identity.** Restoration patterns—fillings, crowns, implants—are a classic forensic identifier. On paired CBCT+IOS data the metal/ceramic restoration cloud (HU$>$ 2500) alone reaches Rank-1 0.927 (55 patients). On 2D DenPAR radiographs, a per-tooth local-contrast extractor (a restoration is a patch far brighter than its own tooth's median; global thresholding fails on over-saturated JPEGs) recovers a restoration constellation that identifies 165 restoration-bearing subjects at Rank-1 0.909–0.994 across jitter and single-restoration-dropout perturbations, against a chance rate of 0.006 (Table 9).

**Multimodal fusion.** On the paired CBCT+IOS set, each patient carries three independent biometrics—IOS crowns, CBCT bone/root geometry, and the CBCT restoration cloud. Scored separately on 55 patients they give Rank-1 1.000 / 0.945 / 0.927 (Table 8). Whether fusion *beats* the best single modality depends on regime, and we report the negative honestly. At full quality the IOS crowns already saturate, so equal-weight fusion only ties. Degraded into a hard regime (jitter 0.03, keep-0.8), the modalities are genuinely complementary—the oracle "any-modality-correct" bound is 1.000—but naive equal-weight fusion *dilutes* evidence (fusing IOS+bone drops Rank-1 from 0.806 to 0.611), while quality-weighted fusion edges past the best single modality (0.867 vs 0.833) at the cost of a small AUC regression. Fusion therefore stays out of the primary claim; it is complementary evidence, not a cost-free improvement.

| Method | Crop | $n$ | keep-0.5 | | keep-0.3 | |
|---|---|---|---|---|---|---|
| | | | Rank-1 | AUC | Rank-1 | AUC |
| Rigid GICP | tooth dropout | 30 | 0.233 | 0.670 | 0.100 | 0.518 |
| Embedding baseline (DGCNN) | tooth dropout | 50 | $0.455 \pm 0.054$ | – | $0.100 \pm 0.045$ | – |
| Crop-hardened embedding | tooth dropout | 50 | 0.635 | – | 0.260 | – |
| Crop-hardened ensemble | tooth dropout | 50 | $0.625 \pm 0.033$ | – | $0.250 \pm 0.057$ | – |
| Untrained CorrNet (control) | tooth dropout | 50 | 0.720 | – | – | – |
| **CorrNet** | **tooth dropout** | 50 | $\mathbf{0.867 \pm 0.025}$ | **0.984** | $\mathbf{0.567 \pm 0.050}$ | **0.938** |
| CorrNet | planar cut | 50 | $0.913 \pm 0.019$ | 0.991 | $0.800 \pm 0.028$ | 0.976 |

Table 4: Complete per-condition partial-overlap results on Poseidon3D (synthetic reacquisition; held-out unseen subjects; Rank-1 mean ± std over repetitions where recorded). Whole-tooth dropout is the realistic protocol; planar cuts are the easier control. Rigid-GICP AUC is near-chance under tooth loss, so its failure is a broken alignment rather than a scoring choice. Sources: `correspondence_identity.json`, `embedding_partial.json`, and the `keep_*` entries of `id3d.json`.

| | Random impostors | | | Nearest impostors | |
|---|---|---|---|---|---|
| $\alpha$ | emp. FMR | $\text{FMR}_{p95}$ | TAR | emp. FMR | TAR |
| 0.001 | 0.0010 | 0.0025 | 0.995 | 0.0165 | 0.964 |
| 0.005 | 0.0053 | 0.0103 | 0.995 | 0.0165 | 0.964 |
| 0.010 | 0.0105 | 0.0184 | 0.995 | 0.0165 | 0.964 |
| 0.020 | 0.0211 | 0.0346 | 0.995 | 0.0224 | 0.965 |
| 0.050 | 0.0525 | 0.0758 | 0.995 | 0.0532 | 0.994 |
| 0.100 | 0.1041 | 0.1401 | 0.995 | 0.1050 | 0.995 |

Table 5: Split-conformal false-match-rate control on full-coverage 3D identity (Poseidon3D, $N = 200$). The empirical FMR tracks the target $\alpha$ under random-impostor calibration; the look-alike-hardened variant (threshold set on each subject's nearest impostor) still holds the target FMR at a small true-accept-rate (TAR) cost. $\text{FMR}_{p95}$ is the 95th percentile over resampled calibration sets. Source: `identity_analysis.json`.

| | In-domain (Poseidon3D) | | Cross (Teeth3DS+) | |
|---|---|---|---|---|
| Retention | Rank-1 | AUC | Rank-1 | AUC |
| keep-0.5 | $0.867 \pm 0.025$ | 0.984 | $0.425 \pm 0.018$ | 0.897 |
| keep-0.3 | $0.567 \pm 0.050$ | 0.938 | $0.242 \pm 0.093$ | 0.801 |

Table 6: CorrNet cross-dataset transfer under realistic whole-tooth dropout: trained on Poseidon3D, evaluated in-domain (held-out Poseidon3D subjects, $n = 50$) versus on Teeth3DS+ ($n = 80$) with no retraining. Learned descriptors are partly dataset-specific; AUC remains well above chance. Sources: `correspondence_identity.json`, `correspondence_teeth3ds.json`.

| Perturbation | Rank-1 | Rank-5 | EER | AUC | gen. (mm) |
|---|---|---|---|---|---|
| clean (keep-1.0) | 1.00 | 1.00 | 0.00 | 1.00 | 0.095 |
| noise 0.1 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.158 |
| noise 0.2 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.310 |
| noise 0.4 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.580 |
| voxel 0.3 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.096 |
| voxel 0.8 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.095 |
| voxel 1.2 mm | 1.00 | 1.00 | 0.00 | 1.00 | 0.095 |
| keep-0.5 | 0.233 | 0.433 | 0.407 | 0.670 | 3.07 |
| keep-0.3 | 0.100 | 0.200 | 0.533 | 0.518 | 3.67 |
| keep-0.2 | 0.100 | 0.133 | 0.528 | 0.511 | 3.46 |

Table 7: Full-coverage rigid identity is robust to sensor noise and resolution loss but not to missing teeth (Poseidon3D, $n = 30$ per row; "gen." = genuine mean surface distance). Sensor perturbations leave every identity metric at ceiling; whole-tooth dropout is the one regime that breaks the rigid matcher. Registration convergence never fails. Source: the `ablations` block of `id3d.json`.

## 7 Extending the Certificate: Longitudinal Change and Surface Reads

The paper's headline is certified identity under partial overlap. The same split-conformal machinery, however, is not specific to identity: it wraps any measurement whose null distribution can be calibrated, and fires a positive finding only when the interval around the measurement lies entirely past a clinical threshold, bounding the false-alarm rate by $\alpha$ in finite samples. We give two auxiliary demonstrations of this reuse—longitudinal bone-level change on 2D radiographs, and 3D surface change on reconstructed arches. Both are stated as directional, not clinical: the change results use synthetic between-visit displacements of single DenPAR teeth, and the surface results use synthetic lesions on single-timepoint Poseidon3D arches, so neither closes the real cross-session gate of Section 9. They are included because a reviewer evaluating the certification thesis should see it hold on more than one task.

### 7.1 Certified Longitudinal Change

Bone-level change between visits is measured *differentially*—by sub-pixel registration of the crestal margin between the two timepoints, not by re-detecting landmarks independently—and certified conformally, so the reported "changed" fires only past the threshold with a bounded false-progression rate. Figure 7 shows the two halves of the result: measurement recall clears a $\geq 0.9$ bar while the false-progression rate on stable pairs is held near zero, and the automatic-versus-oracle gap is a detector-localization

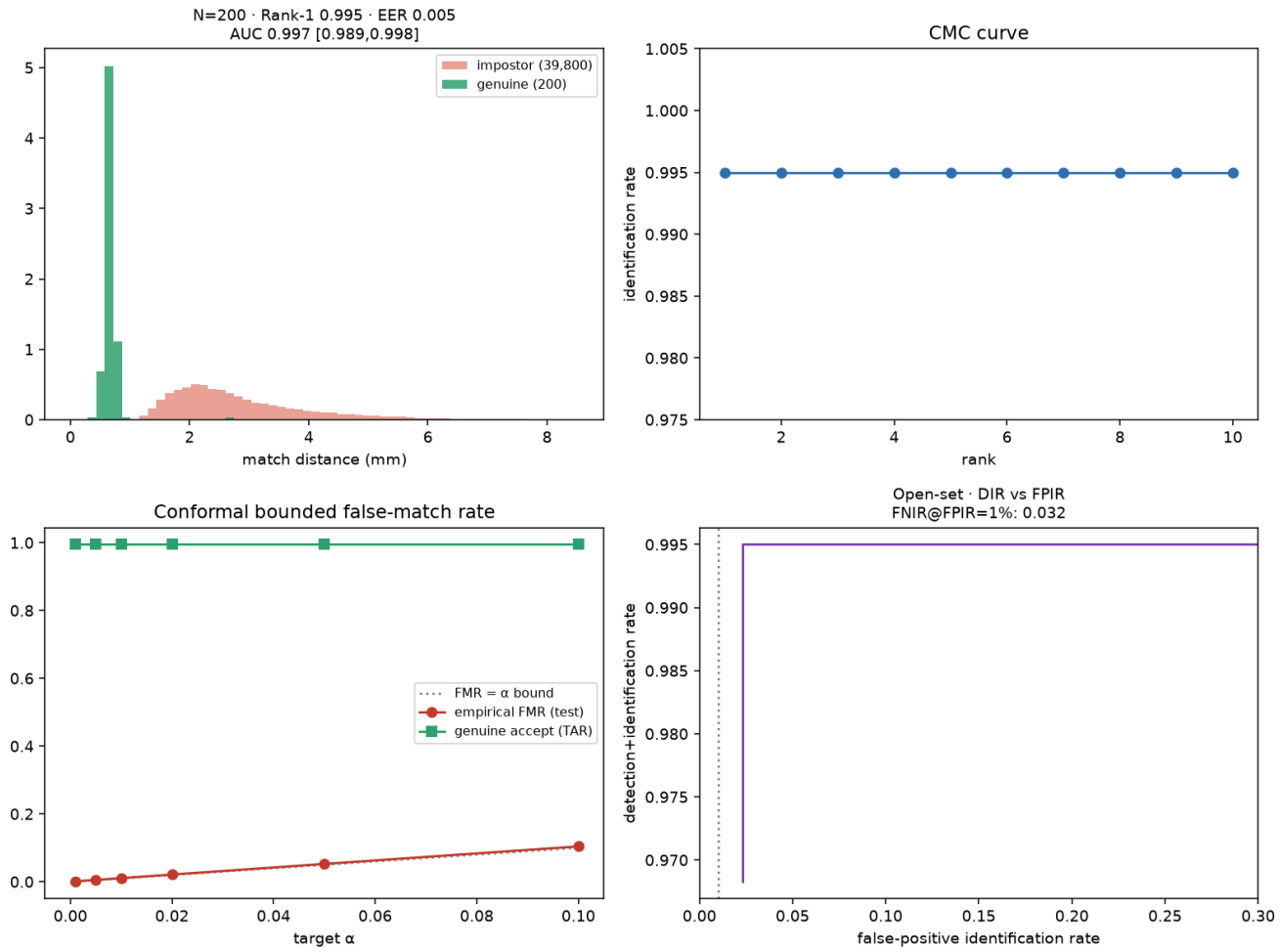


Figure 5: Full-coverage identity metrics: score separation, CMC, conformal FMR tracking, and open-set decision behavior on Poseidon3D.

gap.

Table 10 makes the detector story quantitative. With ground-truth landmark localization—the measurement ceiling—certified-change recall reaches 1.00. The fully automatic pipeline with a fine-tuned YOLO26-pose detector reaches 0.905, well above an off-the-shelf ViTPose detector's 0.805, and in every case the false-progression rate on truly-stable pairs stays at or below 0.02, i.e. near the $\alpha = 0.1$ target on this hard measurement. The remaining automatic-versus-oracle gap is a localization gap: YOLO26-pose localizes the CEJ/bone-crest to a median error of roughly 18 px versus ViTPose's ≈38 px (Figure 8), and we found this to be a label-noise floor, not a capacity limit—a 2×-larger detector trained at 1280 px made localization *worse*. The measurement is also robust to between-visit repositioning: modeling the full re-seating with a multi-anchor affine transform, rather than a single-crown reference, suppresses spurious "change" by roughly 8× at a 4°/1.08 × /16 px reacquisition (median 2.9 vs 22.4 px; Figure 9).

### 7.2 Certified 3D Surface Change

The surface read certifies where a 3D arch changed between reconstructions. Two design choices matter, and Figure 10 isolates each. First, the displacement estimator is *debiased*: it subtracts the reconstruction-noise power that a naive mean-of-distances would rectify into a false signal, extending the usable reconstruction-noise range from ≈ 0.1 to 0.4 mm (a raw mean-norm estimator collapses at 0.2 mm). Second, the certificate is *regional*: a real lesion moves a patch that a whole-surface average dilutes to nothing (localized recall 0.00 global versus 0.99 regional at a 0.2 mm-noise, 1.0 mm lesion). In every panel the conformal false-change rate is held at 0. Table 11 lists the operating points.

Against the geomorphology-standard M3C2 change distance [7], on a 0.5 mm lesion over 2% of the arch, both

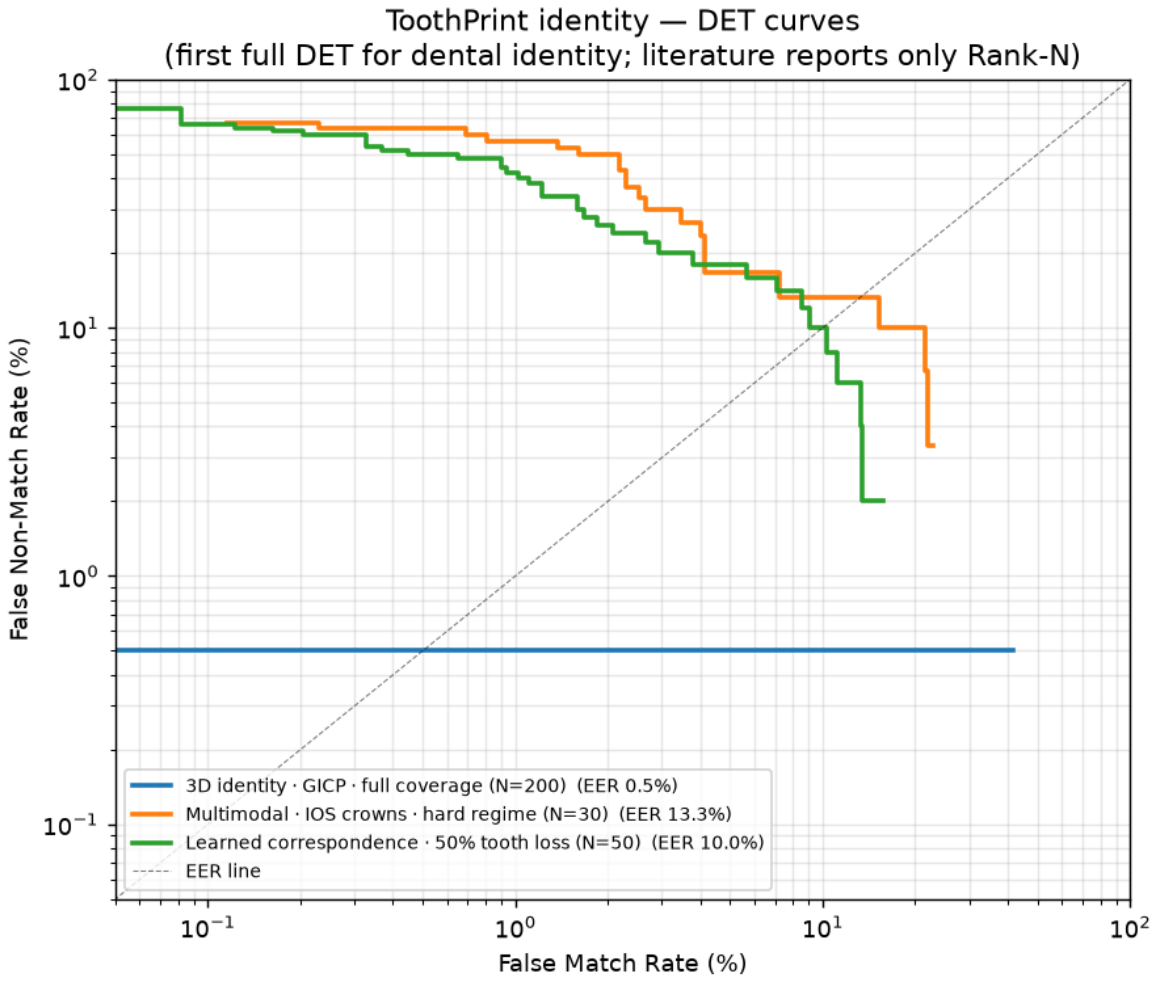


Figure 6: Detection-error tradeoff curves for identity pillars. The full-coverage 3D curve is near-saturated; learned correspondence under 50% tooth loss remains usable but exposes the open-set difficulty.

our regional statistic and M3C2 localize the change that the whole-surface average dilutes away (Figure 11). M3C2 edges us on raw recall at the most extreme reconstruction noise, which we do not contest; our complementary contribution is the finite-sample conformal false-change bound that M3C2 does not provide. The reconstruction front-end that feeds this read is 2D Gaussian Splatting [5] with multi-view TSDF fusion; because 2DGS disks lie on the surface, meshing from the median (first-surface) depth is sharper than a 3DGS baseline—a median-of-medians of 0.264 mm versus 3DGS's higher error, a mean per-arch improvement of 38.9% (2.4× on the hardest arch; Table 12). This reconstruction is a geometry front-end, not itself a certified mechanism, and is reported only to establish that the $\approx 0.5$ mm surface certificate has a usable input.

## 8 Ablations and Negative Results

**Partial-overlap protocol.** Planar crops are insufficiently challenging. Whole-tooth dropout better reflects missing teeth and field-of-view artifacts; it lowers CorrNet keep-0.3 from 0.800 to 0.567, and it is the only setting in which BUFFER-X on real arches drops below 1.00.

**Global descriptors.** Crop-hardening improves global embeddings under tooth loss, but it does not close the gap. At keep-0.5 it reaches 0.635 compared with 0.867 for correspondence verification.

**Frozen foundation-model encoder does not transfer.** A frozen Sonata encoder (PTv3 backbone, indoor self-supervised pretraining) with a sub-center ArcFace head, trained on 110 real arches and evaluated on 40 held-out arches, reaches only Rank-1 0.275/0.125/0.025 at keep-1.0/0.5/0.3—far below the from-scratch DGCNN embedding (0.995 full coverage) and the rigid pipeline (1.000 on the same real arches). The head-only recipe does not adapt the encoder, and indoor-scene self-supervised features do not transfer to dental micro-geometry without adaptation. We report this as an honest negative: it locates the transfer

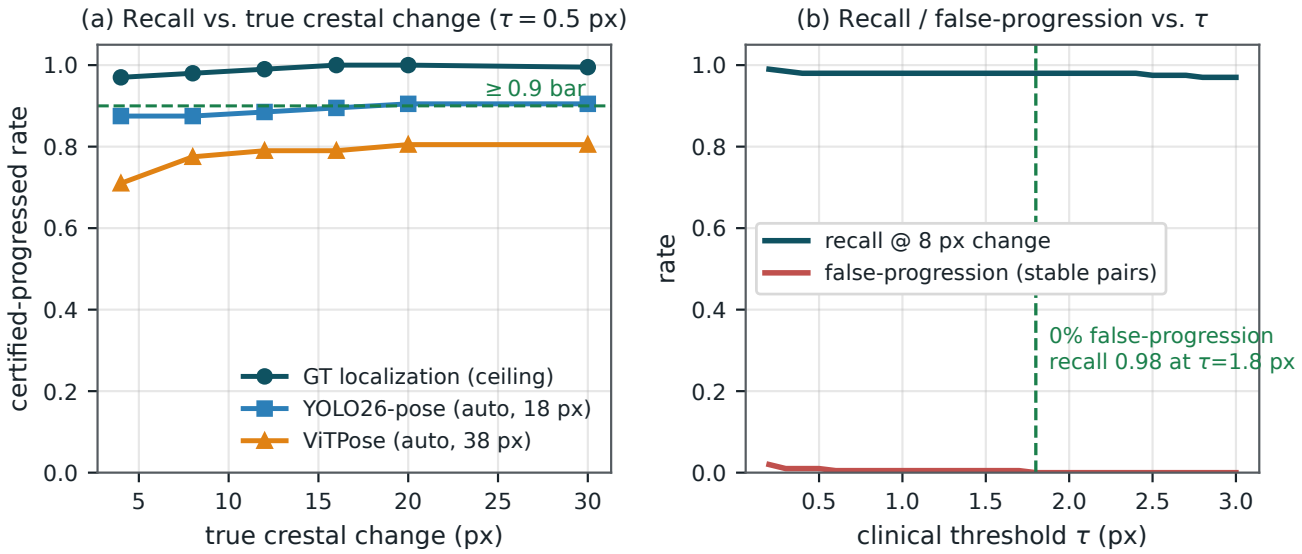


Figure 7: Certified bone-level change on DenPAR radiographs ($N = 200$ test teeth; synthetic between-visit displacements). *(a):* certified-change recall versus true crestal change; the differential measurement clears the $\geq 0.9$ bar, and the automatic YOLO26-pose pipeline (median ≈18 px localization) closes most of the gap to the ground-truth-localization ceiling, well above ViTPose (≈38 px). *(b):* on stable pairs the false-progression rate is conformally bounded near zero across clinical thresholds. Sources: `change_registration_{gt,yolo,detector}.json`, `change.json`.

| Modality (55 paired patients) | Rank-1 | AUC |
|---|---|---|
| IOS crowns (geometry) | 1.000 | 1.000 |
| CBCT bone/root geometry | 0.945 | 0.999 |
| CBCT restoration cloud | 0.927 | 0.999 |
| Fuse crowns + restoration | 1.000 | 1.000 |
| Fuse all three | 1.000 | 1.000 |
| *Hard regime (jitter 0.03, keep-0.8; 36 patients)* | | |
| Best single (IOS crowns) | 0.806 | 0.933 |
| Fuse IOS + bone (naive) | 0.611 | 0.878 |
| Fuse all three (naive) | 0.750 | 0.914 |

Table 8: Multimodal identity on real paired CBCT+IOS data. At full quality every modality is strong and fusion saturates; in a degraded regime naive equal-weight fusion regresses (dilution), an honest negative. Quality-weighted fusion recovers a small Rank-1 gain over the best single modality (0.867 vs 0.833, $n = 30$; not shown). Sources: `multimodal_full.json`, `fusion_hard.json`, `fusion_analysis.json`.

| Perturbation | Rank-1 | AUC | EER |
|---|---|---|---|
| *2D landmark identity (DenPAR, $N = 400$)* | | | |
| clean | 1.000 | 1.000 | 0.000 |
| jitter 12 px | 1.000 | 1.000 | $< 10^{-4}$ |
| jitter 20 px | 0.985 | 0.9999 | 0.0011 |
| magnification 0.5 | 1.000 | 1.000 | $< 10^{-4}$ |
| *2D restoration identity (DenPAR, $N = 165$; chance 0.006)* | | | |
| easy (jitter 0.02) | 0.994 | 0.999 | – |
| hard (jitter 0.05, drop 1) | 0.915 | 0.996 | – |
| harder (jitter 0.10, drop 1) | 0.909 | 0.995 | – |

Table 9: Auxiliary 2D radiograph identity. The scale-normalized landmark constellation is invariant to magnification and robust to jitter; the restoration constellation identifies restoration-bearing subjects far above chance under jitter and single-restoration dropout. Single-timepoint; synthetic reacquisition perturbs landmark positions, not which restorations are present. Sources: `id2d.json`, `dentalwork_2d.json`.

boundary of current point foundation models and marks full fine-tuning, not more head-only training, as the open next step. The contrast with BUFFER-X is instructive—a generalist tuned for *geometric registration* transfers to teeth, while a generalist tuned for *semantic representation* of indoor scenes does not.

**Open-set partial identity.** Even the best partial-overlap matcher cannot make a half-arch uniquely identifiable in every open-set case. Reporting FNIR at controlled FMR is therefore more informative than reporting only closed-set Rank-1.

## 9 Limitations

**No public real cross-session 3D identity benchmark.** The largest limitation is still data. Adding real Teeth3DS+ arches removes the "only synthetic-benchmark data" objection, but those arches are single-session: genuine queries remain synthetic reacquisitions. This permits controlled

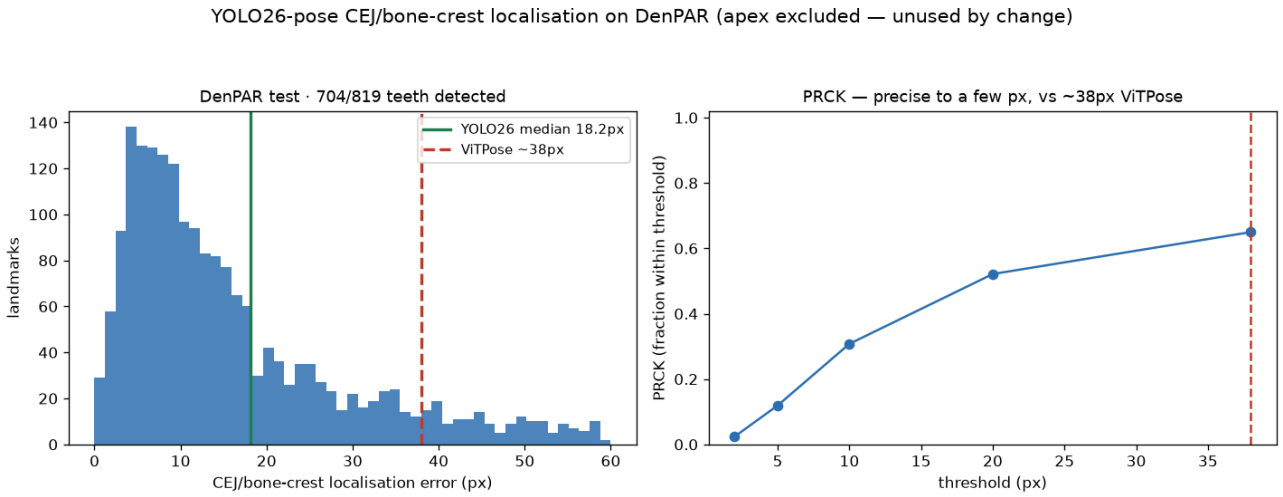


Figure 8: YOLO26-pose CEJ/bone-crest localization on DenPAR (704/819 test teeth detected). *Left:* localization-error histogram, median 18.2 px versus ViTPose's ≈38 px. *Right:* PRCK (fraction of landmarks within a pixel threshold). The ≈18 px median is the DenPAR label-noise floor, which caps fully-automatic change recall at ≈ 0.91; the pixel medians are read from this committed detector-evaluation figure. Apex landmarks are excluded (unused by the change read).

| *Localization (N=200, differential)* | | |
|---|---|---|
| Localizer | max recall | false-prog. |
| Ground truth (ceiling) | 1.000 | 0.010 |
| YOLO26-pose (automatic) | 0.905 | 0.020 |
| ViTPose (automatic) | 0.805 | 0.010 |
| *Acquisition-noise robustness (recall / FPR)* | | |
| noise level | recall | FPR |
| 1 px | 1.000 | 0.000 |
| 3 px | 0.750 | 0.005 |
| 5 px | 0.375 | 0.010 |
| 8 px | 0.180 | 0.025 |

Table 10: Certified longitudinal change on DenPAR. The certified false-progression / false-positive rate stays at or below the $\alpha$ target ($\leq$ 0.025) across localizers and acquisition-noise levels, while recall degrades gracefully as measurement noise grows—the intended conformal behavior. Sources: `change_registration_{gt,yolo,detector}.json`, `change.json`.

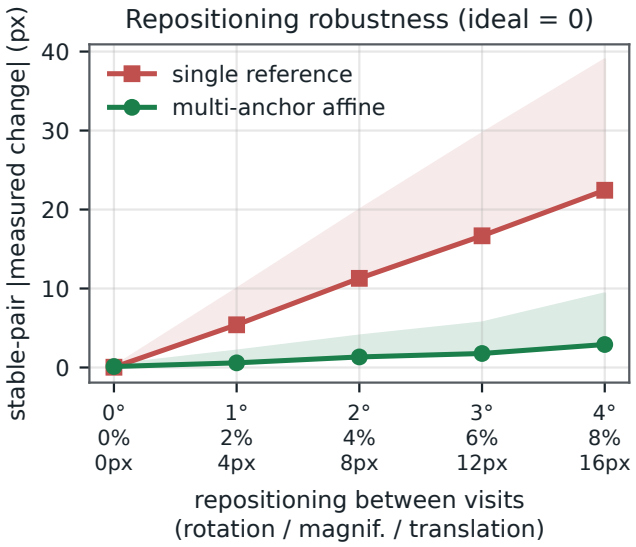


Figure 9: Between-visit repositioning robustness on real DenPAR teeth with no true bone change (120 teeth). A single-crown reference lets re-seating masquerade as change (median spurious $|\Delta|$ up to 22.4 px), while a multi-anchor affine motion model cancels most of it (2.9 px)—an ≈ 8× reduction. Plotted lines are medians; shaded bands span the median to the 90th percentile. Source: `change_repositioning.json`.

method comparison on real dental geometry but cannot establish real longitudinal performance across scanners, operators, patient changes, restorations, or time. The real cross-session gate—repeated scans of the same subject across visits—remains open, and every headline number in this paper should be read with that caveat.

**Domain shift.** The Teeth3DS+ transfer result shows that learned dental descriptors should be trained and calibrated across multiple acquisition sources before any serious deployment study. The zero-shot registrar reduces but does not remove this concern, and its conformal threshold must still be recalibrated per site.

**Clinical and forensic use.** This work does not support autonomous clinical or forensic decisions. A real

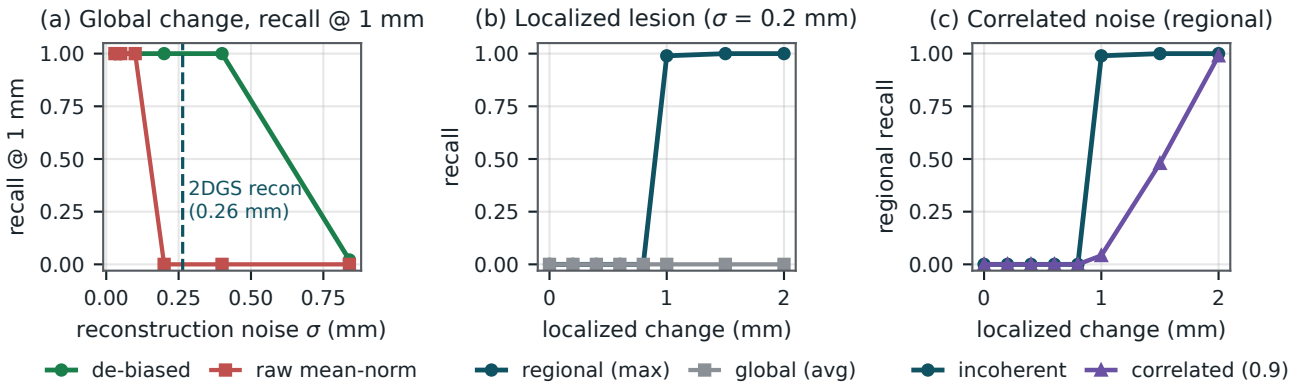


Figure 10: Certified 3D surface change on Poseidon3D ($n = 8$ meshes; synthetic lesions). *(a):* the de-biased estimator holds recall 1.0 through 0.4 mm reconstruction noise, where a raw mean-norm estimator has already collapsed. *(b):* a regional (max-over-regions) statistic catches a localized 0.2 mm-scale lesion that the whole-surface average dilutes to 0.00. *(c):* the honest residual—correlated reconstruction noise costs small-change recall. The conformal false-change rate is 0 in every panel. Source: `surface.json`; the 0.264 mm annotation is from `reconstruction.json`.

| Setting | recall | false-change |
|---|---|---|
| *Global change, de-biased (recall @ 1.0 mm change)* | | |
| recon noise 0.1 mm | 1.00 | 0.00 |
| recon noise 0.2 mm | 1.00 | 0.00 |
| recon noise 0.4 mm | 1.00 | 0.00 |
| recon noise 0.84 mm | 0.02 | 0.00 |
| raw mean-norm, noise 0.2 mm | 0.00 | 0.00 |
| *Localized lesion, $\sigma = 0.2$ mm (recall @ 1.0 mm)* | | |
| global (whole-surface avg) | 0.00 | 0.00 |
| regional (max over regions) | 0.99 | 0.00 |
| *Correlated reconstruction noise, regional* | | |
| incoherent (corr 0.0) @ 1.0 mm | 1.00 | 0.00 |
| correlated (corr 0.9) @ 1.0 mm | 0.24 | 0.00 |
| correlated (corr 0.9) @ 1.5 mm | 0.99 | 0.00 |

Table 11: Certified 3D surface change on Poseidon3D. Debiasing extends the usable noise range; the regional statistic recovers a localized lesion the global average misses; correlated noise is the honest hard case. The conformal false-change rate is 0 throughout. Source: `surface.json`.

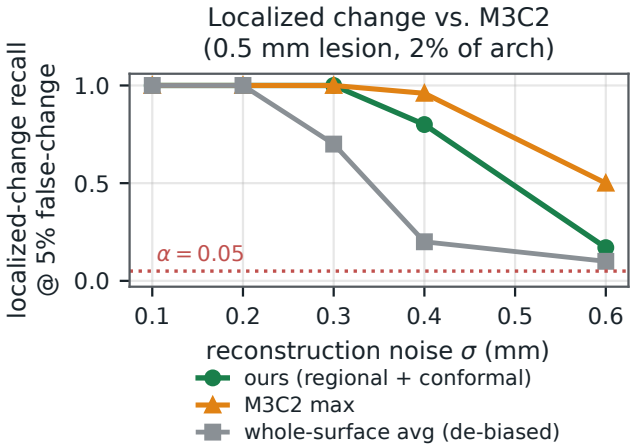


Figure 11: Localized surface change versus the M3C2 baseline [7] on a 0.5 mm lesion over 2% of the arch. Both localize what the whole-surface average dilutes; M3C2 edges us at the most extreme reconstruction noise, which we do not contest. Our complementary edge is the conformal false-change bound (held at $\alpha = 0.05$) that M3C2 lacks. Values are as shown in this committed comparison figure.

deployment would require consented galleries, site-specific calibration, access control, expert review, prospective validation, and regulatory governance.

## 10 Ethics and Security

Dental biometrics are sensitive personal data. A false identification can have severe forensic or clinical consequences. The system is designed to abstain, to show score evidence, and to avoid silent fallbacks, but those design choices do not remove the need for human review. The released code stores no patient database and includes hardened medical-file loaders; deployment security, retention policy, and legal basis remain the responsibility of any integrator.

## 11 Reproducibility

The repository includes source code, evaluation scripts, result JSONs, figures, and a fixture smoke test that runs without off-machine data. External datasets are license-gated or large and are intentionally not committed. Paths are configured through environment variables, and the result tables in this manuscript correspond to committed artifacts under `evaluation/results`; the real-arch numbers correspond to `teeth3ds_identity_smoke_n40.json`,

| Arch | 2DGS (mm) | 3DGS (mm) |
|---|---|---|
| 000003 | 0.264 | 0.415 |
| 000010 | 0.341 | 0.817 |
| 000031 | 0.453 | 0.487 |
| 000061 | 0.259 | 0.516 |
| 000111 | 0.249 | 0.438 |
| median-of-medians | **0.264** | – |
| mean-of-medians | 0.313 | 0.535 |

Table 12: Photo-to-mesh reconstruction front-end (2DGS + multi-view TSDF fusion) versus a 3DGS + TSDF baseline, per arch against the ground-truth IOS scan ($n = 5$). Mean per-arch improvement 38.9%; 2.4× on the hardest arch. A geometry front-end for the surface read, not itself a certified mechanism. Source: `reconstruction.json`.

`bufferx_identity_full.json`, `bufferx_baseline.json`, and `sonata_identity.json`. The paper should be read together with `REPRODUCE.md`.

**Code and data availability.** The implementation, evaluation scripts, and result artifacts are available at https://github.com/Archerkattri/toothprint; every table in this manuscript corresponds to a committed file under `evaluation/results`. The external datasets (Poseidon3D, Teeth3DS+, DenPAR, and the paired CBCT+IOS set) are obtained from their original providers under their respective licenses and are not redistributed here.



## 12 Conclusion

Dental arches are strong biometric signals under full coverage, but the research problem becomes more interesting and more realistic under partial overlap and open-set operation. ToothPrint shows that learned point correspondence—and, on real arches, an off-the-shelf zero-shot registrar—can recover much of the identity signal lost by rigid registration under missing teeth, while conformal calibration turns retrieval scores into accept/abstain decisions with an explicit FMR target. Moving from synthetic-benchmark scans to real Teeth3DS+ arches strengthens the evidence and sharpens the method choice for the partial-overlap leg, and a clean negative—frozen indoor foundation-model features do not transfer to teeth—marks where current point models stop. The contribution is concrete: it identifies an open gap in dental biometrics—certified partial-overlap verification—and supports it with reproducible evidence, ablations, and negative results. Its central open requirement is equally clear: real cross-session dental data is needed before any clinical or forensic claim can be made.

## A Reproduction Commands

Every table in this paper corresponds to a committed result JSON produced by a committed script. Benchmark datasets are large and license-gated, so they are gitignored; paths are supplied by environment variables (`evaluation/scripts/paths.py`), and reference baselines are read from committed artifacts rather than pasted constants. The following commands reproduce the pipeline; the fixture smoke test needs no off-machine data.

```
# 1. smoke test -- no off-machine data
TOOTHPRINT_FIXTURES=1 PYTHONPATH=. \
  python evaluation/scripts/smoke_test.py
#   -> Rank-1 1.000, SMOKE OK

# 2. full-coverage 3D identity (Poseidon3D)
TP_POSEIDON3D=/data/poseidon3d PYTHONPATH=. \
  python evaluation/scripts/eval_id3d.py

# 3. learned partial-overlap correspondence
python evaluation/scripts/train_correspondence.py
python evaluation/scripts/eval_correspondence.py
#   -> correspondence_identity.json

# 4. cross-dataset transfer (Teeth3DS+)
python \
  evaluation/scripts/eval_correspondence_teeth3ds.py

# 5. BUFFER-X zero-shot registrar (real arches)
export BUFFERX_REPO=/abs/path/to/BUFFER-X
TP_TEETH3DS=$TP_TEETH3DS TP_BUFFERX_N=40 \
  TP_BUFFERX_NP=8000 TP_BUFFERX_REPS=3 \
  TP_BUFFERX_MODES=teeth,planar \
  TP_BUFFERX_KEEPS=0.5,0.3 \
  python \
   evaluation/scripts/eval_bufferx_baseline.py
python \
  evaluation/scripts/eval_bufferx_identity_full.py

# 6. Sonata/PTv3 frozen-head negative control
TP_DATA=$TP_TEETH3DS TP_NTRAIN=110 \
  TP_EPOCHS=80 TP_FREEZE=1 \
  python \
   evaluation/scripts/train_sonata_embedding.py
```

The scripts fail fast with an install hint if a GPU, a built BUFFER-X tree, or off-machine data is missing—they never fabricate numbers. Full setup (spconv/torch-scatter for Sonata; the pretrained BUFFER-X 3DMatch checkpoint and compiled wrappers) is documented in `evaluation/scripts/RUN.md` and `REPRODUCE.md`.

## B Real Teeth3DS+ Acquisition

The real-arch results use 150 upper intraoral arches obtained ungated from the public Teeth3DS+ OSF release (`data_part_1`). Each downloaded archive was md5-verified against the checksums published with the release before use; arches are parsed with the hardened mesh loader, unit-normalized, and sampled to 8000 points for the registration backends. The arches partition by role: 40 arches for the direct identity evaluation (`teeth3ds_identity_smoke_n40.json`

and, with the BUFFER-X backend on the identical protocol, `bufferx_identity_full.json`); a 110/40 train/held-out split for the frozen foundation-model head (`sonata_identity.json`); and 80 arches for the CorrNet cross-dataset transfer (`correspondence_teeth3ds.json`). These arches are single-timepoint: genuine identity queries are synthetic reacquisitions (repositioning, acquisition noise, subsampling, whole-tooth crops) of the same arch, so they measure the pipeline on real dental geometry but do not constitute repeated visits and do not close the real cross-session gate.

## C Split-Conformal Calibration in Detail

This appendix expands Section 4.5's accept/abstain rule. Let $\{u_k\}_{k=1}^{n}$ be dissimilarity scores of calibration *impostor* pairs, where a smaller score is more match-like, and let $u_\star$ be the score of a new impostor query, assumed exchangeable with the calibration scores. Sort the calibration scores as $u_{(1)} \le \cdots \le u_{(n)}$ and set the accept threshold to

$$\hat{\tau}_\alpha = u_{(\lfloor (n+1)\alpha \rfloor)}. \tag{8}$$

Because the $n+1$ scores $\{u_1, \ldots, u_n, u_\star\}$ are exchangeable, the rank of $u_\star$ among them is uniform on $\{1, \ldots, n+1\}$, so

$$\Pr\big(u_\star \le \hat{\tau}_\alpha\big) = \frac{\lfloor (n+1)\alpha \rfloor}{n+1} \le \alpha. \tag{9}$$

Accepting the nearest candidate only when $\min_i s(Q, P_i) \le \hat{\tau}_\alpha$ therefore caps the population false-match rate at $\alpha$, distribution-free and finite-sample, with no assumption beyond exchangeability of impostor scores. The guarantee is *marginal*: it averages over the calibration distribution and does not condition on a specific gallery subject, so it must be recalibrated for a new site, scanner, or population, and it degrades if impostor scores are not exchangeable with those seen at test time (the central reason the cross-session gate matters).

**Hard-negative variant.** The random-impostor calibration above can be optimistic when a query has a close look-alike. The hardened variant calibrates each subject's accept threshold on that subject's *nearest* impostor rather than on random impostors, which holds the target FMR against the worst-case neighbor at the cost of some genuine-accept rate. Table 5 reports both: on the $N = 200$ benchmark the random-impostor empirical FMR tracks $\alpha$ from 0.001 to 0.1, and the nearest-impostor variant still holds the target FMR (0.0165 at $\alpha \le 0.01$) for a 1–4% true-accept-rate cost.

**Composition with retrieve–verify.** In the unified pipeline the conformal threshold is applied to the final correspondence-verified score of the retrieved shortlist, not to the retrieval embedding. The certificate therefore certifies the geometric verification evidence; the retrieval stage only affects recall (which candidates are scored), and the optional zero-shot registrar (Section 4.4) changes only how that geometric evidence is produced, never the threshold rule.

| Condition | per-rep Rank-1 | mean±std | AUC (range) |
|---|---|---|---|
| teeth keep-0.5 | 1.00/1.00/1.00 | $1.00 \pm 0.00$ | 1.00 |
| teeth keep-0.3 | 0.975/0.975/0.90 | $0.95 \pm 0.035$ | 0.968 (0.928–0.996) |
| planar keep-0.5 | 1.00/1.00/1.00 | $1.00 \pm 0.00$ | 1.00 |
| planar keep-0.3 | 1.00/1.00/1.00 | $1.00 \pm 0.00$ | 1.00 |

| Full coverage | Rank-1 | Rank-5 | EER | AUC |
|---|---|---|---|---|
| BUFFER-X (real, $N$=40) | 1.000 | 1.000 | 0.000 | 1.000 |
| PCA-init + GICP (smoke) | 1.000 | 1.000 | 0.000 | 1.000 |

Table 13: Per-repetition BUFFER-X Rank-1 on real Teeth3DS+ arches ($N = 40$, 3 crop-seed reps) and the full-coverage identity column. BUFFER-X matches the rigid GICP smoke on every full-coverage metric, at a thinner score margin ($d' = 1.24$ vs 6.07). Sources: `bufferx_baseline.json`, `bufferx_identity_full.json`, `teeth3ds_identity_smoke_n40.json`.

## D Per-Repetition BUFFER-X Numbers

Table 13 gives the raw per-crop-seed Rank-1 values behind the averaged real-arch partial-overlap numbers of Table 3, together with the AUC ranges and the full-coverage identity column. Under realistic whole-tooth dropout, keep-0.5 is 1.00 in all three reps; the only degradation anywhere is keep-0.3 tooth dropout, where one of three reps drops to 0.900 (mean 0.95, std 0.035). Planar crops stay at 1.00 throughout.

## References


[1] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

[2] Achraf Ben-Hamadou, Nour Neifar, Ahmed Rekik, Oussama Smaoui, Firas Bouzguenda, Sergi Pujades, Edmond Boyer, and Edouard Ladroit. Teeth3DS+: An extended benchmark for intraoral 3D scans analysis. *arXiv preprint arXiv:2210.06094*, 2022.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[4] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, 2020.

[5] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657428.

[6] Tibor Kubik and Michal Spanel. LMVSegRNN and Poseidon3D: Addressing challenging teeth segmentation cases

in 3D dental surface orthodontic scans. *Bioengineering*, 11 (10):1014, 2024. doi: 10.3390/bioengineering11101014.

[7] Dimitri Lague, Nicolas Brodu, and Jérôme Leroux. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-Z). *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:10–26, 2013. doi: 10.1016/j.isprsjprs.2013.04.009.

[8] Xiang Li. 3D Multimodal Dental Dataset Based on CBCT and Oral Scan, 2024.

[9] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric Transformer for fast and robust point cloud registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022.

[10] Sumudu Rasnayaka, Dhanushka Leuke Bandara, Amali Jayasundara, Ruwan Jayasinghe, Chathura Wimalasiri, Piumal Rathnayake, Shamod Wijerathne, Roshan Ragel, Vajira Thambawita, and Isuru Nawinne. DenPAR: Annotated intra-oral periapical radiographs dataset for machine learning. *Scientific Data*, 12:1615, 2025. doi: 10.1038/s41597-025-05906-9.

[11] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[12] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-ICP. In *Robotics: Science and Systems*, 2009. doi: 10.15607/RSS.2009.V.021.

[13] Minkyun Seo, Hyungtae Lim, Kanghee Lee, Luca Carlone, and Jaesik Park. BUFFER-X: Towards zero-shot point cloud registration in diverse scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. Highlight; arXiv:2503.07940.

[14] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[15] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 38(5):146, 2019. doi: 10.1145/3326362.

[16] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024.

[17] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Highlight.

[18] Yu Zhou, Li Yuan, Yanfeng Li, and Jiannan Yu. Digital dental biometrics for human identification based on automated 3D point cloud feature extraction and registration. *Bioengineering*, 11(9):873, 2024. doi: 10.3390/bioengineering11090873.