

From Voxels to Views to Reports: A Segmentation Guided VLM Pipeline for CBCT Report Generation

Shivam Singh^{1,2} Chen Lu^{1,2} Benedikt Sagl^{1,2}

¹University Clinic of Dentistry, Medical University of Vienna, Vienna, Austria

²Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Vienna, Austria

{shivam.singh, chen.lu, benedikt.sagl}@meduniwien.ac.at



Introduction

Dental cone-beam computed tomography (CBCT) provides rich 3D anatomical information, yet clinical report generation remains largely manual. Direct 3D vision-language model (VLM) processing is computationally expensive, and volumetric compression can obscure fine anatomical detail. To address this, we leverage anatomical segmentation as a spatial index to derive targeted, clinically meaningful 2D evidence.

Main contribution: An anatomy-grounded, schema-constrained VLM pipeline that combines multi-view visualization, structured visual question answering (VQA), and rule-based report assembly.

Challenge Setup

- Task:** Generate a maxillofacial and surgical-planning report from a single preprocessed CBCT volume.
- Dataset:** 622 CBCT volumes from four multi-center sub-datasets, each with one to three radiological reports.
- Evaluation:** 40 cases for local validation and 50 hidden cases from an independent center for official testing.
- Metrics:** RadFact F1 for clinical factual agreement; BLEU-4 and METEOR for textual similarity.

Methodology

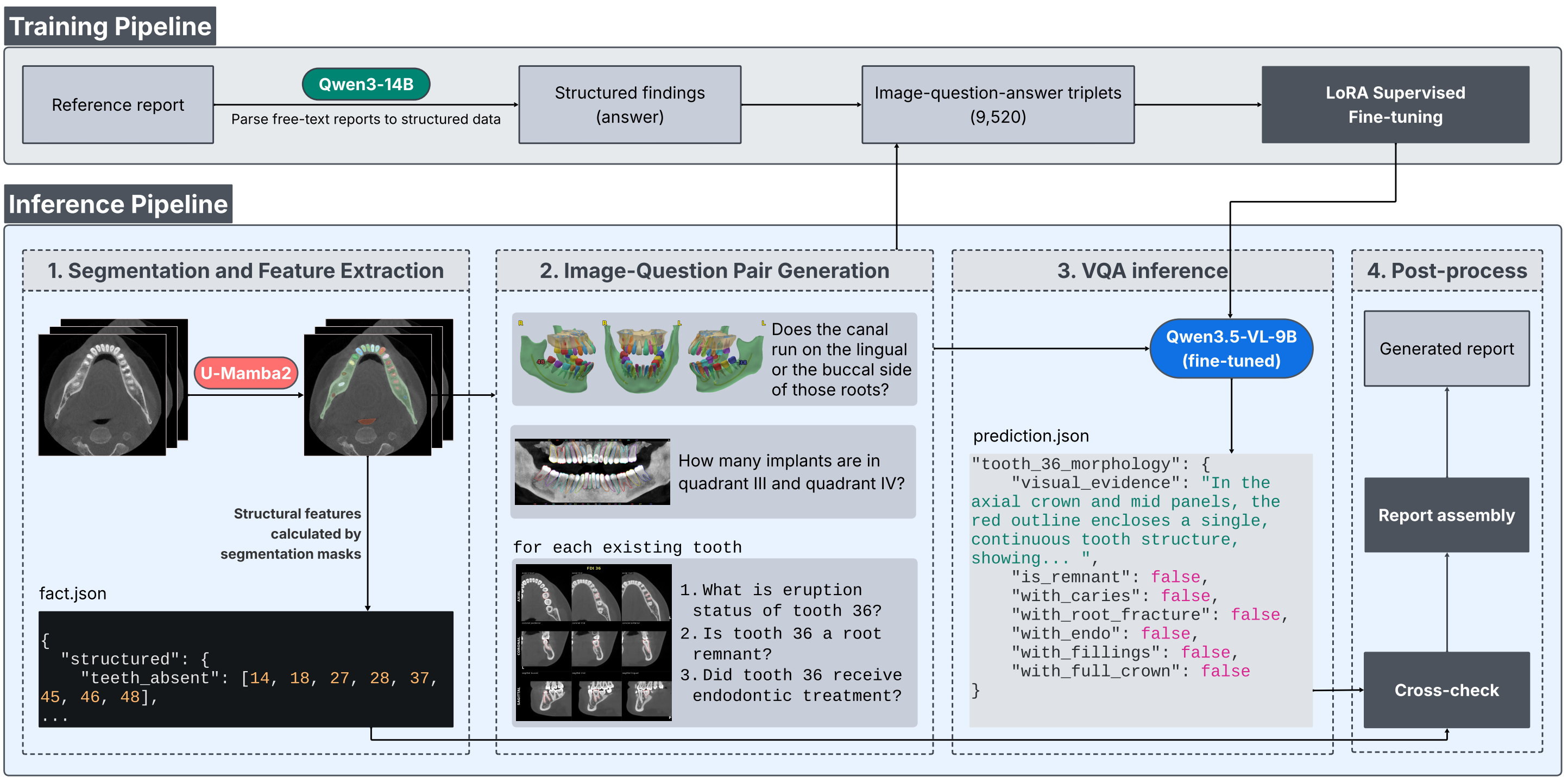


Fig. 1. Overview of the proposed framework.

- Segmentation:** U-Mamba2 generates a dense anatomical parse used as a spatial index.
- View Generation:** Masks guide 3D renderings, panoramic reconstructions, and local tooth/sinus multi-planar views.
- Vision-Language Model:** Qwen3.5-VL-9B is fine-tuned with low-rank adaptation (LoRA) on 9,520 image-question-answer triplets.
- Report Generation:** Logical constraints, source selection, and confidence filtering refine predictions before report assembly.

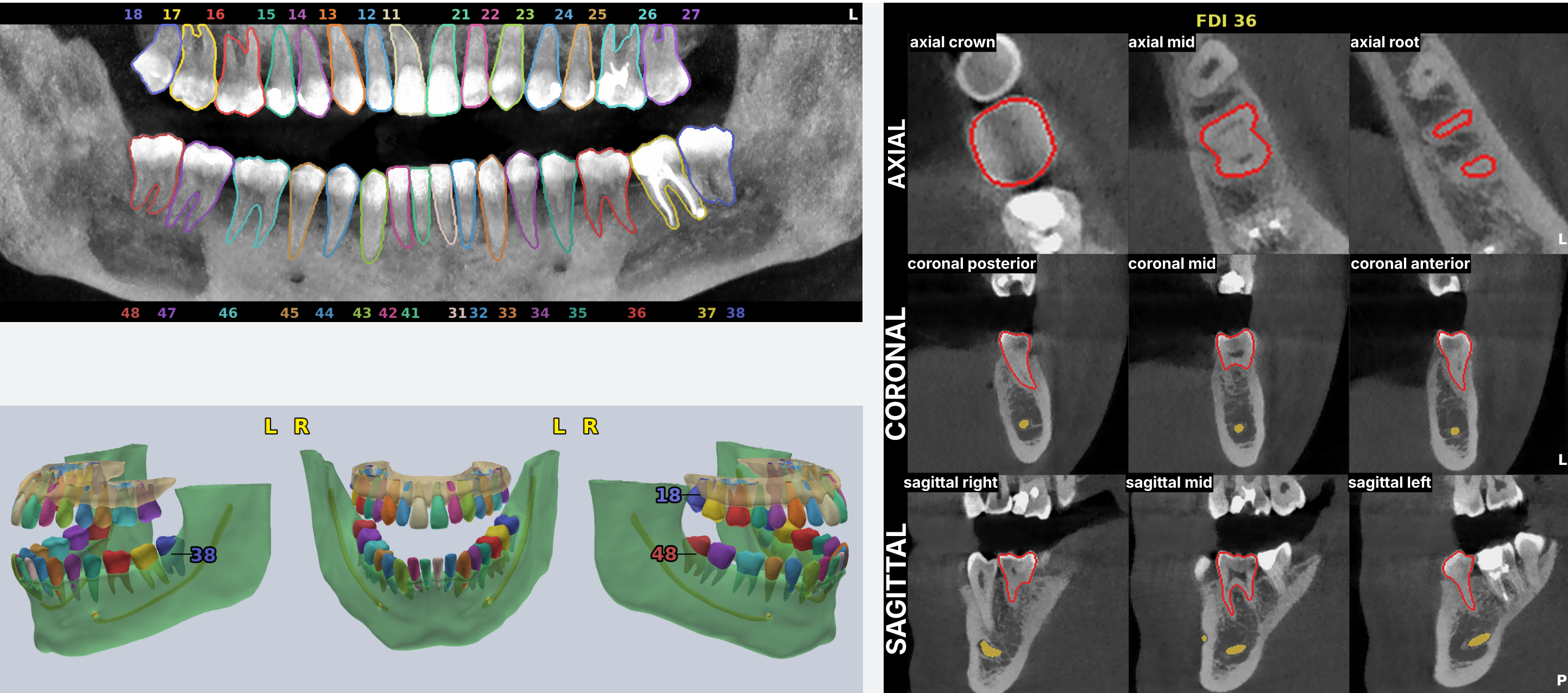


Fig. 2. Visual inputs: Panoramic reconstruction (top left), 3D multi-perspective anatomical renderings (bottom left), and localized orthogonal multi-slice composites (right).

Results

- LoRA fine-tuning improves local **RadFact F1** from **0.353** to **0.444** and **Final Score** from **0.336** to **0.409**.
- Segmentation alone achieves a **Final Score** of **0.277**, showing the benefit of VLM-based reasoning.
- On the hidden test set, the pipeline achieves **RadFact F1 = 0.392** and **Final Score = 0.361**.
- Finding-level F1 is highest for impaction (**0.813**), absent teeth (**0.762**), and endodontic treatment (**0.741**), while fine-grained findings such as caries, root remnants, and post-and-core remain challenging (**F1 $\leq$ 0.080**).

Method	BLEU-4	METEOR	RadFact			Final
			Precision	Recall	F1	
<i>Local validation (N = 40)</i>						
Only segmentation	0.1435	0.3006	0.3443	0.3464	0.2907	0.2770
Qwen3.5-9B	0.1506	0.3824	0.4318	0.4023	0.3530	0.3357
Qwen3.5-9B + LoRA (Ours)	0.1553	0.3848	0.5319	0.4762	0.4442	0.4094
<i>Official hidden test (N = 50)</i>						
Qwen3.5-9B + LoRA (Ours)	0.1476	0.3269	0.5343	0.3316	0.3924	0.3614

Table 1. Quantitative comparison on local validation and official hidden test sets.

Finding	Precision	Recall	F1
Impaction	0.780	0.850	0.813
Absent teeth	0.690	0.850	0.762
Endodontic	0.880	0.640	0.741
Fixed bridges	1.000	0.550	0.710
Canal-adjacent teeth	0.690	0.690	0.690
Implants	0.680	0.530	0.596
Crown	0.290	0.310	0.300
Fillings	0.300	0.240	0.267
Post-and-core	0.060	0.120	0.080
Caries	0.140	0.040	0.062
Root remnant	0.050	0.080	0.062

Table 2. Finding-level precision, recall, and F1 on local validation set.

Generated report

Mandible: partially included in the scan volume. Right mandibular canal with a regular course, predominantly lingual. Left mandibular canal with a regular course, predominantly lingual. Absence of teeth 36, 37, 38, 45, 46, and 48, with associated bone atrophy. Endodontic treatment involving teeth 41. Endosseous implants in position 45, 46. No definite osteolytic or osteocondensing lesions.

Maxilla: partially included in the scan volume. Absence of teeth 15, 18, 27, and 28, with associated bone atrophy. Endodontic treatment involving teeth 25 and 26. Conservative composite restoration(s) on teeth 16 and 17.

Maxillary sinuses: partially included in the acquisition volume, normally aerated.

Mandibular condyles: not included in the scan volume.

Reference report

Mandible: mild atrophy involving the left molar region, characterized by edentulism. Absence of e.d. 46-47-48. Endodontic treatment of e.d. 41, adequate, with endodontic material extending beyond the root apex. Endosseous implants in positions 45-46-47. Prosthetic crown on e.d. 35. Mandibular canal with a predominantly lingual course.

Maxilla: partially included in the scan. Absence of e.d. 15-16-27-28. Endodontic treatment of e.d. 25 and 26, adequate on all roots. Prosthetic crown on e.d. 25. Filling on e.d. 26. Radiolucent area at the apices of the buccal roots of e.d. 26 and at the root apex of e.d. 25, compatible with granulomas. Radiolucent area at the level of the roots of e.d. 16 (absent), compatible with a periodontal pocket.

Maxillary sinuses included in the most caudal portion: mild mucosal hyperemia of the left maxillary sinus.

Mandibular condyles partially included in the scan.

Fig. 3. Generated versus reference reports: green indicates agreement, red conflicting findings, and yellow missed reference findings.

Conclusion

Anatomical segmentation serves as a spatial index that converts volumetric CBCT data into localized visual evidence. Combined with schema-constrained VQA, LoRA fine-tuning, and post-processing, our pipeline substantially improves automated report generation. Nevertheless, subtle, rare, and visually ambiguous findings remain the primary challenge for future work.

Acknowledgments

This work was developed and evaluated within the ODIN 2026 challenge cluster at MICCAI 2026, comprising the STS, ToothFairy, and Bite2Text challenges.