أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

# ANCHOR: Anatomy-Grounded Narrative Report Generation from Maxillofacial Cone-Beam CT

**Hussam Alsuhayman, Ahmed Bajafar, Anas Zafar,**
**Nawaf Almutairi, Saud Alhazmi, Dr. Salman Khan**

hussam.alsuhayman@kaust.edu.sa, ahmed.bajafar@kaust.edu.sa, anas.zafar@kaust.edu.sa,
nawaf.almutairi@kaust.edu.sa, saud.alhazmi@kaust.edu.sa, salmank255@gmail.com

## Abstract

Clinical reporting from cone-beam computed tomography (CBCT) is still written by hand, and the meaning of a dental finding depends on which tooth it is attributed to. We are building **Anchor**, a three-stage pipeline that generates narrative maxillofacial reports from CBCT volumes on the ToothFairy4 (ODIN 2026) corpus: a U-Mamba2 segmentation model recovers 46 anatomical structures, a deterministic extractor turns the label map into a typed findings object, and a language model verbalises that object into prose. The language model never sees the volume, so it cannot invent findings the geometry does not support. This report records the work completed so far. We verified the released corpus (622 cases, 1,000 English and 1,000 Italian reports, 367 cases read by at least two surgeons), aligned the ToothFairy4 volumes to the ToothFairy3 voxel labels bit-exactly, and found that the ToothFairy3 laterality convention is mirrored relative to the reports, which if left uncorrected would swap left and right in every report the system generates. We trained the winning ToothFairy3 segmentation recipe on our own case-level split and validated it on 106 held-out cases: mean Dice is 0.72 over all 46 classes and 0.88 restricted to cases where the structure is inside the field of view, and the apparent upper-arch weakness is truncation, not tooth misidentification. The extractor produced 21,028 schema-valid findings over 530 of 532 labelled cases, and a prompt-constrained generator produced 104 reports with zero forbidden-content flags. We also measured the ceiling this design implies: 52.7% of reference-report findings are groundable in the anatomy labels, so recall is capped near 0.53. No report-quality score has been measured yet; the next milestone is the first score against baselines and a human-agreement ceiling.

**Date:** August 24, 2026
**Code:** https://github.com/anaszafar16/cbct-report.git

# Contents


**1 Introduction** 2

**2 Related Work** 3

**3 Methodology** 4

3.1 System overview . . . 4
3.2 Data . . . 4
3.3 Stage 1: multi-class segmentation . . . 6
3.4 Stage 2: from label map to typed findings . . . 7
3.5 Stage 3: verbalisation . . . 7
3.6 Evaluation procedure . . . 8

**4 Results and Progress to Date** 9

4.1 Corpus verification . . . 9
4.2 Label alignment . . . 9
4.3 The ToothFairy3 label convention is mirrored . . . 10
4.4 Segmentation: training run and fold-0 validation . . . 11
4.5 Findings extraction over the labelled corpus . . . 12
4.6 How much of a reference report can be grounded at all . . . 14
4.7 Split and report generation, arm E1 . . . 15
4.8 Evaluation harness . . . 16
4.9 What has not been measured . . . 17

**5 Discussion** 17

**6 Limitations and Future Work** 19

**7 Conclusion** 21

**References** 22

# 1 Introduction

**Motivation.** Cone-beam computed tomography (CBCT) is routine in implantology and oral and maxillofacial surgery, yet the report that accompanies a scan is still written by a clinician, one volume at a time. Automating that step is attractive, but dental reporting has a property that most radiology report generation work does not have to deal with: the clinical meaning of a finding depends on which of up to 32 near-identical instances it is attributed to. "Periapical radiolucency at tooth 36" and the same sentence with "46" describe different treatments on different sides of the mouth. Prior automation for dental CBCT stops at segmentation and classification, or at fixed-slot templates enumerating tooth conditions, which is not the narrative text a surgeon actually writes [1]. The ODIN 2026 challenge defines the task we care about, generating a narrative report from a maxillofacial CBCT volume, and released the first paired corpus for it, ToothFairy4 [2].

**Problem statement.** Given a CBCT volume, produce a narrative maxillofacial report whose findings are correctly attributed to anatomical instances (FDI tooth numbers, left and right canals and sinuses), under three constraints that shape everything downstream: the paired corpus is small (622 volumes, more than an order of magnitude fewer than chest CT report corpora), the English references are machine translations of Italian dictations, and there is no published inter-observer agreement figure telling us what "good" looks like on this task.

**Assumptions and scope.** Four scope decisions are fixed for the project. First, we do not chase segmentation accuracy: the ToothFairy2 and ToothFairy3 challenges solved and open-sourced that problem, and the failure mode that matters to us is tooth *identity* error rather than boundary error [3, 4]. Second, only findings that can be derived from the anatomy labels are grounded in this phase; caries, periodontal staging, periapical lesions and root fractures have no label source in any ToothFairy edition, so the generator is never told about them and must not invent them. Third, the 95 cases of the held-out acquisition centre (subset A) carry no voxel labels in any edition, so perception quality on that centre cannot be measured without new annotation. Fourth, this is the challenge *task* with our own architecture, scored on our own case-level split; we do not adopt the challenge's offline-Docker constraints, hidden test set or submission format, but we keep its metric so that numbers remain comparable.

**The system in one paragraph.** ANCHOR decouples perception from generation (Figure 1). Stage 1 is a multi-class segmentation model, U-Mamba2 trained with the official ToothFairy3 recipe on our own split, which labels 46 anatomical structures. Stage 2 is deterministic code that converts the label map into a typed findings object: field-of-view coverage, tooth inventory in FDI numbering, edentulism spans, restoration sites, and the course and buccolingual position of the mandibular canals. Stage 3 is a language model that verbalises that object into a report. Two generator arms receive the identical findings object, an API model with a constrained system prompt and, next, a Gemma 4 model adapted with LoRA [5]; neither arm ever sees the volume. That separation is the scientific claim of the project, not an implementation detail: whatever the language model says must be traceable to a structure the segmentation found, and every finding carries the label classes and the rule that produced it, so a wrong sentence can be traced back to a wrong voxel. Generated reports will be scored with the challenge's RadFact-style logical precision, recall and F1, against two floor baselines, against the agreement between the two surgeons who read the same scan, and against the share of reference-report content that anatomy labels can ground at all.

**What this report contains.** This is a progress report, written after all three stages exist and before any report-quality score exists. Sections 2 and 3 record the positioning and the design decisions.

Section 4 reports what has actually been measured: the verified corpus statistics, the label alignment and the mirrored laterality convention, the segmentation model's validation on 106 held-out cases, the findings extractor's run over the labelled corpus, the measured share of a reference report that can be grounded at all, and the first 104 generated reports. Sections 5 to 7 interpret those results and list the concrete next steps. Every number in this document is traced to a file in the project repository; quantities that have not been measured are written as such, and Section 4.9 lists them in one place.

## 2 Related Work

The literature review was run in August 2026 as an adversarial keyword search over 2023 to July 2026 for methods and 2024 to July 2026 for dental vision-language work. It is organised in five streams; the paragraphs below keep only what changed a decision in this project.

**Dental CBCT segmentation.** The ToothFairy challenge series moved from a single structure, the inferior alveolar canal [6], to dense multi-structure parsing of jaws, canals, sinuses, pharynx, restorations and individually FDI-numbered teeth [3]. nnU-Net [7] and its residual-encoder variants won ToothFairy2, and U-Mamba2, which places Mamba-2 state-space blocks at the bottleneck of an nnU-Net residual-encoder U-Net, won both ToothFairy3 tasks with a mean Dice of 0.84 on the held-out test set [4]. The finding we rely on most is in the ToothFairy2 report: assigning correct FDI numbers was harder than delineating the teeth. That is independent evidence that instance identity, not boundary accuracy, is the binding constraint in this anatomy, and it is why our perception stream is scored on FDI identity accuracy and confidence calibration rather than on Dice.

**Report generation from 3D volumes.** CT2Rep established narrative report generation on chest CT with more than 25,000 volumes [8], and generalist 3D multimodal models followed [9]. The corpus scale in this stream is more than an order of magnitude larger than ours, so any end-to-end claim we could make would be underpowered by construction. We treat that as an argument for an architecture that puts a deterministic, inspectable stage between the volume and the language model, rather than as something to hide.

**Grounded and segmentation-assisted generation.** RadGPT shows that per-voxel segmentation substantially improves lesion detection in generated abdominal reports [10], and BTReport separates deterministic feature extraction from LLM narrative formatting [11]. The grounding hypothesis is therefore already supported and we do not claim it as a discovery. All of these systems operate on organ-level anatomy with a small referent vocabulary; none has been tested where the referents are 32 numbered instances of the same object class, which is the regime dental reporting lives in.

**Evaluating generated reports.** Metrics progressed from n-gram overlap [12] to entity-aware scores [13], to LLM-based logical entailment (RadFact, introduced with MAIRA-2 [14]), and to error-taxonomy judges such as GREEN [15], FineRadScore [16] and CRIMSON [17], building on the ReXVal clinical error taxonomy [18] and packaged in toolkits such as RadEval [19]. Two facts from this stream corrected an early assumption of ours. Location errors are already a scored category: GREEN's category (c), anatomic location misidentification, is its weakest, with a reported detection accuracy of 0.60 to 0.65, validated on chest X-ray where the referent vocabulary is roughly a dozen coarse regions. And permuting the referents of a reference report to test a metric is prior methodology, described in the GREEN paper, with ReXamine-Global as precedent that stress-testing existing metrics across sites is a publishable contribution in itself [20]. We

therefore do not propose a new metric. The open question we keep, as an instrument rather than a promised result, is whether that sensitivity to location holds when the referent space grows to FDI numbering.

**Vision-language models in dentistry.** The 2026 wave of dental vision-language work is fast and almost entirely 2D: PanDent releases 9,524 panoramic radiographs with expert-validated tooth-level annotations and reports that fluent generation does not imply clinically consistent tooth-level reasoning [21]; OralGPT-Plus learns tool use for panoramic X-ray analysis [22]; IOSVLM targets intraoral surface scans [23]; and several further panoramic systems appeared in the same months. PanDent's reports are constructed from expert-validated findings using clinician-defined reporting logic, which makes it a rule-generated corpus, whereas ToothFairy4 reports were dictated by surgeons under a fixed taxonomy. Maxillofacial CBCT narrative reporting has no published system.

**Human agreement as a reference point.** Obtaining additional expert readings of ground-truth reports, and treating clinician-versus-clinician agreement as the yardstick for a generator, has precedent in chest radiography [24]. ToothFairy4 gives us that yardstick for free: 367 of its cases were read independently by two surgeons, which is the asset our human-ceiling analysis is built on.

# 3 Methodology

## 3.1 System overview

Figure 1 shows the pipeline. A CBCT volume enters a segmentation model and leaves as a label map. Deterministic code converts the label map into a typed findings object. A language model turns the findings object into a narrative report. The scoring harness compares that report with the surgeon's reference report. Two design rules apply everywhere. The language model never receives image features, only the findings object, so that every sentence it writes can be traced back to a labelled structure. And the two generator arms we compare receive byte-identical input; if one arm got extra context the comparison would be worthless.

The findings object is the whole architecture in one place. It is the only channel between perception and generation, it is typed and versioned, and every finding in it carries the label classes and the rule that produced it. That makes the generator auditable in a way an end-to-end model is not: a wrong sentence can be traced to a wrong finding, and a wrong finding to a wrong voxel.

## 3.2 Data

**ToothFairy4 (ODIN 2026).** The corpus is the ToothFairy4 track of the ODIN 2026 challenge [2]. The team re-hosted the released files as a private Kaggle dataset and verified every count against the file manifest and by reading every report file. The released training corpus contains 622 cases, one CBCT volume each, from four subsets: P (412 cases), A (95), F (63) and S (52). Reports come in two mirrored directories, `reports_en` and `reports_it`, with 1,000 files each. The English reports are GPT translations of what the surgeon dictated in Italian, with only about 10% clinician review on training cases, so Italian is the primary language of the reference text. Volumes are 0.3 mm isotropic. Reports follow a fixed clinical taxonomy with FDI tooth numbering mandated. The challenge's own 50-case test set (Radboud UMC) is never released, so every split we use is our own. Subset A is the held-out acquisition centre. The design document quotes 630 patients and

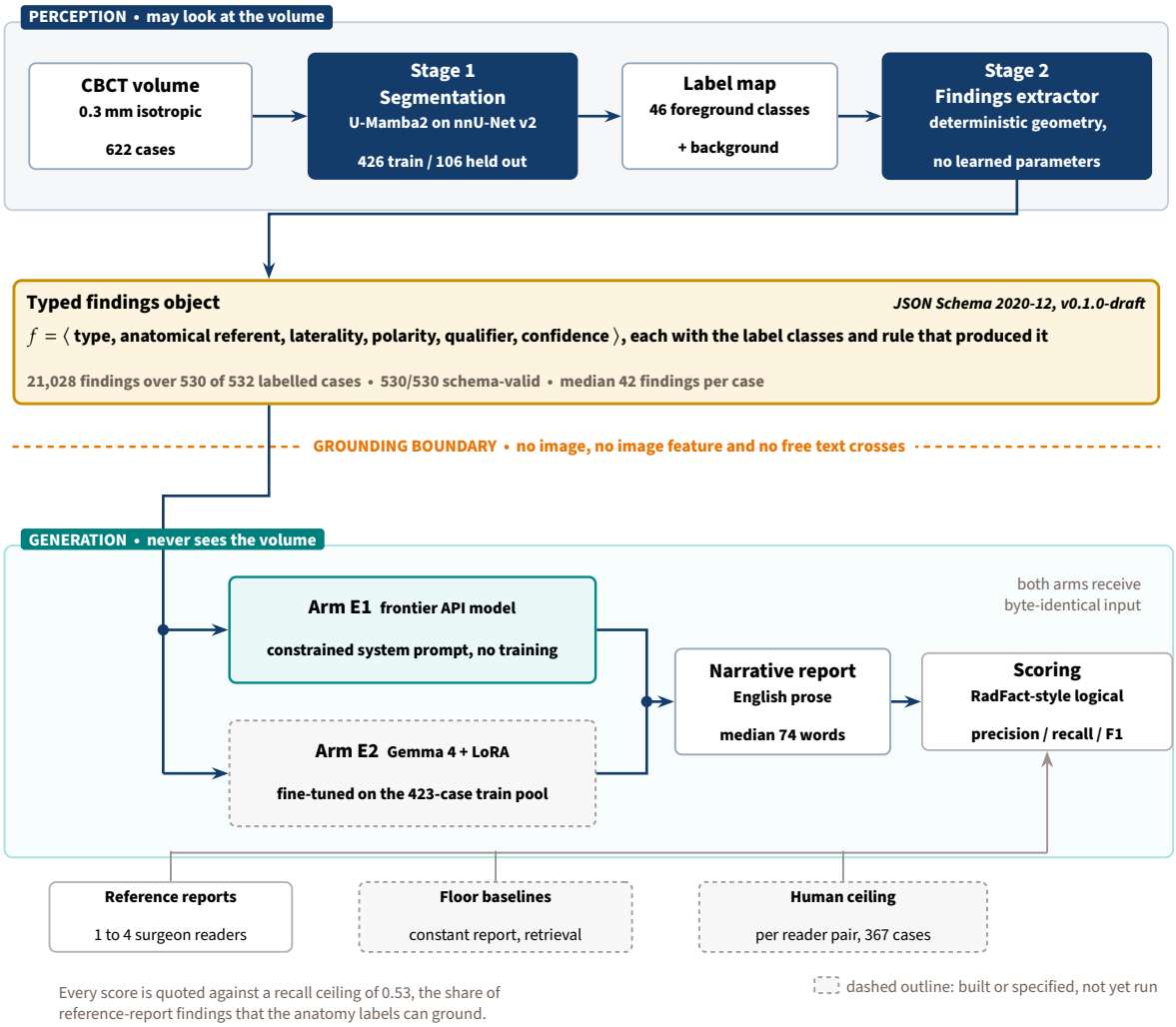



**Figure 1** The Anchor pipeline. Perception (Stage 1) and generation (Stage 3) are decoupled by a typed findings object (Stage 2), which is the only channel between them. Counts are the measured state at the time of writing: 21,028 findings extracted over the labelled corpus (Section 4.5) and 104 reports generated by arm E1 (Section 4.7). Those 104 reports were generated from findings extracted from the ToothFairy3 reference labels rather than from Stage 1's own predictions, so the figure describes an architecture of which one path, segmentation to findings to report, has not yet been run end to end.

about 30% multi-read cases; the released files show 622 and 59.0%, and the release wins where they disagree. Licence: CC BY-NC-SA per the design document.

**ToothFairy3 voxel labels.** Anatomy labels come from ToothFairy3, which covers the same P, F and S volumes with 532 labelled cases (P 417, F 63, S 52). 527 of those overlap with ToothFairy4 by case identifier (for example `F001` is `ToothFairy3F_001`); the 95 subset A cases have no labels in any edition, and five ToothFairy3 cases (P120, P121, P262, P263, P378) have no ToothFairy4 report. So 527 of 622 cases (85%) can carry anatomy grounding and the held-out centre carries none. The raw label schema has 78 classes: lower and upper jawbone; left and right inferior alveolar canal; left and right maxillary sinus; pharynx; bridge, crown and implant; the 32 teeth in FDI numbering (11 to 48); left and right mandibular incisive canal and the lingual canal (103 to 105); and one pulp class per tooth (111 onward). For training, the official U-Mamba2 recipe remaps this to 46 foreground classes plus background by merging the per-tooth pulps into a single class, because nnU-Net rejects non-consecutive label integers.

**Two conventions that have to be fixed before any label is reused.** The two editions disagree twice, in ways that are silent if not tested for, and both corrections are implemented in `src/perception/align_labels.py`.

The first is orientation. The editions store the same voxels under different NIfTI axis codes (ToothFairy4 LAS, ToothFairy3 LPS). The header implies an anterior-posterior flip only; the transform that actually aligns them was determined empirically and also flips the superior-inferior axis: `tf4_volume == tf3_array[:, ::-1, ::-1]`. It is checked per case by `verify()` and was confirmed bitwise on six cases across F, P and S (Section 4.2).

The second is laterality. Every sided code in the ToothFairy3 label space refers to the opposite patient side from the one the ToothFairy4 reports use. `to_clinical_convention()` is an involutive value remap (teeth quadrants 1 to 2 and 3 to 4, pulp likewise, the inferior alveolar canals, the maxillary sinuses and the incisive canals) that must be applied to any label map facing the reports, including Stage 1's predictions after the model-index-to-FDI translation. The evidence is in Section 4.3, and it is the single most consequential measurement in this reporting period: without it every laterality statement in every generated report is wrong.

### 3.3 Stage 1: multi-class segmentation

We use U-Mamba2 [4], the ToothFairy3 winner, built on nnU-Net v2 [7]. The public U-Mamba2 weights were trained on the ToothFairy3 training set, which is exactly our P, F and S cases; using them would leave no labelled case unseen by the model. We therefore re-ran the official recipe on our own case-level split so that a held-out labelled pool exists: 426 training and 106 validation cases (fold 0 of a single split, seed 42; validation F 12, P 82, S 12). The recipe, taken from the authors' documentation and confirmed in the training log, is: trainer `nnUNetTrainer_TF3_Task1_1500ep_accum2`, configuration `3d_fullres_torchres_mambabot2_ps160x288x288_bs1`, plans `nnUNetResEncUNetLPlans_torchres`, network `UMambaBot2` with seven stages and 32 to 320 features per stage, patch size $160 \times 288 \times 288$, batch size 1 with gradient accumulation simulating batch size 2, oversampling 0.33, CT intensity normalisation, initial learning rate 0.01 with the nnU-Net polynomial schedule, and initialisation from the authors' self-supervised checkpoint `ssl_umamba2_3d_depth7_160x288x288.pth`. Training ran on the KAUST Ibex cluster. Two environment facts cost real time and are recorded here: `mamba_ssm` does not build from source on Ibex (no `nvcc`; CUDA modules stop at 12.4 while the driver is 12.8), so the prebuilt `cp312` wheels for `causal_conv1d` 1.5.2 and `mamba_ssm` 2.2.5 must be used and PyTorch pinned to 2.8.0+cu128; and the recipe is CPU-bound on data augmentation, so raising the augmentation worker count is what shortened the epoch time (Section 4.4).

The schedule nominally ran to 1,500 epochs and was stopped at epoch 542, deliberately rather than for lack of compute; Section 4.4 gives the curves that decision rested on. Epoch 542 is the final model, not an interim checkpoint. Validation is a real inference pass over the 106 held-out cases with `nnUNetv2_evaluate_folder`, reproducible through `Test/modal_validation.py`.

The perception stream is not scored on Dice alone. Its metrics are per-tooth presence accuracy, FDI identity accuracy with the confusion matrix, restoration presence per tooth, canal laterality and buccolingual side, and structure-in-field-of-view detection, plus a per-instance identity confidence that will be calibrated (a 0.9-confidence label should be right about 90% of the time). Candidate confidence signals are softmax margin, test-time augmentation agreement and, if a second fold is ever trained, ensemble disagreement. Dice is reported because it is what the field reports and it establishes that the recipe was reproduced; Section 4.4 reports it in the two aggregations that make it interpretable rather than as a single mean.

### 3.4 Stage 2: from label map to typed findings

The extractor is deterministic code, not a learned model, and it is what makes the generator auditable. Its contract is a set of findings $f = \langle$type, anatomical referent, laterality, polarity, qualifier, confidence$\rangle$, where the referent is an FDI index or a named structure, laterality is left, right, bilateral or not applicable, polarity is affirmed, negated or normal, the qualifier carries severity, certainty or a measurement, and confidence is in $[0, 1]$. Every finding also carries a provenance block naming the label classes and the rule that produced it. The schema is JSON Schema 2020-12 at `src/findings/schema/`, version `0.1.0-draft`; the freeze condition was that arm E1 consume the object end to end, which Section 4.7 reports as met.

`src/findings/extract.py` emits nine finding types from geometry alone: field-of-view coverage per structure, dentition inventory (tooth present and tooth absent), edentulism spans, implant, crown and bridge site attribution including abutment and pontic roles, mandibular canal course, and canal buccolingual position on a five-point scale from lingual to buccal. Model indices are mapped back to FDI numbers by name (`src/perception/labels.py`); this matters because the remapped index is not the tooth number (predicted index 27 is FDI 31), and the mapping was checked against all 32 tooth names with zero disagreements. Merged pulp components are re-attributed to teeth by shell-majority overlap, falling back to the nearest tooth, which recovers the attribution that the many-to-one pulp merge destroyed; pulp findings are held in an instrumentation block and are not verbalised. Section 4.5 reports the validation that was the precondition for emitting them; the decision to emit them has not been taken.

Two safety behaviours are worth naming because they shape the output. First, each case runs an orientation self-check before anything is emitted, and a case whose upper jaw is not superior to its lower jaw in the aligned frame is refused rather than guessed at. Second, "not scanned" and "not present" are kept as distinct finding types: the condyles being outside the field of view is a different statement from tooth 46 being absent, and conflating them is the most common way a plausible report becomes a wrong one.

A finding is emitted to the verbaliser only when its confidence exceeds a threshold, and that gate is the axis the planned ablation sweeps. Groundable from the labels: dentition and restoration inventory (and which tooth carries the restoration), canal course, laterality and buccolingual position, incisive and lingual canals, maxillary sinus anatomy, jaw morphology and ridge height at edentulous sites, and field-of-view coverage. Not groundable, because no label source exists in any edition: caries, periodontal staging, periapical lesions, root fracture, peri-implantitis, sinus mucosal thickening and bone quality. Recall against the reference reports is therefore hard-capped at the groundable share of reference findings, which Section 4.6 measures.

### 3.5 Stage 3: verbalisation

The verbaliser takes text, not images, so it does not need to be a vision-language model. Arm E1 is a frontier API model with a constrained system prompt and no training. Arm E2 will be Gemma 4 adapted with LoRA [5] on findings-object to reference-report pairs (Gemma 4 was released on 2 April 2026 under Apache 2.0; MedGemma at matched size is the natural comparison for the uncommitted question of whether medical pre-training helps or hurts grounded verbalisation). Both arms are given the same rendered text, produced by one `render_findings()` function that strips provenance and the instrumentation block. Neither arm can see how a finding was derived, and neither is given anything the other is not.

The E1 system prompt states that the input is the complete record of what was observed and that the model has never seen the images, then forbids by name every ungroundable category: caries, periodontal disease, periapical and intrabony lesions, root fracture, endodontic treatment

and canal filling material, implant osseointegration, sinus mucosal change and bone quality, "not even to state their absence". It mandates FDI two-digit numbering, it requires that coverage and absence stay separate in the prose, and it fixes the report skeleton as `Mandible`, `Maxilla`, `Mandibular condyles`. The prompt and its SHA-256 are stored in the run manifest, and the generated output of every case records the model actually served, so the arm can be reported with its contamination caveat rather than in spite of it.

Few-shot exemplars are two cases drawn from the training pool only (P009 and F006). Their assistant turns are hand-written from those cases' own findings, in corpus style, and restricted to groundable content. This is a deliberate departure from using reference reports as exemplars: a real reference report demonstrates exactly the ungroundable claims the system prompt forbids, so showing one to the model works against the constraint it is being asked to obey.

Reader identity is treated as a variable (Section 4.1 shows why), and the invention rate of the generator is measured separately from the metric, by a regular expression over the forbidden vocabulary that runs on every generated report.

### 3.6 Evaluation procedure

**Split.** There is no split in the released data, and 367 cases carry two or more reports, so a report-level split would leak; all splits are case-level (`src/eval/splits.py`, with a leakage assertion). The split is built against the segmentation split so that any case scored end to end is one the segmentation model never trained on. Three tiers of increasing difficulty are defined: Tier I, matched distribution, the 104 segmentation-validation cases that also carry reports; Tier II, held-out centre, the 95 subset A cases, which have no voxel labels and therefore wait on Stage 1 predicting there; Tier III, cross-corpus, test-only on MMDental as a stretch goal. The remaining 423 cases are the verbaliser training pool. Fourteen of the sixteen reader identifiers appear in both A and P/F/S, so the held-out-centre tier is viable, with a reader-offset check planned.

**Metric.** The primary metric is the challenge's own: the dental adaptation of RadFact logical precision, logical recall and logical F1, as shipped in the public `radfact_lite` package (version 0.1.0, pinned) and called by the challenge harness with `ReportType.TOOTHFAIRY`, narrative-text mode and negative filtering disabled. The default judge is `gpt-4o-mini`, used for phrase parsing, entailment and filtering alike, and any override of the model identifier is recorded. BLEU-4 [12] and METEOR are the offline fallback of the same harness and are reported as secondary. Six observations about the shipped prompts (for example, the entailment prompt contains no identity constraint and its single few-shot example is sinus laterality; the examples use dotted FDI numbers while the reports use plain two-digit ones) are recorded as hypotheses about weakness, not findings; the permutation study below is what would turn them into findings.

**Reference points.** Two floor baselines are run on the same split before any model number is read: a constant report (the same unremarkable-study text for every case) and a retrieval baseline (the most similar training report retrieved from information available without the image). If the pipeline does not beat both, the task framing or the method is wrong, and the project's gate scheme stops the work there rather than reporting the number. The ceiling is the agreement between the two surgeons who read the same scan, scored one against the other with the same metric on the 367 dual-reported cases. It is reported per reader pair, never pooled: there are 26 pairs, the four largest covering 51, 49, 48 and 44 cases, and reader completeness must be characterised first because reader style dominates report content (Section 4.1). The primary ceiling is computed on the Italian pair, the English one is reported alongside, and the gap estimates the translation cost. Above all of these sits the recall ceiling of Section 4.6: no pipeline restricted to groundable

findings can exceed it, and a score quoted without it is not interpretable.

Two instrument-only studies are pre-registered: a metric sensitivity study that permutes FDI indices in the reference reports, with the promotion thresholds fixed in advance (detection accuracy below 0.50 promotes to a named contribution, 0.50 to 0.60 stays instrument-only, 0.60 or above earns one paragraph), and a 3D-dependency partition of finding types whose share of clinically significant findings decides whether that contribution survives (at least 25% strong, 15 to 25% viable, below 15% dropped).

**Tools.** PyTorch 2.8 (CUDA 12.8), nnU-Net v2 inside the U-Mamba2 fork, SimpleITK and SciPy for the extractor, Kaggle for private data and model hosting, the KAUST Ibex cluster (SLURM) for training, Modal for validation and for the corpus-wide extraction run, OpenRouter for the API generator arm, and `radfact_lite` for scoring.

## 4 Results and Progress to Date

This section reports only what has been measured or built. Every number is traced to a file in the project repository. No end-to-end report-quality score exists yet, and no number below should be read as one.

### 4.1 Corpus verification

The first deliverable was to replace the estimates in the challenge design document with counts taken from the released files. Table 1 summarises them. Two findings changed the plan. First, the dual-read subset is twice as large as documented: 367 of 622 cases (59.0%) carry two or more reports, against "about 30%" in the design document, and the challenge harness discards the 378 extra reports. Those 367 cases are the asset the human-agreement ceiling is built on, and they are also the reason every split must be case-level. Second, reader style dominates report content. Report filenames follow `cbct_<A>_<B>.txt`; B is a globally unique report id and A takes 16 distinct values that behave exactly like a reader id (in all 367 dual-read cases the two reports carry different A values, with zero exceptions), although the organisers have not documented this and it must be confirmed with them. Across readers, mean report length varies 3.5-fold (51 to 177 words) and the number of distinct FDI references per report varies 4.2-fold (3.0 to 12.7); one reader uses a fixed mandible / maxilla / condyles skeleton that omits the dentition altogether. Consequently the ceiling will be reported per reader pair, and reader identity enters training and evaluation as a variable. An earlier suspicion of Italian text inside `reports_en` was checked over all 1,000 files and retracted; the cause was a filename collision in the download script.

### 4.2 Label alignment

Header comparison over 15 cases showed identical shapes and spacing but different axis codes (LAS versus LPS), which would suggest an anterior-posterior flip only. A voxel-level comparison on six cases (F001, F002, P001, P002, S0000, S0001) over all eight flip combinations found that the correct transform is `tf3_array[:, ::-1, ::-1]`, and it is bitwise exact on all six (mean absolute difference 0.00). As a sanity check, aligned tooth labels sit on voxels with mean intensity roughly 1,570 to 2,020 against a background mean of −505 to −1. The lesson recorded for the rest of the project is that header-only checks are not sufficient: the header-implied fix would have left every label vertically inverted, putting upper-jaw labels on the mandible, and it would have produced no error and no warning. The transform is self-inverse and is called through `verify()` for every case before any label is reused.

**Table 1** ToothFairy4 corpus as released, verified against the file manifest and by reading every report file (kernel `tf4-extract-reports`, 2026-08-03). The design document's figures are shown where they differ.

| Quantity | Released files | Design document |
|---|---|---|
| Cases (one CBCT volume each) | 622 | 630 |
| Subsets | P 412, A 95, F 63, S 52 | |
| Reports | 1,000 English + 1,000 Italian, mirrored | |
| Cases with 1 / 2 / 3 / 4 reports | 255 / 360 / 3 / 4 | |
| Cases with two or more reports | 367 (59.0%) | about 30% |
| Reader ids (inferred from filenames) | 16 | 13 surgeons |
| Reader pairs on dual-read cases | 26 (largest: 51, 49, 48, 44 cases) | |
| Readers per subset (P / A / F / S) | 14 / 14 / 6 / 3 | |
| Voxel spacing | 0.3 mm isotropic | 0.3 mm |
| Voxel labels (ToothFairy3) | 527 of 622 cases (85%); none in A | |
| Held-out challenge test set | 50 cases (Radboud UMC), never released | |

### 4.3 The ToothFairy3 label convention is mirrored

Geometric alignment is not the same as semantic alignment, and the second disagreement between the editions was found only when the extractor first produced text that could be compared with a report. **Every sided code in the ToothFairy3 label space refers to the opposite patient side from the one the ToothFairy4 reports use**, coherently: tooth quadrants, inferior alveolar canals, maxillary sinuses and incisive canals all agree with each other and all disagree with the reports.

Two independent lines of evidence establish it, both frame-independent.

1. **Against the reports.** Explicit per-tooth absence and implant-position claims were mined from the English reports of the staged fold-0 validation cases and scored against the ground-truth tooth codes. On strictly asymmetric claims, the only decisive kind, the quadrant-mirrored reading wins 69 of 72 (95.8%) and the direct reading wins 3, each of which is parsing noise or a conflict between two reports of the same case on inspection. F005 is typical: the reports place a crown on 36 and record absence of 46 and 48, while the raw labels place the crown on 46 and record absence of 36 and 38. An exact mirror, every time.
2. **Internally.** A side audit over the staged cases asked, for every structure whose class name states a side, whether it sits on the opposite side from the quadrant-1 teeth. It does, in every adjudicable case but one: inferior alveolar canals 76 of 76, mandibular incisive canals 75 of 75 and 74 of 74, maxillary sinuses 9 of 10. Denominators are the cases in which both the structure and quadrant-1 teeth were present and unambiguous, which is why they are smaller than the presence counts in Table 3. The label world is therefore internally coherent and coherently mirrored as a whole, so one global swap fixes all of it.

After applying `to_clinical_convention()`, report agreement flips to 95.8% for the direct reading and the per-case canal laterality self-check passes on all 106 cases, having failed on all 106 before.

Two consequences matter. Fold-0 Dice is unaffected, because predictions and ground truth share the convention, but the sided *class names* in `results/summary.json` are in the ToothFairy3 convention and their values swap within each left/right pair when read clinically; every sided figure in this report has had that swap applied. And Stage 1 learned the ToothFairy3 convention, so its predictions need the same remap after the model-index-to-FDI translation. FDI identity accuracy scored against the reports without it would come out catastrophically and misleadingly

low, which is precisely the failure this project is instrumented to detect. It is worth stating plainly as a general lesson: reusing labels across dataset editions requires adjudication against clinical ground truth, not just against headers, and the check has to be a statement a clinician could falsify.

### 4.4 Segmentation: training run and fold-0 validation

Table 2 lists the configuration and the final state of the run, and Figure 3 shows the curves it was stopped on. Training was stopped at epoch 542 of the nominal 1,500, deliberately: the per-class training curves showed that lower jawbone, the inferior alveolar canals and the 32 teeth had been flat since roughly epoch 350, and the only classes still improving were crown, bridge and the maxillary sinuses, the last of which is outside the acquired volume in over 90% of cases. Epoch 542 is the final model, not a snapshot awaiting a better one. The checkpoint-selection metric, the exponential moving average of the training-time pseudo-Dice, reached 0.847 at epoch 541 and the highest single-epoch mean pseudo-Dice was 0.913 at epoch 494. Against U-Mamba2's published mean Dice of 0.84 on the ToothFairy3 held-out test set [4], that says the recipe, environment and data plumbing were reproduced correctly, which was the purpose of the run. It says nothing else: pseudo-Dice is computed on single patches with foreground oversampling while training proceeds, and it runs optimistic.

The honest figure is the validation pass: a real inference run over all 106 held-out cases, scored with the official `nnUNetv2_evaluate_folder`. Its `foreground_mean` Dice is **0.7246** over 46 classes and all 106 cases, and that number should never be quoted on its own, because it averages in structures that were never inside the field of view. A class that scores 0 on a case where the anatomy was not imaged is not a segmentation failure.

**Two aggregations, one of which is interpretable.** Figure 2 recomputes every class over only the cases where the structure is present in the ground truth, and shows the presence rate alongside. The classes whose two aggregations differ are exactly the classes that are frequently outside the volume. The maxilla is present in 36 of 106 cases (34%) and the maxillary sinuses in 10 and 9 (9% and 8%); restricting to those cases moves the maxilla from 0.45 to 0.76 and the sinuses from 0.53 to 0.79. Averaged over all 46 classes, the correction is 0.7246 to 0.8806.

**The upper arch is truncated, not misidentified.** This distinction is the one that decides whether the model is usable downstream. Averaged over all cases the upper teeth score 0.613 and the lower teeth 0.878, a gap of 0.265 that would be alarming if it were real, because a systematic upper-arch weakness in a model whose job is instance identity would suggest FDI confusion. Restricted to cases where the tooth is present, the two arches score 0.906 and 0.939 and the gap collapses to 0.032. The upper arch looks weak because the maxilla is outside the acquired volume in 66% of these cases, not because the model puts the wrong number on a tooth. The same effect explains why the posterior lower arch scores lower than the anterior one under the all-cases aggregation: third molars are present in about half the cases, because they get extracted.

Presence detection points the same way. Every non-tooth class was detected in every case where it was present. Among the teeth, 2,050 of 2,061 present instances were detected, 11 misses spread over eight classes, the worst of which is 95.5%. Table 3 gives the per-structure detail for everything the findings extractor depends on.

**What is genuinely weak.** Three groups survive the present-only correction and are real limitations. The mandibular incisive canals score 0.63 and 0.61 while being present in 97% and 93% of cases, and the lingual canal 0.69 while present in 98%. These are thin, low-contrast structures, they are not field-of-view artefacts, and the incisive canal is clinically relevant to implant planning in the interforaminal region. The restoration classes are next: bridge 0.68, crown 0.69 and implant

**Table 2** Segmentation run: configuration and final state. Pseudo-Dice values are training-time signals on single patches; the validation figures are a real inference pass over the 106 held-out cases.

| Item | Value |
|---|---|
| Model, recipe | U-Mamba2, official ToothFairy3 Task 1 recipe (nnU-Net v2 fork) |
| Trainer | `nnUNetTrainer_TF3_Task1_1500ep_accum2` |
| Configuration | `3d_fullres_torchres_mambabot2_ps160x288x288_bs1` |
| Plans | `nnUNetResEncUNetLPlans_torchres` |
| Network | `UMambaBot2`, 7 stages, features 32 to 320 |
| Patch, batch | 160 × 288 × 288, batch 1 with gradient accumulation 2 |
| Initialisation | authors' SSL checkpoint `ssl_umamba2_3d_depth7_160x288x288.pth` |
| Labels | 46 foreground + background (per-tooth pulps merged) |
| Split | 426 train / 106 validation (fold 0, seed 42; validation F 12, P 82, S 12) |
| Median image size, spacing | 169 × 349 × 371 voxels, 0.3 mm isotropic |
| Epochs completed | 542, stopped deliberately (nominal schedule 1,500) |
| Best EMA pseudo-Dice | 0.847 at epoch 541 |
| Peak single-epoch mean pseudo-Dice | 0.913 at epoch 494 |
| Published U-Mamba2 test-set Dice | 0.84 |
| Epoch time | about 510 s, then about 300 s after raising the augmentation worker count (1.7× faster) |
| GPU utilisation before the change | oscillating 100% / 0%, 32 of 82 GB VRAM used |
| **Validation, 106 held-out cases** | `nnUNetv2_evaluate_folder`, `results/summary.json` |
| `foreground_mean` Dice, all 46 classes | **0.7246** (all cases) |
| Same 46 classes, present-only mean | **0.8806** |
| Detection given presence | 100% for 38 of 46 classes; 2,050 of 2,061 present tooth instances |
| FDI identity accuracy, confusion matrix | *not yet measured* |

0.77 under present-only aggregation. Since restoration *site attribution* is one of the things the extractor emits, boundary quality there matters less than getting the right tooth, but a 0.68 bridge is a weak basis for describing abutments and pontics. Third, the merged pulp class reaches 0.84 present-only, which is adequate for the shell-majority attribution described below but not for anything measured on pulp geometry itself.

## 4.5 Findings extraction over the labelled corpus

The extractor was run on every ToothFairy3 label map after the laterality correction of Section 4.3. **530 of 532 cases were extracted, all 530 validated against the schema, and 21,028 findings were emitted**, a median of 42 findings per case with a range of 25 to 55.

The two refusals are S0002 and S0015, where the per-case orientation self-check found the upper jaw not superior to the lower jaw in the aligned frame. They are two cases stored outside the corpus-wide convention, caught exactly as the self-check was designed to catch them, and they are excluded rather than repaired: auto-flipping a label map without its image risks reintroducing the silent side swap that Section 4.3 exists to prevent. They are recoverable by downloading their images and running the per-case `verify()` to establish their true transform.

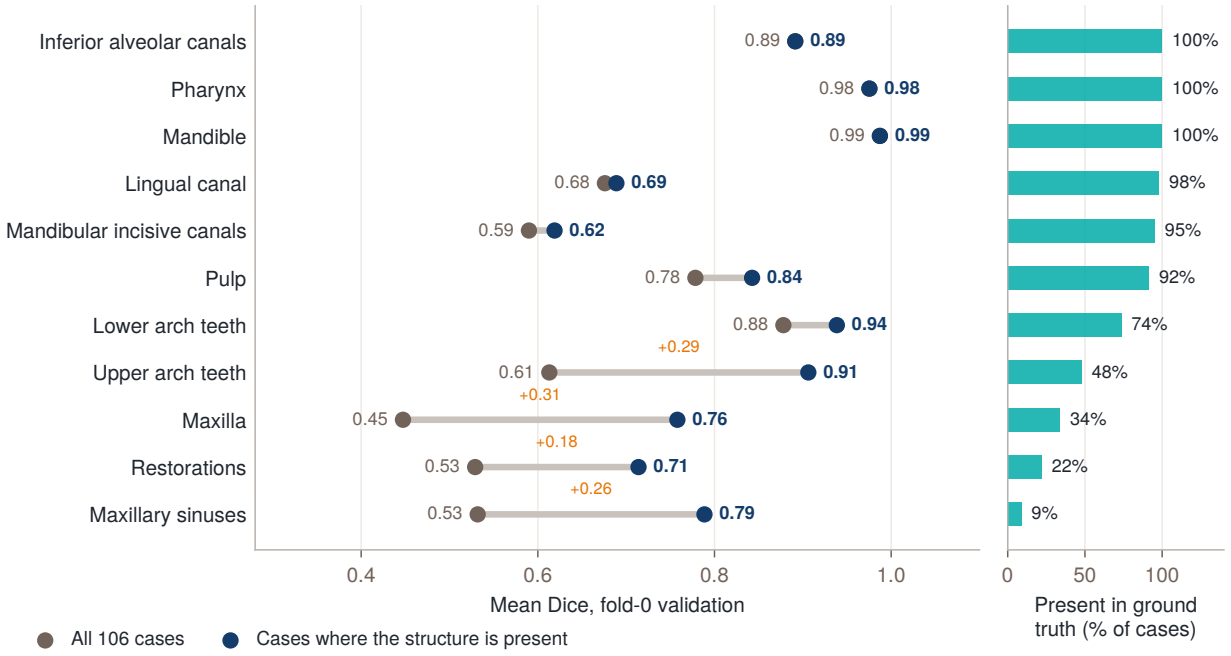



**Figure 2** Fold-0 validation Dice under two aggregations, grouped by anatomy and ordered by how often the structure is in the scan at all. Groups whose two values coincide are the structures that are always present; the gap grows exactly as presence falls, which is the signature of field-of-view truncation rather than of segmentation error. Left panel: mean Dice over all 106 cases (grey) and over only the cases where the structure is present in the ground truth (navy), with the difference annotated where it exceeds 0.10. Right panel: the share of the 106 cases in which the structure is present. Rendered from `results/summary.json`.

**The pulp merge is invertible in practice.** The training remap collapsed the 32 per-tooth pulp classes into one, and that merge is many-to-one, so a predicted pulp voxel carries no tooth attribution. Shell-majority re-attribution was scored against the true per-tooth codes over the whole corpus: **9,759 of 9,763 components correct, 99.96%**, with mean shell agreement 0.9993. All four mismatches are flagged by their own confidence rather than passing silently: an impacted third molar whose pulp is ambiguous with its neighbour's (shell agreement 0.573), one suspected ground-truth typo where the pulp is coded for the mirrored-arch tooth although its shell lies entirely inside one tooth, and two 12-voxel orphan fragments adjacent to no tooth at all (confidence 0.0). Pulp findings remain in the instrumentation block and are not verbalised, but the measurement that would justify emitting them now exists.

**Canal buccolingual position agrees with the reports in 89% of adjudicable cases.** Where a report states a position for the body of the mandibular canal, the extractor agrees in **370 of 418 cases**. Sentences about emergence at the mental foramen are excluded from that comparison, because emergence is anatomically buccal and is not what the extractor measures. The corpus-wide distribution is dominated by central and centro-lingual, which is consistent with mandibular canal anatomy. Two version-1 limitations are logged and both are visible in Section 4.7: two adjacent implants can win the same interpolated tooth slot, and canal position is computed as a single whole-body average, so a report phrase such as "predominantly lingual with buccal emergence" collapses to one descriptor.

**Table 3** Fold-0 validation, per structure. *All* averages over the 106 cases; *present* averages only over cases where the structure appears in the ground truth; *present in* is how many of the 106 cases those are, given as a mean rate for the three multi-class dentition rows. Sided rows are named in the clinical convention of the reports, so their values are swapped relative to the raw class names in `results/summary.json` (Section 4.3).

| Structure | All | Present | Present in | Detected |
|---|---|---|---|---|
| *Always in the volume* | | | | |
| Lower jawbone | 0.987 | 0.987 | 106 (100%) | 106/106 |
| Pharynx | 0.975 | 0.975 | 106 (100%) | 106/106 |
| Inferior alveolar canal, right / left | 0.892 / 0.891 | 0.892 / 0.891 | 106 / 106 | 106 / 106 |
| *Nearly always in the volume, and genuinely weak* | | | | |
| Lingual canal | 0.676 | 0.689 | 104 (98%) | 104/104 |
| Mandibular incisive canal, right / left | 0.611 / 0.568 | 0.629 / 0.609 | 103 / 99 | 103 / 99 |
| Pulp (merged class) | 0.778 | 0.843 | 97 (92%) | 97/97 |
| *Dentition* | | | | |
| Lower teeth (16 classes) | 0.878 | 0.939 | 74% of cases | |
| Upper teeth (16 classes) | 0.613 | 0.906 | 48% of cases | |
| All 32 teeth | 0.745 | 0.922 | 61% of cases | 2,050 / 2,061 |
| *Sparse: present in a minority of cases* | | | | |
| Upper jawbone | 0.447 | 0.758 | 36 (34%) | 36/36 |
| Crown | 0.482 | 0.690 | 37 (35%) | 37/37 |
| Bridge | 0.432 | 0.683 | 19 (18%) | 19/19 |
| Implant | 0.673 | 0.769 | 14 (13%) | 14/14 |
| Maxillary sinus, right / left | 0.498 / 0.565 | 0.698 / 0.880 | 10 / 9 | 10 / 9 |

## 4.6 How much of a reference report can be grounded at all

The recall a grounded pipeline can reach is bounded by the share of what surgeons actually write that anatomy labels can support, and that share is now measured rather than assumed. Sixty reports were sampled stratified across subsets (A 12, P 24, F 12, S 12), one English report per case, decomposed into 916 atomic findings, and synonym-merged into 29 canonical finding types. Each type was judged against the ToothFairy3 label space.

**52.7% of reference-report findings are deterministically groundable, 17.1% are geometric but need work beyond version 1, and 30.1% have no label source in any ToothFairy edition.** So recall for this phase is capped near 0.53, rising to roughly 0.70 if the geometric partials are promoted. This is the number every pipeline score must be quoted against; without it, a score of 0.4 is uninterpretable, and next to it, the same 0.4 is three quarters of what was reachable.

Figure 4 shows where the mass sits. The largest single family is field-of-view coverage (142 findings), which validates treating coverage as a first-class finding type rather than as a caveat, and the canal course and position cluster (92 across course, position and emergence) is both large and fully groundable. The largest ungroundable cluster is endodontic status (84), for which attributed pulp is the only adjacent labelled signal, followed by osteolytic and sclerotic lesions (74) and periodontal disease (34). Negated lesion boilerplate is ungroundable and frequent, which means the constant-report baseline will look strong on exactly those sentences. That is not a problem with the baseline; it is the reason for running it.

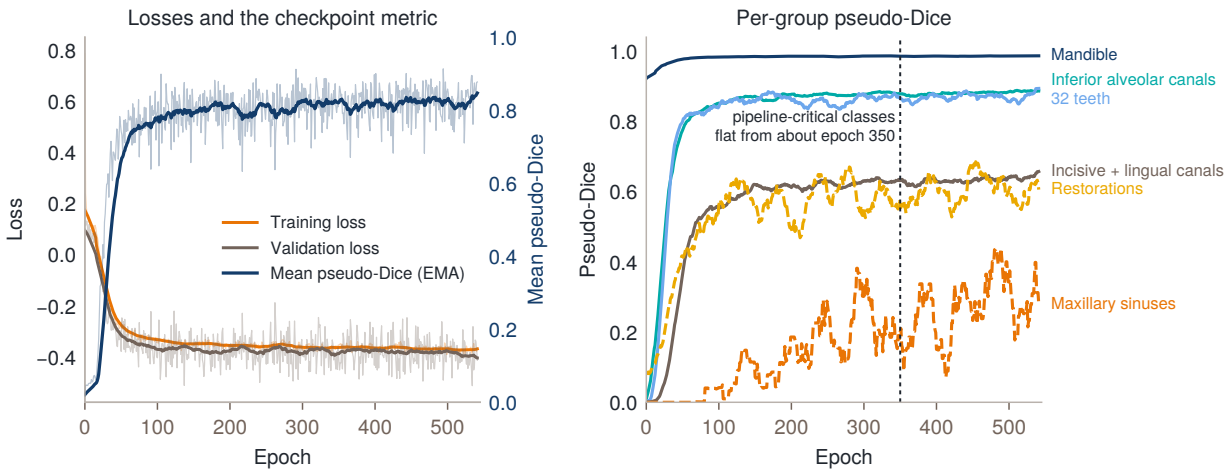



**Figure 3** The segmentation run, parsed from the two nnU-Net training logs (542 epochs). Left: training and validation loss with the mean pseudo-Dice and the exponential moving average that nnU-Net uses to select a checkpoint; the EMA peaks at 0.847 at epoch 541 and the best single epoch reaches 0.913 at epoch 494. Right: pseudo-Dice by anatomical group, which is the evidence the decision to stop at epoch 542 rested on. The mandible, the inferior alveolar canals and the 32 teeth, which is everything the findings extractor depends on, are flat well before epoch 350; only the restorations and the maxillary sinuses are still moving, and the sinuses are outside the acquired volume in over 90% of cases. Pseudo-Dice is a training-time signal computed on single patches with foreground oversampling and runs optimistic against the validation figures of Figure 2.

### 4.7 Split and report generation, arm E1

**The split was built first.** Case-level, leak-free, and partitioned against the segmentation split so that no case is scored on a model that trained on it: Tier I has the 104 segmentation-validation cases that carry reports, Tier II the 95 subset A cases, and the verbaliser training pool the remaining 423, with zero cases unassigned (`results/split.json`).

**The run.** Arm E1 generated reports for all 104 Tier I cases from the findings object alone, through OpenRouter with the model identifier pinned to `anthropic/claude-sonnet-5` and verified against the model listing at startup. Every one of the 104 requests was served by that model, and each case's response, provider, token usage and prompt hash were logged. The findings consumed were those extracted from the ToothFairy3 reference labels, not from Stage 1's own predictions, so this is the oracle-perception condition: it isolates the verbaliser, and it is also the condition the planned ablation needs as its upper bound. The segmentation-to-findings-to-report path has not been run end to end.

Results of the run: 104 reports generated with no API errors, and **zero forbidden-content flags** across all of them. The forbidden-content check is a regular expression over the categories the system prompt bans by name (caries, periodontal, lesions, endodontic material, osseointegration, mucosal change and the rest), and it never fired. Median report length is 74 words, mean 79, range 44 to 154. The run consumed 940,888 prompt and 23,705 completion tokens.

One defect is worth recording rather than rounding away. 103 of the 104 responses terminated naturally; case P030 hit the 900-token completion cap and its report is truncated mid-sentence, because 795 of those 900 tokens went to reasoning rather than to output. That is one incomplete report in 104, it is a budget setting rather than a model failure, and it must be fixed before any scoring run, since a truncated report would depress logical recall for a reason that has nothing to do with the architecture.

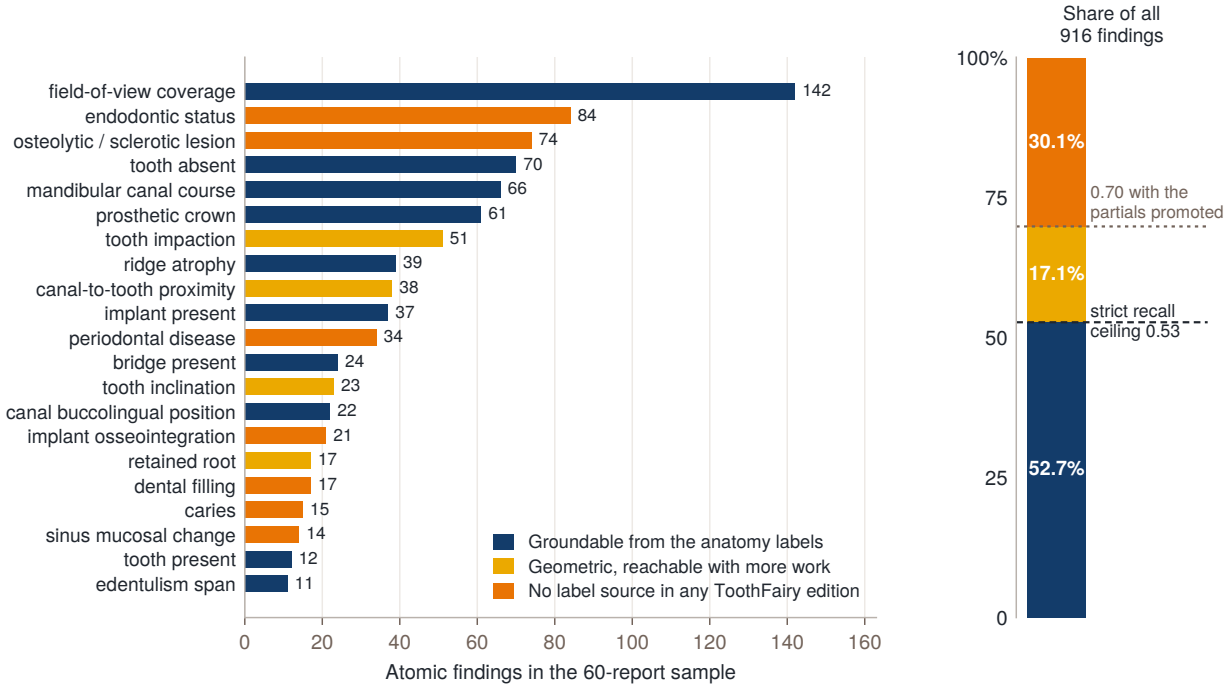



**Figure 4** Groundability of reference-report content. Left: the 21 canonical finding types with 11 or more occurrences, covering 872 of the 916 atomic findings decomposed from the 60-report sample, coloured by whether the ToothFairy3 label space can support them. Right: the adjudicated share of all 916 findings, and the recall ceiling it implies. Rendered from `results/report_vocab_sweep.json`. The per-type tags stored in that file aggregate to 54.9 / 15.0 / 30.1; the 52.7 / 17.1 / 30.1 figures quoted here and in the text are the adjudicated split recorded after borderline types were reviewed, which is why the two do not match and why the ceiling is described throughout as "about 0.53".

**What the output looks like.** Table 4 traces one case, F005, from findings to generated prose to the two surgeons who read the same scan. It shows the three behaviours the architecture is supposed to produce, and it shows them together. Findings that the labels support are verbalised and match a reader, including the distinction between a structure that is absent and a structure that was not scanned. Content with no label source is absent from the generated report, and that absence is correct even though it costs recall against the reference. And the two extractor limitations of Section 4.5 carry through visibly rather than being smoothed over, which is what makes the arm comparison meaningful: whatever E2 does with the same input can be attributed to E2.

The same case also shows why the human ceiling is not optional. The two readers disagree with each other on whether the condyles are in the scan and on the buccolingual course of the mandibular canal, one of them itemises endodontic treatment and individual fillings that the other does not mention, and where one names teeth by FDI number the other describes the same anatomy by quadrant.

## 4.8 Evaluation harness

The challenge harness, `evaluate.py` under `ODIN2026/ToothFairy4/evaluation/` in the organisers' repository, and the `radfact_lite` package it depends on (version 0.1.0, a 13 kB wheel) were read in full, and their behaviour is documented in the project workspace: the dental adaptation lives in three prompt directories (report-to-phrases, entailment, negative filtering), the challenge calls it with narrative-text mode and negative filtering disabled, and the default judge is `gpt-4o-mini`. The split logic exists and asserts no leakage. The floor baselines, the

**Table 4** One case, F005, traced end to end. Findings are from the typed object (Stage 2, extracted from the reference labels); generated text is verbatim from arm E1; reference text is verbatim from the two English reports for this case, by readers 29 and 14. Tooth numbers are in the clinical convention.

| Findings object | Generated (E1) | Reference reports |
|---|---|---|
| **Grounded, verbalised, and matched by a reader** | | |
| `tooth_absent` 46, 48 | "Absence of teeth 46 and 48." | R29: "Absence of teeth 46 and 48." |
| `crown_present` at 36 | "Prosthetic crown on tooth 36." | R29: "…with an overlying prosthetic crown." R14: "…prosthetic reconstruction already performed." |
| `fov_coverage` condyles, coverage `excluded` | "Excluded from the acquisition and not visible." | R14: "Excluded from the acquisition and not visible." R29: "partially included in the scan." |
| `fov_coverage` maxillary sinuses, coverage `partial`, bilateral | "Maxillary sinuses partially included in the acquisition bilaterally." | R29: "Maxillary sinuses partially included in their most caudal portion, clear." |
| **Version-1 extractor limitations, carried through faithfully** | | |
| two `implant_present` findings, *both* at 26; `bridge_present` abutments 25 and 28, pontics 26 and 27 | "Endosseous implant in position 26. Fixed bridge prosthesis with abutments on teeth 25 and 28…" | R29: "Endosseous implants in positions 26 and 27 with an overlying crown." |
| `canal_position` centro-buccal, both sides | "…in a centro-buccal position." | R29: "predominantly lingual course." R14: *[states course, not position]* |
| **No label source: correctly absent from the generated report** | | |
| none | *not written* | R29: "Endodontic treatment of tooth 36 is adequate." |
| none | *not written* | R29: "Fillings on teeth 47, 37, and 38." |
| none | *not written* | R14: "No definite osteolytic or osteocondensing lesions." |
| none | *not written* | R14: "two endosseous implants …radiologically well osseointegrated." |

per-reader-pair ceiling, the permutation study and the first end-to-end score have not been run.

### 4.9 What has not been measured

For the avoidance of doubt: no RadFact-style precision, recall or F1 for any generator, on any tier; no human-agreement ceiling; no constant-report or retrieval baseline score; no FDI identity accuracy and no confusion matrix; no per-instance confidence or calibration curve; no measurement of any kind on subset A, which has no voxel labels; no 3D-exclusive share of finding types; and no result for arm E2, which has not been started. Nothing in this report has been run through the full segmentation-to-findings-to-report path. Each of these has an owner and a next step in Section 6.

## 5 Discussion

**The segmentation recipe is reproduced, and that is where its importance ends.** The training run tracked the published result closely enough to say the setup, environment and data plumbing are right, which was its purpose. The validation pass then says something more useful and less flattering: 0.7246 over 46 classes and all cases, 0.8806 once absent structures stop being counted as failures. Neither number tells us the model is good at what the pipeline needs. Dice rewards boundary overlap; the report needs per-structure identity and presence. A crown segmented at

0.69 Dice but attributed to the correct tooth yields a perfect finding, while a tooth at 0.95 Dice labelled 46 instead of 36 yields a confident, clinically dangerous one. That asymmetry is why FDI identity accuracy, which is still unmeasured, is the metric that will decide whether Stage 1 is good enough, and why the presence-detection figures matter more here than the Dice column next to them. Finding 2,050 of 2,061 present tooth instances, and every non-tooth class in every case where it existed, is what licenses building on top of this model.

**Two aggregations of the same numbers say opposite things.** The upper arch scores 0.613 over all cases and 0.906 over cases where the tooth is present. Reported alone, the first invites the conclusion that the model is confusing maxillary teeth, which for this project would be close to fatal. The second shows the arches nearly level, with the residual gap at 0.032. The difference between the two aggregations is entirely accounted for by the maxilla being outside the acquired volume in 66% of cases. We think the general form of this is worth stating: in a corpus with variable acquisition extent, a class mean over all cases confounds "the model missed it" with "it was never imaged", and the two have opposite implications for a report. Two honest numbers, Dice given presence and presence detection as its own metric, replace one misleading average, and they map directly onto what a report says, because whether the sinus is in the scan at all is itself a reportable statement. Coverage is also the single most frequent thing surgeons write about in this corpus: it is the largest finding family in the 60-report sample, at 142 of 916 atomic findings.

**The mirrored laterality convention is the result of this period that would have been most expensive to miss.** It is a silent failure by construction. Both editions are internally coherent, the volumes align bitwise, the segmentation model trains to a competitive Dice, and every downstream number looks healthy, because predictions and ground truth share the convention. Nothing surfaces until the pipeline emits a sentence that a clinician could read, and at that point every laterality in it is wrong. The transferable lesson is not that this dataset has a quirk; it is that geometric verification and semantic verification are different tests, and that reusing labels across dataset editions needs an adjudication against something a clinician could falsify. We had already learned the weaker version of this lesson once, when the NIfTI headers implied an anterior-posterior flip and the true transform also flipped superior-inferior. Both checks now run per case rather than as a corpus-wide constant, which is also how the two ToothFairy3 cases stored outside the convention were caught rather than silently mislabelled.

**Reader style is a variable, not noise.** A 3.5-fold spread in report length and a 4.2-fold spread in tooth references between readers means a single pooled "human ceiling" would be an average over readers who are not describing the same amount of the scan. It also means a generator trained on all reports learns a mixture of styles that no single reader would produce. Reporting the ceiling per reader pair, and treating reader identity as a conditioning variable or a normalisation step, follows directly. F005 makes it concrete: its two readers disagree about whether the condyles are in the scan, disagree about the buccolingual course of the mandibular canal, one itemises endodontic treatment and individual fillings that the other does not mention, and where one names teeth by FDI number the other describes the same anatomy by quadrant. The same fact softens a comparison we had planned to lean on: ToothFairy4 reports are clinician-authored, but under a fixed taxonomy, so "free narrative versus rule-generated" becomes "clinician-authored under a taxonomy versus rule-generated from annotations".

**The first score will be low, and the ceiling is now a number rather than an excuse.** This phase grounds only geometrically derivable findings. Pathology has no labels, so the generator is never

told about periodontitis, caries, periapical lesions or fractures and is instructed not to invent them. Recall against the references is therefore bounded by the groundable share, which we measured at 52.7%, rising to about 0.70 if the geometric partials are promoted. The value of having measured it before the first score is that "we scored 0.4" becomes "we scored 0.4 against a reachable 0.53", which is a different claim about a different system. It also tells us where the remaining effort is worth spending: endodontic status alone is 84 of 916 sampled findings and has attributed pulp as its only adjacent labelled signal, which is now shown to be recoverable at 99.96% accuracy, so it is the most promising single extension of the groundable set. The two floor baselines must still run on the same day as the first score, and the constant-report baseline in particular will look strong on the negated lesion boilerplate that no grounded system can produce. That is the reason for running it, not an argument against it.

**Zero forbidden-content flags is a real result and a narrow one.** Across 104 generated reports, the regular expression over the banned categories never fired. That is worth stating because the failure mode it tests for, a fluent model filling gaps with clinically plausible pathology, is the specific risk that motivates putting a deterministic stage in front of the language model. But it is a test of vocabulary, not of truth. It does not show that the tooth numbers are right, that the laterality is right, or that a finding was not dropped, and it was run on an arm that received findings derived from reference labels rather than from the segmentation model. What it does establish is that the constraint mechanism works well enough that the remaining errors will be errors of grounding rather than of invention, which is the class of error the metric is designed to see.

**Two engineering decisions, now made rather than deferred.** The first is the training schedule. Stopping at epoch 542 was a decision about what the remaining compute buys, and Figure 3 is the evidence it rested on: the marginal Dice would have landed almost entirely on anatomy this corpus barely contains. The alternative use of that compute, a second fold giving both an ensemble and out-of-fold predictions for the 426 training cases, was considered and declined for this phase; a single fold stands.

The second is the few-shot design for arm E1. Using reference reports as exemplars is the obvious choice and the wrong one, because a real reference report demonstrates exactly the ungroundable claims the system prompt forbids. The exemplars are therefore hand-written from two training-pool cases' own findings, which keeps the demonstration and the constraint pointing the same way.

## 6 Limitations and Future Work

Each limitation is paired with the concrete next step that is already scheduled, in the order the team intends to take them.

1. **Nothing has been run through the full pipeline.** Stage 1 is validated on 106 cases, Stage 2 has run over 530 label maps, and Stage 3 has produced 104 reports, but the reports were generated from findings extracted from the reference labels, not from Stage 1's predictions. That is the oracle-perception condition, which the ablation needs, and it is not an end-to-end result. *Next:* run the extractor on the Stage 1 predictions for the 104 Tier I cases and generate the same reports again, so that the oracle and predicted conditions can be reported side by side and the cost of perception error is a measured quantity rather than an argument.
2. **Dice is not the metric the pipeline needs, and the metric it does need is still unmeasured.** *Next:* FDI identity accuracy with its confusion matrix on the 106 held-out cases, applying

the laterality remap of Section 4.3 before comparing anything to a report, followed by per-instance confidence and a calibration curve. This is the highest-value open item in the perception stream.

3. **No end-to-end score, baseline or ceiling has been run.** *Next:* in this order, fix the completion-token cap that truncated one of the 104 reports, run the constant-report and retrieval baselines, compute the per-reader-pair agreement on the 367 dual-read cases in Italian and English, then score arm E1, and score arm E2 on identical inputs once it exists. Every number is quoted against the 0.53 recall ceiling.
4. **Arm E2 does not exist yet.** The comparison the project is built around, a frontier API model against a fine-tuned open-weight model on byte-identical input, currently has one arm. *Next:* LoRA-adapt Gemma 4 on the 423-case training pool using the same `render_findings()` text the API arm receives, and keep the exemplar policy identical so the two arms differ only in the model.
5. **Three structure groups are genuinely weak, and one of them matters clinically.** The mandibular incisive canals reach 0.63 and 0.61 and the lingual canal 0.69 under present-only aggregation, while being present in 93 to 98% of cases, so these are not field-of-view artefacts. The incisive canal is relevant to interforaminal implant planning. The restoration classes (bridge 0.68, crown 0.69, implant 0.77) are the other group. *Next:* decide whether findings on these structures are emitted with a lower confidence, gated out, or reported with an explicit reliability caveat, and measure site attribution separately from boundary overlap, since attribution is what the report actually asserts.
6. **Only fold 0 was trained.** There is no ensemble and no out-of-fold prediction for the 426 training cases, so those cases can never be scored end to end without leakage. The decision to let a single fold stand has been made for this phase and is recorded in Section 5; it should be revisited if the predicted-input condition in item 1 turns out to be much worse than the oracle condition.
7. **Subset A has no voxel labels in any edition.** Perception quality on the held-out centre is unmeasurable by construction, and Tier II therefore cannot be scored at all until Stage 1 predicts there. *Next:* either annotate a small manual sample or state the limitation explicitly in the held-out-centre results; the choice belongs to whoever owns that tier.
8. **Pathology is not groundable, so recall is capped at about 0.53.** *Next:* the cheapest extension of the groundable set is endodontic status, the largest ungroundable cluster at 84 of 916 sampled findings, for which attributed pulp is now validated at 99.96%. Decide whether to emit pulp-derived findings, and if so with what qualifier, since the presence of a pulp chamber is not the same statement as the adequacy of an endodontic treatment.
9. **Two cases were refused and two extractor limitations are known.** S0002 and S0015 failed the orientation self-check and are excluded rather than repaired. Two adjacent implants can win the same interpolated tooth slot, and canal buccolingual position is a single whole-body average rather than a per-segment profile, so "predominantly lingual with buccal emergence" collapses to one descriptor. *Next:* joint assignment for restoration sites and a segmented canal profile in extractor version 2; recover the two refused cases by downloading their images and running `verify()`.
10. **The English references are translations.** About 10% of the training translations were clinician-reviewed. *Next:* compute the primary ceiling on the Italian pair and report the English one alongside; do not claim multilingual generation.
11. **The API generator arm is not contamination-free.** A frontier model may have seen public

dental text. The model identifier is pinned, every request is logged with the model that served it, and the system prompt is stored with its hash, so the caveat can be stated precisely; it cannot be removed. The fine-tuned open-weight arm is the clean comparison.

12. **Open governance items.** Publication permission from the data providers is reported cleared but the evidence is not yet filed; the 622 versus 630 case-count discrepancy against the design document is unexplained; the reader-identifier convention is inferred, not confirmed by the organisers, and a per-reader analysis built on a misread filename convention would invalidate the human ceiling; ethics approval and reader recruitment for the blinded study (with a lead time landing around November 2026) have not started; and compute beyond Ibex is undecided. *Next:* file the permission evidence, raise the count and reader-id questions with the organisers, and open the ethics application.
13. **Deferred: the 3D-dependency partition.** The lead contribution of the eventual paper, classifying finding types by whether they are derivable from a projection, is deferred for the current sprint. The 29-type canonical vocabulary of Section 4.6 is the input it needs and now exists. *Next:* partition those types by 3D dependency and book the clinician time; the pre-registered thresholds (25% / 15%) are already fixed.

Promising extensions once the baseline exists include the uncertainty-aware findings gate as the ablation axis, the medical versus general pre-training question for the verbaliser (Gemma 4 against MedGemma at matched size, precision reported separately from recall), and cross-corpus testing on MMDental if its text proves to be narrative.

## 7 Conclusion

The project set out to do the ToothFairy4 task, narrative report generation from maxillofacial CBCT, with an architecture that keeps the language model away from the image: segmentation, then a typed findings object, then verbalisation. In this reporting period all three stages were built and measured individually, and the pieces of evidence that decide whether the architecture is worth pursuing were collected before any headline number was chased.

The corpus was verified file by file (622 cases, 1,000 reports per language, 367 dual-read cases, 16 inferred readers whose style varies several-fold). The ToothFairy3 labels were aligned to the ToothFairy4 volumes bit-exactly, and their laterality convention was found to be mirrored relative to the reports, a silent failure that would have swapped left and right in every report the system generates and that no geometric check would have caught. The segmentation model was trained on our own case-level split and validated on 106 held-out cases: 0.7246 mean Dice over 46 classes and all cases, 0.8806 restricted to cases where the structure was actually inside the field of view, with the apparent upper-arch weakness shown to be truncation rather than tooth misidentification and presence detection at 2,050 of 2,061 tooth instances. The findings extractor ran over the labelled corpus and produced 21,028 schema-valid findings across 530 of 532 cases, refusing two rather than guessing at them, and recovered the tooth attribution that the training label merge had destroyed at 99.96% accuracy. A prompt-constrained generator turned those findings into 104 narrative reports with zero forbidden-content flags, one of them cut short by a completion-token budget. And the ceiling this design implies was measured rather than assumed: 52.7% of what surgeons actually write can be grounded in anatomy labels, so recall is capped near 0.53.

Two things a reader should take away. No report-quality score has been measured yet, and nothing has been run through the full segmentation-to-findings-to-report path; the 104 reports were generated from reference labels, which is the oracle-perception condition the ablation needs and not an end-to-end result. And when the first score is read it will be read next to a

constant-report baseline, a retrieval baseline, a per-reader-pair human agreement figure and that 0.53 ceiling, because a report-quality number without those four is not interpretable. Producing them is the milestone that comes next.

## References


[1] Wenbo Zhou et al. "A deep learning framework for automated dental segmentation and diagnostic report generation from cone-beam computed tomography". In: *Head & Face Medicine* 21.1 (2025), p. 80. DOI: 10.1186/s13005-025-00555-0.

[2] Federico Bolelli et al. *ODIN 2026 Challenges: Multimodal Text Report Generation for Oral and Dental Image Analysis*. Zenodo, challenge design document. 2026. DOI: 10.5281/zenodo.19727377.

[3] Federico Bolelli et al. "Multi-structure segmentation in CBCT volumes: The ToothFairy2 challenge". In: *Medical Image Analysis* (2026). Article S1361841526001647.

[4] Zhi Qin Tan et al. "U-Mamba2: Scaling State Space Models for Dental Anatomy Segmentation in CBCT". In: *arXiv preprint arXiv:2509.12069* (2025).

[5] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations (ICLR)*. 2022.

[6] Federico Bolelli et al. "Segmenting the Inferior Alveolar Canal in CBCT Volumes: The ToothFairy Challenge". In: *IEEE Transactions on Medical Imaging* (2024). DOI: 10.1109/TMI.2024.3523096.

[7] Fabian Isensee et al. "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

[8] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. "CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2024, pp. 476–486. DOI: 10.1007/978-3-031-72390-2_45.

[9] Fan Bai et al. "M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models". In: *arXiv preprint arXiv:2404.00578* (2024).

[10] Pedro R. A. C. Bassi et al. "RadGPT: Constructing 3D Image-Text Tumor Datasets". In: *arXiv preprint arXiv:2501.04678* (2025).

[11] Juampablo Heras Rivera et al. "BTReport: Deterministic Feature Extraction for Interpretable Brain Tumor Radiology Report Generation". In: *Medical Imaging with Deep Learning (MIDL)*. arXiv:2602.16006. 2026.

[12] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of ACL*. 2002, pp. 311–318.

[13] Weike Zhao et al. "RaTEScore: A Metric for Radiology Report Generation". In: *Proceedings of EMNLP*. 2024, pp. 15004–15019.

[14] Shruthi Bannur et al. "MAIRA-2: Grounded Radiology Report Generation". In: *arXiv preprint arXiv:2406.04449* (2024).

[15] Sophie Ostmeier et al. “GREEN: Generative Radiology Report Evaluation and Error Notation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 374–390. DOI: 10.18653/v1/2024.findings-emnlp.21.

[16] Alyssa Huang et al. “FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores”. In: *arXiv preprint arXiv:2405.20613* (2024).

[17] Mohammed Baharoon et al. “CRIMSON: A Clinically-Grounded LLM-Based Metric for Generative Radiology Report Evaluation”. In: *arXiv preprint arXiv:2603.06183* (2026).

[18] Feiyang Yu et al. “Evaluating Progress in Automatic Chest X-ray Radiology Report Generation”. In: *Patterns* 4.9 (2023), p. 100802.

[19] Justin Xu et al. “RadEval: A Framework for Radiology Text Evaluation”. In: *arXiv preprint arXiv:2509.18030* (2025).

[20] Oishi Banerjee et al. “ReXamine-Global: A Framework for Uncovering Inconsistencies in Radiology Report Generation Metrics”. In: *arXiv preprint arXiv:2408.16208* (2024).

[21] Xiaohan Li et al. “PanDent: Toward Comprehensive Tooth-Level Structure-Language Consistency in Dental Radiology”. In: *arXiv preprint arXiv:2607.27378* (2026).

[22] Yuxuan Fan et al. “OralGPT-Plus: Learning to Use Visual Tools via Reinforcement Learning for Panoramic X-ray Analysis”. In: *arXiv preprint arXiv:2603.06366* (2026).

[23] Huimin Xiong et al. “IOSVLM: A 3D Vision-Language Model for Unified Dental Diagnosis from Intraoral Scans”. In: *arXiv preprint arXiv:2603.16781* (2026).

[24] Ryutaro Tanno et al. “Consensus, Dissensus and Synergy Between Clinicians and Specialist Foundation Models in Radiology Report Generation”. In: *arXiv preprint arXiv:2311.18260* (2023).