OVERJET-1: A MULTI-MODAL GENERALIST VISION-LANGUAGE MODEL FOR GROUNDED DENTAL IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

A complete dental examination is an inherently multi-modal, multi-view diagnostic task: no single imaging source provides complete insight into underlying pathologies. Clinicians aggregate evidence across sources such as panoramic and intraoral radiographs, color photographs, cephalograms, periodontal charts, and 3D CBCT scans. With the advent of deep learning and vision-language models (VLMs), significant progress has been made in dental image analysis. However, current VLMs and task-specific models process single images in isolation, missing crucial cross-modal context and allowing errors to compound across findings. We present *Overjet-1*, a generalist dental VLM built on Gemma 4 26B-A4B that natively processes multi-image, multi-modal inputs. Overjet-1 is post-trained on 1,631,594 supervised records spanning 122 dental tasks. These records are organized through the Dental Graph, a case-level representation that unifies tooth segmentations, finding regions, clinical measurements, and 3D geometry across all available multi-modal sources. The graph supplies the spatial coordinates for each training target, links observations of the same tooth across views, and constructs chain-of-thought (CoT) reasoning paths through graph traversal. We introduce *Dental Consistency Reinforcement Learning (DC-RL)*, a novel framework extending Group Relative Policy Optimization (GRPO). DC-RL computes verifiable, deterministic reward signals directly from grounded facts in the Dental Graph, simultaneously optimizing diagnostic correctness and spatial localization without relying on extensive manual annotation or CoT distillation. Evaluated on MMOral-OPG-Bench, Overjet-1 achieves a new state of the art of 71.5%, compared to the best baseline of 61.3%. We also introduce *Overjet-Bench*, a more comprehensive benchmark covering 9 task families across single-image, multi-image, and multi-modal setups. Overjet-1 achieves a macro score of 78.8% on Overjet-Bench, outperforming top proprietary and open-source baselines (33.5%) significantly on real-world tasks. To our knowledge, Overjet-1 is the first model natively trained and evaluated on both multi-image and multi-modal dental tasks with grounded supervision. Code is available at `https://anonymous.4open.science/r/Overjet-1-code-D205`; Overjet-Bench will be released upon acceptance.

1 INTRODUCTION

Real-world dental diagnosis is inherently a multi-modal, multi-view reasoning task because no single imaging modality captures the complete clinical picture. Visual examination is the first-line method for caries on accessible surfaces, while intraoral bitewing radiographs provide complementary evidence. Cone-beam computed tomography (CBCT) adds 3D anatomical information beyond 2D radiography in selected endodontic situations and for indications such as impacted teeth and implant placement (Fayad et al., 2015). Periodontal assessment adds structured clinical measurements: conventional 2D radiographs do not capture soft-tissue attachment, so probing and periodontal charting remain necessary alongside radiographic bone assessment (Herrera et al., 2025). Dental clinicians—and dental AI systems—therefore cannot treat image interpretation as an isolated single-view classification task. A generalist system must aggregate heterogeneous 2D, 3D, visual, and structured clinical evidence into a unified tooth-level understanding.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

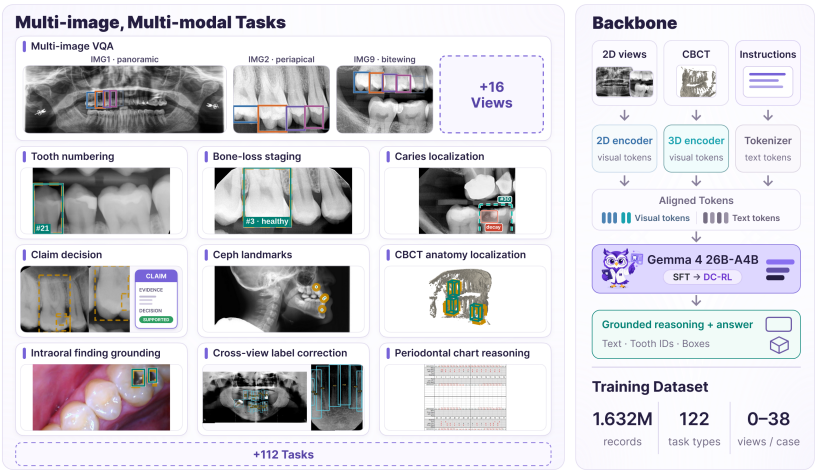


Figure 1: **Overview of Overjet-1 capabilities and architecture.** *Left:* Ten representative tasks spanning multi-image and multi-modal dental inputs. *Right:* 2D images, CBCT volumes, and text instructions are encoded separately and passed to a shared language model, which generates clinical text, tooth identities, and coordinates. Overjet-1 is trained with supervised fine-tuning followed by DC-RL. The lower panel summarizes the training dataset.

CNN- and transformer-based specialized models trained for tooth segmentation, numbering, and pathology detection can perform very well on panoramic and intraoral radiographs (Jader et al., 2018; Hamamci et al., 2023; Budagam et al., 2024; Yarlagadda et al., 2025). However, most of these specialized models accept only one image from a specific modality, operate within a fixed label space, and return labels or segmentation masks. In practice, these tasks are handled by separate models. Even when each model is accurate on its own task, their predictions can disagree about the same tooth across images or modalities. The main limitation is therefore not single-task accuracy, but the lack of one consistent and explainable interpretation of the whole case.

Vision-language models have shown promising results in medical imaging (Li et al., 2023; Moor et al., 2023; Tu et al., 2024; Sellergren et al., 2026), and recent work has begun to extend them to dentistry (Hao et al., 2025; 2026b; Fan et al., 2026). However, even when medical VLMs cover several modalities, their training and evaluation usually treat each image independently (Jha et al., 2026; Wang et al., 2026; Li et al., 2026), leaving case-level reasoning across multiple inputs underexplored. Beyond being underexplored, such cross-image reasoning is also difficult: on general-domain benchmarks, even frontier VLMs fall well below human accuracy and below their own single-image performance (Wang et al., 2025a; Meng et al., 2025; Fu et al., 2024; Zhao et al., 2024). Moreover, language generation does not by itself ensure that an answer is based on the relevant image evidence. In Figure 2, all three frontier models identify caries on tooth #3, yet their predicted bounding boxes miss the annotated lesion.

Crucially, no existing dental benchmark evaluates grounding and reasoning across multiple images within a case, preventing accurate quantification of this performance gap. To address these limitations, we propose a method for constructing high-quality grounded training data, a novel specialized reinforcement learning framework that exploits dental data structure, and a novel multi-modal evaluation benchmark. Our contributions are the following:

- **Large-scale grounded dental supervision.** Training a VLM on multi-image, multi-modal dental cases requires explicit links across the inputs: which observations belong to the same tooth, which finding comes from which source, and where the supporting evidence appears. We construct this supervision from a large, diverse clinical database spanning the full range of dental modalities, including repeated observations of the same teeth across radiographs, photographs, and 3D scans. We generate tooth segmentations, finding regions, and clinical measurements from proprietary specialist systems, such that every coordinate and clinical fact traces to a source annotation and

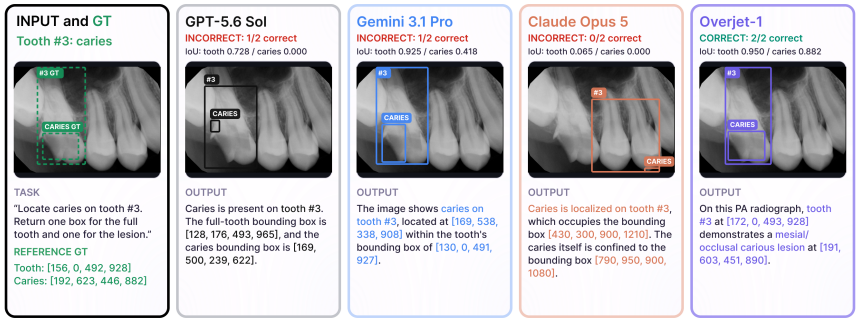


Figure 2: **Caries localization on a periapical radiograph.** Three frontier vision-language models and Overjet-1 receive the same image and prompt. GPT-5.6 Sol, Gemini 3.1 Pro, and Claude Opus 5 all report that caries is present, but their lesion boxes miss the annotated region. Only Overjet-1 correctly localizes both tooth #3 and the carious lesion.

is *grounded* in the images. From this data we build 1,631,594 multi-modal, multi-image records across 122 tasks.

- **Scalable chain-of-thought generation through Dental Graph.** We organize this dataset and its annotations into the Dental Graph, a case-level representation indexed by tooth and view that anchors each training target in image geometry. *Graph traversal* then generates chain-of-thought (CoT) traces that identify the relevant teeth, link each finding to its region, and derive the final answer. This yields reasoning supervision at scale, derived from the graph itself and extending grounding to the intermediate reasoning steps rather than the final answer alone. 54.2% (more than 800,000) of the records include such traces.
- **Dental consistency reinforcement learning (DC-RL).** We introduce *DC-RL*, a novel GRPO framework specialized for dental tasks that computes a deterministic, verifiable reward directly from the task schema and the graph, without a learned reward model or an LLM judge. The highest reward is given when the final answer is clinically correct, the reasoning trace supports that answer, and the grounding references identify the correct teeth and regions; the reward also checks tooth order, spacing, and arch separation within each view, and shared tooth identities across images and modalities. DC-RL improves every reported metric, adding 3.4 points to the Overjet-Bench macro score over a matched control without spatial consistency.
- **State-of-the-art performance on a public dental benchmark.** We fine-tune Gemma 4 26B-A4B on this data with the proposed method. The resulting model, Overjet-1, achieves a new state of the art of 71.5% and 75.1% on the closed and open splits of MMOral-OPG-Bench, exceeding the strongest baselines by 10.2 and 11.5 points and surpassing proprietary frontier models, while zero-shot outperforming a dental VLM trained directly on the benchmark dataset. We also compare Overjet-1 with specialist systems for dental-image modality recognition and end-to-end tooth numbering, and Overjet-1 remains competitive on both (Appendix H.3).
- **Overjet-Bench, a multi-modal VLM benchmark for broad dental image understanding.** To evaluate case-level reasoning with explicit spatial grounding, we construct Overjet-Bench, which evaluates localization, tooth identification, question answering, cross-view reasoning, and report generation on both single- and multi-image cases. Overjet-1 reaches 78.8% on Overjet-Bench, a gain of 45.3 points over the second-best model, reflecting a capability gap: current VLMs lack dental grounding and cannot perform the multi-image, cross-view task. We will release the benchmark publicly.[1]

2 Related Work

Specialist models for dental image analysis. Automated dental image analysis has been extensively investigated (Singh & Raza, 2022; Sivari et al., 2023; Demir et al., 2024; Khattak et al., 2025; Zhou et al., 2025). Representative examples include tooth instance segmentation on panoramic radiographs (Jader et al., 2018; Budagam et al., 2024) and benchmarks for tooth detection, enumer-

[1]Conditioned on acceptance.

ation, and diagnosis (Hamamci et al., 2023). Other extensively studied tasks include caries detection (Mohammad-Rahimi et al., 2022) and radiographic periodontal bone loss assessment (Khubrani et al., 2025). These models can achieve strong performance in the settings for which they are trained, but they usually use a fixed label set, return labels, boxes, or masks, and process one image at a time. Applying them to a complete examination in real-world scenarios therefore often requires several models and an additional procedure to combine their outputs across tasks and views.

Medical and dental vision-language models. Medical vision-language models have been developed for biomedical question answering, few-shot image interpretation, and broader multi-modal medical tasks (Li et al., 2023; Moor et al., 2023; Tu et al., 2024; Li et al., 2025a;b). MMOral introduced 1.3 million instructions from 20,563 panoramic images and used them to train the 7B OralGPT model (Hao et al., 2025). OralGPT-Omni expanded to five modalities and added a reasoning dataset and the MMOral-Uni benchmark (Hao et al., 2026b). OralGPT-Plus introduced visual-tool use and reinforcement learning for panoramic analysis (Fan et al., 2026). DentVLM fine-tunes a general VLM for seven imaging modalities and 36 tasks, but reads each image on its own and names a predefined region instead of returning coordinates (Meng et al., 2026b). OrthoAgent and DentAgent leave the model untrained and route the images through specialist tools, merging the outputs with fixed rules (Hao et al., 2026a) or flagging conflicts among them (Meng et al., 2026a). Neither system is trained to check the evidence against the image, so a tool error that nothing contradicts passes into the answer. In contrast, we organize the training data by examination rather than treating each image in isolation, and train the model itself to ground and check each finding.

Grounding in vision-language models. Grounded vision-language models connect generated text to points, boxes, or masks. Shikra and Kosmos-2 represent spatial references in the language sequence (Chen et al., 2023; Peng et al., 2024), while Ferret, Groma, and GLaMM use region features or segmentation interfaces (You et al., 2024; Ma et al., 2024; Rasheed et al., 2024). In medical imaging, MAIRA-2 and related work ground phrases in radiology reports (Bannur et al., 2024; Zou et al., 2025). Thinking with Visual Primitives extends grounding to intermediate reasoning steps (Lu et al., 2026). Likewise, we include coordinates directly in the reasoning process and the final output. Every piece of evidence is derived from geometry stored in the Dental Graph, so the tooth identity, clinical finding, and supporting image region remain linked to the same source annotation.

Reinforcement learning for medical reasoning. Recent medical vision-language models use GRPO and related methods to improve reasoning from verified examples (Pan et al., 2025). These approaches use rewards designed for answer verification or tool use, whereas grounded dental answers require a score for both clinical content and spatial evidence. Our reward is computed directly from the Dental Graph rather than by a learned judge. The spatial terms contribute only when the clinical answer is correct, so a plausible box cannot compensate for a wrong answer.

3 METHOD

We formulate conditional generation over a set of multi-modal inputs: let $X = (x_1, \ldots, x_m)$ denote the visual or textual inputs available for one dental examination, which we refer to as a case. Given a task prompt q, Overjet-1 models $p_\theta(y \mid X, q)$, where the target y contains the clinical answer and grounding information for the teeth or findings it mentions.

3.1 GROUNDED DENTAL REASONING DATA

Our proprietary database groups all material from the same dental examination. Specialist systems provide tooth segmentations, finding regions, restorations, bone measurements, and cephalometric landmarks; structured records provide periodontal measurements, clinical notes, and treatment information; and human annotations provide precise geometry for tasks such as full-mouth tooth numbering. For each case, we combine these sources in a Dental Graph $\mathcal{G} = (V, E_{\text{within}}, E_{\text{cross}})$. A node $v \in V$ represents one appearance of a tooth in one image and stores its image identifier, tooth number, geometry, findings, restorations, and available measurements. Structured information used by a task but not tied to a specific tooth is stored as a case-level attribute rather than as a node. Edges in E_{within} encode relationships among tooth appearances in the same image, while edges in E_{cross} link appearances of the same physical tooth across images.

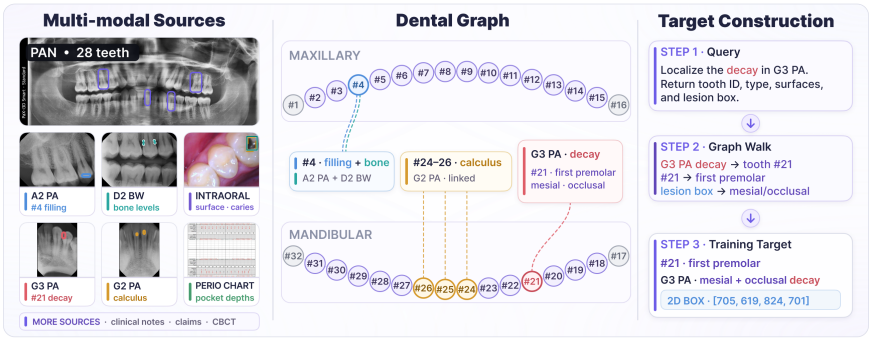


Figure 3: **Multi-modal sources and Dental Graph construction.** Sources including radiographs, intraoral photographs, and periodontal charts from the same examination are linked by tooth in the Dental Graph; the selected subgraph forms the grounded target. A typical radiographic set contains an 18-image full-mouth series and one panoramic image, and additionally it can contain intraoral photographs, CBCT, charts and notes.

Each task type defines a graph query, a prompt generator, and a target schema (Figure 3). The graph query first selects the relevant teeth, findings, relations, and geometry, after which the prompt generator samples the wording. Short targets, including localization, counting, tooth identification, and multiple-choice answers, are generated deterministically. Every coordinate is copied from the geometry returned by the query, so the named tooth or finding and its spatial reference come from the same source annotation. We then use Gemini 3.5 Flash-Lite to diversify the language styles of the reasoning traces from the queried facts (Google DeepMind, 2026b). Its instructions prohibit new findings, restrict clinical terminology, and require coordinates to be kept exactly as provided. For most tasks, the teacher receives only the structured facts and does not receive the source image. Using the Dental Graph, we turn existing annotations into verifiable reasoning supervision at scale, without requiring expensive human annotations. For tasks involving several images, the graph query follows the identity links between appearances of the same tooth. Across the 2D and text dataset, 54.2% of records contain a supported reasoning trace. Further details are provided in Appendix C and Algorithm 2. The 2D and text dataset contains 1,607,528 records across 118 task types, and four CBCT task types contribute another 24,066 records. The complete training set therefore contains 1,631,594 records across 122 task types (Table 6). All data splits used for supervised training, DC-RL, validation, and internal evaluation are patient-disjoint. Overjet-1 fully fine-tunes Gemma 4 26B-A4B (Google DeepMind, 2026d). Supervised training runs for one epoch on eight H200 GPUs (Appendix B). Direct CBCT volumes use a separate 3D encoder that passes semantic slot tokens and a coarse spatial grid to the language model; its design and evaluation are described in Appendix E.

3.2 Dental Consistency Reinforcement Learning (DC-RL)

Supervised fine-tuning teaches the model to produce clinical answers together with spatial references. We then apply DC-RL. All reward components are deterministic and are computed based on the information from the Dental Graph. For a prediction $\hat{y}$, the reward is

$$r(\hat{y}) = c(\hat{y})\,(1 + \lambda_g g(\hat{y}) + \lambda_e e(\hat{y}) + \mu f(\hat{y}))\,, \tag{1}$$

where c measures task-specific correctness, g combines agreement with available reference geometry and tooth-layout consistency, e checks whether the evidence trace supports the predicted entities and claims, and f checks the required answer format. We use $\lambda_g = 0.35$, $\lambda_e = 0.15$, and $\mu = 0.05$. The definition of c follows the target schema. Since c multiplies the complete reward, a completion with $c = 0$ receives no reward for plausible geometry or fluent reasoning. The grounding term g compares predicted regions with the references stored for the structures they describe. Within g, a spatial-consistency score s examines the relationships among predicted tooth regions: anatomical order, inter-tooth spacing, separation of the maxillary and mandibular arches, and reuse of one box for several teeth. These relative checks complement absolute overlap and provide a signal when reference regions are unavailable. Detailed definitions and boundary cases, the DC-RL training setup, and the matched control that removes the spatial-consistency contribution are given in Appendices B and D.3.

Table 1: **Results on MMOral-OPG-Bench.** The closed-ended split contains 491 questions; Avg. is accuracy over all of them, and a question can count toward several dimensions. The open split reports the mean LLM-judge score as a percentage (Appendix F.1). Overall averages the Closed and Open aggregate scores. Proprietary-model split aggregates average three API runs (Table 14). † marks closed scores reported by the MMOral authors (Hao et al., 2025); other filled cells use our evaluations. Bold and underline indicate the best and second-best values in each metric column.

		Closed split (%)						Open split	
Model	Params	Teeth	Path.	Treat.	Bone	Sum.	Avg.	Avg.	**Overall**
Gemma 4 26B-A4B base	26B	43.1	45.3	60.6	55.8	64.5	46.4	61.3	53.9
Overjet-1	26B	**71.0**	**66.2**	<u>73.2</u>	**76.7**	<u>67.7</u>	**71.5**	**75.1**	**73.3**
Medical and dental vision-language models (Li et al., 2023; Pan et al., 2025; Sellergren et al., 2026; Lin et al., 2025; Hao et al., 2025)									
LLaVA-Med	7B	9.0	8.8	13.4	14.7	12.9	10.8	52.3	31.6
MedVLM-R1	2B	31.5	30.4	40.8	55.8	41.9	38.1	42.3	40.2
MedGemma 1.5-4B	4B	28.7	16.9	34.5	58.1	38.7	35.4	45.2	40.3
HealthGPT-XL32	32B	37.2	41.2	39.4	61.2	64.5	43.0	46.7	44.9
OralGPT†	7B	43.2	41.0	43.1	63.6	37.5	46.2†	-	-
General open models (Qwen Team, 2026; Wang et al., 2025b; Meta Superintelligence Lab, 2026)									
Qwen3.6-35B-A3B	35B	39.7	31.1	48.6	67.4	61.3	46.2	57.5	51.9
InternVL3.5-30B-A3B	30B	40.6	43.2	41.5	65.9	54.8	46.2	58.4	52.3
Muse Glimmer-30B	30B	42.5	42.6	60.6	59.7	**74.2**	47.3	50.1	48.7
Frontier proprietary models (OpenAI, 2026; Google DeepMind, 2026a; Anthropic, 2026)									
GPT-5.6 Sol	–	<u>60.0</u>	<u>48.6</u>	**75.4**	67.4	64.5	<u>61.3</u>	60.9	<u>61.1</u>
Gemini 3.1 Pro	–	53.5	43.2	71.8	67.4	54.8	56.6	<u>63.6</u>	60.1
Claude Opus 5	–	49.9	48.0	71.1	<u>69.8</u>	61.3	55.0	59.3	57.1

3.3 OVERJET-BENCH

Overjet-Bench extends dental VLM evaluation beyond isolated images by including case-level tasks that require evidence to be combined across inputs from the same dental examination. It has nine task families from 29 source task types. Because the family metrics measure different quantities, we report each family separately and use their unweighted macro average only as a summary of task coverage. The benchmark contains 825 single-image, 85 multi-image, and 90 multi-modal records over 1,298 unique images. Its composition and scoring rules are summarized in Table 13.

4 EXPERIMENTS

Evaluation sets. MMOral-OPG-Bench contains a 491-question closed-ended split and an open-ended split (Hao et al., 2025). Each item contains one panoramic radiograph. Overjet-Bench evaluates the nine task families defined in Section 3.3.

Comparison models. The comparison set includes OralGPT (Hao et al., 2025), medical vision-language models, and general-purpose open-weight vision-language models. The proprietary frontier models are GPT-5.6 Sol (OpenAI, 2026), Gemini 3.1 Pro (Google DeepMind, 2026a), and Claude Opus 5 (Anthropic, 2026). We also evaluate Muse Glimmer-30B (Meta Superintelligence Lab, 2026) as a general-purpose open-weight vision-language model. The inference settings, the scoring rules, the statistical protocol, and the ablation design are described in Appendix G.

5 RESULTS AND DISCUSSION

MMOral-OPG-Bench. Overjet-1 reaches 71.5% on the 491 closed questions. Unlike OralGPT, which was trained on the MMOral instruction corpus, Overjet-1 was trained **without any in-distribution images or question templates**, hence evaluating the *zero-shot* performance on their dataset distribution. The strongest comparison model, GPT-5.6 Sol, averages $(61.30 \pm 1.43)\%$ over three API runs, giving Overjet-1 a 10.2-point advantage. On the open split, Overjet-1 obtains

Table 2: **Results on Overjet-Bench.** Scores use the task-family metrics in Table 13, scaled to 0–100; Overall is their unweighted mean across nine families. X-input comprises multi-image and multimodal tasks. Proprietary-model Overall values average three API runs (Table 14). Bold and underline indicate the best and second-best values in each metric column.

Model	Ground	Ident.	Numb.	QA	Count	Bone	X-input	Report	Mod.	**Overall**
Gemma 4 26B-A4B base	11.5	14.1	34.3	29.5	22.6	16.8	16.2	15.8	70.0	25.6
Overjet-1	**90.8**	**91.3**	**95.0**	**77.5**	**77.4**	**48.8**	**82.4**	**57.7**	**88.6**	**78.8**
Medical and dental vision-language models										
LLaVA-Med	10.5	2.8	14.0	18.2	9.2	4.0	14.9	14.7	25.7	12.7
MedVLM-R1	10.0	10.0	18.0	21.0	0.0	5.6	0.2	4.2	25.7	10.5
MedGemma 1.5-4B	10.2	7.5	33.3	23.4	9.3	14.8	11.7	14.4	52.9	19.7
HealthGPT-XL32	10.5	5.6	21.4	27.0	18.7	18.8	9.7	6.9	42.9	17.9
General open models										
Qwen3.6-35B-A3B	11.2	26.0	38.7	26.5	20.5	17.2	19.3	17.5	61.4	26.5
InternVL3.5-30B-A3B	10.8	7.4	35.2	15.5	20.1	13.2	10.8	15.4	62.9	21.3
Muse Glimmer-30B	16.5	40.1	43.8	36.3	37.4	16.8	24.1	17.8	68.6	33.5
Frontier proprietary models										
GPT-5.6 Sol	10.5	16.2	67.9	11.3	38.0	14.0	24.3	13.7	72.9	29.9
Gemini 3.1 Pro	10.7	6.6	48.9	7.2	14.6	9.6	9.4	15.3	62.9	20.6
Claude Opus 5	10.5	12.8	50.4	13.3	42.3	14.8	8.9	13.9	68.6	26.2

Table 3: **Ablations.** The MMOral-OPG *Avg.* is the mean of Closed and Open. For Overjet-Bench, *Ground* and *Report* are the grounding and report-generation task families, and *Avg.* is the nine-family macro average. *Overall* is the arithmetic mean of the two Avg. columns. Best and second-best values in each metric column are bolded and underlined.

		MMOral-OPG			Overjet-Bench			
Variant	What changes	Closed	Open	Avg.	Ground	Report	Avg.	**Overall**
Our backbone (Gemma 4 26B-A4B)								
Base	no post-training	46.4	61.3	53.9	11.5	15.8	25.6	39.8
+ SFT, no grounding	coordinates removed	56.6	65.1	60.9	10.5	40.2	49.0	55.0
+ SFT, no reasoning	answer-only targets	60.7	61.9	61.3	83.3	47.8	59.9	60.6
+ SFT (full)	full data in Table 6	65.4	72.2	68.8	85.3	50.6	70.5	69.7
+ DC-RL, no spatial consistency	all uses of s removed	67.0	71.9	69.5	87.2	55.1	75.4	72.5
+ DC-RL	full reward in (1)	**71.5**	**75.1**	**73.3**	**90.8**	**57.7**	**78.8**	**76.1**
OralGPT's backbone (Qwen2.5-VL-7B)								
Base	no post-training	31.2	37.5	34.4	9.9	9.5	17.9	26.2
+ SFT	same data and schedule	59.9	60.3	60.1	71.6	44.6	57.6	58.9
+ DC-RL	same full reward	62.1	63.9	63.0	75.2	45.5	59.2	61.1

75.1% under the LLM-judge protocol. The strongest comparison model, Gemini 3.1 Pro, averages $(63.60 \pm 3.55)\%$, leaving an 11.5-point gap; GPT-5.6 Sol averages $(60.90 \pm 3.29)\%$, while LLaVA-Med scores 52.3%. This judge-based metric does not reproduce the closed-split ordering. Averaging the closed and open scores, Overjet-1 reaches 73.3% Overall, 12.2 points above the strongest comparison model's 61.1%. Per-dimension scores are reported in Appendix H.1. These results show that case-level grounded training transfers to external question answering. External transfer is also observed on COde (Lv et al., 2026), where Overjet-1 reaches 76.63% accuracy without COde fine-tuning, compared with 65.14% for the strongest other zero-shot model (Table 15).

Overjet-Bench. Overjet-1 obtains a macro score of 78.8 on Overjet-Bench, exceeding the strongest comparison model by 45.3 points (paired 95% CI: [41.3, 48.7]; Table 2). It also improves over its 26B base by 53.2 points, demonstrating the contribution of our post-training pipeline. On cross-input tasks, Overjet-1 scores 82.4, exceeding the strongest comparison model by 58.1 points. This gain demonstrates its ability to combine evidence across images and modalities. The qualitative

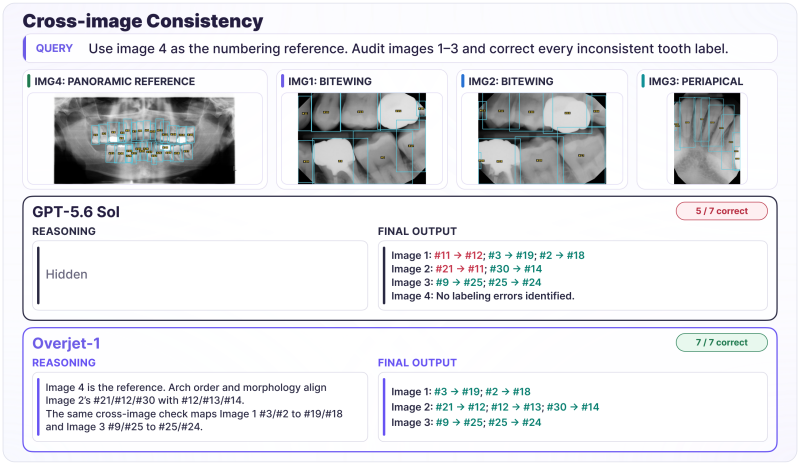


Figure 4: **Cross-view tooth-number consistency comparison with GPT-5.6 Sol and Overjet-1.** Image 4 is the panoramic numbering reference. Each model audits two bitewing views and one periapical view. GPT-5.6 Sol returns five of seven corrections and misses two errors. Overjet-1 returns all seven corrections consistently. GPT-5.6 Sol's reasoning is hidden.

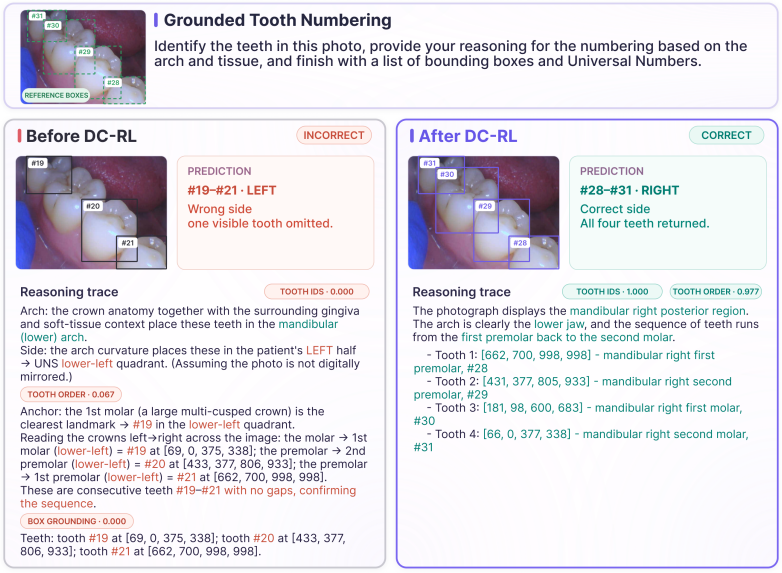


Figure 5: **Comparison of model output before and after DC-RL.** The supervised checkpoint assigns a mandibular-right view to the left quadrant, returns teeth #19–#21, and does not produce a final answer. After DC-RL, the saved final answer recovers the reference sequence #28–#31 with aligned boxes.

comparison in Figure 4 shows a cross-view tooth-number consistency audit from the multi-image family.

Effect of DC-RL. As shown in Table 3, both DC-RL variants improve MMOral closed accuracy and the Overjet-Bench macro over full SFT, and the complete spatial reward improves every reported measure over the no-spatial control. This result separates a general DC-RL gain from the contribution of spatial consistency. Relative to full SFT, the no-spatial control changes MMOral closed, MMOral open, and the Overjet-Bench macro by $+1.6$, -0.3, and $+4.9$ points. Adding spatial consistency contributes a further 4.5, 3.2, and 3.4 points.

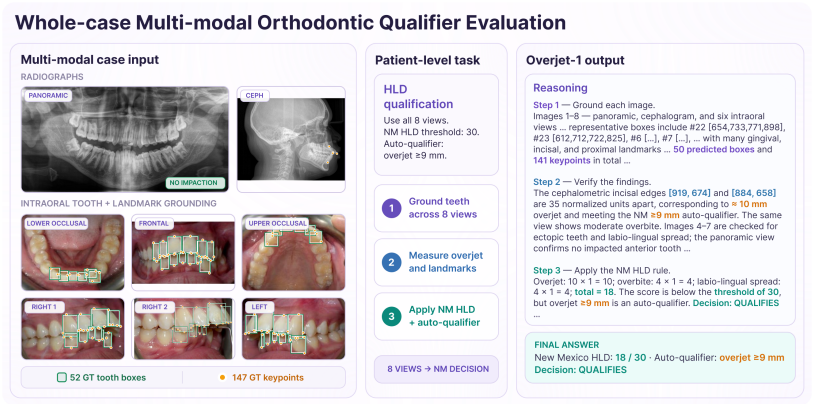


Figure 6: **Whole-case multi-modal orthodontic decision.** Eight views from one patient are interpreted jointly. The model uses the grounded evidence to apply the criteria and predict the decision.

Figure 5 shows the two error types addressed by the reward. The SFT model uses a coherent sequence from the wrong quadrant. Reference grounding detects the absolute error, while arch order alone would not. Full DC-RL recovers both the correct sequence and its supporting regions.

Whole-case multi-modal reasoning. Figure 6 shows a multi-modal orthodontic task that combines a panoramic radiograph, a lateral cephalogram, and six intraoral photographs. The grounded evidence links tooth identity across views, measures the findings, and applies the New Mexico Handicapping Labio-lingual Deviation (HLD) criteria to return the score and decision. On the 200-case orthodontic evaluation, qualification F_1 drops from 88.9 to 24.3 when photographs are shuffled across patients (Table 11). This shows the importance of correctly matching patient inputs across modalities.

Ablations. Full SFT reaches an Overjet-Bench macro score of 70.5, compared with 49.0 without grounding and 59.9 without reasoning. The paired 95% CIs for the corresponding macro gains are [19.0, 24.7] and [8.3, 12.9], respectively. Full DC-RL improves the macro over the no-spatial control by 3.4 points (paired 95% CI: [1.2, 5.6]; Holm-adjusted $p = 0.012$). This matched comparison supports an additional contribution from spatial consistency. Per-metric changes are reported in Appendix H.4. The same ordering holds for Qwen2.5-VL-7B (Bai et al., 2025): its Overjet-Bench macro score rises from 17.9 to 57.6 after SFT and to 59.2 after DC-RL, providing evidence that the training gains can extend to other backbones.

6 Conclusion

We introduced Overjet-1, a generalist dental vision-language model for multi-image and multi-modal case interpretation. Its 1,631,594 training records span 122 task types and are organized through the Dental Graph, which connects tooth segmentations, clinical findings, measurements, and geometry within each case. This representation supports grounded training targets, annotation-derived reasoning traces, and a DC-RL reward that evaluates answer correctness and localization without relying on a learned judge. Overjet-1 achieves 71.5% on the MMOral-OPG closed-ended split, 75.1% on the open-ended split, and 78.8% on Overjet-Bench, exceeding the strongest comparison models by 10.2, 11.5, and 45.3 points, respectively. Ablations show that grounding and supported reasoning traces both improve supervised training, that DC-RL provides a further gain, and that spatial consistency adds another 3.4-point improvement on Overjet-Bench. The same training recipe also improves Qwen2.5-VL-7B, indicating that the gains are not specific to the larger backbone. Overall, Overjet-1 brings multi-modal and multi-image question answering, report generation, localization, and reasoning into one model while keeping its predictions linked to grounded evidence. Limitations are discussed in Appendix A.

AI USE STATEMENT

Generative AI tools were used in this work as follows. Gemini 3.5 Flash-Lite was used to diversify the language of the reasoning traces generated from Dental Graph facts (Section 3.1 and Appendix C.2); its instructions prohibit new findings and require all coordinates to be kept exactly as provided, and every record was validated against the structured facts. Gemini 3.7 Flash (Google DeepMind, 2026c) was used as the judge for the MMOral open-ended split, following the benchmark's original protocol (Section 4 and Appendix F.1). Beyond these two research uses, generative AI tools were used only to polish the wording of the manuscript; they were not used to develop the method, design or run experiments, write code, analyze or interpret results, or prepare figures. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses a proprietary clinical database of dental examinations (Section 3.1). This study was approved by Advarra IRB. All data were de-identified prior to analysis in accordance with institutional and applicable privacy regulations. For privacy and licensing reasons, the training corpus cannot be released (see the Reproducibility statement). Limitations of the training supervision and of the evaluation coverage are discussed in Appendix A.

REPRODUCIBILITY STATEMENT

Training configurations for supervised fine-tuning and DC-RL are listed in Tables 4 and 5 (Appendix B). The Dental Graph construction, record generation, prompt structure, and the corresponding algorithms are described in Appendices C and D, and the complete definition of the DC-RL reward, including its boundary conditions, in Appendix D.3. The composition and scoring rules of Overjet-Bench, the evaluation protocol, and the repeat protocol for proprietary models are given in Appendix F; comparison models, inference settings, the statistical protocol, and the ablation design in Appendix G. Code is available at `https://anonymous.4open.science/r/Overjet-1-code-D205`. Due to privacy and licensing restrictions, we cannot release the whole training corpus; Overjet-Bench and a public demo will be released upon acceptance.

REFERENCES

Anthropic. Claude Opus 5 system card. `https://www.anthropic.com/claude-opus-5-system-card`, 2026. System card published 24 July 2026. Baseline run zero-shot through the API using `claude-opus-5`; accessed 11 August 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Perez-Garcia, Valentina Salvatelli, Harshita Sharma, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Devichand Budagam, Azamat Zhanatuly Imanbayev, Iskander Rafailovich Akhmetov, Aleksandr Sinitca, Sergey Antonov, and Dmitrii Kaplun. OralBBNet: Spatially guided dental segmentation of panoramic X-rays with bounding box priors. *arXiv preprint arXiv:2406.03747*, 2024. Code and the UFBA-425 dataset: `https://github.com/devichand579/Instance_seg_teeth`.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Kubra Demir, Ozlem Sokmen, Isil Karabey Aksakalli, and Kubra Torenek-Agirman. Comprehensive insights into artificial intelligence for dental lesion detection: A systematic review. *Diagnostics*, 14(23):2768, 2024. doi: 10.3390/diagnostics14232768.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research (TMLR)*, 2023.

Yuxuan Fan, Jing Hao, Hong Chen, Jiahao Bao, Yihua Shao, Yuci Liang, Kuo Feng Hung, and Hao Tang. OralGPT-Plus: Learning to use visual tools via reinforcement learning for panoramic X-ray analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.06366.

Mohamed I. Fayad, Madhu Nair, Martin D. Levin, Erika Benavides, Richard A. Rubinstein, Sevin Barghan, Craig S. Hirschberg, and Axel Ruprecht. AAE and AAOMR joint position statement: Use of cone beam computed tomography in endodontics 2015 update. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 120(4):508–512, 2015. doi: 10.1016/j.oooo.2015.07.033.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9.

Google DeepMind. Gemini 3.1 Pro: Model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026a. Model card published 19 February 2026. Baseline run zero-shot through the API using `gemini-3.1-pro-preview`.

Google DeepMind. Gemini 3.5 Flash-Lite: Model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/`, 2026b. Model card published 21 July 2026. Used through the Vertex AI API, model identifier `gemini-3.5-flash-lite`.

Google DeepMind. Gemini 3.7 Flash: Model card. `https://deepmind.google/models/model-cards/gemini-3-7-flash/`, 2026c. Model card published 13 August 2026. Used as the LLM judge for the MMOral open-ended split.

Google DeepMind. Gemma 4: Model card. `https://ai.google.dev/gemma/docs/core/model_card_4`, 2026d. Checkpoint google/gemma-4-26B-A4B-it, `https://huggingface.co/google/gemma-4-26B-A4B-it`.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Ibrahim Ethem Hamamci, Sezgin Er, Omer Faruk Durugol, Gulsade Rabia Cakmak, Ezequiel de la Rosa, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, et al. DENTEX: Dental enumeration and tooth pathosis detection benchmark for panoramic X-ray. *arXiv preprint arXiv:2305.19112*, 2023. Revised challenge report, 2025.

Jin Hao, Xiwei Dai, Zijie Meng, Botang Yuan, Tong Jiao, Bin Feng, Huikai Wu, Yang Feng, Dian Jing, Jun Zhao, Joey Tianyi Zhou, Bing Fang, Zuozhu Liu, and Lunguo Xia. OrthoAgent: A knowledge-enhanced multi-agent framework for multimodal orthodontic diagnosis and treatment planning. *Dental Research*, 1(3):100039, 2026a. doi: 10.1016/j.dtrs.2026.100039.

Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H. Ai, Lun M. Wong, Hao Tang, and Kuo Feng Hung. Towards better dental AI: A multimodal benchmark and instruction dataset for panoramic X-ray analysis. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, Qiankun Li, Hao Tang, James Kit-Hon Tsoi, Linlin Shen, and Kuo Feng Hung. OralGPT-Omni: A versatile dental multimodal large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. arXiv:2511.22055.

David Herrera, Maurizio S. Tonetti, Iain Chapple, Moritz Kebschull, Panos N. Papapanou, Anton Sculean, Loreto Abusleme, Mario Aimetti, Georgios Belibasakis, Juan Blanco, et al. Consensus report of the 20th European Workshop on Periodontology: Contemporary and emerging technologies in periodontal diagnosis. *Journal of Clinical Periodontology*, 52(S29):4–33, 2025. doi: 10.1111/jcpe.14152.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Gil Jader, Jefferson Fontineli, Marco Ruiz, Kalyf Abdalla, Matheus Pithon, and Luciano Oliveira. Deep instance segmentation of teeth in panoramic X-ray images. In *SIBGRAPI Conference on Graphics, Patterns and Images*, 2018.

Debesh Jha, Tapas Kumar Dutta, Roshan Paudel, Onkar Kishor Susladkar, Aashish Ghimire, Anupam Dhakal, Sabin Adhikari, Nand Kumar Yadav, Deepak Ranjan Nayak, Pravin Pawar, William C. W. Chen, and Glenda Reynolds. DentiAsk: A VQA benchmark for multimodal reasoning in panoramic dental radiographs. *arXiv preprint arXiv:2607.08790*, 2026.

Osama Khattak, Ahmed Shawkat Hashem, Mohammed Saad Alqarni, Raha Ahmed Shamikh Almufarrij, Amna Yusuf Siddiqui, Rabia Anis, Shahzad Ahmad, Muhammad Amber Fareed, Osama Shujaa Alothmani, Lama Habis Samah Alkhershawy, Wesam Waleed Zain Alabidin, Rakhi Issrani, and Anshoo Agarwal. Deep learning applications in dental image-based diagnostics: A systematic review. *Healthcare*, 13(12):1466, 2025. doi: 10.3390/healthcare13121466.

Yahia H. Khubrani, David Thomas, Paddy J. Slator, Richard D. White, and Damian J. J. Farnell. Detection of periodontal bone loss and periodontitis from 2D dental radiographs via machine learning and deep learning: systematic review employing APPRAISE-AI and meta-analysis. *Dentomaxillofacial Radiology*, 54(2):89–108, 2025. doi: 10.1093/dmfr/twae070.

Chenjun Li, Laurin Lux, Alexander H. Berger, Martin J. Menten, Mert R. Sabuncu, and Johannes C. Paetzold. Fine-tuning vision language models with graph-based knowledge for explainable medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 198–207. Springer, 2025a. doi: 10.1007/978-3-032-05185-1_20.

Chenjun Li, Cheng Wan, Laurin Lux, Alexander Berger, Richard B. Rosen, Martin J. Menten, and Johannes C. Paetzold. Synthetic vasculature and pathology enhance vision-language model reasoning. *arXiv preprint arXiv:2512.11060*, 2025b.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Meng-Xun Li, Wen-Hui Deng, Zhi-Xing Wu, Chun-Xiao Jin, Jia-Min Wu, Yue Han, James Kit Hon Tsoi, Gui-Song Xia, and Cui Huang. MetaDent: Labeling clinical images for vision-language models in dentistry. *Journal of Dental Research*, 2026. doi: 10.1177/00220345261424242.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.09838.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, et al. Thinking with visual primitives. Technical report, DeepSeek-AI, April 2026. URL `https://huggingface.co/datasets/NodeLinker/deepseek-ai-Thinking-with-Visual-Primitives-deleted-repo`. Released 30 April 2026 and later withdrawn from the official GitHub repository; mirror accessed 18 September 2026.

Haoxin Lv, Ijazul Haq, Jin Du, Jiaxin Ma, Binnian Zhu, Xiaobing Dang, Chaoan Liang, Ruxu Du, Yingjie Zhang, and Muhammad Saqib. A benchmark multimodal oro-dental dataset for large vision-language models. *Scientific Data*, 13:990, 2026. doi: 10.1038/s41597-026-07342-9.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Zijie Meng, Xiwei Dai, Yixuan Tang, Jin Hao, Yang Feng, Fudong Zhu, Xiaoqiang Liu, Shaosheng Cao, and Zuozhu Liu. DentAgent: Evidence-centric multi-agent coordination for multimodal dental reasoning. *arXiv preprint arXiv:2608.18878*, 2026a.

Zijie Meng, Jin Hao, Xiwei Dai, Yang Feng, Jiaxiang Liu, Bin Feng, Huikai Wu, Xiaotang Gai, Hengchuan Zhu, Tianxiang Hu, et al. A multimodal vision-language model for comprehensive dental diagnosis and enhanced clinical practice. *Nature Communications*, 17:8933, 2026b. doi: 10.1038/s41467-026-75718-x.

Meta Superintelligence Lab. Muse Glimmer-30B model card. Official model card for the meta-models/Muse-Glimmer-30B checkpoint, August 2026. URL https://huggingface.co/meta-models/Muse-Glimmer-30B. 30-billion-parameter open-weight vision-language model with a dedicated perception encoder.

Hossein Mohammad-Rahimi, Saeed Reza Motamedian, Mohammad Hossein Rohban, Joachim Krois, Sergio E. Uribe, Erfan Mahmoudinia, Rata Rokhshad, Mohadeseh Nadimi, and Falk Schwendicke. Deep learning for caries detection: A systematic review. *Journal of Dentistry*, 122:104115, 2022. doi: 10.1016/j.jdent.2022.104115.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, 2026. System card published 9 July 2026. Baseline run zero-shot through the API using gpt-5.6-sol.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025. doi: 10.1007/978-3-032-04981-0_32. arXiv:2502.19634. 2B parameters, 600 VQA training samples.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e112a4671e8779aa9f640a0e3f81bd26-Abstract-Conference.html.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Official model card for the Qwen/Qwen3.6-35B-A3B checkpoint, April 2026. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. 35B parameters in total, 3B activated.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. GLaMM: Pixel grounding large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081. Checkpoint: google/medgemma-1.5-4b-it; model card: https://huggingface.co/google/medgemma-1.5-4b-it.

Nripendra Kumar Singh and Khalid Raza. Progress in deep learning-based dental and maxillofacial image analysis: A systematic review. *Expert Systems with Applications*, 199:116968, 2022. doi: 10.1016/j.eswa.2022.116968.

Esra Sivari, Güler Burcu Senirkentli, Erkan Bostanci, Mehmet Serdar Guzel, Koray Acici, and Tunc Asuroglu. Deep learning in diagnosis of dental anomalies and diseases: A systematic review. *Diagnostics*, 13(15):2512, 2023. doi: 10.3390/diagnostics13152512.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations (ICLR)*, 2025a.

Rongyang Wang, Shuang Zhou, Jiashuo Wang, Wenya Xie, and Xiaoxia Che. OralMLLM-Bench: Evaluating cognitive capabilities of multimodal large language models in dental practice. *arXiv preprint arXiv:2605.01333*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://huggingface.co/OpenGVLab/InternVL3_5-30B-A3B. OpenGVLab/InternVL3_5-30B-A3B checkpoint; 30.8B parameters in total, 3B activated.

Sri Kalyan Yarlagadda, Navid Samavati, Mina Ghorbanifarajzadeh, Vlada Levinta, Alireza Sojoudi, Wardah Inam, and Teresa A Dolan. Development and validation of an AI-enabled oral score using large-scale dental data. *Scientific Reports*, 15(1):20398, 2025. doi: 10.1038/s41598-025-07484-7.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations (ICLR)*, 2024.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

Zhenhuan Zhou, Jingbo Zhu, Yuchen Zhang, Xiaohang Guan, Peng Wang, and Tao Li. Deep learning in dental image analysis: A systematic review of datasets, methodologies, and emerging challenges. *arXiv preprint arXiv:2510.20634*, 2025. URL https://arxiv.org/abs/2510.20634.

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, Yih Chung Tham, and Huazhu Fu. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11315–11329, 2025. doi: 10.1109/TPAMI.2025.3596878.

APPENDIX

A LIMITATIONS

The training supervision may inherit errors from its source systems. Much of the geometry comes from production segmentation and detection models, so an incorrect region can affect multiple records generated from the same case. The teacher model may also introduce omissions or unclear phrasing in longer targets. Our validation catches malformed coordinates and inconsistencies with the structured facts, but it cannot identify every clinically plausible error.

The evaluation also has limited external coverage. Although Overjet-Bench is patient-disjoint, it is drawn from the same clinical database as the training corpus. Its results therefore do not establish performance on data from other institutions, imaging devices, acquisition protocols, or patient populations. MMOral-OPG-Bench provides an external evaluation, but it covers only panoramic radiographs and question answering.

B TRAINING DETAILS

Supervised fine-tuning. Overjet-1 is a full fine-tune of Gemma 4 26B-A4B, a mixture-of-experts vision-language model with 25.2B total parameters and 3.8B activated during inference. All parameters are trained with bf16 and DeepSpeed ZeRO-3 on eight H200 GPUs.

Table 4: Supervised fine-tuning configuration.

Component	Configuration
Backbone	Gemma 4 26B-A4B, instruction-tuned; full-parameter training
Hardware and sharding	8×H200, bf16, DeepSpeed ZeRO-3
Objective	Cross-entropy on assistant tokens; prompt and visual tokens masked
Context	20,480 tokens; longest training record 19,623 tokens
Token budget	14,336 unpadded tokens per rank per step
Visual budget	1,120 tokens per view for up to 8 images; 560 for 9–16; 280 for 17 or more
Optimization	Learning rate 1×10^{-5}, cosine schedule, 3% warmup, gradient clip 0.3
Schedule	1 epoch; 23,398 optimizer steps
Training volume	2.68B tokens over 256,600 packed batches; 87% packing rate

DC-RL. DC-RL starts from the final SFT checkpoint and uses 1,120 visual tokens per image.

We apply DC-RL to the full supervised checkpoint. Each step samples 16 prompts and eight completions per prompt. The DC-RL mixture contains 11,679 image-based records, 84% of which include reference boxes. DC-RL runs for 500 steps using the configuration in Table 5. We also train a matched control that removes the spatial-consistency contribution from graph grounding while keeping the checkpoint, training data, random seeds, optimizer, rollout settings, and step budget fixed. Comparing this control with supervised fine-tuning measures the effect of DC-RL without spatial consistency, while comparing it with full DC-RL isolates the contribution of spatial consistency.

C DATA CONSTRUCTION AND PROMPT DESIGN

Grounding and localization account for 34.2% of the records, and visual question answering and classification account for 32.6%; the remaining records cover reports, structured reads, multi-image reasoning, clinical decisions, knowledge, and volumetric CBCT tasks (Table 6).

Table 5: DC-RL configuration.

Component	Configuration
Method	GRPO with DAPO token normalization and batch-scaled rewards
Hardware and sharding	8×H200, bf16, DeepSpeed ZeRO-3
Sampling	16 prompts per step, 8 completions per prompt, temperature 1.0, top-p 0.95
Length	16,384-token context; 4,096-token completion limit
Optimization	500 steps, language learning rate 2×10^{-6}, vision learning rate 5×10^{-7}
Regularization	$\beta = 0.1$, 3% warmup, gradient clip 1.0
Reward	$c(1 + 0.35g + 0.15e + 0.05f)$ from Equation (1)
Generation mode	Evidence trace followed by a separately parsed final answer

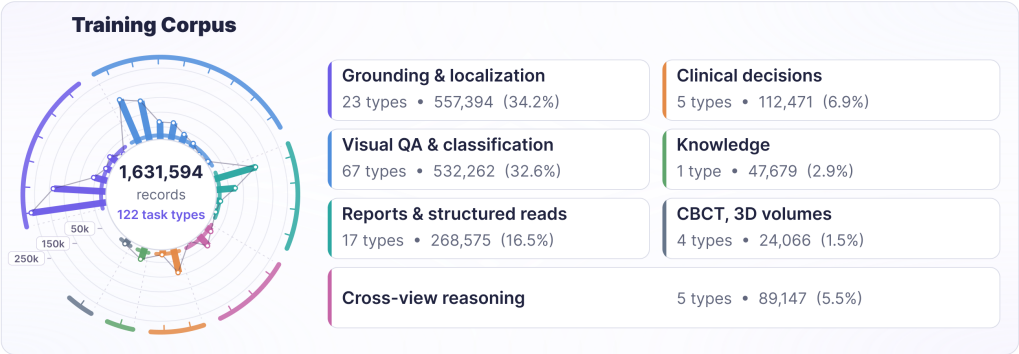


Figure C1: Distribution of 1,631,594 training records over 122 task types.

Table 6: **Training distribution.** The 122 task types are grouped by their target. Each task type appears once, and percentages use all 1,631,594 records. Among the 1,607,528 2D and text records, 54.2% include a supported reasoning trace. Single-view, multi-view, and text-only records account for 91.0%, 5.0%, and 3.9%, respectively. Records contain up to 38 views. Four volumetric task types add 24,066 records for the 3D encoder in Appendix E.

Family and category	Task types	Records	%
Grounding & localization		**557,394**	**34.2**
Tooth grounding	6	261,821	16.0
Pathology localization	6	183,107	11.2
Reverse grounding	1	52,409	3.2
Landmarks & keypoints	3	23,879	1.5
CBCT, 2D views	5	32,885	2.0
Restorations	2	3,293	0.2
Visual QA & classification		**532,262**	**32.6**
Closed questions	2	169,502	10.4
Multiple choice	34	143,584	8.8
Tooth identity	11	65,720	4.0
Bone level	2	66,372	4.1
Modality & position	4	38,285	2.3
Counting & eruption	3	25,115	1.5
Orthodontic measures	6	7,803	0.5
CBCT views	5	15,881	1.0
Reports & structured reads		**268,575**	**16.5**
Whole-image reports	4	155,229	9.5
Per-tooth reports	1	73,843	4.5
Chart transcription	3	23,971	1.5
Text Extraction	9	15,532	1.0
Cross-view reasoning		**89,147**	**5.5**
Paired-view comparison	2	30,530	1.9
Label repair	2	54,703	3.4
Full-mouth reconciliation	1	3,914	0.2
Clinical decisions		**112,471**	**6.9**
Report Validation & planning	4	89,889	5.5
Document pages	1	22,582	1.4
Knowledge		**47,679**	**2.9**
Text-only QA	1	47,679	2.9
CBCT, 3D volumes	4	24,066	1.5
Total	**122**	**1,631,594**	**100.0**

C.1 DENTAL GRAPH CONSTRUCTION

The Dental Graph integrates information such as tooth outlines and numbers, finding regions, restorations, bone measurements, cephalometric landmarks, periodontal charts, and CBCT annotations. Each view-specific tooth node stores identity, geometry, findings, restorations, and available measurements. Within-view edges encode arch adjacency, while cross-view edges connect the same physical tooth.

C.2 RECORD GENERATION

Each task type defines a Dental Graph query, prompt template, and target schema. The query fixes the relevant teeth, findings, relations, and geometry before the prompt wording is rendered. Short targets use task-specific templates. For longer targets, Gemini 3.5 Flash-Lite rewrites the graph-derived facts and reasoning traces in natural language; all supplied coordinates are copied exactly.

For example, the record in Figure 3 queries the decay in the G3 periapical view, follows its finding-to-tooth link to tooth #21, and retrieves the tooth type and lesion surfaces. The rendered prompt requests the tooth ID, type, involved surfaces, and lesion box; the target supplies tooth #21, *first premolar*, *mesial* and *occlusal*, and the normalized box `[705, 619, 824, 701]`.

Verified reasoning traces follow an annotation-derived graph walk from the requested structure to the final answer. Record validation checks the schema, tooth identities, coordinates, clinical claims, and graph alignment. It also removes annotation-only metadata, exact duplicates, and repeated reasoning text. Supported abstention targets are generated only when the source target explicitly rules out every listed tooth. Separately audited narrative and chart tasks retain their existing traces, while records without sufficient evidence contain only the answer.

Table 7: **Record construction pipeline.** Each record is generated from a validated Dental Graph query and its task-specific target schema.

#	Stage	What it does
1	Eligibility	Retain records with an answerable prompt, complete target fields, valid tooth identities, and valid geometry.
2	Clinical normalization	Reconcile source-group facts and apply the task-specific clinical rules.
3	Choice construction	Balance distractors, answer positions, and the required multiple-choice layout.
4	Dental Graph query	Retrieve the relevant teeth, findings, relations, measurements, and geometry.
5	Target rendering	Render the task schema, copy exact coordinates, and compute structured chart totals.
6	Reasoning trace	Attach an annotation-derived graph walk when every step is supported.
7	Validation	Check schema, content, and grounding; deduplicate records; add supported abstention variants (Appendix C.2).

C.3 General prompt structure

Table 8: General prompt structure used to form training records. The system body is fixed; only the optional reasoning prefix and task-specific content vary by record.

Part	Structure shown to the model
Turn order	SYSTEM → USER (ordered media and task instruction) → ASSISTANT (optional verified reasoning and final answer).
System message	*You are an expert dental AI assistant. Review the provided radiographs, photographs, documents, or clinical text and answer the question.* • Refer to teeth by Universal Numbering System (UNS) number, #1–#32. • Give any region as `[x1, y1, x2, y2]` in 0–1000 normalized coordinates. • Ground every claim in the evidence; do not state findings it does not support. • Answer in exactly the format the question asks for. For multiple choice, end with the single letter (A/B/C/D) on its own line.
Reasoning control	A record with a verified reasoning trace prepends `<|think|>` to the otherwise identical system message, which activates the model's reasoning channel. A record without a trace uses no prefix and no empty reasoning block.
User message	One `<image>` marker per ordered visual input, followed by the task-specific question or instruction. The task text may also supply view identifiers, a query region, answer choices, and the required response schema. Text-only records omit the media markers.
Assistant target	Optional verified reasoning in the reasoning channel, followed by the final answer in the requested task schema. Without a verified trace, the final answer follows the user turn directly.

C.4 Prompt vocabulary

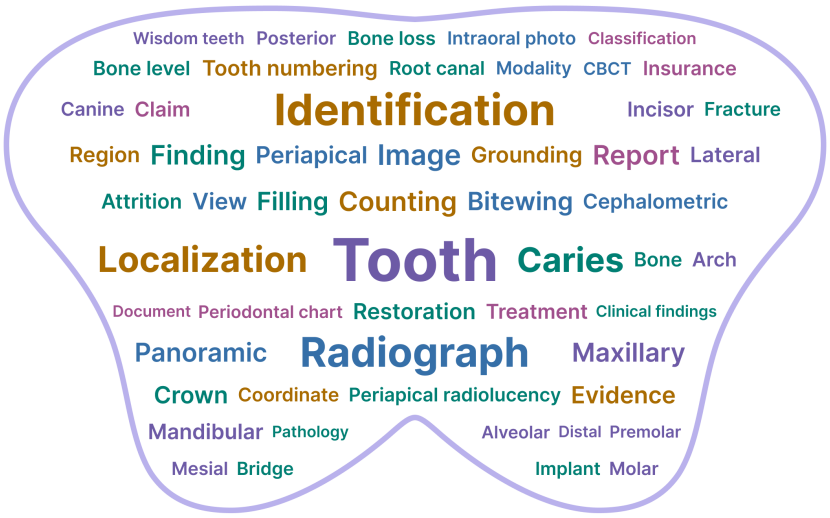


Figure C2: **Prompt word cloud.** Word size represents document frequency across user prompts, with each normalized term counted at most once per prompt. Color denotes semantic family.

D Algorithms and Reward Details

D.1 Record construction

Algorithm 1 Dental Graph record construction.

Input: source record r, Dental Graph $\mathcal{G}$
Returns: validated record set R

if r lacks a supported target **then return** $\emptyset$
$r \leftarrow \textsc{NormalizePromptAndTarget}(r)$
$T \leftarrow \textsc{Query}(\mathcal{G}, r)$
$r \leftarrow \textsc{Render}(r, T)$; $\tau \leftarrow \textsc{None}$
$\tau \leftarrow \textsc{ReasoningTrace}(T)$ when supported
$R \leftarrow \textsc{Validate}(\{(r, \tau)\}, T)$
if T supports explicit abstention **then**
    $R \leftarrow R \cup \{\textsc{Abstention}(r, T)\}$
end if
return $\textsc{Deduplicate}(R)$

D.2 ANNOTATION-DERIVED REASONING

Algorithm 2 Reasoning trace derived from Dental Graph evidence.

Input: question q, answer a, Dental Graph $\mathcal{G}$
Returns: reasoning trace or NONE

$T \leftarrow \text{QUERY}(\mathcal{G}, q, a)$
if T lacks the required clinical evidence **then return** NONE
for all relevant view v in prompt order **do**
    for all tooth $t \in T_v$ in anatomical order **do**
        emit the tooth identity and source region for t
    end for
    for all finding $z \in T_v$ **do**
        emit z, its tooth identity, and its source region
    end for
end for
emit the conclusion specified by T

D.3 DC-RL REWARD COMPONENTS

Target representation and reward. For each training record r, the Dental Graph query produces the typed target

$$T_r = (a_r, \mathcal{T}_r, \mathcal{C}_r, \mathcal{B}_r, \mathcal{R}_r, \mathcal{P}_r), \tag{2}$$

where a_r is the requested answer, $\mathcal{T}_r$ contains tooth identities, $\mathcal{C}_r$ contains tooth–finding–polarity claims, and $\mathcal{B}_r$, $\mathcal{R}_r$, and $\mathcal{P}_r$ contain tooth boxes, finding boxes, and keypoint coordinates. The model receives only the prompt and images, while T_r is used by the deterministic reward.

Multi-image and multi-modal targets retain shared tooth identities and clinical claims from the Dental Graph. Reference geometry remains tied to its source view or modality. The scorer retains one reference occurrence per tooth and evaluates 2D boxes, chart entries, and 3D cuboids in their native spaces.

A prediction is parsed into an evidence trace $\hat{h}$, final answer $\hat{a}$, claims (tooth–finding–polarity tuples), tooth identities, findings, and normalized coordinates. The DC-RL reward is

$$r(\hat{y}) = c(\hat{y})\left[1 + 0.35g(\hat{y}) + 0.15e(\hat{y}) + 0.05f(\hat{y})\right]. \tag{3}$$

Here, every component lies in $[0, 1]$ and $0 \le r \le 1.55$. The multiplicative factor c makes clinical correctness a prerequisite for the grounding g, evidence e, and format terms f. Set-valued targets receive proportional credit through their task-specific F_1 scores.

Clinical correctness c. The target schema selects the correctness function summarized in Table 9. Counts use exact match, multiple-choice questions use the final selected option, tooth sets use set F_1, and per-tooth statements are compared by tooth identity, finding, and polarity. Localization-only tasks use a presence check because their spatial accuracy is measured by g.

Negative claims are scored only when the target explicitly encodes absence.

Grounding score g. Tooth boxes combine mean overlap with one-to-one detection F_1, keypoints use normalized distance, and finding regions combine overlap recall with a penalty for predicting too many or too few regions. Image coordinates are integers on a 0–1000 scale, and valid regions have positive area. For identity-aware tooth boxes, let $\hat{b}_t$ and b_t denote the predicted and reference box for tooth t. The grounding score is evaluated by a function based on intersection-over-union (IoU):

$$\overline{\text{IoU}}_{\text{target}} = \frac{1}{|\mathcal{B}_r|} \sum_{t \in \mathcal{B}_r} \text{IoU}(\hat{b}_t, b_t). \tag{4}$$

Unmatched targets contribute zero to the target-wise mean. A tooth is detected when the identity agrees and $\text{IoU}(\hat{b}_t, b_t) \ge 0.3$. Precision includes unmatched and duplicate predictions. The identity-aware score is

$$i_{\text{tooth-bbox}} = 0.75\overline{\text{IoU}}_{\text{target}} + 0.25F_1^{\text{det}}. \tag{5}$$

Table 9: Task-specific correctness functions used by DC-RL.

Target type	Clinical correctness
Multiple choice	Exact match to the final committed option; an option mentioned only inside the evidence trace is not a verdict.
Count	Exact integer match; an explicit enumeration is also counted when no total is written.
Tooth set or tooth localization	Set F_1 over canonical tooth identities.
Per-tooth clinical claim	$0.55F_1^{\text{tooth}} + 0.45F_1^{\text{claim}}$, where a claim is a tooth, finding, and polarity tuple.
Region or keypoint localization	Presence of the requested coordinate type; geometric accuracy is scored by g.
Jaw or clinical label	F_1 over canonical typed claims.
Other graph query	Equal-weight tooth and claim F_1 when typed entities exist; otherwise content-token F_1 against a_r.

With at least two predicted tooth boxes, spatial consistency s (defined below) is included as

$$g_{\text{tooth-bbox}} = 0.9 i_{\text{tooth-bbox}} + 0.1 s; \tag{6}$$

otherwise $g_{\text{tooth-bbox}} = i_{\text{tooth-bbox}}$.

Reference keypoints are paired greedily to predictions by Euclidean distance, where the distance is denoted d. With $d_{\max} = 150$ normalized-coordinate units and $q = \min(n_{\text{pred}}, n_{\text{ref}}) / \max(n_{\text{pred}}, n_{\text{ref}})$,

$$g_{\text{point}} = 0.8 \, \overline{\max(0, 1 - d/d_{\max})} + 0.2q. \tag{7}$$

Generic finding regions use greedy one-to-one IoU matching. If u is mean IoU recall over the reference regions, their score is

$$g_{\text{findings-bbox}} = 0.8u + 0.2q \min(1, u/0.1). \tag{8}$$

Finally, the score g is the sum of $g_{\text{tooth-bbox}}, g_{\text{point}}$, and $g_{\text{findings-bbox}}$. When reference geometry is unavailable, positive targets use the capped topology term $g = 0.2s$, while fully negative targets use $g = 0$.

Spatial consistency s. For every predicted tooth box, let (x_t, y_t) be its center and w_t its width, and let $\tilde{w}$ be the median positive box width in the prediction. The canonical sweep order is $(1, \ldots, 16)$ for the maxillary arch and $(32, \ldots, 17)$ for the mandibular arch. For successive reported teeth t, u in one arch, DC-RL computes

$$o_{tu} = \mathbf{1}[x_u > x_t], \tag{9}$$

$$p_{tu} = \max\left(0, 1 - \frac{|\,|x_u - x_t|/\tilde{w} - 1\,|}{2}\right), \tag{10}$$

$$a_{tu} = 0.6 o_{tu} + 0.4 p_{tu}. \tag{11}$$

The within-arch term is the mean of a_{tu} over reported successive teeth. When both arches are present, the separation term is $\mathbf{1}[\overline{y}_{\max} < \overline{y}_{\text{mand}}]$. The duplicate term is the fraction of distinct tooth pairs whose boxes have IoU below 0.85. The final s is the mean of all applicable arch, separation, and duplicate terms. Fewer than two valid tooth-box pairs give $s = 0$. These relations use relative order and scale within one 2D view. Note that the cross-image and multi-modal relations are scored through the shared Dental Graph identities in c and e. For cross-image and multi-modal tasks, the Dental Graph supplies the shared tooth identities and typed clinical facts used by the correctness and evidence terms. The scalar geometry target retains one reference occurrence per tooth rather than defining overlap between incommensurate views or between 2D and 3D coordinates.

Evidence and format scores e and f. The evidence trace is parsed separately from the final answer. Tooth evidence uses set F_1, and clinical evidence uses F_1 over tooth–finding–polarity tuples. Their product is used when both are required. Count evidence uses exact agreement. Cross-image and multi-modal records use the shared identities and claims returned by the Dental Graph query.

The format score is 1 for a closed evidence channel followed by a nonempty answer, 0.25 for a valid direct or separately recoverable answer, and 0 when no answer is present. Its small weight distinguishes clinically correct predictions with different response structures.

Boundary conditions. Table 10 summarizes reward behavior across task and annotation conditions.

Table 10: DC-RL reward behavior across task and annotation conditions.

Case	Reward behavior
Incorrect exact-match answer	$c = 0$, so the complete reward is zero.
Set-valued answer	Set F_1 provides proportional c, which scales all remaining terms.
Unavailable or invalid coordinates	Unmatched targets contribute zero overlap; out-of-range boxes are excluded.
Single tooth box	Reference overlap is scored, while $s = 0$.
One visible arch	Available arch terms are renormalized without an arch-separation term.
Positive target without reference geometry	Grounding uses the capped topology term $g = 0.2s$.
Unspecified annotation	The claim is omitted from the typed target.
Cross-image input	Shared identities and relations are scored through c and e; geometry remains view-specific.
Cross-modal or 2D–3D input	Shared semantics are scored through c and e; each geometry type uses its native coordinate space.

No-spatial-consistency ablation. The control run starts from the same full-SFT checkpoint and uses the same records, data order, seeds, optimizer, rollout configuration, and step budget. Its reward is

$$r_{\neg s}(\hat{y}) = c(\hat{y})\,(1 + 0.35 g_{\neg s}(\hat{y}) + 0.15 e(\hat{y}) + 0.05 f(\hat{y}))\,. \tag{12}$$

For tooth boxes, $g_{\neg s} = i_{\text{tooth-bbox}} + g_{\text{point}} + g_{\text{findings-bbox}}$. Keypoint distance, region overlap, clinical correctness, evidence, format, and absolute grounding remain unchanged.

E Direct CBCT Encoding and Cross-modal Evaluation

The main image encoder processes CBCT-derived 2D views. Direct CBCT volumes are processed by a dedicated encoder comprising a frozen 3D vision backbone, a projector, and language adapters. The adapters are trained with paired 2D and 3D records from the same examination.

The vision backbone is a hierarchical 3D ConvNeXt (Liu et al., 2022). Volumes are resampled to $240 \times 272 \times 336$ voxels. A stride-4 stem feeds four stages with depths $(3, 3, 9, 3)$ and widths $(64, 128, 256, 512)$. A feature pyramid predicts anatomy segmentation, and three-axis sinusoidal embeddings encode position. The language model receives fixed semantic slots followed by a coarse spatial grid. Slots represent teeth, pulp, named structures, and per-tooth restoration states; tooth geometry also locates each restoration.

Paired cross-modal records connect a 2D box and 3D cuboid through a shared tooth identity. The model identifies the tooth from either endpoint and returns its geometry in the paired modality. Figure E1 shows both directions for tooth #15 in one panoramic–CBCT examination.

Algorithm 3 Deterministic DC-RL reward. Cross-view and cross-modal relations use typed Dental Graph targets, while geometry remains view-specific.

Input: prediction $\hat{y}$, graph-derived target T_r
Returns: scalar reward

$(\hat{h}, \hat{a}, \hat{\mathcal{T}}, \hat{\mathcal{C}}, \hat{\mathcal{Z}}) \leftarrow \textsc{Parse}(\hat{y})$
$c \leftarrow \textsc{TypedCorrectness}(\hat{a}, T_r)$
if $c = 0$ **then**
  return 0
end if
$s \leftarrow \textsc{SpatialConsistency}(\hat{\mathcal{Z}})$; $g_{\text{tooth-bbox}} \leftarrow 0$, $g_{\text{point}} \leftarrow 0$, $g_{\text{findings-bbox}} \leftarrow 0$
if T_r has identity-aware reference tooth boxes **then**
  $i_{\text{tooth-bbox}} \leftarrow 0.75\,\overline{\text{IoU}}_{\text{target}} + 0.25F_1^{\text{det}}$
  $g_{\text{tooth-bbox}} \leftarrow 0.9 i_{\text{tooth-bbox}} + 0.1s$ if $n_{\text{box}}^{\text{pred}} \geq 2$; otherwise $g_{\text{tooth-bbox}} \leftarrow i_{\text{tooth-bbox}}$
else if T_r has reference keypoints **then**
  $g_{\text{point}} \leftarrow 0.8\,\overline{\max(0, 1 - d/d_{\max})} + 0.2q$
else if T_r has reference regions **then**
  $u \leftarrow \text{IoURecall}$; $g_{\text{findings-bbox}} \leftarrow 0.8u + 0.2q\min(1, u/0.1)$
end if
if T_r is a fully negative target or has no spatial reference **then**
  $g \leftarrow 0$ for a fully negative target; otherwise $g \leftarrow 0.2s$
else
  $g \leftarrow g_{\text{tooth-bbox}} + g_{\text{point}} + g_{\text{findings-bbox}}$
end if
$e \leftarrow \textsc{EvidenceConsistency}(\hat{h}, T_r)$
$f \leftarrow \textsc{FormatScore}(\hat{y})$
return $c(1 + 0.35g + 0.15e + 0.05f)$

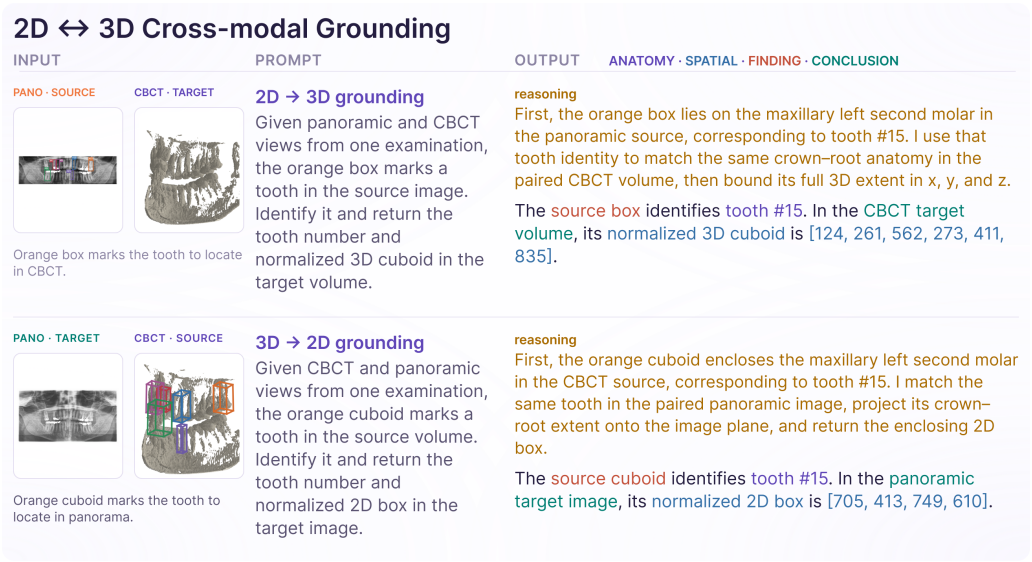


Figure E1: **Bidirectional 2D–3D cross-modal grounding.** Both rows use a panoramic radiograph and CBCT volume from the same examination. The first row maps the panoramic box for tooth #15 to its CBCT cuboid. The second maps the cuboid back to the panoramic box. Only the source modality contains the orange geometry.

Evaluation of the multi-modal utilization For multi-modal tasks, the optimal answer should be derived by aggregating information from all sources, rather than providing shortcut solutions from a single modality. To explicitly test if the model can utilize multi-modal information, we ran the following experiment. We define two evaluation slices of 200 cases that correspond directly to

Table 11: **Evaluation of the multi-modal utilization.** Primary F_1 denotes orthodontic-qualification F_1 or 2D–3D tooth-identity F_1, depending on the slice. Evidence F_1 is computed over task-specific finding or rule-evidence tuples. 2D R@.50 is recall of required evidence regions at IoU ≥ 0.50; 3D R@.25/.50 is CBCT cuboid recall at the two IoU thresholds. Predictions for a withheld target modality are scored against reference annotations; "–" denotes a metric not evaluated for that input condition. For Pano only and CBCT only experiments, only 2D $\rightarrow$ 3D and 3D $\rightarrow$ 2D directions are evaluated, respectively.

Input condition	Primary F_1	Evidence F_1	2D R@.50	3D R@.25/.50
Radiographs + intraoral photographs: orthodontic qualification ($n = 200$)				
Full evidence	**88.9**	**80.3**	**87.5**	–
Radiographs only	46.1	31.4	52.5	–
Photographs only	41.8	28.8	47.5	–
Shuffled Radiographs	28.6	16.6	44.5	–
Shuffled photographs	24.3	14.3	42.0	–
Panoramic radiograph + direct CBCT volume: bidirectional tooth grounding ($n = 200$)				
Full pair	**90.2**	**79.9**	85.5	**83.5/71.0**
Panoramic radiograph only	24.7	13.8	–	12.0/2.5
CBCT only	26.1	15.2	12.5	–
Shuffled CBCT	13.6	6.9	11.0	9.0/1.5
Shuffled Pano	12.4	5.9	9.5	8.0/1.0

multi-source tasks represented during training: grounded orthodontic qualification from radiographs and intraoral photographs, and bidirectional tooth grounding between a panoramic radiograph and a direct CBCT volume. The two slices are patient-disjoint from training and from each other.

For orthodontic qualification, policy criteria and structured clinical fields are fixed across input conditions, and every case contains scored evidence from both the radiographs and photographs. The 2D–3D tasks match a tooth between the panoramic radiograph and CBCT and require both a 2D box and a 3D cuboid.

Each full-evidence condition is compared with either source alone and with an auxiliary input shuffled from another patient with the intention of distracting the model. Shuffled photographs are matched by acquisition composition, and shuffled CBCT volumes by arch and approximate tooth coverage. All controls retain the same prompt and output schema as the full-evidence condition.

Across both slices in Table 11, the full evidence (original) gives the highest primary and evidence F_1. Orthodontic qualification reaches 88.9 primary F_1 and 80.3 evidence F_1; shuffling the photographs reduces these scores to 24.3 and 14.3. This demonstrates that the performance is reduced when only a single image source is available. The performance is also compromised when the wrong information is injected in one of the modalities. This demonstrates that the model consumes both modalities to provide the answer. In the 2D–3D slice, the full pair achieves 90.2 primary F_1, 79.9 evidence F_1, and 83.5/71.0 cuboid recall. This signifies that the model indeed extracts necessary information from all sources to provide the optimal answer, hence demonstrating the multi-modal reasoning capability.

F Overjet-Bench Specification

F.1 Composition and scoring

Overjet-Bench contains 1,000 records. Sampling is stratified by source task, modality, number of images, clinical category, and answer length, with all records from the same patient assigned to one split. Each record stores a stable identifier, ordered views, prompt, target, task family, modality, answer schema, and primary scorer. The benchmark contains 1,356 ordered view references to 1,298 unique images, with 825 single-image, 85 multi-image, and 90 multi-modal records. Single-image items test local recognition and grounding, multi-image items test tooth matching and rea-

soning across images, and multi-modal items combine 2D images with other modalities. The multi-modal subset includes radiograph–intraoral grounding and synthesis, cross-modal numbering, and orthodontic qualifier evaluation.

Table 12: **Detailed Overjet-Bench composition.** S, MI, and MM denote single-image, multi-image, and multi-modal records. BW, PA, PAN, CEPH, and IO denote bitewing, periapical, panoramic, lateral cephalometric, and intraoral photographic inputs. "Mixed" denotes a set with more than one radiographic modality. The final column gives the scorer for each task; family totals appear in Table 13.

Benchmark task	Rows	Views	Set	Modality	Scoring rule
Bone assessment (125 rows)					
Global bone-loss assessment	40	40	S	PAN	severity–distribution semantic match
Bone-loss multiple choice	40	40	S	PAN	committed-letter accuracy
Local bone-severity assessment	45	45	S	BW, PA, PAN	severity semantic match
Closed QA (100 rows)					
Radiographic closed QA	30	30	S	BW, PA, PAN	normalized closed-answer match
Intraoral-photograph closed QA	20	20	S	IO	normalized closed-answer match
Closed multiple-choice questions	50	50	S	BW, PA, PAN	committed-letter accuracy
Counting (70 rows)					
Third-molar eruption count	35	35	S	PAN	third-molar count and status
Teeth per arch	35	35	S	BW, PA, PAN	exact maxillary/mandibular counts
Localization (220 rows)					
Requested-tooth localization	55	55	S	BW, PA, PAN	reference-box recall at IoU ≥ 0.5
All-visible-teeth localization	55	55	S	BW, PA, PAN	reference-box recall at IoU ≥ 0.5
Cross-radiograph tooth localization	35	128	MI	BW/PA/PAN, mixed	per-view box recall at IoU ≥ 0.5
Radiograph–intraoral tooth localization	25	85	MM	PAN/PA + IO	per-input box recall at IoU ≥ 0.5
Radiographic pathology localization	30	30	S	BW, PA	finding-box recall at IoU ≥ 0.5
Intraoral finding localization	20	20	S	IO	finding-box recall at IoU ≥ 0.5
Modality and position (70 rows)					
Full-mouth-series position	35	35	S	PA, PAN	normalized choice or task-exact match
Radiograph modality identification	20	20	S	BW, PA, PAN	normalized choice or task-exact match
Intraoral view identification	15	15	S	IO	normalized choice or task-exact match
Cross-input reasoning (60 rows)					
Cross-radiograph comparison	20	45	MI	BW/PA/PAN, mixed	tooth-set F_1; token F_1 for free text
Radiograph–intraoral evidence synthesis	20	40	MM	PAN/PA + IO	tooth–finding tuple F_1

Continued on next page

Table 12: **Detailed Overjet-Bench composition (continued).**

Benchmark task	Rows	Views	Set	Modality	Scoring rule
Orthodontic claim qualification	20	60	MM	PAN, CEPH, IO	decision accuracy; HLD-factor F_1
Report generation (135 rows)					
Whole-image report	45	45	S	BW, PA, PAN	ROUGE-L
Open clinical synthesis	45	45	S	BW, PA, PAN	token F_1
Per-tooth report	45	45	S	BW, PA, PAN	ROUGE-L
Tooth identity (120 rows)					
Tooth identity from a marked region	40	40	S	BW, PA, PAN	exact tooth or tooth-set F_1
Highlighted-tooth identification	40	40	S	BW, PA, PAN	normalized choice or task-exact match
Neighboring-tooth discrimination	40	40	S	BW, PA, PAN	committed-letter accuracy
Tooth numbering (100 rows)					
Cross-radiograph numbering correction	30	98	MI	BW/PA/PAN, mixed	label-pair or tooth-set F_1
Radiograph–photo numbering transfer	25	75	MM	PAN/PA + IO	cross-modal tooth-set F_1
Single-view numbering correction	45	45	S	BW, PA	label-pair or tooth-set F_1
Total	**1,000**	**1,356**	**825 S / 85 MI / 90 MM**		
1,298 unique images; 29 source tasks					

Each family uses a scoring rule matched to its output. Exact answers are normalized and matched deterministically, tooth sets use set F_1, and predicted regions are matched one-to-one with reference boxes using intersection over union (IoU). Localization is scored at IoU ≥ 0.50. Structured outputs are parsed into tooth identities, findings, and polarity. Whole-image and per-tooth reports use ROUGE-L, while open clinical synthesis uses token F_1.

For scoring, exact answers, committed options, and counts use deterministic matching. Tooth sets and numbering corrections use exact match or set F_1. Localization uses one-to-one matching within each view and recall at IoU 0.5. Report tasks use ROUGE-L or token F_1. The macro score is therefore a summary of task coverage rather than a single accuracy measure. Scores are averaged within each source task and then within each family. The reported overall score is the unweighted mean of the nine family scores.

Evaluation protocol. The closed-ended split reports accuracy for teeth, pathology, treatment, bone, summary, and overall performance. Primary evaluation uses fixed decoding and task-specific length limits. Each run stores the model identifier, prompt, decoding parameters, prediction, and scorer version. Localization matches predicted and reference regions within each view under a one-to-one constraint. Proprietary-model evaluations use dated model identifiers, identical prompts, and the same scorer implementations. The open-ended split follows the original LLM-as-a-judge protocol: Gemini 3.7 Flash (Google DeepMind, 2026c) receives the benchmark's nine-example few-shot prompt together with the question, reference answer, and prediction, and assigns a score from 0 to 1. The prompt construction, score levels, and aggregation otherwise follow the original protocol (Hao et al., 2025). Each proprietary model is evaluated in three independent complete calls for each MMOral split and for Overjet-Bench; Table 14 reports the individual scores and their sample variation.

F.2 External COde evaluation

We evaluate six-class anomaly classification on the external COde dataset (Lv et al., 2026). We follow the original task definition: the model receives the available demographics, chief complaint,

Table 13: **Overjet-Bench evaluation set.** The 1,000 records include 825 single-image, 85 multi-image, and 90 multi-modal cases over 1,298 unique images. Sampling is stratified by source task, modality, number of views, category, and answer length. Exact targets use deterministic metrics, structured targets use deterministic parsers, and reports use task-specific automatic content metrics. The final case-disjoint evaluation set retains these families and scoring rules.

Family	Rows	Signal
Localization	220	IoU recall
Report generation	135	ROUGE-L
Bone assessment	125	graded token
Tooth identity	120	exact tooth
Closed QA	100	committed letter
Tooth numbering	100	tooth-set match
Counting	70	exact integer
Modality & position	70	exact choice
Cross-input	60	tuple/decision match
Total	**1,000**	**29 tasks**

Table 14: **Repeatability of proprietary-model evaluations.** MMOral Closed cells show correct/491 with accuracy in parentheses. Open uses Gemini 3.7 Flash under the benchmark's original nine-example few-shot rubric on 578 questions and reports the mean 0–1 score as a percentage. Overjet-Bench reports the unweighted nine-family macro. Summary statistics use the unrounded scorer outputs. SD and s^2 are the sample standard deviation and variance over three independent complete API calls; Range is max–min.

Model	Split	Run 1	Run 2	Run 3	Mean $\pm$ SD	s^2	Range
MMOral-OPG-Bench							
GPT-5.6 Sol	Closed	304/491 (61.9)	293/491 (59.7)	306/491 (62.3)	61.30 $\pm$ 1.43	2.03	2.65
	Open	64.7	59.0	59.0	60.90 $\pm$ 3.29	10.81	5.73
Gemini 3.1 Pro	Closed	276/491 (56.2)	289/491 (58.9)	269/491 (54.8)	56.62 $\pm$ 2.07	4.27	4.07
	Open	65.4	59.5	65.9	63.60 $\pm$ 3.55	12.61	6.35
Claude Opus 5	Closed	282/491 (57.4)	284/491 (57.8)	244/491 (49.7)	54.99 $\pm$ 4.59	21.07	8.15
	Open	58.7	64.4	54.9	59.30 $\pm$ 4.78	22.80	9.49
Overjet-Bench macro							
GPT-5.6 Sol	Overall	26.63	30.55	32.47	29.88 $\pm$ 2.98	8.86	5.84
Gemini 3.1 Pro	Overall	19.81	24.62	17.46	20.63 $\pm$ 3.65	13.34	7.17
Claude Opus 5	Overall	27.34	21.10	30.17	26.20 $\pm$ 4.64	21.51	9.06

radiographs, and intraoral photographs and predicts one of six anomaly categories. We report accuracy, weighted precision, and weighted F_1.

Overjet-1 was not fine-tuned on COde. It reaches 76.63% accuracy and 76.35% weighted F_1, matching the COde-fine-tuned 3B reference model and approaching the fine-tuned 7B model. All rows use the complete 1,793-example evaluation set.

Table 15: **Six-class anomaly classification on COde (%).** We follow the task definition and metrics of the original study: accuracy, weighted precision (wP), and weighted F_1 (wF_1). All rows use the complete evaluation set; only the two marked reference models were fine-tuned on COde. † denotes COde-fine-tuned results reported by Lv et al. (2026). Best and second-best values are bolded and underlined. All other models were evaluated zero-shot, including our model Overjet-1. Overjet-1 not only achieved the best zero-shot accuracy, but also outperformed the 3B model fine-tuned on the reference dataset.

Model	Params	Accuracy	wP	wF_1
Gemma 4 26B-A4B base	26B	52.04	62.97	49.56
Overjet-1 (zero-shot)	26B	<u>76.63</u>	<u>80.24</u>	<u>76.35</u>
COde-fine-tuned reference models				
Qwen-VL-7B†	7B	**78.92**	**85.06**	**79.39**
Qwen-VL-3B†	3B	74.90	79.97	76.19
Medical and dental vision-language models				
MedVLM-R1 (Pan et al., 2025)	2B	31.79	34.54	25.90
MedGemma 1.5-4B (Sellergren et al., 2026)	4B	27.77	46.64	25.15
LLaVA-Med (Li et al., 2023)	7B	26.83	56.12	21.72
HealthGPT-XL32 (Lin et al., 2025)	32B	17.96	19.55	15.50
General open models				
Muse Glimmer-30B (Meta Superintelligence Lab, 2026)	30B	57.33	50.51	53.01
InternVL3.5-30B-A3B (Wang et al., 2025b)	30B	47.85	63.21	47.02
Qwen3.6-35B-A3B (Qwen Team, 2026)	35B	46.01	58.21	42.50
Frontier proprietary models				
GPT-5.6 Sol (OpenAI, 2026)	–	65.14	70.47	63.74
Gemini 3.1 Pro (Google DeepMind, 2026a)	–	53.93	68.77	51.71
Claude Opus 5 (Anthropic, 2026)	–	51.65	57.73	49.20

G Experimental Setup

G.1 Evaluation sets

For comparison with specialist models, we use two shared held-out image sets for five-class dental-image modality recognition and PA/BW classification within full-mouth series. We also evaluate end-to-end tooth numbering on 170 cases containing 849 boxes.

MMOral uses FDI tooth identifiers, whereas Overjet-1 is trained with Universal numbers. We therefore apply a fixed FDI-to-Universal translation to every Overjet-1 variant before inference.

G.2 Comparison models

For MMOral and Overjet-Bench, every model that we evaluate directly receives the same case inputs, task content, and answer schema. We save each raw prediction before normalization and apply the same scorer implementation to all models. Published results are included only when the benchmark and metric match ours exactly.

The specialist comparisons use task-specific production models for modality recognition, PA/BW classification, and tooth numbering. Only the final DC-RL-trained Overjet-1 checkpoint is included. Test-set sizes and metrics are specified in Table 16.

G.3 Inference and scoring

Locally hosted models use the fixed decode saved with each evaluation: MMOral uses temperature 0.1, repetition penalty 1.05, and a fixed seed, while the task benchmarks use greedy decoding with task-specific output-length limits. The proprietary models are evaluated in separate API calls using the default reasoning setting. For each MMOral split and for Overjet-Bench, we run every propri-

etary model three times and report the mean and sample standard deviation across the three complete API calls.

The MMOral closed-ended split uses inclusive accuracy for each clinical dimension and for the overall score. Counts use exact integer match, and tooth predictions use set F_1. Localization is evaluated with pooled precision, recall, and F_1 after greedy one-to-one IoU matching within each image. We report every IoU threshold used for a task.

Structured answers are parsed into tooth, finding, and polarity fields before scoring. A parsing failure receives zero for the affected fields. Whole-image and per-tooth reports use ROUGE-L, while open clinical synthesis uses token F_1.

We compute 95% confidence intervals using case-level bootstrap resampling, with all inputs from the same case kept together. Paired system comparisons reuse the same bootstrap samples. We apply Holm correction to the prespecified ablation comparisons. The complete statistical protocol and the repeat protocol for proprietary models are provided in Appendix F.1.

G.4 Ablation design

The main ablation series uses the 26B backbone and contains six variants: the base model, SFT without grounding, SFT without reasoning traces, full SFT, DC-RL without spatial consistency, and full DC-RL. The grounding ablation removes coordinates from the training targets, while the reasoning ablation removes supported reasoning traces and retains the corresponding answers. All other data and optimization settings are kept fixed.

Both DC-RL variants start from the same full-SFT checkpoint and use the same records, data order, random seeds, optimizer, rollout configuration, and number of training steps. The matched control removes every contribution from spatial consistency: it removes the topology mixture from identity-aware tooth grounding and disables the capped topology branch when reference geometry is unavailable, while leaving absolute grounding, evidence, and format unchanged.

A second ablation series uses Qwen2.5-VL-7B, the backbone used by OralGPT. It includes the base model, full SFT, and full DC-RL using the same training data and schedule. This experiment tests whether the proposed data and reinforcement learning procedure transfer across model scale rather than being specific to the 26B backbone.

H Additional Results

H.1 MMOral-OPG-Bench

DC-RL raises closed-split accuracy from 65.4% after full SFT to 71.5%, a 6.1-point gain with a paired 95% interval of [3.3, 9.0]. Both variants use the same backbone, 491 questions, decode, and scorer. DC-RL improves all five closed clinical dimensions over SFT and gives the best listed result for teeth, pathology, and bone. It scores 71.0% for teeth, 66.2% for pathology, and 73.2% for treatment. Its bone and summary scores are 76.7% and 67.7%. The corresponding gains over SFT are 5.9, 5.4, 0.7, 9.3, and 3.2 points, so the overall gain is not driven by a single category.

H.2 Effect of DC-RL

The displayed task-family measures show the same overall progression. Grounding rises from 85.3% after SFT to 87.2% without spatial consistency and 90.8% with the full reward. Report generation rises from 50.6% to 55.1% and then 57.7%. The cross-benchmark Overall score follows the same progression, rising from 69.7% after SFT to 72.5% without spatial consistency and 76.1% with the full reward. Full DC-RL corrects geometry without changing an already correct clinical answer on 16.6% of paired records. It corrects both answer and geometry on another 8.4%. Correct-answer regressions occur on 2.0%, and unsupported spatially plausible predictions occur on 0.4%. The low regression rate is consistent with the multiplicative gate. A sampled prediction receives no geometric credit after a wrong clinical decision. Spatial shaping therefore concentrates on predictions that first satisfy the task's correctness rule.

Table 16: **Comparison with specialist models (%).** The specialist systems are task-specific production models, and Overjet-1 is the final DC-RL-trained checkpoint. FMX denotes a full-mouth series; PA and BW denote periapical and bitewing radiographs. Δ is the percentage-point difference, Overjet-1 minus the specialist model. The end-to-end tooth-numbering metric is measured on 170 cases containing 849 boxes.

Task	Test items	Metric	Specialist	Overjet-1	Δ
Five-class modality recognition	44,346 images	Accuracy	**99.58**	98.29	-1.29
FMX PA/BW classification	20,970 images	Accuracy	98.71	**99.82**	$+1.11$
End-to-end tooth numbering	170 cases 849 boxes	F_1 at IoU ≥ 0.1	**97.4**	95.3	-2.1

H.3 Comparison with specialist models

We compare the final Overjet-1 checkpoint with specialist models for dental-image modality recognition, FMX PA/BW classification, and end-to-end tooth numbering. These comparisons assess narrow task performance and are not controlled backbone ablations.

Table 16 compares the final Overjet-1 checkpoint with our in-house specialist models on three tasks. On the shared 44,346-image set, five-class modality accuracy is 98.29%, compared with 99.58% for the specialist model. On the shared 20,970-image FMX set, Overjet-1 reaches 99.82% PA/BW accuracy, 1.11 points above the specialist model. End-to-end tooth-numbering F_1 is 95.3% for Overjet-1 and 97.4% for the specialist model on 170 cases containing 849 boxes.

We additionally evaluate external transfer to COde (Lv et al., 2026). Without fine-tuning on COde, Overjet-1 reaches 76.63% accuracy and 76.35% weighted F_1, matching the performance of the in-domain fine-tuned model (Table 15).

H.4 Ablations

Full SFT improves over the untrained 26B base on every aggregate metric. Closed accuracy increases from 46.4% to 65.4%, and the open score increases from 61.3% to 72.2%. The Overjet-Bench macro score rises from 25.6% to 70.5%. Removing grounding lowers the closed, open, and macro scores by 8.8, 7.1, and 21.5 points. Among the displayed task families, Ground drops by 74.8 points and Report by 10.4 points. Coordinates therefore provide the direct supervision needed to attach findings to image regions. Removing supported reasoning traces also reduces all three aggregate scores. The losses are 4.7 points closed, 10.3 points open, and 10.6 points on the macro score. Ground and Report fall by 2.0 and 2.8 points, respectively.

The reinforcement-learning ablation isolates the contribution of spatial consistency. Relative to full SFT, DC-RL without spatial consistency improves MMOral closed accuracy, the Overjet-Bench macro, Ground, and Report, while the MMOral open score changes by -0.3 point. Sampled optimization with typed correctness and reference-linked overlap is therefore useful on its own. The full reward improves every measure again and is the best 26B variant. Compared with the no-spatial control, full DC-RL gains 3.6 points on Ground and 2.6 points on Report. The gain is largest on multi-tooth answers and records without complete reference geometry. That pattern matches the proposed signal. Spatial consistency evaluates relations among regions, including cases where absolute overlap cannot be computed for every tooth.

The same ordering transfers to Qwen2.5-VL-7B. Full SFT improves its MMOral closed, open, and Overjet-Bench macro scores by 28.7, 22.8, and 39.7 points. Full DC-RL adds another 2.2, 3.6, and 1.6 points. The 7B DC-RL model also improves Ground by 3.6 points and Report by 0.9 points over 7B SFT. Its cross-benchmark Overall score rises from 26.2% at base to 58.9% after SFT and 61.1% after DC-RL.

Table 17: **Generality of the consistency reward.** Two non-dental tasks, each trained from the same Qwen2.5-VL-7B checkpoint with reward-ranked fine-tuning; the reward settings differ only in the shaping terms. *Answer* is task-answer accuracy; *Ground* is absolute grounding (order-blind box F_1 for the synthetic task, box F_1 for GQA); *Consist.* is the consistency score (fraction of adjacent predicted boxes in increasing size order for the synthetic task, fraction of predictions whose stated relation matches the scene-graph edge for GQA); *Joint* requires a correct answer and, for the synthetic task, every predicted box matching the size-sorted gold box at the same position at IoU 0.5; for GQA, box $F_1 \geq 0.5$ and the correct relation. Zero-shot Gemini 3.1 Pro is listed for reference. The best fine-tuned value in each metric column within a block is bolded.

Reward	Answer	Ground	Consist.	Joint
Synthetic ordering (smallest to largest)				
Gemini 3.1 Pro (zero-shot)	100.0	99.3	100.0	99.3
Answer only	99.5	77.9	49.7	1.1
+ grounding	**99.6**	84.2	49.8	1.6
+ grounding + consistency	99.5	**90.9**	**98.7**	**71.6**
GQA (scene-graph relation)				
Gemini 3.1 Pro (zero-shot)	71.4	48.5	69.1	30.9
Answer only	85.5	46.9	75.4	44.7
+ grounding	85.0	**50.9**	73.5	**45.2**
+ grounding + consistency	**85.9**	48.0	77.8	**45.2**
Answer only + consistency	**85.9**	47.5	**78.0**	45.0

I Generality of the Consistency Reward

The spatial-consistency term in Equation (1) is defined for dental tooth layouts. We evaluate the same idea on non-dental images with different consistency rules. Two experiments use Qwen2.5-VL-7B as the base model and compare reward settings that differ only in the shaping terms: final-answer correctness c, correctness with grounding, and correctness with grounding and a consistency term; on GQA we also train correctness with the consistency term alone. Reinforcement learning uses reward-ranked fine-tuning (Dong et al., 2023; Gulcehre et al., 2023). The settings share the base checkpoint, the training records, and the schedule. We also report zero-shot Gemini 3.1 Pro on the same test sets for reference (Google DeepMind, 2026a).

The first experiment uses synthetic images. Each image contains several objects with distinct sizes, and the task asks for the objects and their boxes ordered from smallest to largest. Absolute grounding scores the predicted boxes against the references as a set and ignores their order. The consistency score is the fraction of adjacent predicted boxes that increase in size. Joint accuracy requires the count to be correct and every predicted box to match the size-sorted reference at the same position. The base model localizes the objects but lists them in an arbitrary order, so a prediction can score high on grounding and near zero on the joint metric. The ordering signal comes from the consistency reward.

The second experiment uses GQA (Hudson & Manning, 2019). We build the task from the GQA scene graphs. Each image has a reference object and several instances of a query class. Some instances stand in one semantic relation to the reference (holding, riding, sitting on, carrying, and others). The relation is not named in the question. The model identifies it, counts the instances that hold it, and boxes them. The relation is a scene-graph edge. It cannot be read off the box coordinates, so grounding cannot supply it. The count depends on it: the image also contains query instances that do not stand in the relation. Consistency is the fraction of predictions whose stated relation matches the scene-graph edge. It is read from the text and needs no boxes. The answer is the count; the grounding is the box F_1. The test set has 427 held-out images.

Table 17 reports the results. The consistency reward gives the highest consistency score in both experiments. On the synthetic task it also gives the highest joint accuracy. Figure I1 shows one example per task. On the synthetic task the consistency term raises the order-sensitive joint accuracy from 1.6% to 71.6%, because the size order is not determined by absolute box positions. It also raises

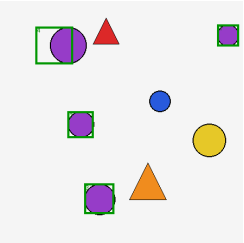




Figure I1: **One example per task**, with boxes from the consistency model. *Left, synthetic: List all purple circles from smallest to largest.* The model returns {`"count":4,"boxes":[...]`} with boxes ordered $1 \rightarrow 4$ by size. *Right, GQA:* the image has one umbrella and several women. The model identifies the relation, counts the women that hold the umbrella, and boxes them: {`"relation":"holding","count":1,...`}. One woman holds the umbrella; the others do not. Both cases score 1.0 on the joint metric.

absolute grounding from 84.2% to 90.9%: fitting the size order sharpens the boxes. On GQA the consistency term raises relation consistency from 75.4% to 78.0%. The grounding term does not: it moves relation consistency to 73.5%. The relation is a scene-graph edge, independent of the box coordinates, so grounding cannot supply it. Consistency read from the text alone gives the same relation gain as consistency with grounding.

Zero-shot Gemini 3.1 Pro reaches 99.3% joint accuracy on the synthetic task and 30.9% on GQA. The synthetic ordering is within reach for a frontier model, and the consistency reward moves the 7B model toward it on the part that absolute grounding leaves unsolved. On GQA the frontier model names the correct scene-graph relation on 69.1% of images, below the fine-tuned 7B model. It also localizes weakly under zero-shot transfer, so its grounding and joint stay low.

The consistency reward can improve absolute grounding, as the synthetic grounding gain shows. The largest gains appear when the consistency rule is part of the correct solution and is not recoverable from the boxes alone, as with size order here or with tooth ordering in the dental task. On GQA the relation is a scene-graph edge that grounding cannot supply, so the consistency reward is the signal that improves graph faithfulness; the gain is smaller because recognizing the relation is bounded by perception. In both cases grounding alone does not secure the consistency, and the consistency reward improves it. This holds beyond dental tasks.

J Dentist-Verified Model Output Case Studies

Figures J1–J6 present case studies from real Overjet-1 model outputs. Dentists verified every displayed answer against the corresponding clinical evidence. Each case includes the inputs, prompt, and model output.

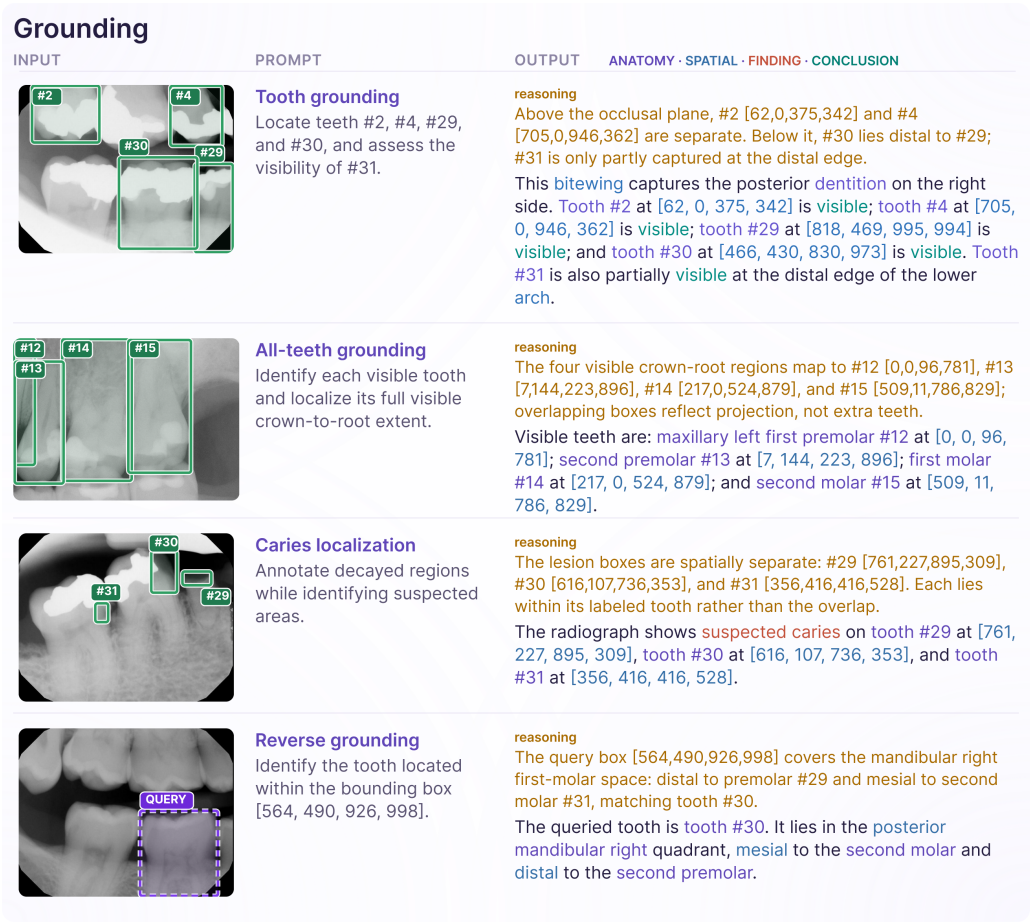


Figure J1: **Grounding and localization.** This set of tasks evaluates the ability of the model to ground the solution in the image. It includes tooth grounding, pathology localization, grounding of multiple objects, and queries about grounded information.

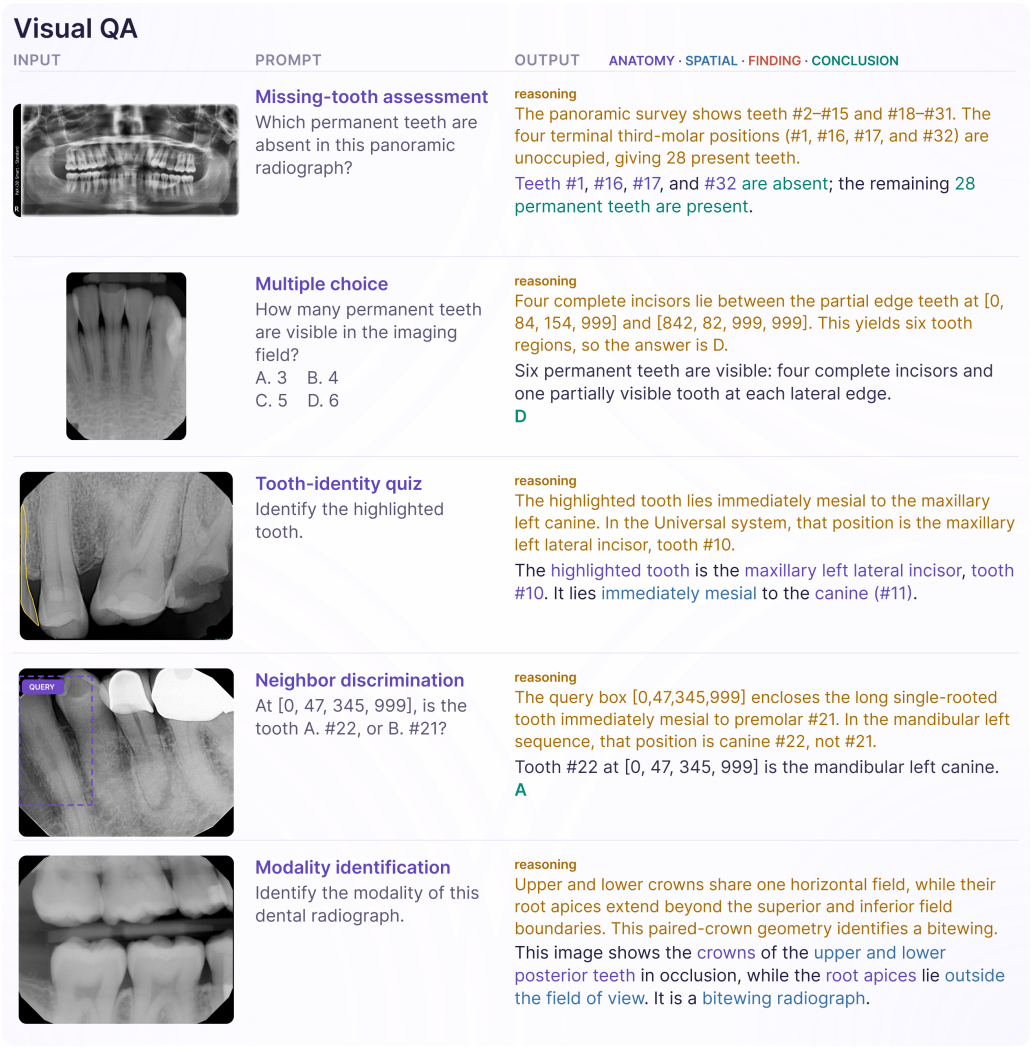


Figure J2: **Visual question answering and classification.** Examples include tooth identification, counting, closed questions, and modality recognition.

Reports

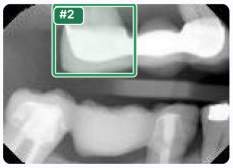
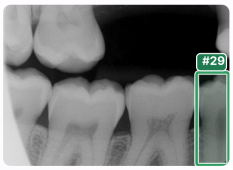
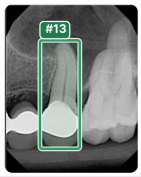
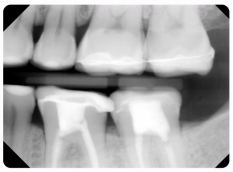
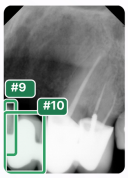
INPUT	PROMPT	OUTPUT ANATOMY · SPATIAL · FINDING · CONCLUSION
	**Whole-image report** Walk me through what this radiograph shows.	reasoning Box [224,0,583,445] anchors restored #2, connected to #3. Below, #29–#30 form another segment; #30 has root-canal material and #31 a filling. This right posterior bitewing shows a connected restorative segment involving maxillary teeth #2 and #3; tooth #2 is localized at [224, 0, 583, 445]. In the opposing arch, teeth #29 and #30 form another connected restorative segment. Tooth #30 has root-canal treatment, and #31 has a filling. No carious lesion or implant is visible in this crop.
	**Per-tooth report** How is the bone around tooth #29?	reasoning Box [855,420,999,999] captures #29 at the right edge. Its distal crest is visible and maintained; the mesial landmark lies outside the crop and is not graded. Tooth #29 at [855, 420, 999, 999], the mandibular right second premolar, presents with healthy bone levels on the distal aspect. The mesial bone level cannot be assessed due to the unavailability of the landmarks.
	**Bone-level severity** Evaluate the bone level on the mesial side of tooth #13.	The interproximal bone level on the mesial side of tooth #13 at [265, 213, 564, 858] is well-maintained, with the crestal margin situated close to the cemento-enamel junction. The mesial alveolar bone level is healthy.
	**Inventory by arch** Count visible teeth by arch and list them.	reasoning Above the occlusal plane, four distinct regions follow #12→#15; below it, three follow #18→#20. No additional crown enters either image edge, giving 4 maxillary and 3 mandibular teeth. Within this bitewing, the maxillary arch shows four teeth (#12–#15), and the mandibular arch shows three teeth (#18–#20).
	**Open synthesis** I am looking at my X-ray, could you explain what is going on here?	reasoning Boxes [0,568,101,888] and [0,636,335,999] anchor #9 and #10 at the left edge. Full-coverage radiopacity continues through #11, forming one anterior crown segment. The image shows full-coverage crowns on the maxillary left central incisor #9 at [0, 568, 101, 888], the lateral incisor #10 at [0, 636, 335, 999], and the canine #11. Together, they form the visible anterior crown-restoration segment; no implant is visible in the displayed field.

Figure J3: **Reports and structured reads.**



Figure J4: **Cross-view reasoning.** Cross-view reasoning tasks involve multiple full-mouth X-ray (FMX) views as well as a panoramic radiograph, and the model is required to aggregate the information across multiple views to provide a consistent answer.

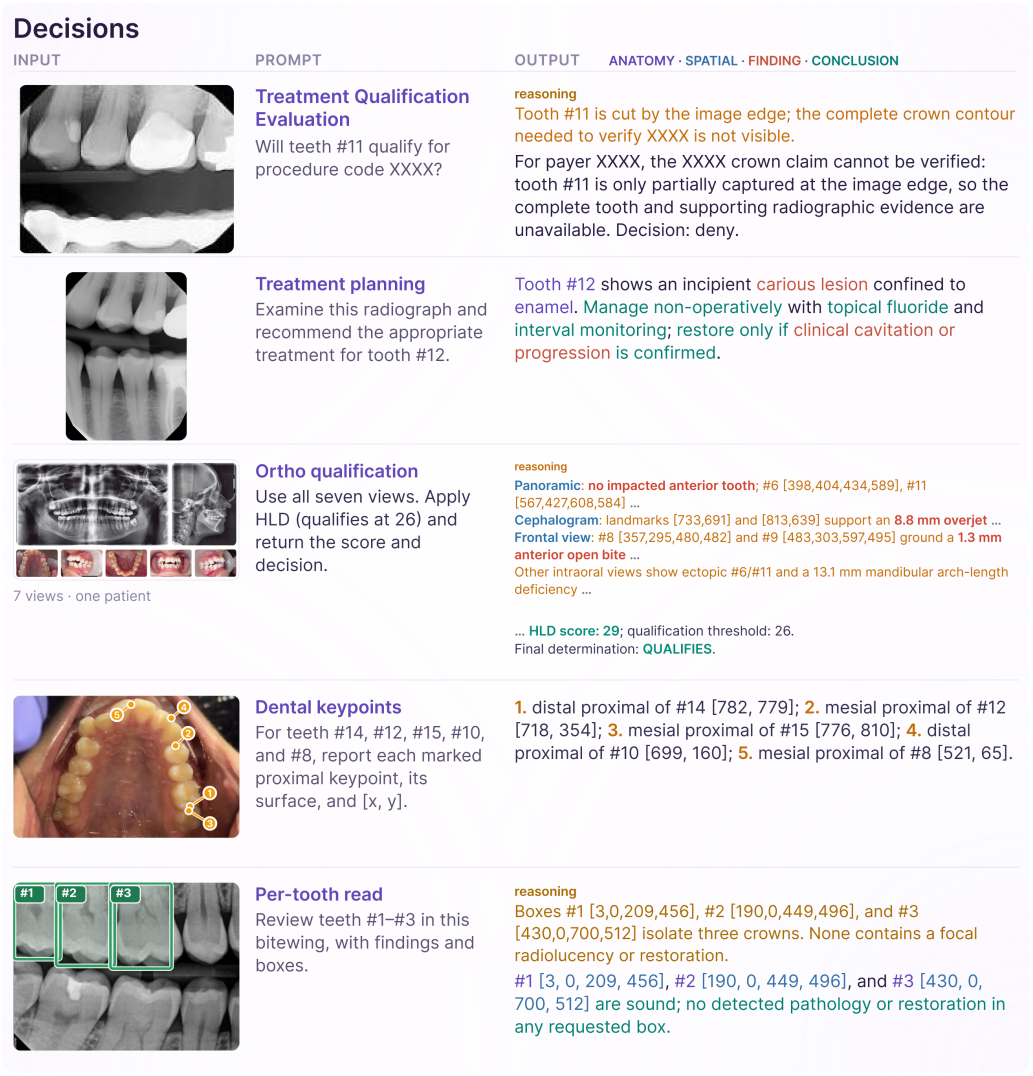


Figure J5: **Clinical decisions and structured records.** Decision and per-tooth examples connect each conclusion to its supporting evidence.

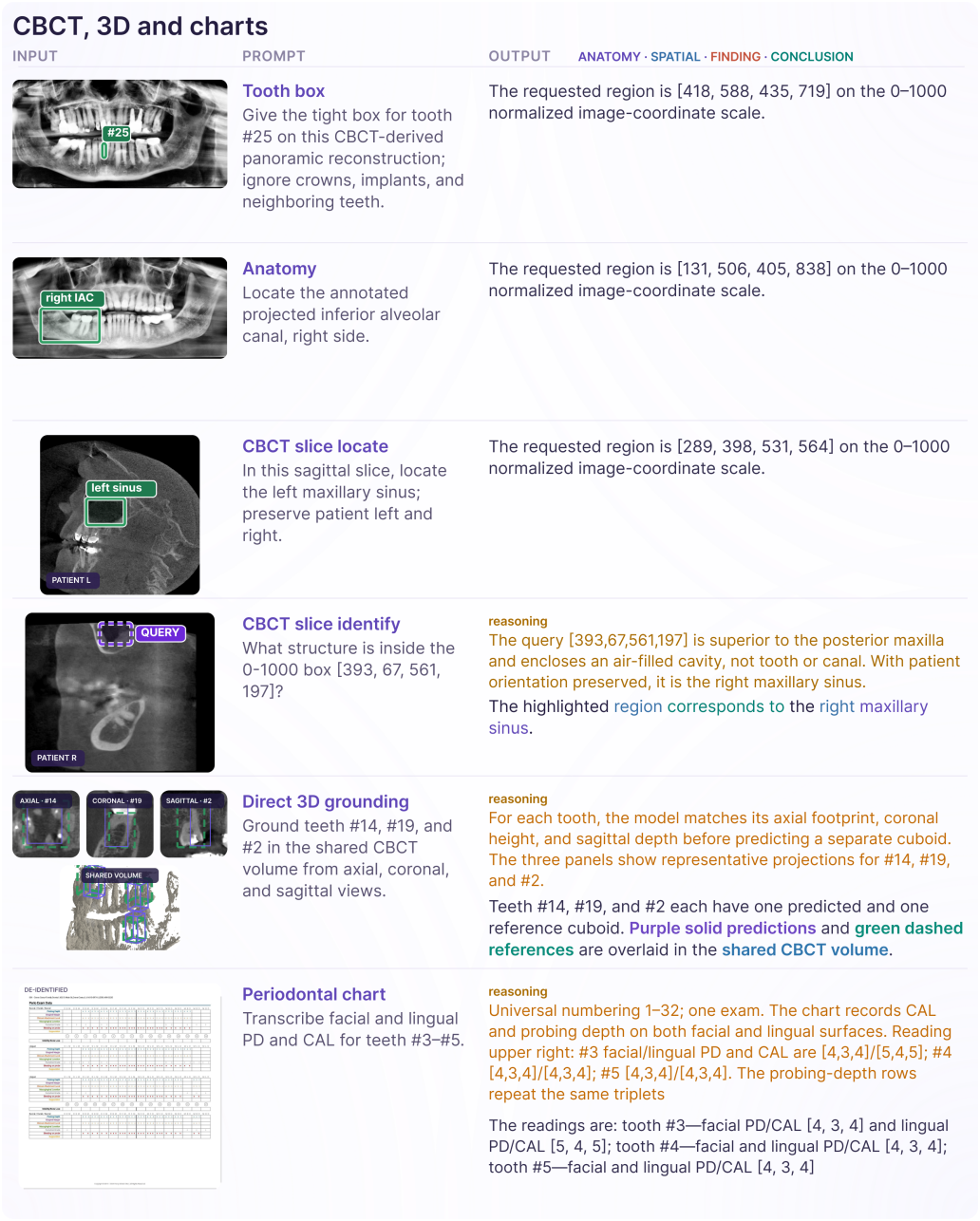


Figure J6: **CBCT and Periochart.** Coordinates on panoramic reconstructions use the same geometry that produced the rendered view. Direct-volume examples overlay predictions (purple, solid) and references (green, dashed) on axial, coronal, and sagittal slices and in the shared volume.