CARIESBENCH: A HIERARCHICAL MULTI-VIEW BENCHMARK FOR DENTAL CARIES SCREENING AND DIAGNOSTIC REASONING FROM 3D INTRAORAL SCANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dental caries assessment is a structured clinical process requiring lesion localization, identification of the affected tooth and surface, and severity grading rather than simple image-level recognition. 3D intraoral scans (IOS) preserve detailed dental geometry and surface morphology and support standardized rendering from complementary viewpoints, enabling fine-grained visual assessment while introducing a distinct need for anatomically grounded multi-view reasoning. Despite recent advances in multimodal large language models (MLLMs) and dental vision-language benchmarks, existing evaluations do not jointly assess hierarchical arch-to-tooth-to-surface reasoning, cross-view evidence integration, and surface-level caries assessment. To address this gap, we introduce CariesBench, a hierarchical benchmark constructed from 5,932 3D IOS, comprising 85,455 standardized multi-view renderings and 133,994 clinically grounded visual question answering instances with expert quality control. CariesBench comprises three complementary task families: Arch-level Clinical Screening (ACS), Tooth-level Diagnostic Reasoning (TDR), and Structured Clinical Decision-Making (SCD). Using the CariesBench training set, we further fine-tune Qwen3-VL-8B-Instruct to obtain IntraoralGPT. Extensive evaluation of 24 representative MLLMs reveals a pronounced global-to-local reasoning gap, with substantially stronger performance on arch-level screening than on fine-grained tooth- and surface-level reasoning. CariesBench-specific supervision substantially improves the Qwen3-VL-8B-Instruct baseline, increasing the 19-task macro-average clinical accuracy from 52.7% to 69.5% and TDR Overall from 34.4% to 44.7%. Despite these gains, fine-grained TDR remains challenging, highlighting persistent difficulty in anatomical localization and cross-view evidence integration. CariesBench thus provides a clinically grounded testbed for evaluating hierarchical multi-view reasoning in digital dentistry.

1 INTRODUCTION

Dental caries is a common oral disease whose clinical management depends on accurate lesion localization and severity assessment. Caries diagnosis is therefore not merely a binary recognition task. Clinicians must identify the affected tooth and surface and grade lesion severity according to standardized diagnostic criteria. The International Caries Detection and Assessment System (ICDAS) (Ismail et al., 2007; Dikmen, 2015), for example, defines caries status at the tooth-surface level and stratifies lesion severity according to visible structural changes. Reliable caries assessment therefore requires visual evidence to be anatomically localized and interpreted at the appropriate clinical severity, making the task inherently tooth-specific, surface-aware, and fine-grained.

3D intraoral scans (IOS) provide a particularly informative representation for such assessment. By preserving detailed dental geometry and surface morphology, IOS enable individual teeth and surfaces to be examined from complementary viewpoints, with previous clinical studies demonstrating their potential for detecting and classifying carious lesions at tooth and surface levels (Ntovas et al., 2023; Eggmann & Blatz, 2024). However, this richer representation also introduces a distinct reasoning challenge: clinically relevant evidence may be distributed across complementary views of the same anatomy. Reliable interpretation therefore requires localizing findings to the correct tooth and surface, maintaining anatomical correspondence, and integrating evidence across views before

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.

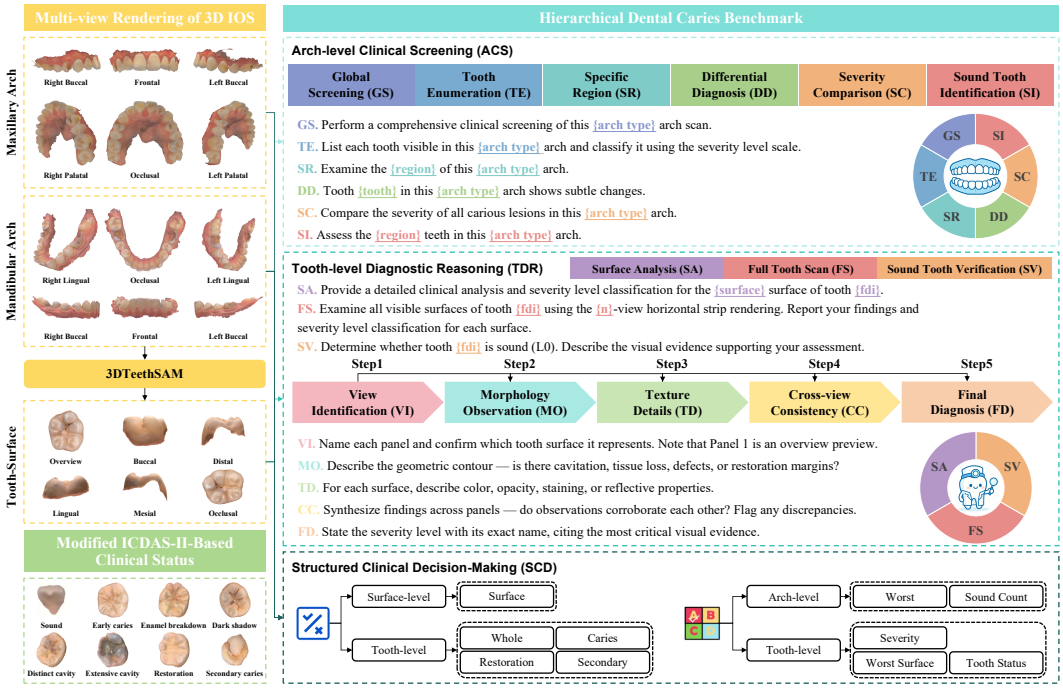

Figure 1: Overview of CariesBench. Starting from complete 3D intraoral scans, standardized arch-level rendering and tooth-level segmentation produce anatomically aligned multi-view representations. CariesBench organizes evaluation hierarchically into Arch-level Clinical Screening (ACS), Tooth-level Diagnostic Reasoning (TDR), and Structured Clinical Decision-Making (SCD), with TDR incorporating a five-step progressive reasoning chain. The benchmark comprises 5,932 3D IOS, 85,455 multi-view images, and 133,994 quality-controlled VQA instances.

severity assessment. IOS-based caries assessment is therefore not merely a single-image recognition problem, but an *anatomically grounded multi-view reasoning* problem.

Recent advances in multimodal large language models (MLLMs) have demonstrated strong capabilities in medical visual understanding and reasoning across diverse imaging modalities and clinical tasks (He et al., 2020; Liu et al., 2021; Hu et al., 2024; Li et al., 2023), motivating growing research interest in dentistry. Dental datasets and benchmarks have accordingly evolved from 2D pathology recognition and localization toward visual question answering, anatomical reasoning, and multimodal interpretation. Traditional resources such as DENTEX (Hamamci et al., 2023) primarily focus on recognition or localization from 2D dental images, whereas recent vision-language benchmarks extend to dental VQA (Li et al., 2026), arch- and tooth-level understanding (Hao et al., 2026a), multimodal dental reasoning (Meng et al., 2026), and native 3D IOS modeling (Xiong et al., 2026). Nevertheless, these benchmarks individually cover only subsets of the combined capabilities targeted by IOS-based caries assessment. As summarized in Table 1, none of the compared resources jointly evaluates hierarchical arch-to-tooth-to-surface reasoning, anatomically corresponding multi-view evidence integration, and surface-level caries assessment within a unified benchmark. Consequently, it remains unclear whether current MLLMs can maintain anatomical correspondence and integrate complementary evidence as diagnosis progresses from global screening to localized tooth- and surface-level assessment.

To address this gap, we introduce CariesBench, a hierarchical benchmark for evaluating clinically grounded multi-view reasoning over observations derived from 3D IOS. Constructed from 5,932 3D IOS, CariesBench comprises 85,455 standardized multi-view 2D renderings and 133,994 clinically grounded VQA instances. A stratified 20% subset of benchmark instances underwent expert validation by two dentists with 8 and 3 years of clinical experience, respectively, under a standardized protocol designed and supervised by a senior dentist with 15 years of clinical experience; disagreements were resolved through adjudication under senior clinical supervision. Unlike conventional image-level VQA, CariesBench derives complementary observations from the same underlying 3D anatomy, thereby preserving correspondence across views. Evaluation spans three complementary task families:

Arch-level Clinical Screening (ACS) for global dentition-level understanding, Tooth-level Diagnostic Reasoning (TDR) for localized reasoning over individual teeth and surfaces, and Structured Clinical Decision-Making (SCD) for objectively scored clinical decisions across anatomical granularities. This hierarchy enables systematic evaluation of MLLMs as the diagnostic scope narrows from the entire arch to individual teeth and surfaces.

Extensive evaluation of 24 representative baseline MLLMs reveals that the transition from global arch-level understanding to localized tooth-level reasoning remains highly challenging. Across model families, performance is substantially higher on arch-level screening than on localized tooth-level diagnosis, exposing a pronounced global-to-local reasoning gap. The most challenging results occur on TDR tasks requiring fine-grained anatomical localization and cross-view surface interpretation. For example, GPT-5.5 achieves 76.9% on Sound Tooth Identification and 65.9% on Tooth Enumeration in ACS, but only 31.8% on Surface Analysis and 30.5% on Full Tooth Scan in TDR. These results indicate that strong global dental understanding does not readily translate into reliable localization and integration of evidence across anatomically corresponding views.

To investigate whether these limitations can be mitigated through benchmark-specific supervision, we fine-tune Qwen3-VL-8B-Instruct on the CariesBench training set to obtain IntraoralGPT. Under the same patient-level held-out protocol, IntraoralGPT improves the 19-task macro-average clinical accuracy of Qwen3-VL-8B-Instruct from 52.7% to 69.5% (+16.8 points). On TDR, its overall accuracy increases from 34.4% to 44.7%, with Surface Analysis improving from 24.3% to 30.6% and Full Tooth Scan from 23.0% to 29.1%. These gains demonstrate the value of CariesBench-specific supervision for hierarchical intraoral reasoning; however, the resulting 44.7% TDR Overall remains slightly below the strongest zero-shot baseline of 45.9%, underscoring the continued difficulty of fine-grained anatomical localization and cross-view evidence integration.

To summarize, our contributions are threefold:

1) We introduce CariesBench, a large-scale benchmark built from 5,932 3D intraoral scans, 85,455 standardized multi-view renderings, and 133,994 clinically grounded VQA instances, enabling hierarchical evaluation from arch-level screening to tooth- and surface-level reasoning over anatomically corresponding views.
2) We comprehensively benchmark 24 representative MLLMs and identify a pronounced global-to-local reasoning gap, in which strong global dental understanding does not reliably translate into fine-grained anatomical localization and cross-view reasoning.
3) We introduce IntraoralGPT, obtained by fine-tuning Qwen3-VL-8B-Instruct with CariesBench supervision, and demonstrate substantial improvements in tooth-level multi-view reasoning and structured clinical decision-making.

2 Related Work

2.1 General Medical Multimodal Benchmarks

General medical multimodal benchmarks have established visual question answering as a widely used paradigm for evaluating medical image understanding. PathVQA (He et al., 2020) focuses on pathology images with open-ended questions grounded in pathological findings, whereas OmniMedVQA (Hu et al., 2024) broadens evaluation across diverse imaging modalities and anatomical regions. These benchmarks provide important testbeds for medical visual understanding, but their evaluation is primarily organized around individual image–question pairs rather than anatomically linked multi-view observations. They therefore do not explicitly capture the hierarchical anatomical structure or cross-view evidence dependencies required for dental diagnosis, in which assessment may progress from the entire dentition to individual teeth and specific tooth surfaces.

2.2 Dental Multimodal Benchmarks

Dental benchmarks have progressively evolved from 2D pathology recognition and localization to 3D anatomical modeling and vision-language reasoning. Traditional datasets such as DENTEX (Hamamci et al., 2023), CariesDet (Tan et al., 2025), AlphaDent (Sosnin et al., 2025), and DentalAI (Valluri, 2023) primarily support recognition, localization, or segmentation of dental ab-

Table 1: Comparison of representative medical and dental multimodal datasets and VQA benchmarks. CariesBench uniquely combines hierarchical anatomical reasoning (Arch → Tooth → Surface), multi-view IOS observations, and explicit diagnostic reasoning among the compared resources.

Benchmark / Dataset	Domain	Input Modality	Scale	Arch	Tooth	Surf	MV	R/CoT	Closed	QC
General Medical VQA Benchmarks										
PathVQA (He et al., 2020) *arXiv'20*	Pathology	2D Pathology Images	4.9K imgs / 32.7K VQA	✗	✗	✗	✗	✗	✓	✗
OmniMedVQA (Hu et al., 2024) *CVPR'24*	General	2D Medical Images	118.0K imgs / 127.9K VQA	✗	✗	✗	✗	✗	✓	✗
Traditional 2D Dental Pathology & Caries Datasets										
DENTEX (Hamamci et al., 2023) *MICCAI'23*	Dental	2D Panoramic X-ray	2.3K labeled / 1.5K unlabeled imgs	✗	✓	✗	✗	✗	✗	✓
CariesDet (Tan et al., 2025) *Caries Res.'25*	Dental (Caries)	2D Bitewing X-ray	1.0K imgs / 4 classes	✗	✗	✗	✗	✗	✗	✓
AlphaDent (Sosnin et al., 2025) *arXiv'25*	Dental (Caries)	2D Intraoral Photo	1.3K imgs / 9 classes	✗	✗	✗	✗	✗	✗	✓
DentalAI (Valluri, 2023) *Kaggle'23*	Dental (Caries)	2D Intraoral Photo	2.4K imgs / 28.9K instances	✗	✗	✗	✗	✗	✗	✗
3D Dental Anatomy & Segmentation Benchmarks										
3DTeethSeg (Ben-Hamadou et al., 2023) *MICCAI'22*	Dental	Native 3D IOS	1.8K scans / 23.9K teeth	✗	✓	✗	✗	✗	✗	✓
ToothFairy2 (Bolelli et al., 2025) *CVPR'25*	Dental	3D CBCT	530 vols / 42 classes	✗	✓	✗	✗	✗	✗	✓
Specialized Dental VLM Resources & VQA Benchmarks										
MetaDent (Li et al., 2026) *JDR'26*	Dental	2D Dental Images	60.6K imgs / 18.4K VQA	✗	✗	✗	✗	✗	✓	✓
MMOral-OPG (Hao et al., 2026a) *NeurIPS'25*	Dental	2D Panoramic X-ray	20.5K imgs / 1.3M instr. / 1.0K eval. VQA	✓	✓	✗	✗	✗	✓	✓
MMOral-Uni (Hao et al., 2026b) *CVPR'26*	Dental	5 Dental Modalities	2.8K open VQA	✗	✓	✗	✗	✗	✗	✓
DentVLM (Meng et al., 2026) *Nat. Commun.'26*	Dental	7 Dental Image Modalities	110.4K imgs / 2.4M VQA	✓	✓	✗	✗	✓	✓	✓
IOSVLM (Xiong et al., 2026) *arXiv'26*	Dental	Native 3D IOS	19.0K scans / 249.0K VQA	✓	✗	✗	✗	✗	✓	✓
CariesBench (Ours)	**Dental (Caries)**	**3D IOS → MV 2D**	**5.9K scans → 85.4K imgs / 133.9K VQA**	✓	✓	✓	✓	✓	✓	✓

Criteria: Arch/Tooth/Surf denote explicit tasks or evaluation at the corresponding anatomical granularity; **MV** denotes complementary views jointly used within one task instance; **R/CoT** denotes explicit diagnostic rationales or reasoning traces; **Closed** denotes closed-ended VQA or structured fixed-choice questions; **QC** denotes documented expert annotation or verification.

normalities from panoramic radiographs, bitewing X-rays, or intraoral photographs. In parallel, 3DTeethSeg (Ben-Hamadou et al., 2023) and ToothFairy2 (Bolelli et al., 2025) introduce native 3D IOS or CBCT data, but primarily target anatomical segmentation rather than clinical vision-language reasoning.

More recent dental vision-language benchmarks extend evaluation toward clinical question answering and increasingly diverse multimodal reasoning settings. MetaDent (Li et al., 2026) introduces dental VQA over 2D dental images; MMOral-OPG (Hao et al., 2026a) focuses on panoramic X-ray understanding; MMOral-Uni (Hao et al., 2026b) further extends evaluation across five dental imaging modalities and five task types; DentVLM (Meng et al., 2026) broadens dental reasoning across seven imaging modalities and 36 clinical tasks; and IOSVLM (Xiong et al., 2026), the closest setting to ours, investigates generative VQA directly over native 3D IOS. However, as summarized in Table 1, these benchmarks cover complementary aspects of dental multimodal understanding, but none of the compared resources jointly evaluates arch-, tooth-, and surface-level reasoning, multi-view evidence integration, and surface-level caries assessment.

CariesBench targets this missing intersection by combining hierarchical anatomical evaluation with multi-view observations derived from the same underlying 3D anatomy. This design enables systematic assessment of anatomical correspondence, cross-view evidence integration, and clinically consistent decisions across arch, tooth, and surface granularities.

3 Construction of CariesBench

3.1 Benchmark Overview

Figure 1 summarizes the construction of CariesBench, which converts 3D IOS into standardized arch- and tooth-level multi-view 2D representations for existing MLLMs.

Based on these representations, CariesBench organizes evaluation into three complementary task families: Arch-level Clinical Screening (ACS) for global understanding of the dentition, Tooth-level Diagnostic Reasoning (TDR) for fine-grained reasoning over individual teeth and their surfaces, and Structured Clinical Decision-Making (SCD) for objectively scored clinical decisions across anatomical granularities. Together, these task families form a hierarchical evaluation from global arch-level screening to localized tooth- and surface-level reasoning and structured clinical decision-making.

The 5,932 IOS were obtained from 2,966 patients, each contributing one maxillary and one mandibular scan. Using a fixed patient-level split, the 2,966 patients were partitioned into 2,372/297/297 patients for training, validation, and testing, respectively. All IOS and their derived arch-level renderings, tooth-level views, surface observations, and VQA instances from the same patient were confined to

the same split before downstream VQA sampling or training-data rebalancing. This patient-disjoint partition prevents patient-level information leakage across training and evaluation sets.

3.2 3D-TO-2D MULTI-VIEW REPRESENTATION

Arch-level Multi-view Rendering. Each complete maxillary or mandibular IOS is rendered into six standardized views covering complementary anatomical directions. Maxillary scans include right buccal, frontal, left buccal, right palatal, occlusal, and left palatal views. Mandibular scans include right lingual, occlusal, left lingual, right buccal, frontal, and left buccal views. The six views are arranged in a fixed arch-specific 2×3 layout, providing consistent anatomical orientation and complementary observations of the same dental arch.

Tooth-level Multi-view Rendering. Each complete IOS is first segmented using the pretrained 3DTeethSAM (Lu et al., 2026) model, and each segmented tooth is rendered into a standardized multi-view panel. Panel 1 provides an overview of the target tooth, followed by anatomically defined surface-specific views. Incisors contain labial, mesial, inner, distal, and incisal views; posterior teeth contain buccal, mesial, inner, distal, and occlusal views; and canines use four corresponding surface-specific views without an additional incisal or occlusal view. Inner surfaces are designated as palatal for maxillary teeth and lingual for mandibular teeth, while anterior and posterior outer surfaces are termed labial and buccal, respectively. Mesial and distal orientations are mirrored between the right and left quadrants to preserve tooth-relative anatomical orientation.

Detailed rendering layouts and camera configurations are provided in Appendix C.

Anatomical Correspondence. Unlike independently acquired 2D dental images, all views associated with a given arch or tooth are rendered from the same underlying 3D anatomical instance. CariesBench therefore preserves explicit anatomical correspondence across views, enabling evaluation of whether an MLLM can associate findings with the correct tooth and surface, integrate complementary evidence, and maintain a consistent diagnosis across viewpoints.

3.3 MODIFIED ICDAS-II-BASED CLINICAL STATUS TAXONOMY

Based on the ICDAS-II caries coding scheme, we define a task-oriented clinical status taxonomy for multi-view IOS assessment. ICDAS-II codes 1 and 2 are merged into a unified early-caries category to reduce ambiguity between subtle early-stage manifestations in rendered IOS views, while the remaining caries categories follow the ordinal progression of ICDAS-II codes 3–6. Because restoration-related conditions are also clinically relevant to dental screening, we additionally introduce two categories representing restored teeth without and with secondary caries. The resulting taxonomy contains eight clinical states, $\mathcal{L} = \{\text{L0,L1,L2,L3,L4,L5,L6,L7}\}$. Specifically, L0–L5 denote Sound, Early Caries, Localized Enamel Breakdown, Underlying Dark Shadow, Distinct Cavity, and Extensive Cavity, corresponding to ICDAS-II codes 0, 1–2, 3, 4, 5, and 6, respectively; L6–L7 represent restoration without and with secondary caries and are categorical extensions rather than ordinal severity levels. This eight-state clinical status taxonomy serves as the standardized clinical label space for benchmark construction and evaluation. Detailed ICDAS-II correspondence and operational visual criteria are provided in Appendix B.

3.4 HIERARCHICAL CLINICAL TASK CONSTRUCTION

Arch-level Clinical Screening (ACS). ACS evaluates global understanding of a complete dental arch through six task types: Global Screening (GS), Tooth Enumeration (TE), Specific Region (SR), Differential Diagnosis (DD), Severity Comparison (SC), and Sound Tooth Identification (SI).

Tooth-level Diagnostic Reasoning (TDR). TDR evaluates fine-grained reasoning over an individual tooth through Surface Analysis (SA), Full Tooth Scan (FS), and Sound Tooth Verification (SV). These tasks require models to associate observations with the correct anatomical surfaces, integrate complementary views of the same tooth, and derive a clinically consistent diagnosis. For reasoning-oriented instances, the diagnostic process follows a five-step progression: View Identification (VI), Morphology Observation (MO), Texture Details (TD), Cross-view Consistency (CC), and Final Diagnosis (FD).

Structured Clinical Decision-Making (SCD). SCD complements open-ended reasoning with objectively scored true/false (TF) and multiple-choice (MC) questions spanning surface, tooth, and arch granularities. These tasks evaluate caries and restoration verification, severity classification, and identification of clinically relevant surfaces or teeth.

Representative examples are illustrated in Fig. 1, while complete task definitions, representative question templates, and the five-step TDR reasoning protocol are provided in Appendix E.

3.5 VQA Generation and Quality Control

We construct CariesBench using a structured-annotation-grounded, semi-automated pipeline that explicitly separates structured clinical annotations from subsequent language realization. The structured clinical annotations were derived from epidemiological surveys, oral examinations, and clinical data curation at a university-affiliated stomatological hospital. Annotation and data entry were conducted by five postgraduate dental trainees, three standardized-residency trainees, and one attending dentist under senior clinical supervision. The resulting annotations were organized by patient, tooth, anatomical surface, and clinical status and served as the reference clinical labels for benchmark construction and evaluation. For the open-ended ACS and TDR tasks, GPT-4o was used only for language realization of annotation-grounded questions and responses. Gemini-2.5-Pro subsequently audited each generated instance for consistency with the structured annotations, anatomical correspondence, and unsupported clinical claims. For SCD, questions and target answers were instantiated directly from the structured annotations rather than generated by an LLM. Across all task families, evaluation targets are derived from the structured clinical annotations rather than from LLM-generated reference prose.

Expert validation was further performed on a stratified subset comprising 20% of the full benchmark, covering all clinical-status categories and task families. Two dentists with 8 and 3 years of clinical experience, respectively, reviewed the selected instances using a standardized five-point clinical evaluation rubric under a protocol designed and supervised by a senior dentist with 15 years of clinical experience. Overall response quality was assessed along five dimensions: Tooth/Surface Mapping, Evidence Grounding, Target Completeness, Cross-view Consistency, and Conclusion Support. For TDR, the five-step diagnostic explanations were additionally assessed in terms of VI, MO, TD, CC, and FD. Across the reported evaluation criteria, mean expert ratings ranged from 4.26 to 4.95 on the five-point scale. Of the expert-reviewed subset, 800 stratified instances spanning all task families and clinical-status categories were independently assessed by both dentists for inter-rater reliability analysis. Inter-rater agreement was quantified using linearly weighted Cohen's κ, yielding $\kappa = 0.86$. Disagreements were resolved through adjudication and consensus under senior clinical supervision; 5.2% of reviewed instances required adjudication, while 1.8% were excluded because of unsupported visual claims or unresolved clinical ambiguity.

3.6 Benchmark Statistics

Figure 2 summarizes the benchmark statistics and expert quality assessment. The benchmark contains 133,994 VQA instances, including 5,861 ACS, 31,545 TDR, and 96,588 SCD instances, with detailed distributions across their respective question types. It covers all eight modified ICDAS-II clinical states and spans complementary screening, diagnostic reasoning, and structured decision formats. Expert evaluation by two dentists yielded consistently high scores for both generated responses and five-step TDR explanations.

4 Experiments

4.1 Experimental Setups

Benchmarked MLLMs. We evaluate 24 representative multimodal large language models (MLLMs) grouped into three categories: commercial, open-source, and medical-specialized models. Commercial models include GPT-5.5 (OpenAI, 2026b), GPT-5.4-Nano (OpenAI, 2026a), GLM-5.2 (Zeng et al., 2026), Claude-Sonnet-5 (Anthropic, 2026), Claude-Haiku-4.5 (Anthropic, 2025), Gemini-3.6-Flash (Google, 2026b), and Gemini-3.1-Pro-Preview (Google, 2026a). Open-source models include LLaVA-1.5-7B (Liu et al., 2024), ShareGPT4V-7B (Chen et al., 2024a), InternVL2-8B (Chen

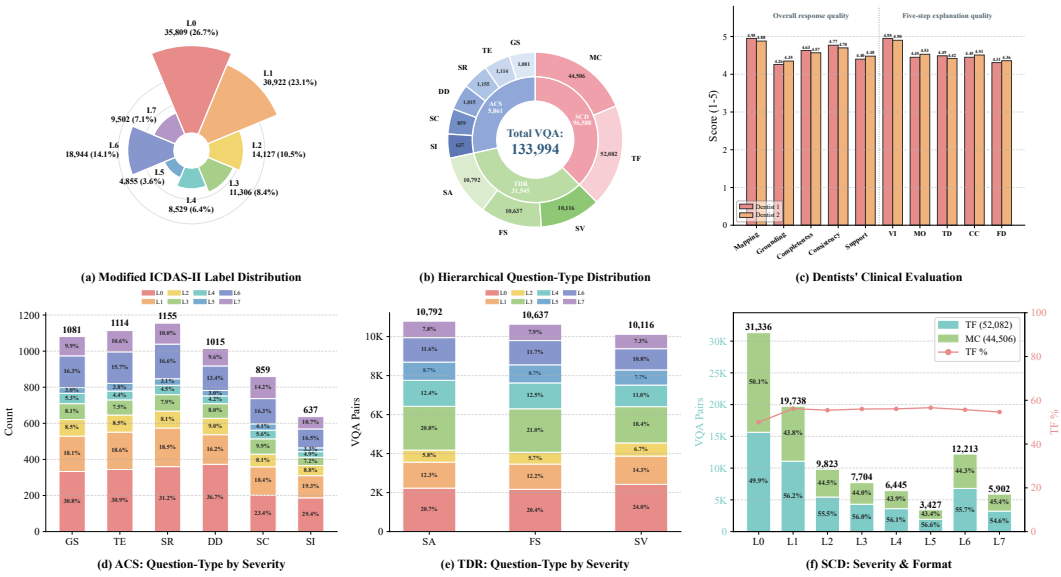

Figure 2: Statistical composition and expert quality assessment of CariesBench.

et al., 2024b), InternVL3.5-8B (Wang et al., 2025), Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a), Qwen3.5-9B (Qwen Team, 2026), DeepSeek-VL-7B (Lu et al., 2024), Kimi-VL-A3B-Thinking (Team et al., 2025), and MiMo-VL-7B-RL (Yue et al., 2025). Medical-specialized models include HuatuoGPT-Vision-7B (Zhang et al., 2023), LLaVA-Med-7B (Li et al., 2023), DentalGPT-Mid (Cai et al., 2025), HealthGPT-Pro-8B (Lin et al., 2025), Lingshu-7B (Xu et al., 2026), OralGPT-Omni-7B (Hao et al., 2026b), and DentVLM (Meng et al., 2026).

Supervised Fine-Tuning. We further fine-tune Qwen3-VL-8B-Instruct on the CariesBench training set using LoRA implemented with LLaMA-Factory (Zheng et al., 2024) for two epochs, with a learning rate of 2×10^{-5}, rank $r = 16$, scaling factor $\alpha = 32$, and BF16 precision. We refer to the resulting model as IntraoralGPT.

Evaluation Metrics. Consistent with the hierarchical structure of CariesBench, our primary metric is Clinical Accuracy, which evaluates structured diagnostic correctness after task-specific response parsing, reducing sensitivity to linguistic variation. For ACS, model responses are parsed according to the expected output structure of each question type, with clinical predictions compared against the corresponding structured reference annotations. For TDR, Clinical Accuracy is computed at the task-appropriate anatomical granularity: exact clinical-status matching for the queried surface in SA, mean surface-level exact-match accuracy across the target tooth in FS, and binary sound/non-sound correctness in SV. For SCD, true/false responses are evaluated by binary correctness, whereas multiple-choice responses are evaluated by exact class match. Each task-family Overall is the macro-average of its question types, and Avg. is the macro-average across all 19 question types. Detailed metric definitions and additional IntraoralGPT training settings are provided in Appendices K and L.2.

Implementation Details. For all benchmarked baseline models, we use zero-shot prompting with deterministic decoding (temperature $= 0.0$, top-$p = 1.0$, and repetition penalty $= 1.0$). Commercial models are evaluated through their official APIs using equivalent decoding settings where supported. All models receive the same task-specific instructions and corresponding rendered inputs. Open-source and medical-specialized models are evaluated locally using PyTorch 2.4.1, Transformers 5.4.0, and CUDA 11.8 with FP16 inference on a server equipped with eight NVIDIA A40 46 GB GPUs.

4.2 Evaluation Results

Based on Tables 2 and 3, we summarize four key findings on CariesBench.

Finding 1. Surface-specific tooth-level reasoning is the most pronounced bottleneck observed on CariesBench. The largest performance limitations occur on Surface Analysis (SA) and Full Tooth Scan (FS), which require surface-specific localization and multi-view integration. For example, GPT-5.5 achieves 31.8% on SA and 30.5% on FS, despite reaching 75.4% on Sound Tooth Verification

Table 2: Evaluation of representative MLLMs on CariesBench across ACS, TDR, SCD-TF, and SCD-MC. Clinical accuracy (%) is reported for all 19 sub-tasks; each Overall is the macro-average over its constituent sub-tasks, and Avg. is the macro-average over all 19 sub-tasks. Best and second-best results across all evaluated models are shown in **bold** and underlined, respectively. IntraoralGPT is reported separately as a CariesBench-supervised model.

		ACS							TDR				SCD-TF						SCD-MC						Avg.
	Model	GS	TE	SR	DD	SC	SI	Overall	SA	FS	SV	Overall	Surf	Whl	NoC	Rst	Sec	Overall	Wor	SC	Sev	WS	TS	Overall	
Commercial	GPT-5.5 (OpenAI, 2026b)	67.9	65.9	67.0	69.6	55.9	76.9	67.2	**31.8**	**30.5**	**75.4**	**45.9**	63.8	61.2	70.4	79.1	86.5	72.2	58.7	57.4	63.8	61.5	65.2	61.3	63.6
	GPT-5.4-Nano (OpenAI, 2026a)	70.5	66.6	67.2	69.3	56.5	75.6	67.6	21.8	22.4	67.2	37.1	50.8	48.9	60.2	72.4	78.6	62.2	45.2	43.7	49.6	46.8	52.1	47.5	56.1
	GLM-5.2 (Zeng et al., 2026)	69.6	69.5	67.7	63.8	53.2	76.3	66.7	24.7	23.9	68.6	39.1	51.5	49.8	62.1	73.8	79.0	63.2	46.3	45.1	50.8	48.0	53.5	48.7	56.7
	Claude-Sonnet-5 (Anthropic, 2026)	72.3	67.6	67.8	70.1	55.8	76.8	68.4	30.1	29.5	73.8	44.5	61.9	59.7	68.9	77.8	84.2	70.5	56.1	54.8	59.5	55.4	58.9	56.9	62.2
	Claude-Haiku-4.5 (Anthropic, 2025)	71.7	61.1	68.4	70.5	60.2	70.2	67.0	26.9	26.1	69.7	40.9	55.4	53.7	64.8	74.5	80.6	65.8	49.7	48.5	54.9	52.6	56.7	52.5	58.7
	Gemini-3.6-Flash (Google, 2026b)	70.5	64.4	62.2	69.1	57.0	76.6	66.6	29.4	28.8	73.1	43.8	60.2	58.6	67.1	77.0	83.5	69.3	54.8	53.2	59.9	57.7	61.4	57.4	61.3
	Gemini-3.1-Pro-Preview (Google, 2026a)	70.2	66.6	64.1	67.4	60.2	77.2	67.6	26.6	25.7	70.2	40.8	58.2	56.7	65.3	81.4	82.0	68.7	52.9	51.5	54.8	52.2	66.7	55.6	60.5
Open-Source	LLaVA-1.5-7B (Liu et al., 2024)	67.8	66.1	68.7	71.7	52.7	75.8	67.1	5.2	25.5	71.2	34.0	41.1	47.0	40.6	76.7	29.8	47.0	32.4	31.2	29.5	60.3	21.6	35.0	48.2
	ShareGPT4V-7B (Chen et al., 2024a)	69.2	67.0	**70.8**	70.2	57.6	75.6	68.4	19.5	24.9	72.5	39.0	32.6	55.5	64.7	15.0	5.7	34.7	38.6	36.9	31.1	67.7	13.8	37.6	46.8
	InternVL2-8B (Chen et al., 2024b)	67.4	69.2	64.2	68.4	59.3	76.1	67.4	19.4	22.0	71.9	37.8	38.3	41.7	64.7	73.1	5.8	44.7	44.1	42.6	24.1	45.6	66.1	44.5	50.7
	InternVL3.5-8B (Wang et al., 2025)	67.2	62.7	68.7	69.3	59.7	76.0	67.3	24.9	25.0	72.0	40.6	51.7	45.0	64.7	15.2	5.7	36.5	49.5	47.8	29.3	38.7	51.8	43.4	48.7
	Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	**72.5**	59.5	69.5	70.1	61.7	76.5	68.3	16.2	24.0	72.1	37.4	40.8	47.6	51.2	80.0	92.5	62.4	39.8	37.6	15.8	52.7	4.2	30.0	51.8
	Qwen3-VL-8B-Instruct (Bai et al., 2025a)	65.2	66.4	69.9	69.9	62.2	75.1	68.1	24.3	23.0	55.9	34.4	49.1	39.3	64.7	68.6	81.9	60.7	54.0	51.8	20.7	43.9	15.4	37.2	52.7
	Qwen3.5-9B (Qwen Team, 2026)	70.7	67.0	65.1	68.0	63.5	76.5	68.5	27.8	29.8	71.8	43.1	11.2	60.9	63.8	2.2	0.4	27.7	52.6	50.4	4.6	3.4	74.0	37.0	45.5
	DeepSeek-VL-7B (Lu et al., 2024)	71.7	67.5	68.4	68.6	52.6	76.7	67.6	9.0	29.2	69.9	36.0	60.6	51.4	64.7	19.2	5.8	40.3	42.8	40.6	8.7	19.7	20.0	26.4	44.6
	Kimi-VL-A3B-Thinking (Team et al., 2025)	67.0	63.6	69.5	68.8	62.2	76.2	67.9	22.6	25.6	71.9	40.0	60.8	46.0	64.9	70.1	82.5	64.9	51.4	49.8	33.7	50.1	25.1	42.0	55.9
	MiMo-VL-7B-RL (Yue et al., 2025)	71.8	68.9	66.9	67.9	**65.7**	76.8	**69.7**	28.1	29.4	73.0	43.5	59.6	51.2	64.7	19.4	5.8	40.1	54.0	52.1	24.0	15.0	67.5	42.5	50.6
Medical-Specialized	HuatuoGPT-Vision-7B (Zhang et al., 2023)	69.7	69.1	70.7	67.9	52.7	75.8	67.7	12.8	27.6	70.3	36.9	59.3	51.4	64.7	14.9	5.7	39.2	43.2	41.7	13.6	50.2	19.9	33.7	46.4
	LLaVA-Med-7B (Li et al., 2023)	71.7	69.4	68.7	64.2	52.6	75.5	67.0	16.8	18.9	66.4	34.0	40.8	48.6	35.3	85.1	**94.3**	60.8	28.4	27.1	0.0	0.2	0.0	11.1	45.5
	DentalGPT-Mid (Cai et al., 2025)	66.9	62.7	65.3	71.5	54.9	75.9	66.2	8.2	10.7	72.7	30.5	49.5	48.6	63.4	81.6	31.3	54.9	38.6	37.2	14.1	48.2	7.8	29.2	47.8
	HealthGPT-Pro-8B (Lin et al., 2025)	68.1	27.5	68.7	71.3	56.5	75.7	61.3	17.4	26.4	72.1	38.6	55.0	45.8	64.7	78.7	60.1	60.9	49.3	47.8	51.5	55.4	29.8	46.8	53.8
	Lingshu-7B (Xu et al., 2026)	68.3	62.6	68.1	68.3	57.0	75.8	66.7	22.7	12.1	71.5	35.4	58.1	51.0	64.7	24.9	31.5	46.0	46.8	45.2	39.6	59.8	48.8	48.0	51.4
	OralGPT-Omni-7B (Hao et al., 2026b)	66.6	69.2	61.5	71.6	57.9	72.9	66.6	10.5	17.4	39.5	22.5	61.7	51.4	64.7	84.0	94.0	71.2	55.6	53.8	12.7	52.7	21.6	39.3	53.6
	DentVLM (Meng et al., 2026)	72.4	**69.7**	68.7	72.1	52.6	75.0	68.4	23.6	26.9	71.8	40.8	50.8	48.7	52.1	47.8	93.1	58.5	57.9	56.3	23.2	24.6	28.5	38.1	53.5
	IntraoralGPT (ours)	70.5	68.9	68.1	**72.4**	53.2	**78.0**	68.5	30.6	29.1	74.5	44.7	**78.1**	**80.8**	**72.7**	**86.1**	**94.3**	**82.4**	**70.5**	**68.9**	**77.0**	**73.1**	**74.2**	**72.7**	**69.5**

Table 3: Effect of CariesBench-specific supervision across the hierarchical clinical evaluation. Values denote mean clinical accuracy (%) over question types within each task group. Δ denotes the absolute improvement of IntraoralGPT over its Qwen3-VL-8B-Instruct base model.

Task Group	Types	Qwen3-VL-8B-Instruct	IntraoralGPT	Δ
ACS	6	68.1	**68.5**	+0.4
TDR	3	34.4	**44.7**	+10.3
SCD-TF	5	60.7	**82.4**	+21.7
SCD-MC	5	37.2	**72.7**	+35.5
All Tasks	19	52.7	**69.5**	**+16.8**

(SV). Across all 24 baselines, SA and FS peak at only 31.8% and 30.5%, respectively, whereas SV reaches 75.4%. This contrast suggests that the largest performance degradation is associated with tasks requiring surface-specific localization and cross-view integration rather than tooth-level input alone.

Finding 2. MLLMs exhibit a pronounced global-to-local reasoning gap. Within CariesBench, models consistently achieve lower performance on TDR than on ACS, revealing a benchmark-level global-to-local reasoning gap. Across the 24 baselines, the mean ACS Overall is 67.3%, compared with only 38.2% for TDR Overall, corresponding to a gap of 29.1 percentage points. This pattern is observed across all three model families. Although ACS and TDR differ in task format as well as anatomical granularity, the consistent gap reveals substantially weaker performance under the localized tooth- and surface-level reasoning demands evaluated by TDR.

Finding 3. Dental or medical specialization does not consistently translate into stronger multi-view intraoral reasoning. Under the same zero-shot evaluation protocol, medical- and dental-specialized models do not consistently outperform general-purpose MLLMs on CariesBench. DentVLM, for example, achieves competitive results on TE (69.7%) and FS (26.9%), but its TDR Overall remains 40.8%, below several general-purpose commercial models. OralGPT-Omni-7B and DentalGPT-Mid achieve 22.5% and 30.5% TDR Overall, respectively, whereas GPT-5.5 achieves the strongest zero-shot baseline TDR performance at 45.9%. These results suggest that domain specialization alone is insufficient to ensure strong performance on the multi-view reasoning requirements evaluated by CariesBench, and that targeted supervision explicitly representing anatomical correspondence and cross-view evidence may be beneficial.

Finding 4. CariesBench-specific supervision substantially improves hierarchical clinical reasoning. To investigate whether the limitations exposed by CariesBench can be mitigated through benchmark-specific supervision, we compare IntraoralGPT with its Qwen3-VL-8B-Instruct base model under the same evaluation protocol. IntraoralGPT improves the 19-task macro-average clinical accuracy from 52.7% to 69.5%, corresponding to a gain of 16.8 percentage points. The gains vary

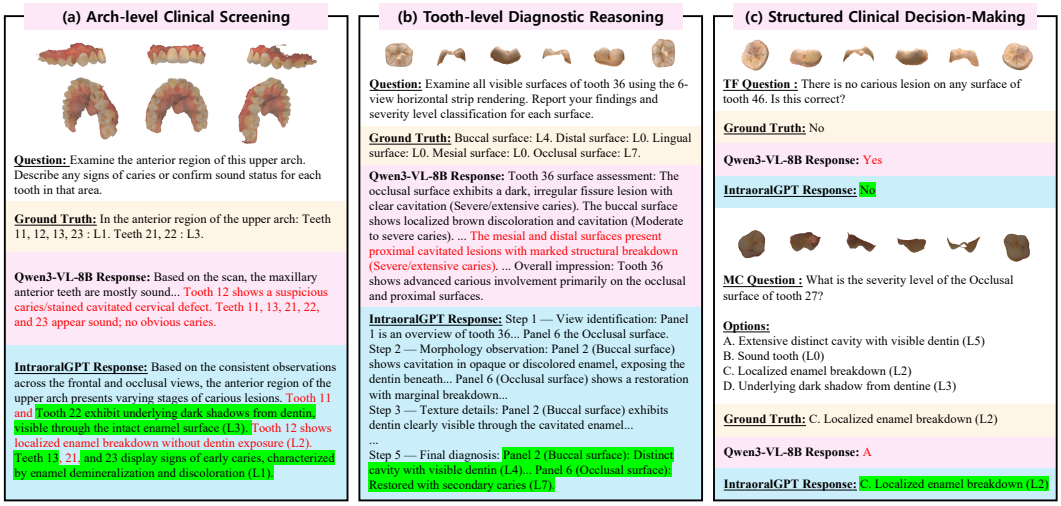

Figure 3: Representative case studies on CariesBench across (a) ACS, (b) TDR, and (c) SCD. Comparisons between Qwen3-VL-8B-Instruct and IntraoralGPT illustrate their differences in hierarchical dental reasoning. Correct predictions are highlighted in green and incorrect predictions in red.

substantially across task groups: ACS improves by 0.4 points, TDR by 10.3 points, SCD-TF by 21.7 points, and SCD-MC by 35.5 points. Within TDR, SA improves from 24.3% to 30.6%, while FS improves from 23.0% to 29.1%. These results show that CariesBench-specific supervision yields particularly large gains on structured clinical decisions while also improving tooth-level multi-view reasoning. Nevertheless, IntraoralGPT reaches 44.7% TDR Overall, remaining slightly below the strongest zero-shot baseline, GPT-5.5, at 45.9%, highlighting the remaining difficulty of fine-grained anatomical localization and cross-view evidence integration.

4.3 Case Study

Figure 3 qualitatively compares Qwen3-VL-8B-Instruct with its CariesBench-supervised counterpart, IntraoralGPT, across the three evaluation levels. In the ACS example, IntraoralGPT produces a more accurate tooth-specific clinical assessment, although residual grading errors remain. In the TDR example, IntraoralGPT more accurately associates the observed abnormality with the corresponding tooth surface, consistent with its improved quantitative performance on tooth-level multi-view tasks. In the illustrated SCD examples, IntraoralGPT correctly answers both structured decisions whereas Qwen3-VL-8B-Instruct fails both. These examples qualitatively mirror the task-wise gains in Table 3; consistent with the quantitative results, fine-grained surface-level reasoning nevertheless remains challenging.

5 Conclusion

In this work, we introduce CariesBench, a hierarchical benchmark for evaluating clinically grounded multi-view reasoning from 3D intraoral scans in multimodal large language models (MLLMs). By organizing evaluation into Arch-level Clinical Screening (ACS), Tooth-level Diagnostic Reasoning (TDR), and Structured Clinical Decision-Making (SCD), CariesBench evaluates the progression from global arch-level screening to localized tooth- and surface-level reasoning and structured clinical decisions. Extensive evaluation of 24 representative MLLMs reveals a pronounced global-to-local reasoning gap: strong arch-level performance does not reliably translate into fine-grained anatomical localization and cross-view evidence integration. CariesBench-specific supervision substantially improves hierarchical clinical reasoning, increasing the 19-task macro-average clinical accuracy of Qwen3-VL-8B-Instruct from 52.7% to 69.5% in IntraoralGPT. Nevertheless, the remaining limitations on TDR highlight the continued difficulty of fine-grained anatomical localization and cross-view diagnosis. CariesBench therefore provides a clinically structured testbed for evaluating MLLMs across anatomical granularities in digital dentistry. Future work should further examine native 3D reasoning and generalization across acquisition settings and clinical populations.

AI USE STATEMENT

Generative AI tools were used during benchmark construction and manuscript preparation. During benchmark construction, GPT-4o was used to verbalize annotation-grounded questions and responses for the open-ended ACS and TDR tasks, while Gemini-2.5-Pro was used to audit generated instances for consistency with structured clinical annotations, anatomical correspondence, and unsupported clinical claims. These tools were not used to determine the underlying clinical reference labels, which were derived from structured clinical annotations rather than generated by an LLM; expert evaluation and adjudication were conducted according to the clinical quality-control protocol described in Section 3.5. For SCD, questions and target answers were instantiated deterministically from structured ground truth rather than generated by an LLM.

Generative AI tools were additionally used for language polishing and improving the clarity and readability of the manuscript. All AI-assisted revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts.

ETHICS STATEMENT

Before data collection, informed consent was obtained from all adult participants and from the parents or legal guardians of minor participants, with assent obtained from minors where required by the approved protocol. All data used for benchmark construction were de-identified before research processing. The study protocol was reviewed and approved by the institutional ethics committee; identifying committee and approval information is withheld during double-blind review. The study was conducted in accordance with applicable regulations governing biomedical research involving human participants and with the principles of the Declaration of Helsinki.

REPRODUCIBILITY STATEMENT

We provide detailed specifications of the CariesBench construction and evaluation pipeline to facilitate reproducibility. Section 3 describes the patient-level data partition, multi-view rendering protocol, clinical-status taxonomy, hierarchical task construction, and quality-control procedure, while Section 4 reports the evaluation protocol and IntraoralGPT training configuration. The Appendix further documents clinical label definitions (Appendix B), multi-view rendering (Appendix C), task and reasoning construction (Appendices E–H), quality control (Appendix J), evaluation metrics (Appendix K), and training details (Appendix L). All reported validation and test results are computed on patient-disjoint splits that retain the naturally occurring clinical-status distribution.

REFERENCES

Anthropic. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, 2025. Accessed: 2026-09-08.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, 2026. Accessed: 2026-09-08.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Achraf Ben-Hamadou, Oussama Smaoui, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Hoyeon Lim, Minchang Kim, Minkyung Lee, Minyoung Chung, Yeong-Gil Shin, et al. 3dteethseg'22: 3d teeth scan segmentation and labeling challenge. *arXiv preprint arXiv:2305.18277*, 2023.

Federico Bolelli, Kevin Marchesini, Niels Van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. Segmenting maxillofacial structures in cbct volumes. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248. IEEE, 2025.

Zhenyang Cai, Jiaming Zhang, Junjie Zhao, Ziyi Zeng, Yanchao Li, Jingyi Liang, Junying Chen, Yunjin Yang, Jiajun You, Shuzhi Deng, et al. Dentalgpt: Incentivizing multimodal complex reasoning in dentistry. *arXiv preprint arXiv:2512.11558*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Benin Dikmen. Icdas ii criteria (international caries detection and assessment system). *Journal of Istanbul University Faculty of Dentistry*, 49(3):63–72, 2015.

Florin Eggmann and MB Blatz. Recent advances in intraoral scanners. *Journal of Dental Research*, 103(13):1349–1357, 2024.

Google. Gemini api documentation: Gemini 3.1 Pro Preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`, 2026a. Accessed: 2026-09-08.

Google. Gemini api documentation: Gemini 3.6 Flash. `https://ai.google.dev/gemini-api/docs/models/gemini-3.6-flash`, 2026b. Accessed: 2026-09-08.

Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, Hongwei Bran Li, Sarthak Pati, Bernd Stadlinger, et al. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. *arXiv preprint arXiv:2305.19112*, 2023.

Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lin Lizhuo, Jinrong Yang, Qiyong Ai, Lun Wong, Hao Tang, and Kuo Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. *Advances in Neural Information Processing Systems*, 38, 2026a.

Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, et al. Oralgpt-omni: A versatile dental multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38509–38519, 2026b.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Amid I Ismail, Woosung Sohn, Marisol Tellez, Ashley Amaya, Ananda Sen, Hana Hasson, and Nigel Berry Pitts. The international caries detection and assessment system (icdas): an integrated system for measuring dental caries. *Community dentistry and oral epidemiology*, 35(3):170–178, 2007.

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. *Advances in neural information processing systems*, 36: 28541–28564, 2023.

M-X Li, W-H Deng, Z-X Wu, C-X Jin, J-M Wu, Yue Han, James Kit Hon Tsoi, G-S Xia, and Cui Huang. Metadent: Labeling clinical images for vision-language models in dentistry. *Journal of Dental Research*, pp. 00220345261424242, 2026.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Zhiguo Lu, Jianwen Lou, Mingjun Ma, Hairong Jin, Youyi Zheng, and Kun Zhou. 3dteethsam: Taming sam2 for 3d teeth segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7609–7617, 2026.

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*, 2026.

P Ntovas, S Michou, AR Benetti, A Bakhshandeh, K Ekstrand, C Rahiotis, and A Kakaboura. Occlusal caries detection on 3d models obtained with an intraoral scanner. a validation study. *Journal of Dentistry*, 131:104457, 2023.

OpenAI. Models api reference - GPT-5.4-nano. `https://developers.openai.com/api/docs/models/gpt-5.4-nano`, 2026a. Accessed: 2026-09-08.

OpenAI. Models api reference - GPT-5.5. `https://developers.openai.com/api/docs/models/gpt-5.5`, 2026b. Accessed: 2026-09-08.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Evgeniy I Sosnin, Yuriy L Vasilev, Roman A Solovyev, Aleksandr L Stempkovskiy, Dmitry V Telpukhov, Artem A Vasilev, Aleksandr A Amerikanov, and Aleksandr Y Romanov. Alphadent: A dataset for automated tooth pathology detection. *arXiv preprint arXiv:2507.22512*, 2025.

Zhi Qin Tan, Ilana Felix Pinho, Ryan Banks, Xiongjie Chen, Laura Regina Antunes Pontes, Mariana Minatel Braga, Fausto Medeiros Mendes, and Yunpeng Li. Caries detection in primary molars with bitewing radiographs through deep learning based-object detectors. *Caries research*, 2025.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Pawan Valluri. Dentalai computer vision project, 2023. URL `https://www.kaggle.com/datasets/pawanvalluri/dental-segmentation`.

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*, 2025.

Huimin Xiong, Zijie Meng, Tianxiang Hu, Chenyi Zhou, Yang Feng, and Zuozhu Liu. Iosvlm: A 3d vision-language model for unified dental diagnosis from intraoral scans. *arXiv preprint arXiv:2603.16781*, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 10859–10885, 2023.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

A Additional Details of CariesBench

This appendix provides additional details of the clinical status taxonomy, standardized multi-view rendering protocol, patient-level data partitioning and sampling strategy, hierarchical task construction, VQA generation pipeline, structured clinical decision construction, language realization and automated auditing, expert quality control, evaluation protocol, and implementation details used in CariesBench. These specifications complement the benchmark construction and experimental settings described in the main paper and are provided to facilitate reproducibility.

B Clinical Status Taxonomy and Diagnostic Criteria

CariesBench adopts a task-oriented clinical status taxonomy derived from the visual progression defined by ICDAS-II (Ismail et al., 2007; Dikmen, 2015). The benchmark contains eight clinical states, $\mathcal{L}$ ={L0,L1,L2,L3,L4,L5,L6,L7}. To reduce ambiguity between subtle early manifestations in rendered intraoral-scan views, ICDAS-II codes 1 and 2 are merged into a unified early-caries state. The subsequent caries states preserve the visual progression of ICDAS-II codes 3–6. Restoration without secondary caries and restoration with secondary caries are included as clinically relevant categorical extensions. The complete correspondence and operational visual criteria are summarized in Table 4.

C Detailed 3D-to-2D Multi-view Rendering Protocol

C.1 Mesh Preprocessing

Each complete intraoral scan is represented as a 3D mesh. For tooth-level rendering, individual tooth meshes are first obtained using the pretrained 3DTeethSAM model (Lu et al., 2026). Before rendering, each mesh is translated to its geometric centroid and isotropically normalized according to its largest spatial extent. For a mesh with vertices $\mathbf{v}$, geometric centroid $\mathbf{c}$, and maximum axis extent $e_{\max}$, the normalized coordinates are defined as

$$\mathbf{v}' = \frac{2(\mathbf{v} - \mathbf{c})}{e_{\max}}. \tag{1}$$

This normalization places different arches and segmented teeth in a consistent coordinate scale while preserving their relative surface geometry. All rendered views are subsequently generated from the normalized mesh using fixed anatomical camera configurations.

Table 4: Clinical status taxonomy used in CariesBench. L0–L5 follow the visual progression of ICDAS-II after merging ICDAS codes 1–2, whereas L6–L7 are restoration-related categorical extensions.

Label	Clinical Status	ICDAS-II	Morphological Definition	Appearance / Texture Cues
L0	Sound tooth	0	Sound tooth surfaces show no visible evidence of caries, with no or only questionable change in enamel translucency.	Developmental defects such as enamel hypomineralization, tooth wear, and extrinsic or intrinsic stains are recorded as sound in the absence of visible caries.
L1	Early caries	1–2	A carious opacity or discoloration, typically white or brown, is present and is inconsistent with the clinical appearance of sound enamel, without visible structural breakdown.	White or brown opacity or discoloration compatible with early enamel demineralization.
L2	Localized enamel breakdown	3	Distinct localized loss of enamel surface integrity is present. The base and walls of the defect remain within enamel, with no visible dentin exposure.	Localized enamel breakdown without visible dentin exposure.
L3	Underlying dark shadow from dentin	4	A shadow of discolored dentin is visible through the enamel surface and may occur with or without localized enamel breakdown.	An intrinsic dark shadow that may appear grey, blue, or brown beneath the enamel.
L4	Distinct cavity with visible dentin	5	Cavitation in opaque or discolored enamel exposes the underlying dentin and involves less than approximately half of the affected tooth surface.	Dentin is clearly visible through the cavitated enamel.
L5	Extensive distinct cavity with visible dentin	6	Extensive cavitation exposes the underlying dentin and involves approximately half or more of the affected tooth surface, with substantial loss of tooth structure.	Extensive visible dentin exposure with marked structural loss involving the cavity walls or base.
L6	Restored without secondary caries	Extension	A restored tooth surface shows no visible evidence of caries adjacent to the restoration margin.	No visible carious lesion or compatible marginal breakdown surrounding the restoration.
L7	Restored with secondary caries	Extension	A new carious lesion or restoration-associated marginal breakdown compatible with secondary caries is present adjacent to an existing restoration.	Restoration-associated structural change, discoloration, or cavitation compatible with secondary caries.

Table 5: Fixed spatial layout of the six arch-level views in CariesBench. Maxillary and mandibular arches use arch-specific row arrangements while preserving consistent right–center–left anatomical organization within each row.

Arch / Row	Left Panel	Center Panel	Right Panel
Maxillary – Top	Right Buccal	Frontal	Left Buccal
Maxillary – Bottom	Right Palatal	Occlusal	Left Palatal
Mandibular – Top	Right Lingual	Occlusal	Left Lingual
Mandibular – Bottom	Right Buccal	Frontal	Left Buccal

C.2 Arch-level Multi-view Rendering

Each complete maxillary or mandibular IOS is rendered into six standardized views covering complementary anatomical directions. The views are arranged in a fixed 2×3 spatial layout. For maxillary arches, the top row consists of right buccal, frontal, and left buccal views, while the bottom row consists of right palatal, occlusal, and left palatal views. For mandibular arches, the top row consists of right lingual, occlusal, and left lingual views, while the bottom row consists of right buccal, frontal, and left buccal views. This convention provides consistent anatomical orientation across cases while exposing complementary observations of the same underlying dental arch.

C.3 Tooth-level Multi-view Rendering

Each segmented tooth is rendered as a horizontal multi-view panel consisting of an overview followed by anatomically defined surface-specific views. The rendering configuration is conditioned on dental arch, laterality, and tooth type. For anterior teeth, the outer surface is designated as *labial*, whereas for posterior teeth it is designated as *buccal*. The inner surface is designated as *palatal* for maxillary teeth and *lingual* for mandibular teeth. Incisors additionally include an *incisal* view, posterior teeth include an *occlusal* view, and canines contain four surface-specific views without an additional incisal or occlusal view.

Mesial and distal orientations are defined relative to tooth anatomy rather than by a globally fixed image direction. Accordingly, the surface sequence is mirrored between the right and left dental quadrants. For teeth in the right quadrants, the horizontal sequence follows Outer–Mesial–Inner–Distal–Incisal/Occlusal where applicable, whereas for teeth in the left quadrants the mesial and distal locations are mirrored as Outer–Distal–Inner–Mesial–Incisal/Occlusal. The complete panel ordering is summarized in Table 6.

Table 6: Tooth-level panel ordering used in CariesBench. Panel 1 provides an overview of the target tooth and is followed by surface-specific observations. Mesial and distal positions are mirrored between the right and left quadrants to preserve tooth-relative anatomical orientation.

Tooth Position	**Panel 1**	**Panel 2**	**Panel 3**	**Panel 4**	**Panel 5**	**Panel 6**
Upper Right Incisor (11–12)	Overview	Labial	Mesial	Palatal	Distal	Incisal
Upper Right Canine (13)	Overview	Labial	Mesial	Palatal	Distal	–
Upper Right Posterior (14–17)	Overview	Buccal	Mesial	Palatal	Distal	Occlusal
Upper Left Incisor (21–22)	Overview	Labial	Distal	Palatal	Mesial	Incisal
Upper Left Canine (23)	Overview	Labial	Distal	Palatal	Mesial	–
Upper Left Posterior (24–27)	Overview	Buccal	Distal	Palatal	Mesial	Occlusal
Lower Right Incisor (41–42)	Overview	Labial	Mesial	Lingual	Distal	Incisal
Lower Right Canine (43)	Overview	Labial	Mesial	Lingual	Distal	–
Lower Right Posterior (44–47)	Overview	Buccal	Mesial	Lingual	Distal	Occlusal
Lower Left Incisor (31–32)	Overview	Labial	Distal	Lingual	Mesial	Incisal
Lower Left Canine (33)	Overview	Labial	Distal	Lingual	Mesial	–
Lower Left Posterior (34–37)	Overview	Buccal	Distal	Lingual	Mesial	Occlusal

Table 7: Rendering and post-processing parameters used to construct standardized multi-view images.

Parameter	**Arch**	**Tooth**
Projection	Perspective	Perspective
Tile resolution	3072×3072	4096×4096
SSAA factor	$2\times$	$2\times$
Background	White	White
Surface interpolation	Smooth	Smooth
Ambient light	0.5 upper / 0.4 lower	0.2
Base light intensity	3.5	2.0
Illumination	Single directional	Key–fill–rim directional
Bilateral-filter d	5	9
σ_{color}	40	75
σ_{space}	40	75

Incisors and posterior teeth therefore contain six panels in total, including one overview and five anatomical surface views, whereas canines contain five panels consisting of one overview and four surface-specific views. Because all views are rendered from the same segmented 3D tooth mesh, the panels preserve a common anatomical identity while exposing complementary surface-level evidence for tooth-centric multi-view reasoning.

C.4 Rendering Configuration

Perspective projection is used for both arch- and tooth-level rendering. Each camera pose is parameterized by elevation, azimuth, camera distance, field of view, and an optional up vector. Camera configurations are fixed for the corresponding anatomical views and shared across cases. The principal rendering and post-processing parameters are reported in Table 7.

For arch-level rendering, a single directional light is combined with arch-specific ambient illumination. For tooth-level rendering, a three-light key–fill–rim configuration is used to improve the visibility of local surface morphology while maintaining a consistent appearance across views. After supersampled rendering, each panel is downsampled to its target resolution and processed using bilateral filtering to suppress rasterization noise while preserving major surface boundaries. The complete camera configurations are fixed before benchmark construction and indexed by arch or FDI tooth number. Each configuration specifies elevation, azimuth, camera distance, field of view, and an up vector; the exact per-view configurations are provided with the rendering implementation.

D Dataset Partitioning and Sampling Protocol

D.1 Patient-level Split and Leakage Prevention

CariesBench contains 5,932 IOS scans from 2,966 patients, with one maxillary and one mandibular scan per patient. Patients are partitioned into 2,372 training patients, 297 validation patients, and 297 test patients. Let $\mathcal{P}_{\text{train}}$, $\mathcal{P}_{\text{val}}$, and $\mathcal{P}_{\text{test}}$ denote the corresponding patient sets. We enforce

$$\mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{val}} = \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \mathcal{P}_{\text{val}} \cap \mathcal{P}_{\text{test}} = \varnothing. \tag{2}$$

All data derived from the same patient—including maxillary and mandibular meshes, arch-level renderings, segmented tooth meshes, tooth-level views, surface observations, and generated VQA instances—are restricted to the same split. Importantly, the patient partition is applied before downstream VQA sampling or training-data rebalancing, preventing leakage through shared anatomy, repeated renderings, or multiple questions generated from the same patient.

D.2 Natural Validation and Test Distribution

The full benchmark is constructed from the available structured clinical annotations without imposing per-class generation quotas on the validation or test partitions. Consequently, these evaluation sets retain the naturally occurring clinical-status distribution of the underlying cohort rather than an artificially class-balanced distribution. This design is intended to characterize model behavior under a distribution closer to the clinical data from which CariesBench was constructed.

D.3 Training-set Rebalancing

Clinical data are dominated by common findings, whereas several clinically important non-sound states occur less frequently. After the patient-level split is fixed, we independently rebalance the ACS, TDR, and SCD training sets toward the same target clinical-status distribution over L0–L7: 25.5%, 17.5%, 12.5%, 10.5%, 10.5%, 5.5%, 12.5%, and 5.5%, respectively. This training-only procedure changes sampling frequency without modifying the underlying clinical annotations or patient assignments. Validation and test remain unrebalanced and retain their naturally occurring clinical-status distributions.

E Detailed Hierarchical Task Definitions

CariesBench contains 19 question types organized into four reported evaluation groups: six Arch-level Clinical Screening (ACS) tasks, three Tooth-level Diagnostic Reasoning (TDR) tasks, five Structured Clinical Decision true/false (SCD-TF) tasks, and five Structured Clinical Decision multiple-choice (SCD-MC) tasks. ACS and TDR use open-ended clinically grounded responses that are parsed into structured predictions, whereas SCD questions are constructed deterministically from structured tooth- and surface-level annotations. Table 8 summarizes the complete task taxonomy.

F VQA Construction Protocol

CariesBench uses a ground-truth-driven construction pipeline that explicitly separates structured clinical annotations from subsequent natural-language realization. Structured annotations are organized by patient, tooth, anatomical surface, and clinical status and serve as the reference source throughout benchmark construction and evaluation. The clinical labels themselves are not generated, revised, or inferred by the language models used during VQA construction.

For the open-ended ACS and TDR tasks, GPT-4o is used only to realize annotation-grounded questions and responses in natural language. The model receives the task-specific clinical context and corresponding multi-view observation and is instructed to describe only evidence supported by the supplied information. The underlying structured clinical labels remain unchanged. For SCD, both questions and target answers are instantiated deterministically from the structured annotations without using an LLM to determine the correct answer.

Table 8: Complete task taxonomy of CariesBench. Question patterns summarize the semantic structure of each task rather than a single mandatory wording. Evaluation targets are derived from structured clinical annotations.

Group	Task	Abbr.	Input Scope	Question Pattern	Evaluation Target
ACS	Global Screening	GS	Arch multi-view	Perform comprehensive screening of the complete arch and report visible caries- and restoration-related findings with their clinical status.	Correct identification of relevant teeth and their corresponding clinical states across the arch.
	Tooth Enumeration	TE	Arch multi-view	Enumerate visible teeth, assign a clinical state to each target tooth, and identify the primary affected surface when supported by the observations.	Correct tooth identification and tooth-level clinical-status assignment, together with the relevant surface when required.
	Specific Region	SR	Arch multi-view	Examine a specified anterior or posterior region and describe caries-related findings or confirm sound status for teeth in that region.	Correct localization and clinical assessment of teeth within the specified anatomical region.
	Differential Diagnosis	DD	Arch multi-view	Given a tooth with subtle visual changes, distinguish between clinically adjacent candidate states using localized visual evidence.	Correct discrimination between the candidate clinical states for the target tooth.
	Severity Comparison	SC	Arch multi-view	Compare eligible carious findings within the arch according to the ordered caries progression represented by L0–L5.	Correct relative ordering of the compared carious findings according to the benchmark caries progression.
	Sound Tooth Identification	SI	Arch multi-view	Assess teeth in a specified region and identify those that are clinically sound.	Correct identification or confirmation of sound (L0) teeth within the queried region.
TDR	Surface Analysis	SA	Tooth multi-view	Analyze a specified anatomical surface of the target tooth and assign its clinical-status label.	Exact clinical-status classification (L0–L7) of the queried tooth surface.
	Full Tooth Scan	FS	Tooth multi-view	Examine all surface-specific views of the target tooth and report the clinical status of each evaluated surface.	Surface-wise clinical-status prediction across all evaluated anatomical surfaces of the target tooth.
	Sound Tooth Verification	SV	Tooth multi-view	Determine whether the target tooth is completely sound and describe the supporting visual evidence.	Binary sound/non-sound correctness for the complete target tooth.
SCD-TF	Surface Verification	Surf	Tooth / surface	Verify whether a statement assigning a clinical state to a sampled tooth surface is correct.	Binary correctness against the exact structured surface annotation.
	Whole-Tooth Soundness	Whl	Tooth	Verify whether the target tooth is completely sound across all annotated surfaces.	True iff every annotated surface of the target tooth is L0.
	No-Caries Verification	NoC	Tooth	Verify whether the target tooth contains no L1–L5 carious state.	True iff none of the annotated surfaces belongs to L1–L5.
	Restoration Verification	Rst	Tooth	Verify whether a restoration-related state is present on the target tooth.	True iff at least one surface is annotated as L6 or L7.
	Secondary-Caries Verification	Sec	Tooth	Verify whether secondary caries is present around an existing restoration.	True iff at least one surface is annotated as L7.
SCD-MC	Worst Tooth Identification	Wor	Arch multi-view	Select the tooth associated with the most advanced eligible carious finding within the queried arch context.	Exact selection of the target tooth or teeth determined from the structured clinical annotations.
	Sound Count	SC	Arch multi-view	Determine the number of completely sound teeth in the complete arch.	Exact match with the number of teeth for which all annotated surfaces are L0.
	Severity Classification	Sev	Tooth / surface	Select the clinical-status label of a sampled anatomical surface from multiple candidate states.	Exact multiple-choice match with the structured surface annotation.
	Worst Surface	WS	Tooth multi-view	Identify the surface containing the most advanced eligible carious finding within the target tooth.	Exact identification of the target surface or surfaces determined from the structured annotations.
	Tooth Status	TS	Tooth multi-view	Select the overall clinical category of the target tooth from *Sound*, *Carious*, *Restored*, or *Restored with secondary caries*.	Exact match with the deterministically derived tooth-level clinical category.

Table 9: Representative ACS question templates used during language realization. Variables in braces are instantiated from structured clinical annotations.

Task	Representative Template
GS	Perform a comprehensive clinical screening of this {arch_type} arch scan. Report the visible caries-related and restoration-related findings and their corresponding clinical status.
TE	List each tooth visible in this {arch_type} arch and classify it using the severity level scale. For non-sound teeth, identify the primary affected surface when supported by the observations.
SR	Examine the {region} of this {arch_type} arch. Describe caries-related findings or confirm sound status for the teeth in that region.
DD	Tooth {FDI} in this {arch_type} arch shows subtle changes. Compare {state A} and {state B} and determine the more appropriate clinical interpretation using visible evidence.
SC	Compare the severity of all carious lesions in this {arch_type} arch and order them according to the benchmark caries progression, with supporting visual evidence.
SI	Assess the {region} teeth in this {arch_type} arch and identify which teeth are sound, supported by visible evidence.

F.1 ACS Question Selection and Templates

For each eligible arch, one ACS question type is selected from the six ACS task types after applying task-specific eligibility constraints. The candidate order is randomized using a fixed seed, and the first eligible task is instantiated. Global Screening, Tooth Enumeration, and Specific Region can be instantiated for valid arches. Differential Diagnosis requires an eligible target exhibiting a subtle or clinically adjacent distinction, Severity Comparison requires multiple eligible carious findings, and Sound Tooth Identification requires the presence of sound teeth. No class-dependent quota is imposed during full benchmark generation.

Table 10: Five-step progressive reasoning protocol used for reasoning-oriented TDR instances. The protocol progresses from anatomical view identification to local visual observation, cross-view integration, and final clinical-status prediction.

Step	Abbr.	Stage	Objective	Expected Output
1	VI	View Identification	Establish the anatomical identity of each panel according to the target tooth's arch, laterality, and tooth type.	Correct correspondence between each panel and its represented anatomical surface.
2	MO	Morphology Observation	Examine visible structural characteristics including contour, enamel integrity, cavitation, tissue loss, exposed dentin, and restoration margins.	Surface-specific description of visible morphological changes without prematurely assigning the final clinical state.
3	TD	Texture Details	Characterize surface appearance cues including color, opacity, discoloration, staining, and reflective properties.	Visually grounded description of appearance and texture cues associated with each relevant surface.
4	CC	Cross-view Consistency	Integrate evidence across the overview and surface-specific views and assess whether the observations provide a consistent interpretation.	Cross-view synthesis identifying corroborating, conflicting, or insufficient evidence.
5	FD	Final Diagnosis	Synthesize anatomically localized evidence and map the findings to the benchmark clinical taxonomy.	Final clinical-status prediction (L0–L7) for the queried surface or target tooth.

F.2 TDR Question Selection and Templates

At the tooth level, one eligible TDR task type is selected for each target tooth. Non-sound teeth are used to construct Surface Analysis or Full Tooth Scan questions, whereas sound teeth are used for Sound Tooth Verification. For TDR generation, the structured surface-level annotation of the target tooth is supplied as clinical context. The realized response is required to preserve correspondence between panel identity, anatomical surface, visual evidence, and final clinical-status assignment.

Representative TDR question templates are as follows:

- **SA:** "Provide a detailed clinical analysis and severity level classification for the {surface} surface of tooth {FDI}."
- **FS:** "Examine all visible surfaces of tooth {FDI} using the {n}-view horizontal strip rendering. Report your findings and severity level classification for each surface."
- **SV:** "Determine whether tooth {FDI} is sound (L0). Describe the visual evidence supporting your assessment."

G Five-Step Progressive Reasoning Protocol for TDR

Reasoning-oriented TDR instances follow a five-step progression from anatomical localization to final clinical-status prediction. The purpose of this structure is to make surface correspondence and cross-view integration explicit rather than allowing the final prediction to be detached from the observed anatomy. The five stages are summarized in Table 10.

View identification follows the tooth-specific anatomical surface mapping defined in Table 6.

H Deterministic Construction of Structured Clinical Decisions

Unlike the open-ended ACS and TDR tasks, SCD target answers are derived directly from structured clinical annotations. This design ensures that correctness labels do not depend on an LLM-generated reference response. Let $y_{t,s} \in \mathcal{L}$ denote the annotated clinical state of surface s of tooth t. The principal construction rules are summarized in Table 11.

For multiple-choice questions, incorrect options are sampled from alternative clinically valid states or anatomically valid candidate surfaces or teeth, depending on the task. The options are randomly ordered, and the correct answer index is stored directly from the randomized option list.

I Language Realization and Automated Auditing

GPT-4o is used only for natural-language realization of the open-ended ACS and TDR instances. The generation input contains the corresponding multi-view observation, structured clinical context,

Table 11: Deterministic rules used for SCD construction. "Carious" denotes caries states L1–L5, whereas L6 and L7 are treated as restoration-related clinical states.

Subtype	Target	Ground-truth Rule
TF-Surf	Surface verification	A statement assigning a clinical state to a sampled surface is evaluated against its exact structured surface annotation.
TF-Whl	Whole-tooth soundness	True iff every annotated surface of the target tooth is L0.
TF-NoC	No caries	True iff no annotated surface belongs to L1–L5. Restoration-related states alone do not satisfy the definition of an L1–L5 carious lesion.
TF-Rst	Restoration	True iff at least one surface is labeled L6 or L7.
TF-Sec	Secondary caries	True iff at least one surface is labeled L7.
MC-Wor	Worst tooth	The correct option is the tooth or set of teeth containing the most advanced eligible carious finding according to the ordered L0–L5 progression used by the comparison task.
MC-SC	Sound count	The correct option equals the number of teeth for which all annotated surfaces are L0.
MC-Sev	Surface status	The correct option is the exact clinical state assigned to the queried surface.
MC-WS	Worst surface	The correct option is the surface or set of surfaces containing the most advanced eligible carious finding within the target tooth according to the ordered L0–L5 progression.
MC-TS	Tooth status	The overall tooth category is derived deterministically from its structured surface annotations as *Sound*, *Carious*, *Restored*, or *Restored with secondary caries*.

Table 12: Configuration used for VQA language realization.

Setting	Value
Language model	GPT-4o
Temperature	0.3
Random seed	42
Response format	JSON
ACS maximum output tokens	8,192
TDR maximum output tokens	4,096
Maximum image side for API input	1,024 px
API image encoding	JPEG
JPEG quality	85
Image detail mode	High

task-specific question, and explicit instructions restricting the response to findings supported by the supplied visual and clinical information. For arch-level instances, the prompt specifies the fixed six-view spatial layout. For tooth-level instances, the prompt explicitly associates each panel with its anatomical surface and, for reasoning-oriented TDR instances, requires the five-stage sequence defined in Appendix G. The response-generation model does not modify the underlying structured clinical annotation.

After language realization, Gemini-2.5-Pro is used as a secondary automated auditor. Each generated ACS or TDR instance is checked for: (1) consistency with the structured clinical annotation; (2) consistency between referenced tooth/surface anatomy and the supplied multi-view observations; and (3) unsupported or contradictory clinical claims. This stage functions as a quality-control filter rather than as a source of reference labels. The structured clinical annotations remain the benchmark ground truth throughout construction.

J Clinical Annotation and Expert Quality Control

J.1 Structured Clinical Annotation

The structured clinical annotations originate from epidemiological surveys, oral examinations, and clinical data curation performed at a university-affiliated stomatological hospital. Annotation and

data entry were conducted by five postgraduate dental trainees, three standardized-residency trainees, and one attending dentist under senior clinical supervision. The annotations are organized at the patient, tooth, anatomical-surface, and clinical-status levels and constitute the reference labels used for benchmark construction, VQA generation, and evaluation. GPT-4o and Gemini-2.5-Pro are not used to determine or revise these clinical labels.

J.2 Expert Review

Expert validation is performed on a stratified subset comprising 20% of the complete benchmark and covering all task families and clinical-status categories. Two dentists with 8 and 3 years of clinical experience, respectively, review the selected instances using a standardized protocol designed and supervised by a senior dentist with 15 years of clinical experience.

General response quality is evaluated along five dimensions:

1. **Tooth/Surface Mapping:** whether the response associates clinical findings with the correct tooth and anatomical surface;
2. **Evidence Grounding:** whether diagnostic statements are supported by observable multi-view evidence;
3. **Target Completeness:** whether the response addresses the complete clinical target required by the question;
4. **Cross-view Consistency:** whether evidence from complementary observations is integrated without internal contradiction;
5. **Conclusion Support:** whether the final clinical conclusion is supported by the preceding observations.

For reasoning-oriented TDR responses, the five reasoning components VI, MO, TD, CC, and FD are additionally assessed. Across the reported quality-control criteria, mean expert ratings range from 4.26 to 4.95 on a five-point scale. Of the expert-reviewed subset, 800 stratified instances spanning all task families and clinical-status categories were independently assessed by both dentists for inter-rater reliability analysis. Agreement was quantified using linearly weighted Cohen's κ, yielding $\kappa = 0.86$. Disagreements were resolved through adjudication and consensus under senior clinical supervision. Among reviewed instances, 5.2% required adjudication, while 1.8% were excluded because of unsupported visual claims or unresolved clinical ambiguity.

K Detailed Evaluation Protocol

K.1 Clinical Accuracy

Clinical Accuracy evaluates task-specific structured diagnostic correctness after parsing each model response into the expected prediction format. For a task containing N independently scored targets, accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right), \tag{3}$$

where $\hat{y}_i$ denotes the parsed model prediction, y_i denotes the corresponding structured reference target, and $\mathbb{I}(\cdot)$ is the indicator function.

ACS. The required structured predictions depend on the ACS question type. Model outputs are parsed into the corresponding tooth identities, clinical-status predictions, regional targets, comparison outcomes, or sound-tooth predictions specified by the task. Parsed predictions are then compared with the corresponding structured reference annotations at the task-appropriate granularity. A prediction is counted as correct only when the required structured target matches the reference; unsupported free-form explanation does not alter the structured correctness score.

TDR-SA. A prediction is correct when the parsed clinical status of the queried anatomical surface exactly matches the corresponding structured surface label.

TDR-FS. Clinical Accuracy is computed over all evaluated surfaces of the target tooth:

$$\mathrm{Acc}_{\mathrm{FS}} = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \mathbb{I}\left(\hat{y}_{t,s} = y_{t,s}\right), \tag{4}$$

where $\mathcal{S}_t$ denotes the set of evaluated anatomical surfaces of tooth t, $\hat{y}_{t,s}$ is the predicted status for surface s, and $y_{t,s}$ is the corresponding structured reference label.

TDR-SV. The response is mapped to a binary sound/non-sound decision and compared with the structured tooth-level target.

SCD-TF. Each response is evaluated by binary correctness against the deterministically constructed true/false target.

SCD-MC. Each response is evaluated by exact match with the deterministically defined correct option or class.

K.2 Overall and Macro-average Scores

For a reported task group g containing K_g question types, the group-level Overall score is defined as

$$\mathrm{Overall}_g = \frac{1}{K_g} \sum_{k=1}^{K_g} \mathrm{Acc}_{g,k}. \tag{5}$$

CariesBench reports four group-level Overall scores corresponding to ACS, TDR, SCD-TF, and SCD-MC. The final Avg. reported in the main benchmark table is the macro-average over all 19 individual question types:

$$\mathrm{Avg.} = \frac{1}{19} \sum_{k=1}^{19} \mathrm{Acc}_k. \tag{6}$$

This formulation assigns equal weight to each question type and prevents task groups containing more VQA instances from dominating the final benchmark score solely because of dataset size.

L Additional Experimental Details

L.1 Baseline Evaluation

All benchmarked baseline MLLMs are evaluated under zero-shot prompting. Deterministic decoding is used where supported, with temperature $= 0.0$, top-p $= 1.0$, and repetition penalty $= 1.0$. Commercial models are evaluated through their official APIs using equivalent decoding parameters when the corresponding settings are available. Open-weight and medical/dental-specialized models are evaluated locally using PyTorch 2.4.1, Transformers 5.4.0, and CUDA 11.8. Local inference uses FP16 precision on a server equipped with eight NVIDIA A40 GPUs with 46 GB memory per GPU. All models receive the same task-specific instructions and corresponding rendered inputs for a given benchmark task.

L.2 IntraoralGPT Supervised Fine-Tuning

IntraoralGPT is obtained by supervised fine-tuning Qwen3-VL-8B-Instruct on the rebalanced CariesBench training partition using LoRA implemented with LLaMA-Factory (Zheng et al., 2024). The principal training configuration is summarized in Table 13. Unless otherwise specified, the remaining training hyperparameters follow the default settings of the corresponding LLaMA-Factory configuration. No preference-optimization or reinforcement-learning stage is applied after supervised fine-tuning.

The supervised training data use the training-only class-composition adjustment described in Appendix D.3. Validation and testing are always performed on the untouched patient-disjoint partitions retaining the naturally occurring clinical-status distribution.

Table 13: Supervised fine-tuning configuration of IntraoralGPT.

Setting	**Value**
Base model	Qwen3-VL-8B-Instruct
Training framework	LLaMA-Factory
Adaptation	LoRA
LoRA rank r	16
LoRA scaling α	32
LoRA dropout	0.05
Learning rate	2×10^{-5}
Optimizer	AdamW
LR scheduler	Cosine
Warm-up ratio	0.05
Per-device batch size	1
Gradient accumulation steps	14
Number of GPUs	7
Effective global batch size	98
Training epochs	2
Numerical precision	BF16
GPU	NVIDIA A40
Optimization objective	Supervised fine-tuning