

From Voxels to Views to Reports

A segmentation-guided VLM pipeline for CBCT report generation

Shivam Singh, Chen Lu, Benedikt Sagl (Team Canal-di-la)

University Clinic of Dentistry, Medical University of Vienna

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

THE PROBLEM

The input is volumetric. The output is clinical prose.

The specificity about dental CBCT requires:

1. Accurately identifying tooth positions
2. Balancing the overall view with tiny details
3. Understanding three-dimensional spatial relationships
4. Knowledge about dentistry-specific terms and imaging features

Current state of the art VLM often lacks domain specific knowledge

Why direct volume-to-report is hard ?

1. 3D CBCT volumes are large and expensive to represent directly
2. Feeding the 3D volume directly usually requires compressing, leading to loss of detail
3. Uniform slice sampling does not identify the relevant tooth or structure

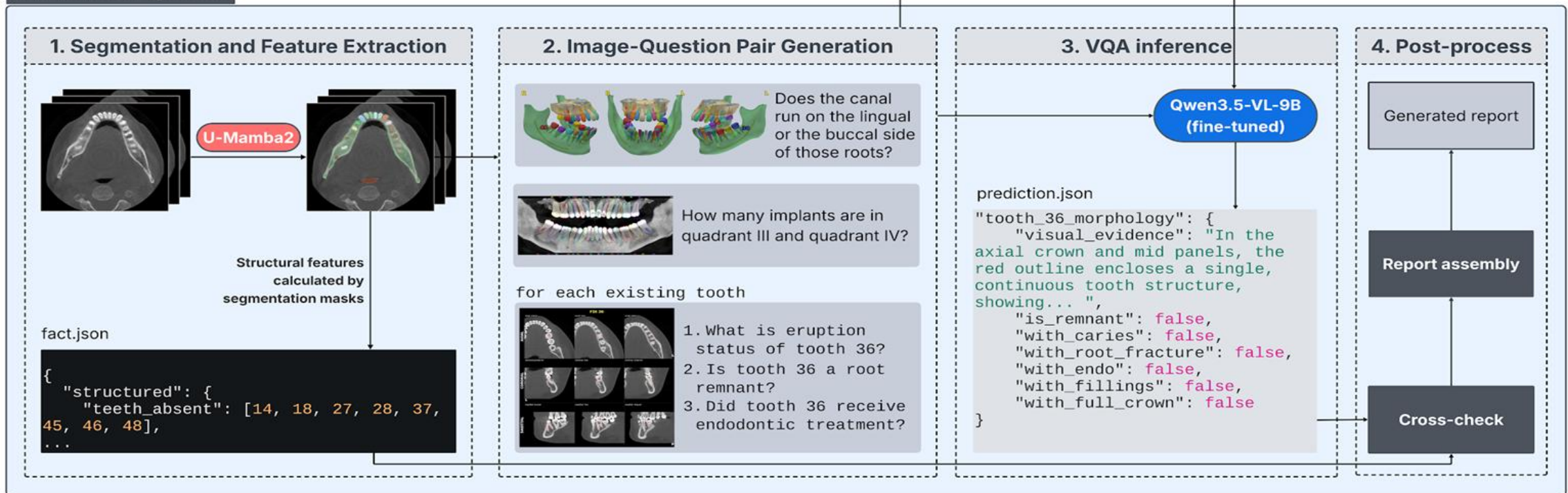
How do we turn one volume into the right visual evidence?

Methodology Overview

Training Pipeline



Inference Pipeline



STAGE 1: Segmentation

Build the spatial index before asking clinical questions

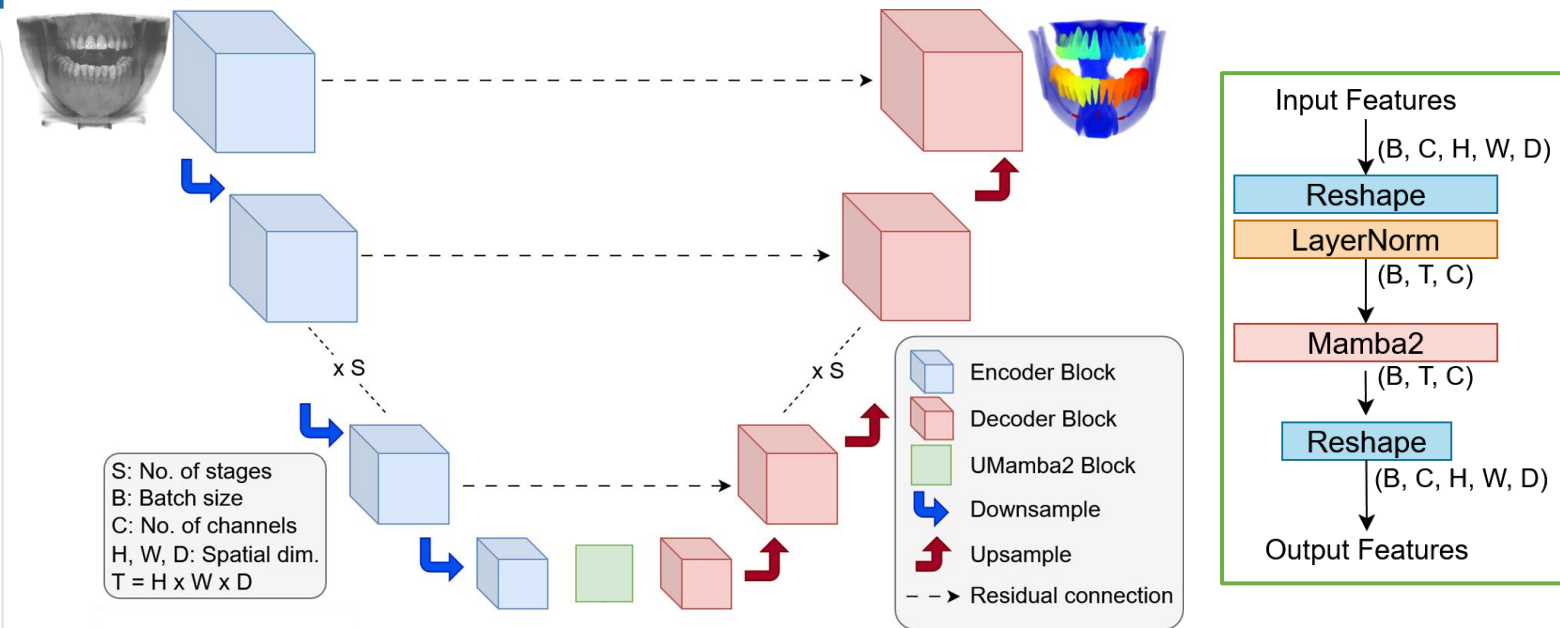
Multiclass segmentation

U-Mamba2 produces a dense anatomical parse.

77 ToothFairy3 classes remapped to 47 task-relevant classes.

Teeth, pulp, bone, mandibular canals, and sinuses become addressable.

Masks provide geometry for cropping, outlining, labels, and cross-checks.

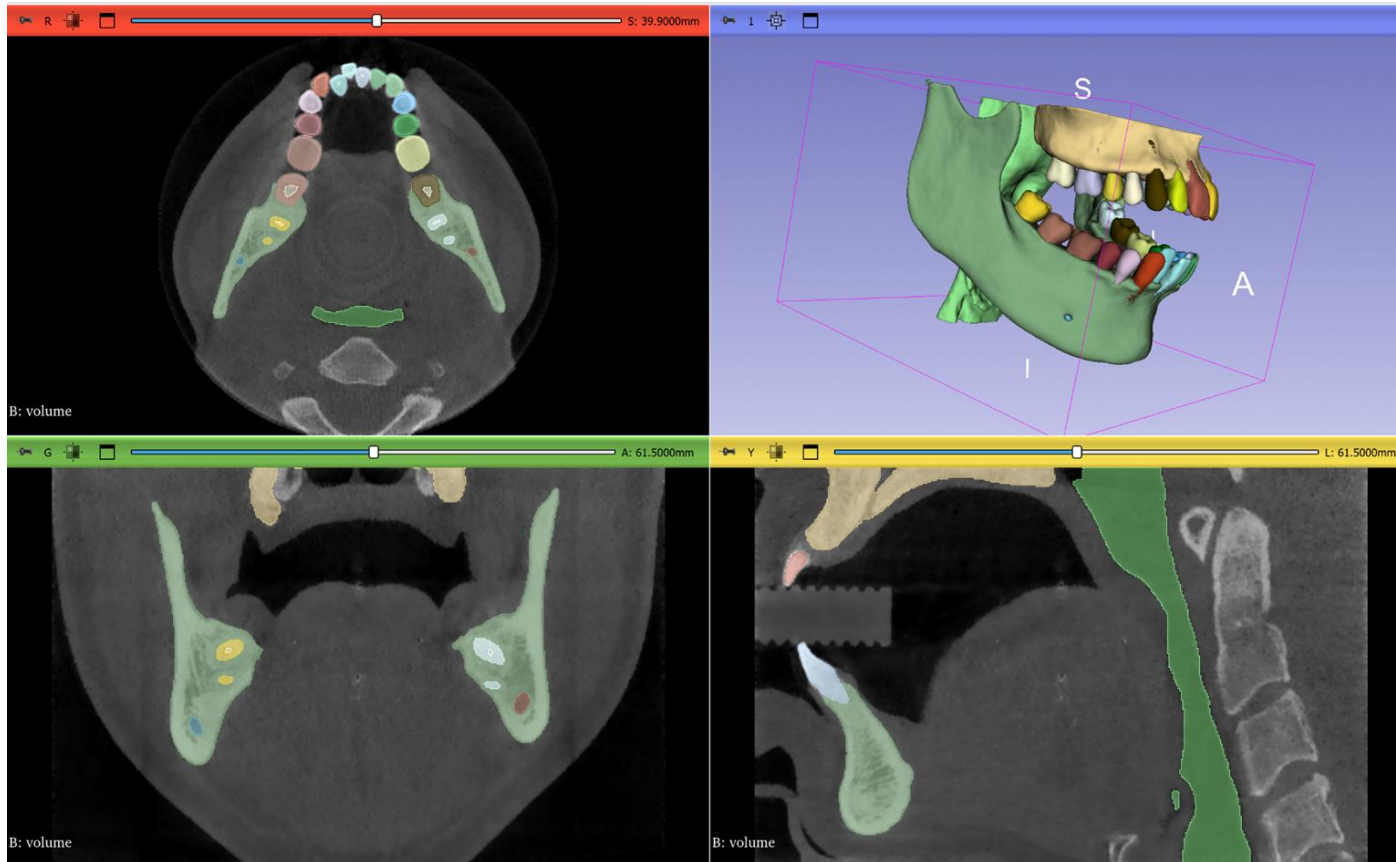


U-mamba2

Tan, Z.Q. et al. (2026). U-Mamba2: Scaling State Space Models for Dental Anatomy Segmentation in CBCT. LNCS 16473, Springer.

STAGE 2: Image-Question Pairs

When a dentist is reading a CBCT...

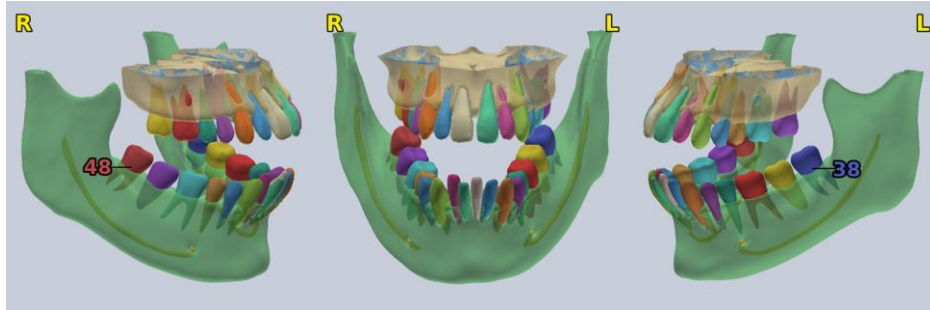


- Scrolling through the **multi-view slices** for general idea
- Go to specific slice **with questions** in head
- 3D construction and panoramic for **easy overview**

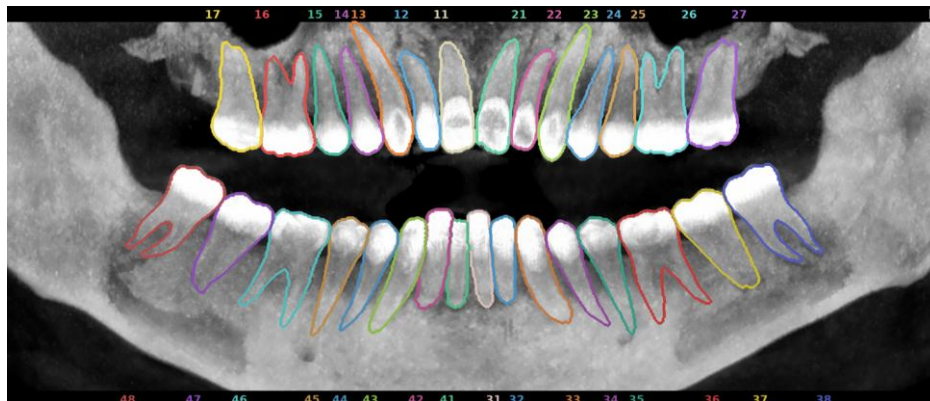
Generate 2D visual evidence mimicking what clinicians use for reading dental CBCT

STAGE 2: Image-Question Pairs

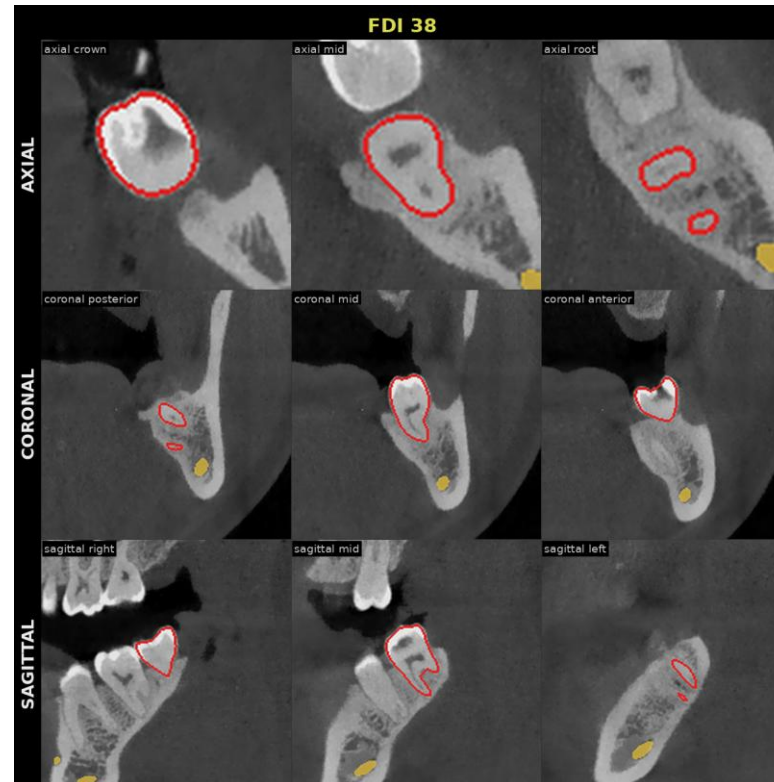
Ask the right question of the right view



Q: Is 48 impacted? If so, which direction?



Q: Going tooth by tooth, what is the single most notable finding you can identify?



Q: What is the eruption state of 38?

3D renderings

scope and spatial info

- Gaussian/Laplacian smoothing
- Marching-cubes isosurface

Panoramic

dentition status

- Cubic spline fitting
- Curved-slab MIP
- CLAHE intensity blend

Multi-view composites

tooth and sinus detailed questions

- Percentile slice sampling
- ROI cropping

With outlines and annotations to guide VLM

The format of the answer matters



Example structured outputs

```
"tooth_37_morphology": {  
  "visual_evidence": "Tooth 37 shows significant restorative work. In the axial 'axial mid' cut, the crown appears filled with material of density higher than dentin but not as high as a metal post (likely a core material). A very bright, metallic density representing a post is clearly visible within one of the root canals extending to the apex. In the coronal and sagittal views (specifically 'coronal anterior/mid' and 'sagittal mid/left'), the bright post is confirmed extending the full length of the canal within the root. ... No dark areas indicative of caries are visible within the outlined tooth substance.",  
  "is_remnant": false,  
  "with_caries": false,  
  "with_root_fracture": false,  
  "with_endo": true,  
  "with_fillings": false,  
  "with_full_crown": true,  
  "with_post_and_core": false  
}
```

STAGE 4: POST-PROCESS

3 post-process rules

	N	Precision / Recall					Decision
		Masks	3d	Panoramic	Composite	Post-process	
Impacted teeth	33	—	0.62/0.48	0.67/0.06	0.79/0.82	0.78/0.85	✓
Absent teeth	232	0.55/0.85	0.54/0.94	0.53/0.67	—	0.69/0.85	✓
Endodontic	90	—	—	0.52/0.13	0.85/0.74	0.88/0.64	✓
Fixed bridges	11	1.00/0.55	—	1.00/0.09	—	1.00/0.55	✓
Canal-adjacent	26	0.69/0.69	0.70/0.27	—	0.17/0.32	0.69/0.69	✓
Implants	32	0.68/0.53	—	0.44/0.12	—	0.68/0.53	✓
Fillings	74	—	—	—	0.31/0.26	0.30/0.24	✓
Crown	49	0.33/0.29	—	—	0.32/0.32	0.29/0.31	—
Post and core	17	—	—	—	0.08/0.09	0.06/0.12	—
Caries	26	—	—	0.00/0.00	0.14/0.05	0.14/0.04	—
Root remnants	13	—	—	—	0.05/0.17	0.05/0.08	—

Constraint rule

Remove logical impossible findings
e.g.: A tooth cannot both be absent and with endodontic treatment at the same time

Selection rule

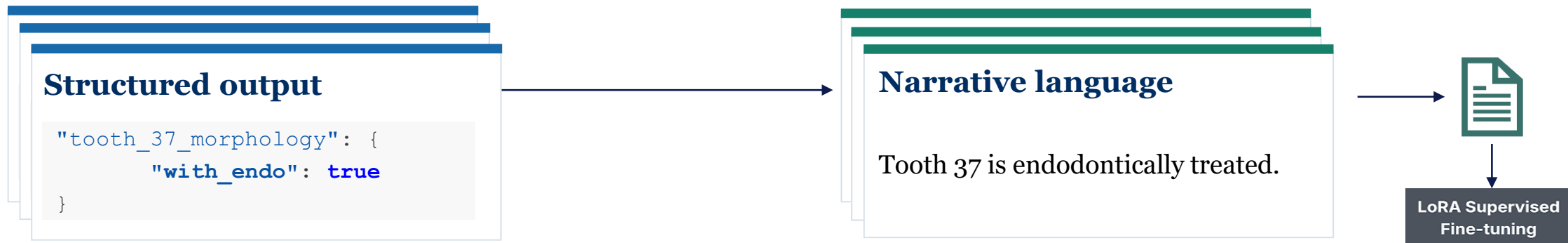
Best source wins

Silence rule

Unreliable findings (precision < 0.30) will not be reported

STAGE 4: REPORT ASSEMBLY

Turn structured data into narrative reporting language



Reference Report Analysis

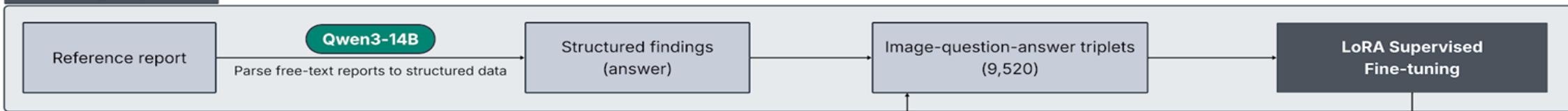
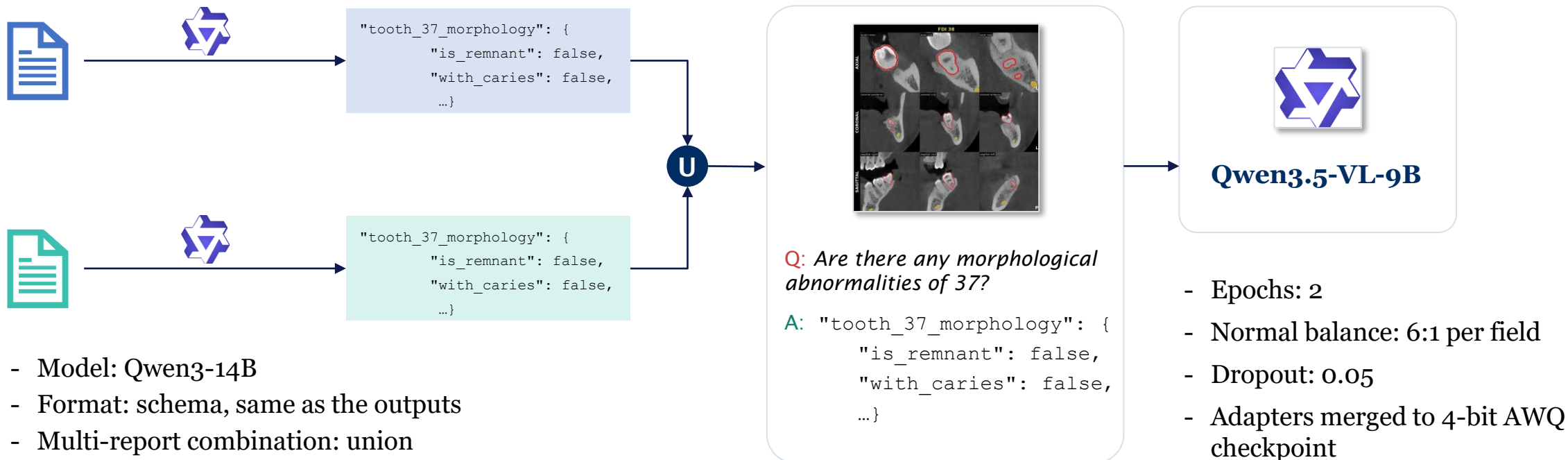
Across the 482 endodontic sentences in the 582 training reports:

Expression	Count
<teeth> endodontically treated (incl. is / are / has been / have been)	129
Endodontic treatment(s) of / involving / on <teeth>	67
<tooth> shows / has undergone endodontic treatment	47
<tooth> with endodontic material ...	33
Presence of / are present on / Sequelae of endodontic treatment	17
Root canal treatment(s) ...	5

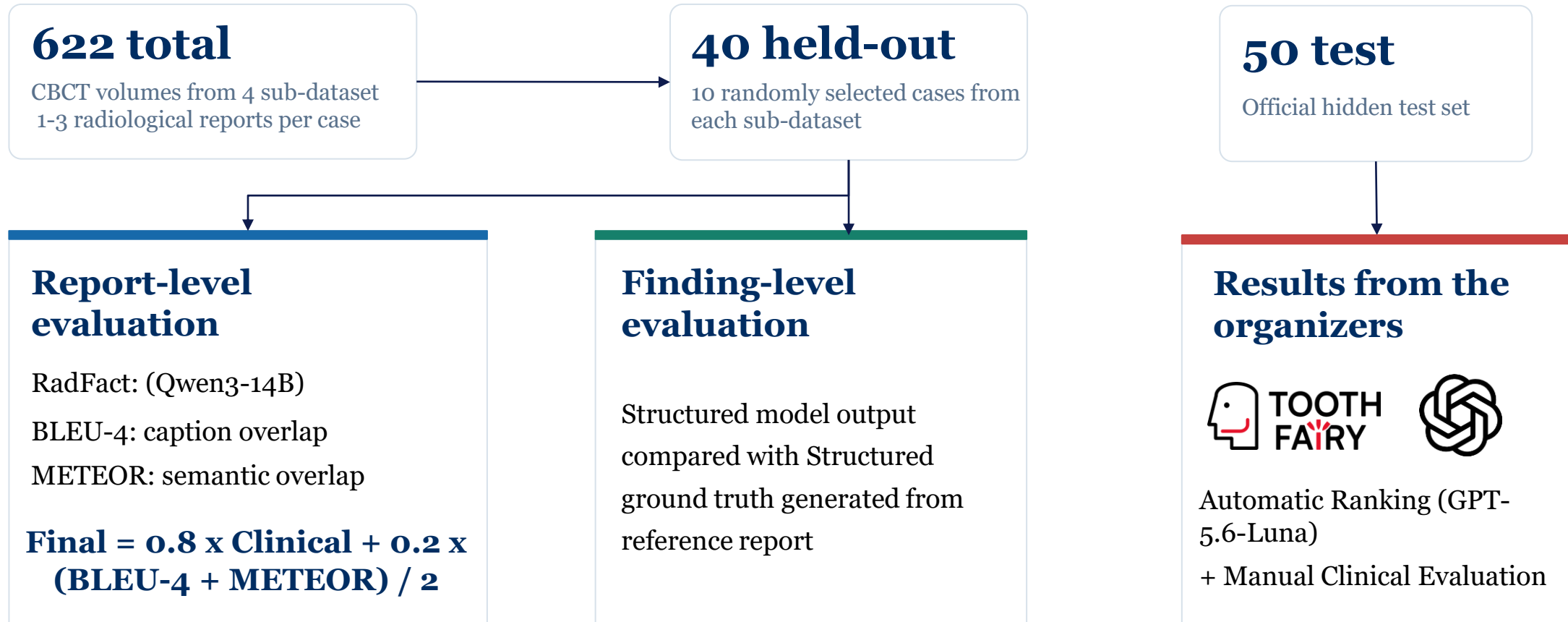
TRAINING

Turn reports into image-question-answer triplets

Training Pipeline



ToothFairy4 / ODIN 2026 benchmark



RESULTS

Ablation study and Finding-level F1 score

Report-level evaluation (Ablation study)

Method	BLEU-4	METEOR	RadFact			Final
			Precision	Recall	F1	
Local validation (N=40)						
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	0.353	0.3357
Qwen3.5-9B + LoRA (Ours)	0.1553	0.3848	0.5319	0.4762	0.4442	0.4094
Official hidden test (N=50)						
Qwen3.5-9B + LoRA (Ours)	0.1476	0.3269	0.5343	0.3316	0.3924	0.3614

Each stage contributes to performance gains, lifting the Final Score from 0.277 to 0.336 and ultimately 0.409.

Finding-level evaluation

Strong Findings

- Impacted teeth: 0.813
- Absent teeth: 0.762
- Endodontic: 0.741

Still Difficult

- Caries: 0.062
- Root remnants: 0.062
- Post-and-core: 0.080

DISCUSSION

Why it is difficult for certain findings?

Inter-agreement per clinical finding for cases with multiple reports (367 cases, 734 reports).

Finding	Both	One only	Both / ≥ 1	Jaccard	κ
Endodontic	160	22	87.9%	0.794	0.732
Implant	50	9	84.7%	0.692	0.535
Impacted teeth	122	26	82.4%	0.805	0.656
Absent teeth	161	100	61.7%	0.639	0.313
Root remnant	20	22 42	47.6%	0.688	0.320
Crowns	69	97	41.6%	0.556	0.236
Fillings	38	91	29.5%	0.642	0.202
Caries	19	65 84	22.6%	0.736	0.177
Post-and-core	0	20 20	0.0%	n/a	0.000

Both / ≥ 1 = Both $\div$ (Both + One only), computed by formula.

Finding-level evaluation

Strong Findings

- Impacted teeth: 0.813
- Absent teeth: 0.762
- Endodontic: 0.741

Still Difficult

- Caries: 0.062
- Root remnants: 0.062
- Post-and-core 0.080
- Crowns: 0.300
- Fillings: 0.267

TAKE-HOME MESSAGE

Using segmentation as a spatial index, we convert 3D CBCT volumes into localized 2D visual evidence for schema-constrained VQA and report generation

OUR HIGHLIGHTS

- Segmentation enables anatomy-grounded 2D image generation
- Schema-constrained VQA with LoRA fine-tuning focuses learning on clinically relevant findings

HONEST LIMITATION

- LoRA fine-tuning improves precision but recall stays the harder problem
- Rare findings (root remnants, post-and-cores) suffer from data sparsity and label noise

Thank you | Questions?

github.com/CC-AIDent-Vienna/v2v2R

