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

Shivam Singh[1,2[0009−0005−6334−0664]⋆], Chen Lu[1,2[0000−0002−4621−9417]⋆], and Benedikt Sagl[1,2[0000−0002−6739−4222]†]

[1] Competence Center Artificial Intelligence in Dentistry, Medical University of Vienna, Vienna, Austria
[2] Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Vienna, Austria
{shivam.singh,chen.lu,benedikt.sagl}@meduniwien.ac.at

This document provides additional methodological, implementation, and analysis details for the accompanying paper. Section, table, and figure numbering is prefixed with "S" to distinguish supplementary material from the main manuscript.

S1 Post-processing Rules and Source Selection

The final structured prediction is produced by applying three rule families described in the main paper: logical constraints, source selection for fields predicted from multiple image types, and a silence rule for low-performing fields.

Table S1. Per-finding validation results across data sources and post-processing (Precision/Recall). Bold indicates the best individual source; status indicates inclusion in the final report.

Finding	*N*	Facts	3D	Panoramic	Detail	Processed	Status
Absent teeth	232	**0.55/0.85**	0.54/0.94	0.53/0.67	–	0.69/0.85	on
Impaction	33	–	0.62/0.48	0.67/0.06	**0.79/0.82**	0.78/0.85	on
Endodontic	90	–	–	0.52/0.13	**0.85/0.74**	0.88/0.64	on
Post-and-core	17	–	–	–	0.08/0.09	0.06/0.12	**off**
Crown	49	**0.33/0.29**	–	–	0.32/0.32	0.29/0.31	**off**
Fillings	74	–	–	–	0.31/0.26	0.30/0.24	on
Caries	26	–	–	0.00/0.00	0.14/0.05	0.14/0.04	**off**
Root remnants	13	–	–	–	0.05/0.17	0.05/0.08	**off**
Implants	32	**0.68/0.53**	–	0.44/0.12	–	0.68/0.53	on
Fixed bridges	11	**1.00/0.55**	–	1.00/0.09	–	1.00/0.55	on
Canal-adjacent	26	**0.69/0.69**	0.70/0.27	–	0.17/0.32	0.69/0.69	on

⋆ These authors contributed equally. † Corresponding author.

S2 VQA Prompt

Each VQA call sends a fixed system prompt, a user message with the image caption (sent before the image) and reading instructions plus the JSON output schema and per-field questions (sent after the image), and returns a JSON object matching that schema. All calls share the same structure, `visual_evidence` first, then closed-ended fields, with an explicit instruction to answer `null` rather than infer unseen anatomy, but the system prompt, image caption, and schema differ by call type (global/whole-arch, per-tooth, per-sinus). We give one representative example of each of the two call types below; per-view and per-tooth calls are structurally identical to these, differing only in which image and schema fields are sent (e.g., the six wisdom-tooth axes reuse the tooth-level template with a different composite image and FDI number).

Global-level example (`global/3d_frontal`, case F022):

```
[SYSTEM]
You are an expert dental/maxillofacial radiologist interpreting CBCT
↪  scans.
Answer ONLY in valid JSON -- no prose, no markdown fences, no extra keys.
Every field in OUTPUT SCHEMA must appear as a key, even if null or empty.
Use exactly the choice strings given; never invent values.
Every answer object starts with "visual_evidence": describe what you
actually see first, then the other fields -- all of which must agree
with it.

REPORT ONLY WHAT IS IN THE IMAGE.
Never infer, assume, or imagine anatomy you cannot point to in a
specific image. If a structure is not shown -- outside the acquisition,
not segmented, or hidden behind something else -- say exactly that in
visual_evidence and answer null, rather than describing what would
normally be there. A structure you expect but cannot see is evidence
of ABSENCE, not permission to assume presence.

[USER, before image]
Image 1 (3d_frontal): 3D reconstruction from CBCT segmentation masks:
camera positioned at the patient's front, bringing the patient's
FRONTAL-view mandible (green), maxilla (tan), mandibular canals
(yellow), and maxillary sinuses (blue) into view. Color key only --
a structure absent from the segmentation is simply not drawn.
Acquisition FOV: maxilla partially included, mandibular condyles
excluded. Every segmented tooth is drawn opaque inside the translucent
bone, so a tooth you cannot see in this image is not present. Tooth
absent from the maxilla: 28. All 16 mandibular teeth are present.

<IMAGE: F022_3d_frontal.png>

[USER, after image]
You are given ONE image for this case: a 3D surface reconstruction
built from this case's CBCT segmentation masks -- a model of the
segmented structures, not a grayscale slice. It is preceded by a
```

```
caption giving its view and what every marker in it means -- read
the caption before the image.

NOTE ON IMAGE QUALITY: this image is generated -- this render IS the
segmentation, drawn as a surface, with no CBCT grayscale underneath
it -- and an automatic segmentation can be wrong. If a structure is
hard to recognise, or what you see does not make anatomical sense, do
not rationalise it into something plausible: say "unable to evaluate"
in visual_evidence and answer null for that finding's other fields.

FDI: 11-18 upper right, 21-28 upper left, 31-38 lower left,
41-48 lower right.

OUTPUT SCHEMA (return a JSON object with exactly these keys)
{
  "maxilla_scope": {
    "visual_evidence": "string -- ...",
    "maxilla_included": "enum:included|not_included"
  },
  "alveolar_bone_atrophy_mandible": { ... },
  "alveolar_bone_atrophy_maxilla": { ... },
  ...   // remaining global facts follow schema.json
}
```

Tooth-level example (`dental_elements/tooth_36`, case F022):

```
[SYSTEM]
You are an expert dental radiologist interpreting per-tooth CBCT
images. Answer ONLY in valid JSON -- no prose, no markdown fences, no
extra keys. Every field in OUTPUT SCHEMA must appear as a key, even
if null or empty. Use exactly the choice strings given; never invent
values. Every answer object starts with "visual_evidence": describe
what you actually see first, then the other fields -- all of which
must agree with it.

REPORT ONLY WHAT IS IN THE IMAGE. [as above]

[USER, before image]
Image 1 (tooth_36_composite): CBCT detail view of tooth FDI 36: a 3x3
grid of grayscale panels, one plane per row, all centred on the tooth.
[... full layout caption, see Sec. 3.2 Key-Slice Selection] Tooth 36
-- and only tooth 36 -- is outlined in red in every panel. The
mandibular canal is filled with a translucent yellow wash wherever it
crosses a panel [...].

<IMAGE: F022_tooth36_composite.png>

[USER, after image]
ALWAYS FOCUS ON TOOTH 36 -- the tooth outlined in RED in every panel.
Locate that outline in each panel before you answer. [...]
```

```
HOW TO READ A TOOTH
Normal structure -- none of this is a finding on its own: enamel
(bright cap over the crown only), dentin (mid-gray bulk), pulp
chamber / root canal (dark central line, crown to apex). A BRIGHT
area can be enamel or restorative material, and a DARK one can be
the normal root canal or a defect; tell them apart by brightness,
shape, and position together [...].

OUTPUT SCHEMA (return a JSON object with exactly these keys)
{
  "tooth_36_morphology": {
    "visual_evidence": "string -- describe the crown surface and root
       outline for this tooth across the available views, or note if
       no crown remains at all",
    "is_remnant": "bool",
    "with_caries": "bool",
    "with_root_fracture": "bool",
    "with_endo": "bool",
    "with_fillings": "bool",
    "with_full_crown": "bool",
    "with_post_and_core": "bool"
  },
  "tooth_36_eruption": { ... },
  "tooth_36_endodontic_treatment": { ... },
  "tooth_36_periodontal_status": { ... },
  "tooth_36_bone_quality": { ... },
  "tooth_36_mandible_canal": { ... }
}

QUESTIONS
1. [tooth_36_morphology] Judge tooth 36's state -- only mark a flag
   true if you're confident, not on a maybe. [...]
2. [tooth_36_eruption] First describe how much of the crown sits
   above versus below the bone line for tooth 36, then answer the
   eruption state. [...]
...
```

S3 Reference Report Analysis

Table S2. Uncertainty frequency of post-and-core restoration across all reference reports (corpus: 1,000 reports, 622 cases).

Term / Search Pattern	Sentence Level		Report Level	
	Total	Uncertain (%)	Total	Uncertain (%)
`post-and-core`	43	23 (53.5%)	30	15 (50.0%)
`post` (bare)	12	7 (58.3%)	11	7 (63.6%)
`endocanal post`	10	**10 (100.0%)**	7	**7 (100.0%)**
`endodontic post`	8	**8 (100.0%)**	8	**8 (100.0%)**
`crown on a core/stump`	4	0 (0.0%)	2	0 (0.0%)
`intracanal post`	2	**2 (100.0%)**	2	**2 (100.0%)**
ANY (deduplicated)	**79**	**50 (63.3%)**	**54**	**33 (61.1%)**

Table S3. Inter-rater 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
Impaction	122	26	82.4%	0.805	0.656
Absent teeth	161	100	61.7%	0.639	0.313
Root remnant	20	22	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	22.6%	0.736	0.177
Post-and-core	**0**	**20**	**0.0%**	**n/a**	**0.000**

Table S3 presents the inter-rater agreement per clinical finding evaluated across 367 cases (totaling 734 reports). Endodontic treatment, implants, and impaction, exhibit substantial agreement (e.g., Cohen's $\kappa > 0.50$), reflecting consistent clinical documentation. Conversely, findings subject to clinician reporting habits or diagnostic ambiguity, such as crowns, fillings, and caries, show markedly lower agreement ($\kappa \leq 0.236$). Notably, post and core demonstrates a complete lack of agreement ($\kappa = 0.000$ with zero concurrent mentions), highlighting the intrinsic difficulty and ambiguity in distinguishing it from standard endodontic treatments. Furthermore, rare conditions like root remnants show limited reliability due to sparse occurrences.

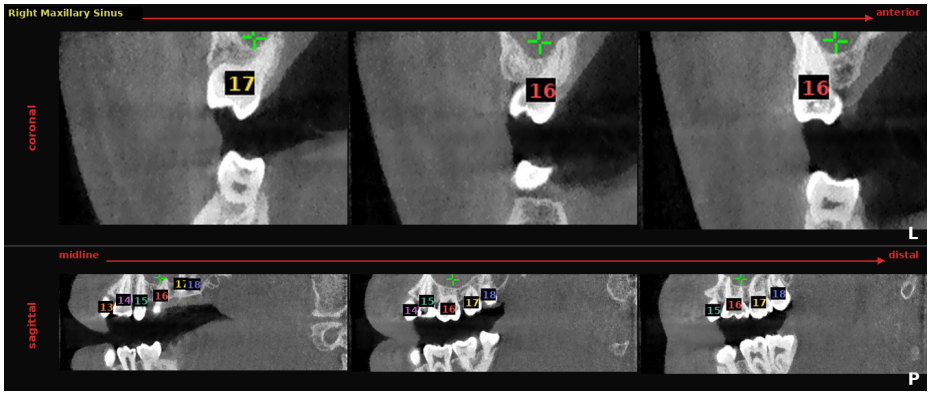

Fig. S1. Coronal and Sagittal view of Right Maxillary Sinus. Sinus is marked, and teeth are tagged with FDI index