# OralGPT-Omni: A Versatile Dental Multimodal Large Language Model

Jing Hao[1*♠] Yuci Liang[2*] Lizhuo Lin[1*] Yuxuan Fan[3] Wenkai Zhou[1] Kaixin Guo[1]
Zanting Ye[4] Yanpeng Sun[5] Xinyu Zhang[6] Yanqi Yang[1] Qiankun Li[7]
Hao Tang[8] James Kit-Hon Tsoi[1] Linlin Shen[9†] Kuo Feng Hung[1†]
[1]Faculty of Dentistry, The University of Hong Kong
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]The Hong Kong University of Science and Technology (GZ)
[4]School of Biomedical Engineering, Southern Medical University
[5]Singapore University of Technology and Design [6]University of Auckland
[7]University of Science and Technology of China
[8]School of Computer Science, Peking University
[9]College of Artificial Intelligence, Shenzhen University

## Abstract

*Multimodal Large Language Models (MLLMs) have exhibited immense potential across numerous medical specialties; yet, dentistry remains underexplored, in part due to limited domain-specific data, scarce dental expert annotations, insufficient modality-specific modeling, and challenges in reliability. In this paper, we present OralGPT-Omni, the first dental-specialized MLLM designed for comprehensive and trustworthy analysis across diverse dental imaging modalities and clinical tasks. To explicitly capture dentists' diagnostic reasoning, we construct TRACE-CoT, a clinically grounded chain-of-thought dataset that mirrors dental radiologists' decision-making processes. This reasoning supervision, combined with our proposed four-stage training paradigm, substantially strengthens the model's capacity for dental image understanding and analysis. In parallel, we introduce MMOral-Uni, the first unified multimodal benchmark for dental image analysis. It comprises 2,809 open-ended question–answer pairs spanning five modalities and five tasks, offering a comprehensive evaluation suite to date for MLLMs in digital dentistry. OralGPT-Omni achieves an overall score of 51.84 on the MMOral-Uni benchmark and 45.31 on the MMOral-OPG benchmark, dramatically outperforming the scores of GPT-5. Our work promotes intelligent dentistry and paves the way for future advances in dental image analysis. All code, benchmark, and models will be made publicly available.*

* Equal Contribution.
♠ Project Leader.
† Corresponding Authors: hungkfg@hku.hk, llshen@szu.edu.cn

## 1. Introduction

Multimodal Large Language Models (MLLMs) have exhibited remarkable capabilities in open-world visual understanding and reasoning in natural domains [1, 4, 5, 13, 21, 29, 35, 44, 51], and they also demonstrate immense potential across medical specialties, including dermatology [64], ophthalmology [24], chest radiology [19], pathology [26], and pediatrics [54]. However, recent studies conclude that existing MLLMs still face notable limitations in dentistry, including insufficient consistency, completeness, and clarity of generated outputs, as well as the observation of hallucinated responses, which prevents their application in real clinical applications [30, 31]. Due to the lack of deep modeling of dentistry-specific expertise and modality-specific features, existing general-purpose and medical-purpose MLLMs struggle to provide reliable support for complex dental imaging analysis and highly specialized clinical requirements. Although some preliminary studies [34, 60] have explored the application of MLLMs to dental imaging, their performance remains far from satisfactory. These limitations largely stem from the substantial heterogeneity across dental imaging modalities, the intrinsic complexity of clinical diagnostic workflows, and the lack of transparency and reliability in model responses. Progress in this domain is further constrained by the scarcity and inconsistent quality of dental imaging datasets, which result from strict privacy concerns, limited data sharing, and the high cost of expert annotation [11, 12, 14, 16]. At the same time, explainable decision-making is indispensable in the medical field, as clinicians and patients must understand not only the final diagnostic conclusion but also the reason-

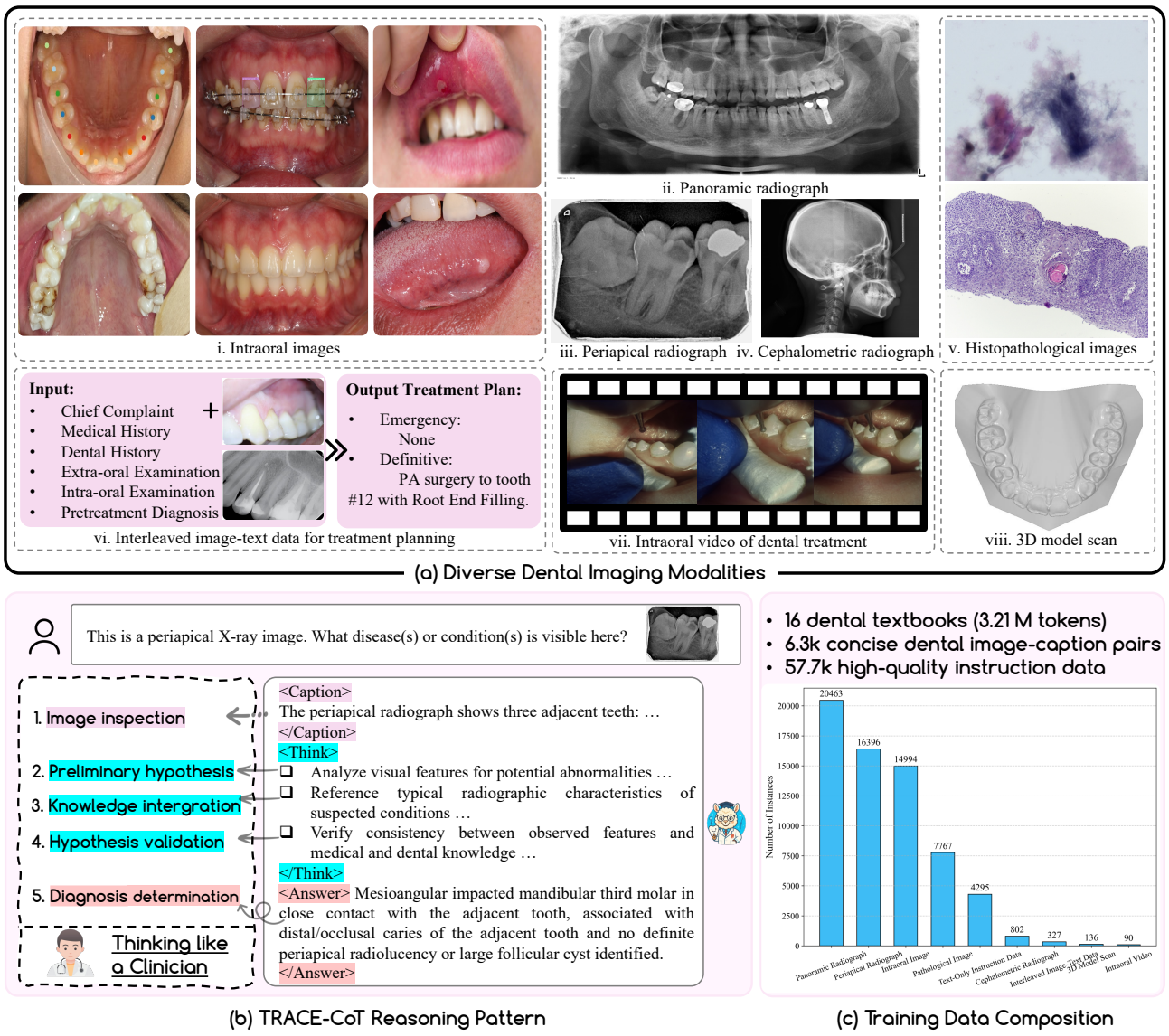
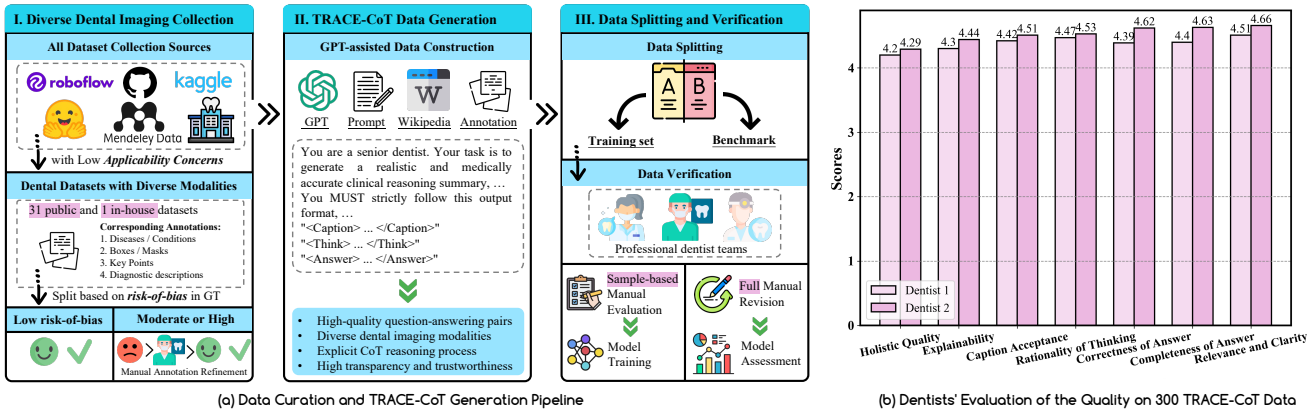



Figure 1. Overview of diverse dental-specialized corpus. (a) Eight types of widely used dental imaging modalities. (b) Introduction of our proposed TRACE-CoT reasoning pattern that enhances the reliability of MLLM's response. (c) The composition of the training corpus for OralGPT-Omni. The bar chart shows the distribution of various dental modalities.

ing process that leads to it. Yet, this critical aspect has been largely overlooked in existing research.

To bridge these gaps in digital and intelligent dentistry, we aim to develop a dental-specialized MLLM capable of robust and comprehensive multimodal imaging analysis. To this end, we introduce OralGPT-Omni, a versatile MLLM that facilitates comprehensive analysis across eight dental imaging modalities and five dental-related tasks, as illustrated in Figure 1(a). Importantly, for the diagnosis of abnormalities, OralGPT-Omni does more than produce final answers. It reveals its diagnostic process by producing explicit chain-of-thought (CoT) rationales that mirror real clinical workflows. This capability substantially improves the model's transparency and interpretability. To build the OralGPT-Omni, we systematically prepare the recipe around three key aspects: diverse dental imaging curation, TRACE-CoT reasoning data construction, and a four-stage training strategy. First, we curate a comprehensive multimodal dental imaging dataset by aggregating data from 31 public sources and one dental hospital. Next, we design the TRACE-CoT (**T**ransparent **R**adiologic **A**nalysis with **C**linical **E**vidence), a reasoning pattern that mirrors radiologists' diagnostic decision-making, which is demonstrated in Figure 1(b). Each CoT instance explicitly outlines the intermediate reasoning processes involved in the diagnosis of abnormalities, encompassing detailed descriptions of visual appearances, the rationale behind diagnostic hypotheses, the integration and verification of domain-




(a) Data Curation and TRACE-CoT Generation Pipeline

(b) Dentists' Evaluation of the Quality on 300 TRACE-CoT Data

Figure 2. (a) The dental imaging data curation and TRACE-CoT data generation pipeline. It involves curating diverse imaging modalities from public datasets and dental hospitals. TRACE-CoT data is then generated using GPT, Wikipedia, and various annotations. Finally, the data is split into a training set and a benchmark, with professional dentists assessing the training samples and a thorough manual correction conducted on the benchmark. (b) Results from two dentists evaluating the quality of 300 TRACE-CoT data from the training set.

specific knowledge, and a final evidence-informed diagnosis. Lastly, leveraging our meticulously curated large-scale, multimodal, dental-specific dataset, we adopt a four-stage training paradigm that enhances OralGPT-Omni's integration of visual comprehension, explicit reasoning, and the ability to follow complex instructions.

Currently, there is only one public dental-related benchmark, MMOral-OPG, for panoramic X-ray analysis [15]. The absence of benchmarks that include various dental imaging modalities hampers the systematic evaluation of MLLMs in dentistry. To fill this gap and guide future optimization, we present MMOral-Uni, the first unified benchmark dedicated to dental multimodal imaging analysis. It comprises 2,809 open-ended question–answer pairs that emulate realistic user interactions. All pairs are validated and refined by two experienced dentists to ensure clinical correctness. The benchmark spans five modalities and five tasks, enabling a comprehensive and rigorous evaluation of MLLMs in the dental domain.

We assess the performance of OralGPT-Omni on the MMOral-OPG [15] and MMOral-Uni benchmarks to thoroughly evaluate its capability for dental applications. OralGPT-Omni achieves an impressive overall score of 51.84 on the Moral-Omni benchmark and 45.31 on the MMOral-OPG benchmark, significantly surpassing the scores of GPT-5. It also markedly outperforms existing medical MLLMs, emphasizing its unique contributions to the dental field. A clinical validity assessment conducted by a radiologist with over ten years of experience on three leading MLLMs indicates that our OralGPT-Omni demonstrates outstanding accuracy and potential clinical utility. Overall, our contributions are summarized as follows:

- We propose OralGPT-Omni, the first dental-specialized MLLM for comprehensive dental imaging analysis across diverse imaging modalities and tasks.
- We present MMOral-Uni, the first unified benchmark for dental multimodal imaging analysis, covering five imaging modalities and five tasks, with 2,809 open-ended VQA pairs that together offer a comprehensive evaluation suite for existing MLLMs.
- Extensive experiments demonstrate that OralGPT-Omni delivers superior performance across diverse dental imaging modalities, achieving the superior overall scores on both the MMOral-OPG and Moral-Omni benchmarks, highlighting its potential in digital dentistry.

## 2. OralGPT-Omni

### 2.1. Training Corpus Construction

We meticulously curate dental-specific multimodal datasets, including text-only corpora and imaging datasets, sourced from trusted public data platforms and dental hospitals. For the text corpora, we compile materials from undergraduate dental textbooks, dialogue data from dental community forums, and QA pairs on oral diseases and diagnoses derived from clinical guidelines. For imaging datasets, we prioritize image quality and label reliability, incorporating 31 public datasets with low applicability concerns, as defined by an authoritative systematic review [16]. The dataset having "low" applicability concern indicates it reports both ethical approval as well as licensing requirements for its reuse. Additionally, we include one expert-annotated dataset sourced from a dental hospital in Hong Kong. A comprehensive list of all collected datasets is provided in the *Appendix*. To further ensure the accuracy and reliability of annotations in the public datasets, we conduct manual post-processing guided by the risk-of-bias rating assigned in the systematic review [16]. This rating quantifies the reliability of the ground truth and is widely used in diagnostic imaging research [49]. For datasets with

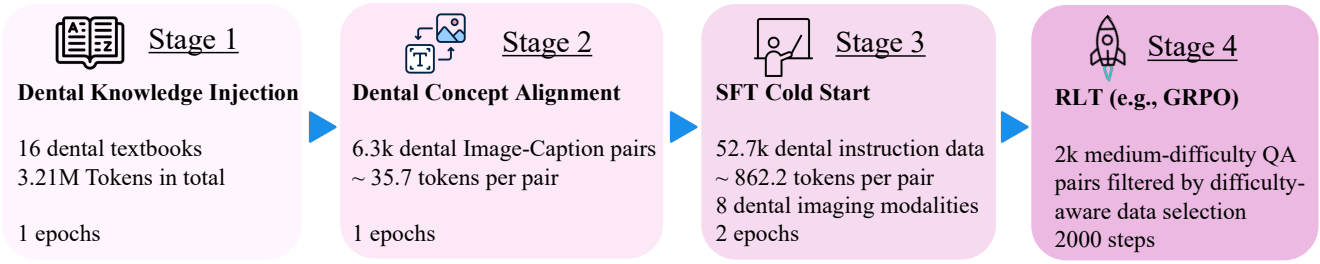



Figure 3. There are four stages for training OralGPT-Omni, and only the first stage is training in the single modality.

moderate or high risk-of-bias ratings, professional dentists manually inspect and refine the annotations, correcting any identified errors. By contrast, datasets with a low risk-of-bias rating do not undergo further manual validation due to high confidence in their annotation quality.

The curated training dataset comprises approximately 3.21 million text tokens, 59,658 images, and 90 videos. It supports five primary tasks: abnormality diagnosis, CVM stage prediction, recommended treatment planning, video understanding, and tooth localization and counting. The dataset encompasses eight dental imaging modalities, as illustrated in Figure 1(c). It includes intraoral photographs and videos, panoramic, periapical, and cephalometric radiographs, pathology images, 3D model scans, and interleaved image-text data. In addition, it provides broad demographic coverage, with samples from at least ten countries. It also offers rich structured annotations, such as abnormality labels, bounding boxes, segmentation masks, keypoints, and diagnostic descriptions. Building on this high-quality, heterogeneous resource, we convert the original annotations into chain-of-thought reasoning data, as detailed in Sec. 2.2.

## 2.2. TRACE-CoT Data Generation

Unlike black-box predictions [2], CoT explicitly reveals the intermediate reasoning steps, thereby enhancing the transparency and trustworthiness of MLLMs. However, two mainstream approaches for generating reasoning chains, CoT prompting strategies [20, 28, 36, 45] and human-verified annotation [8], exhibit notable limitations. The former relies heavily on the reasoning capability of the base model, while the latter is difficult to scale up due to intensive expert involvement.

To overcome these challenges, we propose TRACE-CoT (**T**ransparent **R**adiologic **A**nalysis with **C**linical **E**vidence), a reasoning pattern that mirrors the diagnostic decision-making process of radiologists. This process is clinically coherent and closely aligns with routine radiologic practice, involving image inspection, hypothesis generation, reference to medical expertise, and verification to reach a final diagnosis. The reasoning chains proceed as follows:
**(1) Image inspection:** conduct a thorough examination of the image, carefully describing salient structures, visual appearances, and notable patterns.
**(2) Hypothesis generation:** propose plausible abnormalities based on the observed features.
**(3) Medical expertise reference:** refer to authoritative clinical guidelines and well-established knowledge regarding the suspected abnormalities and their characteristic imaging signatures.
**(4) Feature-based verification:** compare the observed image features against knowledge-based standards to identify and resolve inconsistencies.
**(5) Evidence-informed conclusion:** aggregate the accumulated evidence and finalize the diagnostic findings.

We construct the TRACE-CoT data for abnormality diagnosis across three widely used clinical dental imaging modalities: intraoral images, periapical radiographs, and pathology images. The pipeline can be found in Figure 2(a). We first prompt GPT-5-mini to generate detailed descriptions of visual appearances and patterns, then we treat the sparse annotations for each image as initial diagnostic hypotheses. Guided by these hypotheses, we retrieve characteristic imaging patterns, expressed in natural language, from authoritative clinical guidelines and Wikipedia. Next, we instruct GPT-5-mini to organize these elements into complete five-step reasoning chains, and we ultimately generate 36,777 TRACE-CoT reasoning chains. Because abnormality taxonomies differ across datasets, we design dataset-specific prompts tailored to each taxonomy to ensure the quality of the chains; the full prompts are provided in the *Appendix*. To validate the quality of the generated TRACE-CoT data, we invited two dentists to evaluate 300 instances from seven perspectives, and the results are shown in Figure 2(b), demonstrating its high quality and reliability. Clinically grounded reasoning chains make MLLMs more transparent, interpretable, and reliable. They also mitigate the black-box nature of MLLMs, bolster clinician trust, and facilitate adoption in high-stakes medical settings.

## 2.3. Model & Training Strategy

OralGPT-Omni is built upon the Qwen2.5-VL-7B model [5] due to its superior generalization and instruction-following capability. To further adapt the model's capacity to dental scenarios, we employ a four-stage training strategy that

progressively strengthens its multimodal understanding and reasoning abilities, as shown in Figure 3. In the first stage, we perform dental knowledge injection using the corpus composed of 16 professional dental textbooks. This stage aims to inject and consolidate fundamental dental knowledge into the model, during which only the language model is updated. In the second stage, to guide the initial alignment between dental concepts and visual representations, we employ 6,318 dental image–caption pairs extracted from dental textbooks to optimize only the vision–language projector. In the third stage, we conduct supervised fine-tuning (SFT) on the entire OralGPT-Omni architecture to further enhance its instruction-following, multimodal comprehension, and explicit reasoning capabilities. The fine-tuning process uses 52,725 high-quality dental instruction pairs, comprising 31,777 CoT reasoning pairs and additional pairs without explicit reasoning patterns. The training corpus spans eight imaging modalities as well as text-only dialogue data, comprehensively covering a variety of dental imaging and diagnostic scenarios.

In the final stage, we apply reinforcement learning tuning (RLT) within the Group Relative Policy Optimization (GRPO) [41] framework to further incentivize OralGPT-Omni's reasoning ability. We introduce two components: a difficulty-aware data selection strategy and a TRACE-based reward closely aligned with the TRACE-CoT reasoning pattern. For data selection, we assess the difficulty of each instance relative to the preceding SFT model without the TRACE-CoT pattern and retain only medium-difficulty cases, as shown in Figure 4. This is because extremely easy or extremely hard instances provide limited learning signals [56, 59]. The rationale for selecting a model without the TRACE-CoT pattern is that we anticipate RLT can further enhance the model's performance and generalization by stimulating TRACE-CoT reasoning patterns in these medium-difficulty cases. Concretely, we perform $N$ rollouts per instance, compute the score $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_N\}$, and retain only those that meet two criteria: (1) $0.2 \leq \mathcal{S}_{avg} \leq 0.8$, (2) $Max(\mathcal{S}) - Min(\mathcal{S}) \geq 0.4$. We set $N = 5$ and select 2,000 medium-difficulty cases from 5,000 instruction data. We also introduce a TRACE-based reward $\mathcal{R}_{\text{trace}}$ that comprehensively evaluates reasoning quality with LLM-based judge models. This reward is integrated into RLT to guide models toward producing higher-quality and more reliable reasoning paths across three aspects, including factual knowledge soundness, logical coherence, and answer consistency. Following conventional RTL reward computation, both the answer reward $\mathcal{R}_{\text{answer}}$ and the format reward $\mathcal{R}_{\text{format}}$ are considered. The $\mathcal{R}_{\text{trace}}$ and $\mathcal{R}_{\text{answer}}$ are provided by GPT-5-nano, acting as a reward-judger. The values for these three reward components range from 0 to 1. The overall reward signal unifies these components into a single formulation:

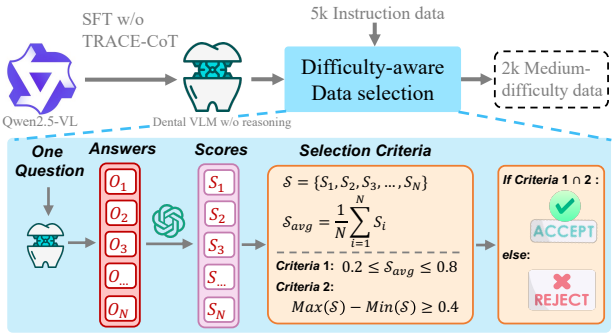



Figure 4. The difficulty-aware data selection strategy for RLT.

$$\mathcal{R}_{\text{total}} = \alpha \cdot \mathcal{R}_{\text{answer}} + \beta \cdot \mathbb{I}_{\mathcal{R}_{\text{answer}}>0} \cdot \mathcal{R}_{\text{trace}} + \gamma \cdot \mathcal{R}_{\text{format}}, \quad (1)$$

where $\alpha + \beta + \gamma = 1$. Note that the $\mathcal{R}_{\text{trace}}$ is ineffective when the $\mathcal{R}_{\text{answer}}$ is completely wrong in order to ensure the consistent correctness of the reasoning and answers. More details of TRACE-based reward and RLT are illustrated in the *Appendix*.

## 3. MMOral-Uni Benchmark

### 3.1. Benchmark Composition

We introduce MMOral-Uni, the first unified benchmark for dental multimodal imaging analysis, spanning five modalities and five tasks. It includes 2,809 open-ended QA pairs to simulate realistic user interactions. The modalities cover intraoral photographs, periapical and cephalometric radiographs, pathological images, intraoral videos, and interleaved image–text inputs. The tasks include abnormality diagnosis, cervical vertebral maturation (CVM) stage prediction, treatment planning, tooth localization and counting, and dental treatment video comprehension. The combination of diverse imaging modalities and task types allows for a thorough evaluation of MLLMs in the dental field. The distribution of modalities and tasks is shown in Fig. 5(a).

To ensure the quality and reliability of the benchmark, we implement strict data selection and extensive manual validation. All images in the MMOral-uni benchmark are sourced from publicly available datasets that had been assessed as low applicability concern [16]. We design the questions to align with specific task types and anticipated user intents. For example, the question is "This is an intraoral photograph of the oral cavity. Identify any oral diseases present." for abnormality diagnosis, and "What is the estimated CVM stage based on this cephalometric radiograph?" for CVM staging. Answers are generated by converting sparse annotations into clear textual descriptions with assistance from GPT-5-mini, after which two experienced dentists validated and refined all QA pairs to further

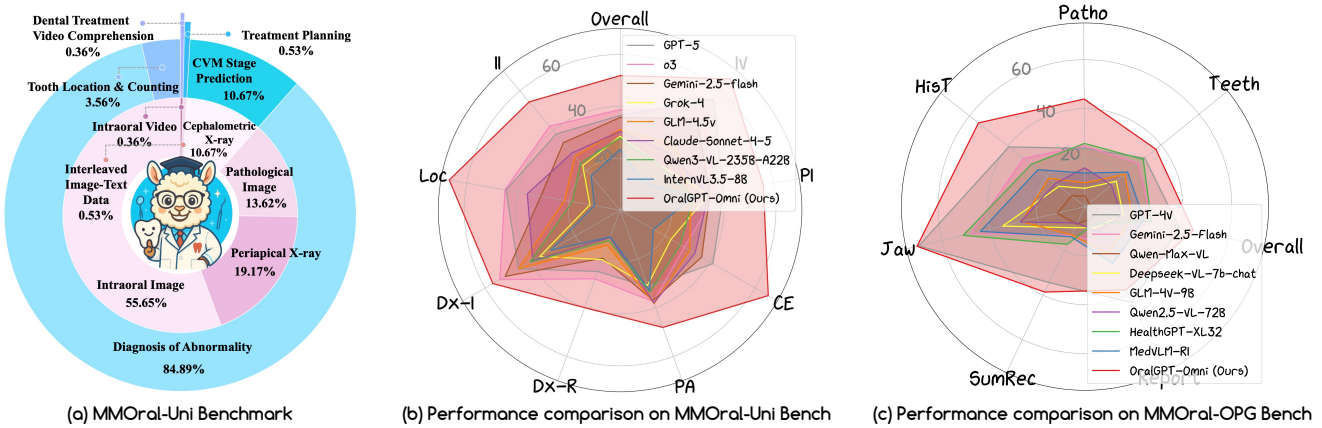



Figure 5. (a) The distribution of the MMOral-Uni benchmark, spanning five dental imaging modalities and covering five tasks. (b) Performance comparison on the MMOral-Uni benchmark. (c) Performance comparison on the MMOral-OPG benchmark.

ensure correctness. More examples can be found in the *Appendix*. The MMOral-Uni benchmark will be released publicly to facilitate comprehensive assessment of current MLLMs and to inform future research in multimodal AI for digital and intelligent dentistry.

### 3.2. Evaluation Metrics

Following well-established benchmarks [15, 57, 58], we meticulously design a few-shot prompt and use GPT-5-mini to conduct the open-ended evaluation. The prompt incorporates five in-context examples with free-form answers, covering fully correct, partially correct, and incorrect cases. GPT-5-mini assigns a score ranging from 0 to 1 for each sample based on the input question, ground truth, and model prediction. We report the evaluation scores for each modality as well as the overall performance. Comprehensive ablation experiments have validated the feasibility and stability of using LLMs as judges; the full few-shot prompt and further details are provided in the *Appendix*. We have integrated the MMOral-Uni evaluation into the VLMEvalKit framework [9] to facilitate subsequent capability assessments of newly developed MLLMs.

## 4. Experiments

### 4.1. Experimental Setup

**Benchmarked MLLMs.** We systematically evaluate OralGPT-Omni on the MMOral-Uni benchmark to provide a comprehensive assessment of its performance on multimodal dental imaging analysis. We conduct evaluations of 27 representative MLLMs: 7 proprietary systems accessed via API [3, 6, 10, 37, 38, 46, 50], 12 general-purpose models [1, 4, 5, 21, 29, 33, 35, 47, 48, 51, 53, 55], and 8 medical MLLMs [7, 20, 22, 27, 39, 43, 52]. We also evaluate OralGPT-Omni on the MMOral-OPG benchmark [15] to assess its capability for panoramic X-ray analysis.

**Implementation Details.** We utilize the LLaMA-Factory framework [63] to train OralGPT-Omni for the first three stages: DKI, DCA, and SFT, followed by the RLT stage using the ms-swift framework [62]. OralGPT-Omni is initialized with the Qwen2.5-VL-7B-Instruct pre-trained model [5] due to its exceptional instruction-following capabilities. The training of OralGPT-Omni is conducted on 2 $\times$ NVIDIA A100 80G GPUs over approximately 90 hours. For the RLT stage, the GRPO policy generates six candidate rationales per sample, with a sampling temperature of $\tau = 0.8$. For further details on the hyperparameters employed in model training, please refer to the *Appendix*.

### 4.2. Comparisons on MMOral-Uni Benchmark

The performance of various MLLMs on the MMOral-Uni benchmark is summarized in Table 1 and Fig. 5(b). In general, our OralGPT-Omni achieves the highest overall score of 51.84, significantly surpassing GPT-5's score of 15.42, attesting to its strong capabilities in multimodal imaging analysis within the dental domain. However, it is noteworthy that OralGPT-Omni underperforms compared to proprietary MLLMs in the recommended treatment planning task. We attribute this discrepancy to differences in task objectives. The treatment planning task requires the formulation of subsequent strategies based on a final diagnosis. This necessitates a deeper understanding of medical expertise related to surgical procedures and postoperative recovery, which is an area where proprietary MLLMs particularly excel. In contrast, OralGPT-Omni is primarily specialized in dental imaging analysis and abnormality diagnosis, with treatment planning data comprising only 0.006% of the whole training dataset. This limitation contributes to its lower performance in this specific task. Nevertheless, effective treatment planning depends critically on

Table 1. Results on the MMOral-Uni for various existing LVLMs across five dental imaging modalities. The abbreviations are defined as follows: II - Intraoral Image, PA - Periapical Radiograph, CE - Cephalometric Radiograph, PI - Pathological Image, TP - Treatment Planning, IV - Intraoral Video. The best-performing model in each category is highlighted **in-bold**, while the second-best is underlined.

| Model | II | | | PA | CE | PI | TP | IV | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | Loc | Dx-I | Dx-R | | | | | | |
| *Proprietary LVLMs* | | | | | | | | | |
| GPT-5 [37] | 44.60 | 45.24 | 25.16 | 31.43 | 41.27 | 40.52 | **80.67** | 56.00 | 36.42 |
| o3 [38] | 45.00 | 53.49 | 28.19 | 37.48 | 28.60 | 37.52 | 75.33 | 50.00 | 38.70 |
| Grok-4 [50] | 11.80 | 35.71 | 20.27 | 30.89 | 22.60 | 34.73 | 72.67 | 23.00 | 28.47 |
| Claude-Sonnet-4-5-20250929 [3] | 36.30 | 38.85 | 11.28 | 38.03 | 33.33 | 34.52 | 79.33 | 29.00 | 30.32 |
| Doubao-1.5-vision-pro-32k [6] | 26.90 | 43.33 | 16.97 | 40.48 | 18.20 | 28.49 | 68.00 | 45.00 | 30.67 |
| Gemini-2.5-flash [46] | 31.00 | 51.10 | 19.70 | 37.01 | 36.30 | 35.09 | 73.33 | 46.00 | 35.72 |
| GLM-4.5v [47] | 21.70 | 45.10 | 14.37 | 38.37 | 31.00 | 27.60 | 70.00 | 33.00 | 31.05 |
| *Open-Source LVLMs* | | | | | | | | | |
| Qwen2.5-VL-7B [5] | 11.20 | 25.40 | 23.84 | 19.72 | 13.70 | 30.73 | 36.67 | 12.00 | 22.88 |
| Qwen3-VL-8B-Instruct [53] | 14.30 | 27.17 | 16.95 | 33.54 | 4.97 | 30.68 | 54.00 | 18.00 | 23.45 |
| Qwen3-VL-235B-A22B [53] | 18.70 | 40.00 | 12.53 | 33.62 | 26.70 | 28.75 | 56.00 | 23.00 | 27.83 |
| GLM-4.1V-9B-Thinking [47] | 15.50 | 37.78 | 18.14 | 37.57 | 13.53 | 27.13 | 62.00 | 43.00 | 27.86 |
| InternVL3.5-8B [48] | 10.10 | 30.85 | 11.02 | 32.95 | 14.80 | 28.67 | 64.67 | 14.00 | 23.39 |
| LLaVA-v1.6-Mistral-7B [29] | 4.80 | 20.57 | 14.03 | 11.61 | 2.50 | 12.06 | 34.67 | 18.00 | 13.54 |
| LLaVA-OneVision [21] | 7.50 | 23.13 | 17.59 | 14.34 | 0.00 | 11.49 | 40.67 | 5.00 | 15.39 |
| MiMo-VL-7B [51] | 25.40 | 36.28 | 22.51 | 34.62 | 18.70 | 31.91 | 62.00 | 43.00 | 29.62 |
| Phi-4-multimodal-instruct [1] | 3.70 | 18.32 | 7.86 | 18.31 | 1.67 | 9.63 | 34.67 | 7.00 | 12.12 |
| Mistral-Small-3.1-24B [35] | 16.30 | 28.14 | 18.77 | 29.15 | 11.50 | 27.02 | 62.00 | 28.00 | 23.69 |
| R-4B [55] | 3.80 | 32.29 | 19.16 | 28.96 | 16.60 | 27.89 | 50.67 | 19.00 | 24.94 |
| Ovis2.5-9B [33] | 15.40 | 37.47 | 19.76 | 36.70 | 18.13 | 34.33 | 66.00 | 61.00 | 29.60 |
| *Medical Specific LVLMs* | | | | | | | | | |
| LLaVA-Med [22] | 5.40 | 3.99 | 5.81 | 25.31 | 1.50 | 16.08 | 28.67 | 14.00 | 10.16 |
| HuatuoGPT-Vision-7B [7] | 11.90 | 30.19 | 23.10 | 25.79 | 19.57 | 28.02 | 41.33 | 19.00 | 25.41 |
| Lingshu-7B [52] | 12.00 | 30.58 | 25.77 | 27.48 | 20.50 | 30.94 | 48.00 | 20.00 | 27.08 |
| MedVLM-R1 [39] | 10.10 | 23.88 | 14.30 | 16.25 | 0.00 | 21.28 | 24.00 | 26.00 | 16.50 |
| Med-R1-Diagnosis [20] | 6.30 | 22.93 | 8.91 | 20.13 | 2.00 | 19.01 | 16.00 | 22.00 | 15.29 |
| Chiron-o1-8B [43] | 4.60 | 20.94 | 17.37 | 34.95 | 2.03 | 31.20 | 42.00 | 23.00 | 21.61 |
| MedGemma-27B [40] | 11.90 | 20.44 | 17.09 | 21.65 | 22.17 | 32.69 | 68.67 | 11.00 | 21.56 |
| HealthGPT [27] | 18.00 | 20.43 | 8.84 | 19.44 | 7.50 | 31.02 | 54.67 | 14.00 | 17.32 |
| OralGPT-Omni (Ours) | **66.80** | **56.60** | **39.99** | **48.11** | **65.90** | **56.01** | 47.33 | **65.00** | **51.84** |

accurate diagnostic findings, underscoring the importance of OralGPT-Omni. Furthermore, we observe that existing medical MLLMs exhibit no substantial advantage over general MLLMs in the field of dentistry, further emphasizing the unique contribution of OralGPT-Omni in multimodal dental image analysis and reasoning.

### 4.3. Comparisons on MMOral-OPG Benchmark

We evaluate OralGPT-Omni on the MMOral-OPG benchmark [15], specifically focusing on open-ended VQA for panoramic X-ray analysis. We assess performance across six clinically grounded dimensions, as shown in Table 2 and Fig. 5(c). OralGPT-Omni outperforms existing medical MLLMs, achieving an overall score of 45.31. Its scores on the the "Teeth", "Patho", "His", "Jaw", and "Summ" dimensions notably exceed those of GPT-4V, underscoring its strength in panoramic X-ray interpretation. By contrast, its report generation performance lags behind GPT-4V. We hypothesize that this gap arises from the intrinsic complexity of panoramic radiographs, which contain dense anatomical structures and multi-dimensional conditions that OralGPT-Omni may not fully capture when generating comprehensive reports.

Table 2. Performance on the MMOral-OPG benchmark for various LVLMs on open-ended VQA tasks. The best-performing model is highlighted **in-bold**, while the second-best is underlined.

| **Model** | **Open-ended VQA** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Teeth** | **Patho** | **His** | **Jaw** | **Summ** | **Report** | **Overall** |
| *General-purpose LVLMs* | | | | | | | |
| GPT-5 [37] | **39.77** | 29.32 | 44.05 | **78.56** | **40.12** | 28.20 | 42.42 |
| GPT-4V [18] | 31.46 | 23.79 | 39.51 | 69.81 | 34.29 | **43.70** | 39.38 |
| Gemini-2.5-Flash [46] | 28.04 | 24.77 | 31.90 | 47.81 | 12.98 | 16.70 | 27.84 |
| Qwen-Max-VL [4] | 2.10 | 4.47 | 7.06 | 11.62 | 7.98 | 5.50 | 5.29 |
| Deepseek-VL-7b-chat [32] | 16.48 | 7.50 | 13.44 | 34.56 | 9.52 | 9.60 | 15.95 |
| GLM-4V-9B [10] | 20.94 | 9.70 | 18.77 | 26.62 | 12.74 | 21.30 | 19.74 |
| Qwen2.5-VL-72B [5] | 13.90 | 15.83 | 15.40 | 27.12 | 7.38 | 11.50 | 15.38 |
| *Medical Specific LVLMs* | | | | | | | |
| LLaVA-Med [22] | 0.91 | 1.52 | 0.00 | 0.00 | 0.00 | 24.50 | 4.76 |
| HealthGPT-XL32 [27] | 30.64 | 25.83 | 27.98 | 51.12 | 17.02 | 8.00 | 27.80 |
| MedVLM-R1 [39] | 22.42 | 13.71 | 24.42 | 43.88 | 13.57 | 25.80 | 24.70 |
| MedDr [17] | 22.99 | 32.58 | 29.57 | 52.44 | 20.95 | 8.70 | 26.20 |
| OralGPT-Omni (Ours) | 37.26 | **43.94** | **55.34** | 70.50 | 38.57 | 37.90 | **45.31** |

### 4.4. Ablation and In-Depth Study

**Effect of Four-Stage Training Strategy.** We conduct ablation studies on MMOral-Uni to evaluate the effectiveness of

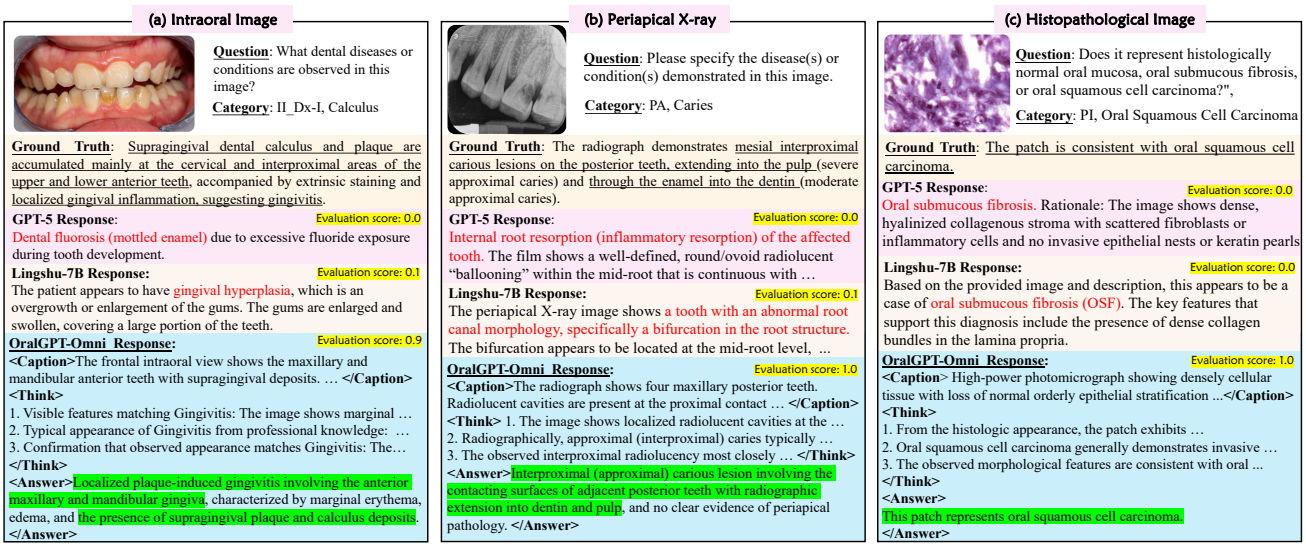



Figure 6. Three modalities of case studies on the MMOral-Uni benchmark are presented, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

Table 3. Ablation experiments of the four-stage training strategy and the TRACE-CoT reasoning pattern.

| No. | Model | II | | | PA | CE | PI | TP | IV | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Loc | Dx-I | Dx-R | | | | | | |
| 1 | Baseline (Qwen2.5-VL-7B) [5] | 11.20 | 25.40 | 23.84 | 19.72 | 13.70 | 30.73 | 36.67 | 12.00 | 22.88 |
| 2 | +1st stage (DKI) | 12.50 | 24.43 | 22.96 | 21.32 | 14.37 | 36.63 | 38.67 | 18.00 | 23.66 |
| 3 | +2nd stage (DCA) | 11.90 | 25.44 | 23.22 | 24.38 | 17.70 | 29.19 | 50.00 | 31.00 | 24.00 |
| 4 | +3rd stage (SFT) | 64.40 | 52.61 | 38.02 | 43.97 | 63.30 | 53.89 | 35.33 | 33.00 | 48.67 |
| 5 | +4th stage (RLT) | 66.80 | 56.60 | 39.99 | 48.11 | 65.90 | 56.01 | 47.33 | 65.00 | 51.84 |
| 6 | SFT wo/ TRACT-CoT | 60.60 | 44.26 | 33.73 | 48.00 | 58.93 | 44.67 | 36.00 | 31.00 | 44.31 |
| 7 | SFT w/ TRACT-CoT | 64.40 | 52.61 | 38.02 | 43.97 | 63.30 | 53.89 | 35.33 | 33.00 | 48.67 |
| - | $\Delta\uparrow$ | +3.80 | +8.35 | +4.29 | -4.03 | +4.37 | +9.22 | -0.67 | +2.00 | +4.36 |

the four-stage training strategy, with results presented in Table 3. Overall, this strategy progressively improves performance in dental image analysis. The first two stages, DKI and DCT, achieve a modest improvement over the baseline model (Qwen2.5-VL-7B). By incorporating a domain-specific dental knowledge corpus and concise descriptions of diverse dental imaging modalities, the overall score on MMOral-Uni increases from 22.88 to 24.00. The SFT stage, with carefully curated high-quality instruction data, yields a substantial improvement, raising the overall score from 24.00 to 48.67. SFT markedly strengthens instruction following and provides a solid foundation for explicit, transparent reasoning in dental image analysis, thereby preparing the model for the subsequent RLT stage. RLT further improves the overall score by 3.17 points and incentivizes stronger reasoning capabilities.

**Effect of TRACE-CoT Reasoning Data.** Our proposed TRACE-CoT pattern mirrors the diagnostic decision-making process of radiologists and is designed to explicitly improve the transparency and trustworthiness of MLLMs. To assess whether explicit reasoning chains can enhance diagnostic accuracy, we perform an ablation study during the SFT phase. We compare the model trained with TRACE-CoT data against the model trained without it, and the results are shown in Table 3 (rows 6-7). Incorporating TRACE-CoT data during SFT increases the overall score by 4.36 points relative to training on answer-only data that excludes reasoning chains. Notably, in our instruction data, TRACE-CoT annotations are included only for the II-Dx-I, II-Dx-R, PA, and PI modalities, and the most significant improvements appear in II-Dx-I, II-Dx-R, and PI. These findings provide concrete evidence that TRACE-CoT data could strengthen diagnostic performance.

**Case Study & Clinical Feedback.** We conduct an in-depth case study of the OralGPT-Omni model, focusing on its performance across various dental imaging modalities, as illustrated in Fig. 6. In comparison to GPT-5 and the medical-specific MLLM Lingshu-7B [52], our OralGPT-Omni demonstrates superior abnormality diagnosis capabilities in three modalities: intraoral images, periapical X-rays, and histopathological images. Additionally, it provided explicit TRACE-CoT reasoning patterns, which enhance its transparency and reliability. To further evaluate the clinical validity of OralGPT-Omni, we invite a radiologist with

over ten years of experience to perform comprehensive clinical assessments of the responses from three top-performing models, the details of which can be found in the *Appendix*.

## 5. Conclusion

In this paper, we introduce OralGPT-Omni, the first dental-specialized MLLM for comprehensive analysis across diverse dental imaging modalities and tasks. The success of our OralGPT-Omni relies on high-quality, large-scale, dental-specific multimodal datasets; the explicit TRACE-CoT reasoning pattern and data construction pipeline; and the four-stage training strategy. In addition, the proposed MMOral-Uni is the first unified benchmark dedicated to multimodal dental imaging analysis, which offers a comprehensive evaluation suite for digital dentistry.

Experiments show that OralGPT-Omni outperforms current state-of-the-art MLLMs on the MMOral-Uni and MMOral-OPG benchmarks by a large margin. It also provides explicit reasoning chains that explain how the final diagnosis is reached. This work paves the way for future advancements in digital and intelligent dentistry by enabling natural language interaction, multimodal dental imaging analysis, and enhanced explainability in next-generation dental artificial intelligence.

## References


[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 1, 6, 7

[2] Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*, 2025. 4

[3] Anthropic. Claude-sonnet-4-5. *https://www.anthropic.com/news/claude-sonnet-4-5*, 2025. 6, 7

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 6, 7

[5] 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. 1, 4, 6, 7, 8

[6] ByteDance. Doubao-1.5-vision-pro. *https://www.volcengine.com/product/doubao*, 2025. 6, 7

[7] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 6, 7, 1, 4, 5

[8] Chao Ding, Mouxiao Bian, Pengcheng Chen, Hongliang Zhang, Tianbin Li, Lihao Liu, Jiayuan Chen, Zhuoran Li, Yabei Zhong, Yongqi Liu, et al. Building a human-verified clinical reasoning dataset via a human llm hybrid pipeline for trustworthy medical ai. *arXiv preprint arXiv:2505.06912*, 2025. 4

[9] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 6

[10] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 6, 7

[11] Jing Hao, Moyun Liu, Lei He, Lei Yao, James Kit Hon Tsoi, and Kuo Feng Hung. Semit-sam: Building a visual foundation model for tooth instance segmentation on panoramic radiographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–121. Springer, 2024. 1

[12] Jing Hao, Lun M Wong, Zhiyi Shan, Qi Yong H Ai, Xieqi Shi, James Kit Hon Tsoi, and Kuo Feng Hung. A semi-supervised transformer-based deep learning framework for automated tooth segmentation and identification on panoramic radiographs. *Diagnostics*, 14(17):1948, 2024. 1

[13] Jing Hao, Yuxiang Zhao, Song Chen, Yanpeng Sun, Qiang Chen, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Fullanno: A data engine for enhancing image comprehension of mllms. *arXiv preprint arXiv:2409.13540*, 2024. 1

[14] Jing Hao, Yonghui Zhu, Lei He, Moyun Liu, James Kit Hon Tsoi, and Kuo Feng Hung. T-mamba: a unified framework with long-range dependency in dual-domain for 2d & 3d tooth segmentation. *arXiv preprint arXiv:2404.01065*, 2024. 1

[15] 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. *arXiv preprint arXiv:2509.09254*, 2025. 3, 6, 7, 1

[16] Jing Hao, Andrew Nalley, Andy Wai Kan Yeung, Ray Tanaka, Qi Yong H Ai, Walter Yu Hang Lam, Zhiyi Shan, Yiu Yan Leung, Abeer AlHadidi, Michael M Bornstein, et al. Characteristics, licensing, and ethical considerations

of openly accessible oral-maxillofacial imaging datasets: a systematic review. *npj Digital Medicine*, 8(1):412, 2025. 1, 3, 5

[17] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR*, 2024. 7, 1

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[19] Shehroz S Khan, Petar Przulj, Ahmed Ashraf, and Ali Abedi. Chestgpt: Integrating large language models and vision transformers for disease detection and localization in chest x-rays. *arXiv preprint arXiv:2507.03739*, 2025. 1

[20] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025. 4, 6, 7, 1

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 6, 7

[22] 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. 6, 7, 1

[23] Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, et al. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *arXiv preprint arXiv:2505.02830*, 2025. 1

[24] Sijing Li, Tianwei Lin, Lingshuai Lin, Wenqiao Zhang, Jiang Liu, Xiaoda Yang, Juncheng Li, Yucheng He, Xiaohui Song, Jun Xiao, et al. Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. *arXiv preprint arXiv:2504.13650*, 2025. 1

[25] Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:2411.14522*, 2024. 1

[26] Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: a multimodal large language model for whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22718–22727, 2025. 1

[27] 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. 6, 7

[28] Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *arXiv preprint arXiv:2509.15279*, 2025. 4

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 6, 7

[30] Zekai Liu, Qi Yong H Ai, Andy Wai Kan Yeung, Ray Tanaka, Andrew Nalley, and Kuo Feng Hung. Performance of a vision-language model in detecting common dental conditions on panoramic radiographs using different tooth numbering systems. *Diagnostics*, 15(18):2315, 2025. 1

[31] Zekai Liu, Andrew Nalley, Jing Hao, Qi Yong H Ai, Andy Wai Kan Yeung, Ray Tanaka, and Kuo Feng Hung. The performance of large language models in dentomaxillofacial radiology: a systematic review. *Dentomaxillofacial Radiology*, page twaf060, 2025. 1

[32] 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. 7

[33] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 6, 7

[34] Zijie Meng, Jin Hao, Xiwei Dai, Yang Feng, Jiaxiang Liu, Bin Feng, Huikai Wu, Xiaotang Gai, Hengchuan Zhu, Tianxiang Hu, et al. Dentvlm: A multimodal vision-language model for comprehensive dental diagnosis and enhanced clinical practice. *arXiv preprint arXiv:2509.23344*, 2025. 1

[35] mistralai. Mistral-small-3.1-24b-instruct-2503. *https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503*, 2025. 1, 6, 7

[36] Chee Ng, Liliang Sun, and Shaoqing Tang. X-ray-cot: Interpretable chest x-ray diagnosis with vision-language models via chain-of-thought reasoning. *arXiv preprint arXiv:2508.12455*, 2025. 4, 1

[37] OpenAI. Gpt-5. *https://openai.com/zh-Hans-CN/index/introducing-gpt-5*, 2025. 6, 7, 4, 5, 10

[38] OpenAI. o3. *https://openai.com/zh-Hans-CN/index/introducing-o3-and-o4-mini*, 2025. 6, 7

[39] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025. 6, 7, 1

[40] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 7

[41] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5, 2

[42] Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025. 1

[43] Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 6, 7

[44] Yanpeng Sun, Jing Hao, Ke Zhu, Jiang-Jiang Liu, Yuxiang Zhao, Xiaofan Li, Gang Zhang, Zechao Li, and Jingdong Wang. Descriptive caption enhancement with visual specialists for multimodal perception. *arXiv preprint arXiv:2412.14233*, 2024. 1

[45] Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025. 4

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6, 7

[47] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. *https://arxiv.org/abs/2507.01006*, 2025. 6, 7

[48] 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. 6, 7

[49] Penny F Whiting, Anne WS Rutjes, Marie E Westwood, Susan Mallett, Jonathan J Deeks, Johannes B Reitsma, Mariska MG Leeflang, Jonathan AC Sterne, Patrick MM Bossuyt, and QUADAS-2 Group*. Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 155(8):529–536, 2011. 3

[50] xAI. grok4. *https://x.ai/grok*, 2025. 6, 7

[51] LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025. 1, 6, 7

[52] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 6, 7, 8, 4, 5, 10

[53] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6, 7, 4, 5

[54] Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems*, 37:138632–138662, 2024. 1

[55] Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. *arXiv preprint arXiv:2508.21113*, 2025. 6, 7

[56] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 5

[57] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[58] Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. 6

[59] Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *arXiv preprint arXiv:2508.12680*, 2025. 5

[60] Jia Zhang, Bodong Du, Yitong Miao, Dongwei Sun, and Xiangyong Cao. Oralgpt: A two-stage vision-language model for oral mucosal disease diagnosis and description. *arXiv preprint arXiv:2510.13911*, 2025. 1

[61] Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*, 2025. 1

[62] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 29733–29735, 2025. 6, 2

[63] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. 6, 2

[64] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024. 1

# OralGPT-Omni: A Versatile Dental Multimodal Large Language Model

## Supplementary Material

## Contents



## 1. Related Works

**Medical Large Vision-Language Models.** Medical MLLMs have shown great potential to serve as valuable assistants for clinicians, researchers, and trainees by providing an interactive natural language interface for analyzing medical images across diverse modalities. Recent efforts [7, 17, 22, 25] aim to develop medical general-purpose MLLMs that are capable of simultaneously analyzing and responding to images and instructions from multiple medical disciplines. Meanwhile, more researchers are dedicated to building discipline-specialized medical MLLMs, such as dermatology [64], ophthalmology [24], chest [19], pathology [26], and pediatrics [54]. These specialized models demonstrate superior imaging diagnostic performance within their respective fields compared to those medical general-purpose MLLMs. However, dentistry-specific MLLMs remain largely underexplored. OralGPT [15], trained on a large-scale instruction dataset, represents the first vision-language model specialized for panoramic X-ray analysis. Jia Z. *et al*. [60] finetuned the Qwen2.5-VL model using less than 2k intraoral images to achieve diagnosis of four types of oral mucosal diseases. DentVLM [34] supports basic oral disease diagnosis on three imaging modalities—panoramic, lateral, and intraoral images—but lacks the capability to provide detailed explanations. In contrast, our OralGPT-Omni offers comprehensive analysis across seven dental imaging-based modalities as well as text-only interactions. It also generates explicit chain-of-thoughts, significantly enhancing the model's transparency and trustworthiness.

**Medical Chain-of-Thought Reasoning.** Existing medical MLLMs have made significant progress through supervised fine-tuning (SFT) on specific instruction datasets. However, these models [17, 22, 25, 34, 60] primarily generate final answers without revealing the underlying reasoning processes, due to the bias towards memorizing task-specific shortcuts rather than learning generalizable reasoning in the SFT stage. Prior studies, including MedVLM-R1 [39], Med-R1 [20], and MED-RLVR [61], have attempted to enhance the reasoning abilities of medical MLLMs via reinforcement learning. Nevertheless, the quality of the generated reasoning chains heavily relies on the base policy model. To address this limitation, GMAI-VL-R1 [42] designs a reasoning data synthesis approach that employs GPT-4o for step-by-step reasoning generation through rejection sampling. However, it still suffers from the hallucination sourced from GPT-4o. To alleviate hallucination issues, Q. Li *et al*. [23] constructed the AOR-Instruction reasoning dataset by incorporating explainable region-level visual content based on anatomical regions and ontologies. Similarly, X-Ray-CoT [36] created reasoning data for chest X-rays by integrating general medical knowledge and visual concept descriptions using the CoT prompting strategy. In contrast, we propose the TRACE-CoT reasoning pattern that closely mirrors the diagnostic decision-making process of radiologists, which is presented in Section 2.2.

## 2. Details on Training Data Construction

A comprehensive list of all data sources utilized during the training phase is presented in Table 4. The training corpus consists of approximately 3.21 million text tokens, 59,658 images, and 90 videos. It includes various types of dental-specific data, such as intraoral images and videos, panoramic radiographs, periapical radiographs, histopathological images, cephalometric radiographs, 3D model scans, text-only data, and interleaved image-text data. Additionally, we provide detailed descriptions of abnormalities and clinically relevant tasks associated with each modality, highlighting the extensive coverage of our OralGPT-Omni in dental imaging analysis. Regarding TRACE-CoT data generation, we construct an automatic pipeline utilizing sparse label categories, the visual appearance of abnormalities, and dental knowledge from academic publications and Wikipedia, instructing GPT-5-mini to generate 36,777 TRACE-CoT reasoning chains. Since abnormality taxonomies differ across datasets, we craft dataset-specific prompts tailored to each taxonomy to ensure the quality of the reasoning chains. The details of these prompt designs are presented in Figures 9, 10, 11, 12, 13. The dental expertise used in the prompts is summarized in Table 7. Besides, we also provide some data used in the training

Table 4. Comprehensive list of dental-specific data sources used in the training phase, including corresponding abnormalities and tasks.

| No | Modality | Abnormality / Subject | Source |
|---|---|---|---|
| 1.1 | Text-only data | 16 Undergraduate textbooks in dentistry | / |
| 1.2 | Text-only data | Dialogue data from dental forums | URL |
| 1.3 | Text-only data | Open-domain oral disease QA dataset | URL |
| 2.1 | Intraoral image | Teeth location and counting | URL<br>URL |
| 2.2 | Intraoral image | Image-level analysis:<br>Calculus, Caries, Gingivitis, Ulcer<br>Tooth discoloration, Defective dentition<br>Cancer, Normality, Orthodontics | URL<br>URL<br>URL<br>URL<br>URL |
| 2.3 | Intraoral image | Region-level analysis:<br>Abrasion, Filling, Crown, Caries, Gingivitis, Deep Overjet,<br>Fenestration and Dehiscence, Tooth Torsion, Tooth emergence,<br>Invisible orthodontic attachment, Fixed orthodontic device,<br>Case fixed orthodontic appliances, Tooth misalignment,<br>Mandibular retrusion, Orthodontic Brace, Dental plaque | URL<br>URL<br>URL<br>URL<br>URL<br>One in-house dataset |
| 3.0 | Periapical radiograph | Impacted tooth, Pulpitis, Caries, Periodontitis, Crown,<br>Apical periodontitis, Bone loss, Root canal treatment,<br>Restoration, Mixed dentition | URL<br>URL<br>URL<br>URL<br>URL<br>URL<br>URL<br>URL<br>URL |
| 4.0 | Cephalometric radiograph | 29 cephalometric landmarks detection,<br>Cervical vertebral maturation (CVM) stage prediction | URL<br>URL |
| 5.0 | Histopathological image | Leukoplakia without dysplasia, Leukoplakia with dysplasia,<br>Oral squamous cell carcinoma, Oral submucous fibrosis,<br>Healthy epithelial nucleus, Abnormal epithelial nucleus,<br>Blood cell nucleus, Reactive cell nucleus, Dividing nucleus | URL<br>URL<br>URL<br>URL |
| 6.0 | Intraoral video | Intraoral videos of real dental treatments | URL |
| 7.0 | Interleaved image-text data | 6 Clinical Cases Guidelines in Dentistry for treatment planning | / |
| 8.0 | 3D model scan | Image description | URL |
| 9.0 | Panoramic radiograph | Open-Ended VQA for panoramic X-ray analysis | URL |

stage for an intuitive understanding, which can be found in Figure 14 - Figure 22.

## 3. Implementation Details of Model Training

### 3.1. Hyperparameters of four-stage training paradigm

We employ the four-stage training strategy that progressively strengthens the multimodal understanding and reasoning ability of the OralGPT-Omni model. The hyperparameters used in each training stage are summarized in Table 5. We utilize the LLaMA-Factory framework [63] to train OralGPT-Omni for the first three stages: DKI, DCA, and SFT, followed by the RLT stage using the ms-swift framework [62]. The "GPU Hour" for each stage is estimated based on the NVIDIA A100 80G GPU. We also provide the word clouds of training datasets used in DKI, DCA, and SFT stages, which are depicted in Figure 8.

### 3.2. Reinforcement Learning Tuning

To further incentivize the reasoning capability of OralGPT-Omni, we implement reinforcement learning tuning within the Group Relative Policy Optimization (GRPO) [41] algorithm on the OralGPT-Omni model. GRPO calculates a group-relative advantage from multiple samples of the same prompt instead of learning a separate value function, making it particularly effective for training tasks where correctness can be verified. During the training, we optimize the OralGPT-Omni with the GRPO loss $\mathcal{J}_{\text{GRPO}}(\theta)$:

Table 5. The hyperparameters used in the four-stage training strategy. The "GPU Hours" are estimated based on the NVIDIA A100 80G GPU.

| **Hyperparameters** | **DKI** | **DCA** | **SFT** | **RLT** |
|---|---|---|---|---|
| Batch Size | 1 | 1 | 1 | 4 |
| Gradient Accumulation Step | 4 | 4 | 4 | 3 |
| Trainable Component | Language model | Projector | Visual encoder, Projector, Language model | Language model |
| LoRA Rank | 8 | 8 | 8 | 8 |
| # Generations | / | / | / | 6 |
| Learning Rate | 1.0e-5 | 1.0e-5 | 1.0e-4 | 1.0e-6 |
| Warmup Ratio | 0.1 | 0.1 | 0.1 | 0.05 |
| bf16 | True | True | True | True |
| # Epoch / Step | 1 epoch | 1 epoch | 1 epoch | 2000 steps |
| $\sim$ GPU Hours | 10 | 20 | 50 | 100 |

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim\mathcal{P}(Q),\\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}} \Bigg[ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Bigg( \frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\Big(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon,\, 1+\epsilon\Big)\hat{A}_{i,t}\Bigg) - \beta\, D_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\Bigg], \tag{1}$$

where $\pi_\theta$ and $\pi_{\theta_{\text{old}}}$ are the current and old policy. $\pi_{\text{ref}}$ is the reference model, which, in this case, is the preceding SFT model with the TRACE-COT pattern. The $q$ and $o$ are the questions and outputs sampled from our dataset and the old policy $\pi_{\theta_{\text{old}}}$, respectively. The $\epsilon$ and $\beta$ are hyperparameters for stabilizing training. $\hat{A}_{i,t}$ is the advantage of the relative rewards of the outputs in each group. For each response, we use an LLM-based judge model to evaluate its correctness reward and our proposed TRACE-based reasoning reward. The detailed reward computation will be discussed in Sec. 3.3. The reward $r_i$ is defined within the range [0, 1]. We use the normalized reward as the advantage:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(r)}{\text{std}(r)}. \tag{2}$$

Besides, we use an unbiased estimator to estimate the KL divergence $D_{\text{KL}}$:

$$D_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}] = \frac{\pi_{\text{ref}}(o_{i,t}\mid q,o_{i,<t})}{\pi_\theta(o_{i,t}\mid q,o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}\mid q,o_{i,<t})}{\pi_\theta(o_{i,t}\mid q,o_{i,<t})} - 1. \tag{3}$$

### 3.3. Reward Computation for GRPO

In conventional RL-based tuning paradigms, supervision is applied only to the final <answer>, while the intermediate reasoning process <think> remains unregulated. This often leads to cases where the final output appears correct, yet the underlying reasoning is flawed or logically unsound. Such discrepancies pose substantial safety risks in medical AI, where unreliable or hallucinated reasoning may still produce a superficially plausible answer but lead to harmful or unsafe clinical interpretations. To address this issue, we introduce the TRACE-based reward $\mathcal{R}_{\text{trace}}$ , a medically aligned reward framework that uses an LLM-based judge model to comprehensively evaluate the quality of the reasoning trace. This framework provides fine-grained supervision over the model's diagnostic reasoning process, ensuring that the generated clinical rationale is logically coherent, medically accurate, and consistent with the final prediction, rather than supervising only the end result. To operationalize this framework, the evaluation is decomposed into three key dimensions:

**Factual Knowledge Soundness.** This dimension measures the correctness and clinical reliability of domain knowledge referenced in the <Think> section, including oral medicine expertise and pathological criteria. Incorrect medical statements receive $d_1 = 0$; partially correct or incomplete knowledge receives $d_1 = 1$; and fully accurate, evidence-based knowledge receives $d_1 = 2$.

**Logical Coherence.** This aspect assesses whether the diagnostic reasoning path presented in the <Think> section is logically consistent with the visual appearance descriptions in the <Caption>. The judge model evaluates whether the reasoning contains logical gaps, unsupported inferences, or clinically implausible conclusions. Fully coherent and clinically sound reasoning receives a score of $d_2 = 1$, whereas reasoning that is logically inconsistent or contradictory receives a score of $d_2 = 0$.

**Answer Consistency.** This reward evaluates whether the gold-standard answer can be directly justified by the reasoning trace and the feature-based verification. The judge model ensures that the conclusion is supported by stated ev-

idence without contradictions, hallucinated features, or reliance on unstated assumptions. Unsupported conclusions receive $d_3 = 0$; partially supported conclusions receive $d_3 = 1$; and fully justified, clinically consistent conclusions receive $d_3 = 2$.

The TRACE-based reward $\mathcal{R}_{\text{trace}}$ is then computed as the average of these three dimensions to get the normalized reward. Following traditional RTL reward computation, both the answer reward $\mathcal{R}_{\text{answer}}$ and the format reward $\mathcal{R}_{\text{format}}$ are considered. The $\mathcal{R}_{\text{trace}}$ and $\mathcal{R}_{\text{answer}}$ are provided by GPT-5-nano, acting as a reward-judger. The prompts for TRACE-based reward and answer reward are provided in Figure 23 and Figure 24, respectively.

The values for these three reward components range from 0 to 1. The overall reward signal unifies these components into a single formulation:

$$\mathcal{R}_{\text{total}} = \alpha \cdot \mathcal{R}_{\text{answer}} + \beta \cdot \mathbb{I}_{\mathcal{R}_{\text{answer}}>0} \cdot \mathcal{R}_{\text{trace}} + \gamma \cdot \mathcal{R}_{\text{format}}, \quad (4)$$

where $\alpha + \beta + \gamma = 1$. Note that the $\mathcal{R}_{\text{trace}}$ is ineffective when the $\mathcal{R}_{\text{answer}}$ is completely wrong in order to ensure the consistent correctness of the reasoning and answers.

## 4. MMOral-Uni Benchmark

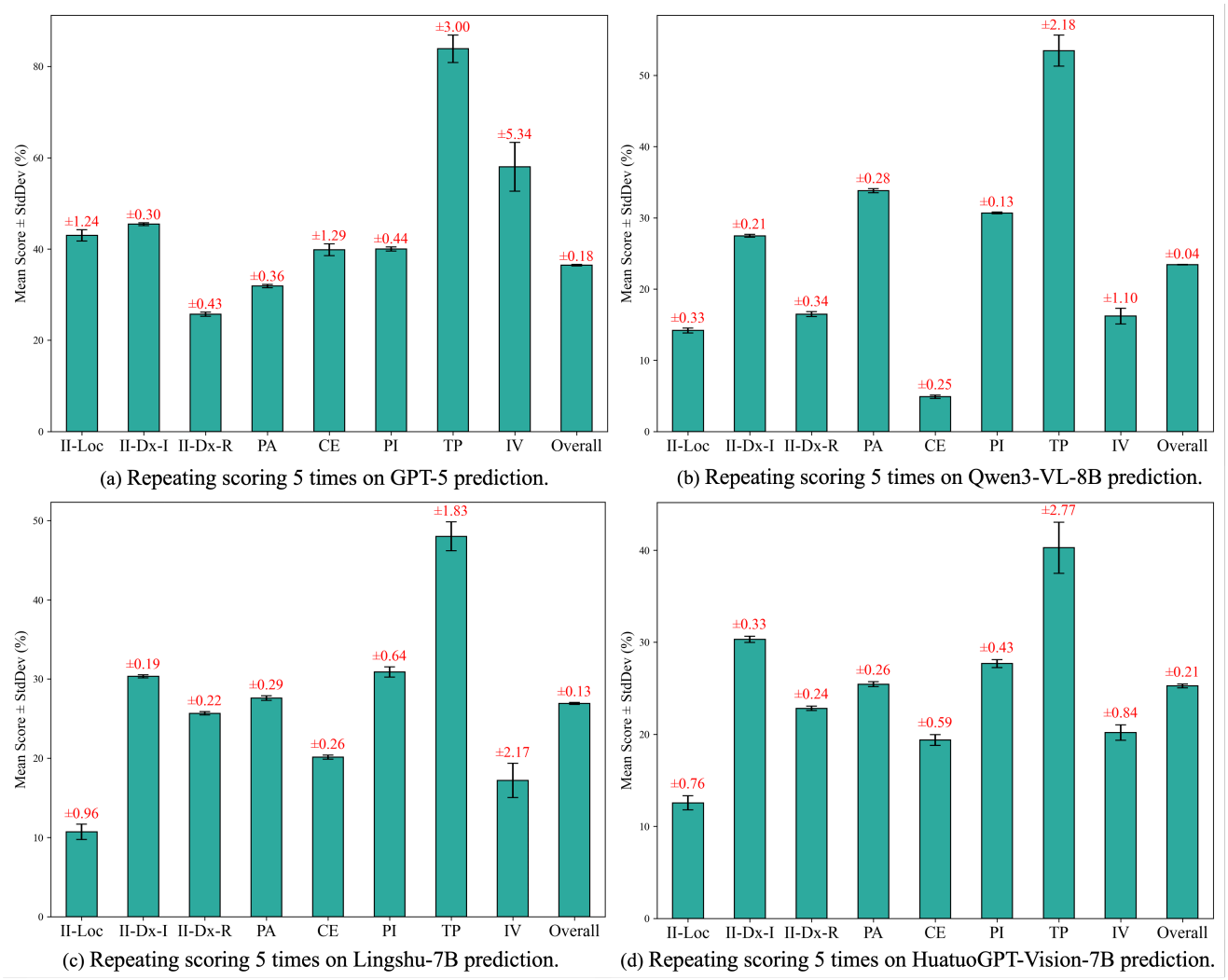
Our MMOral-Uni benchmark is the first unified benchmark for dental multimodal imaging analysis, spanning five modalities and five tasks. It includes 2,809 open-ended QA pairs to simulate realistic user interactions. The modalities cover intraoral photographs, periapical and cephalometric radiographs, pathological images, intraoral videos, and interleaved image–text inputs. The tasks include abnormality diagnosis, cervical vertebral maturation (CVM) stage prediction, treatment planning, tooth localization and counting, and dental treatment video comprehension. The specific quantities for each task and modality are presented in Table 10. Regarding the abnormality diagnosis, the MMOral-Uni benchmark includes 40 categories of abnormalities, which are summarized in Table 8. We also provide some examples in the MMOral-Uni benchmark for an intuitive understanding of this benchmark, which can be found in Figure 25 - Figure 29.

## 5. LLM as the judges for MMOral-Omni: A Feasibility Analysis

**Effectiveness.** To verify the effectiveness of LLM-based evaluation for the MMOral-Omni benchmark, we invite two professional dentists to objectively score the outputs of different LVLMs. We calculate the absolute difference between the evaluators' scores and the human-annotated scores. Specifically, the few-shot prompts designed for LLM-based evaluation are presented to the dentists to guide the evaluation criteria. The two dentists then independently scored the predictions of GPT-5 and Lingshu-7B on 300 cases from the MMOral-Omni benchmark based on these criteria. The absolute differences between human scores and evaluators' scores are shown in Table 9, represented as $\Delta$.

Overall, the absolute differences of the "Overall" metric given by dentists fluctuate by approximately 2 points in comparison to the LLM-based evaluations for the predictions of both LVLMs (GPT-5 and Lingshu-7B). This observation indicates that human scoring preferences generally align with the trends observed in LLM-based evaluations. However, it also suggests that there are subjective differences in the dentists' interpretations of the evaluation criteria outlined in the few-shot prompts. For each subcategory, Dentist A shows smaller differences in scores compared to the LLM-based evaluation for questions in the "II-Loc", "II-Dx-I", "II-Dx-R", "PA", "CE", and "PI", whereas the differences are larger for the "TP" and "IV" categories. Although Dentist B exhibits slightly larger differences with LLM-based scoring across all subcategories, their "overall" score difference is only 2.43 points. This indicates that LLM-based scoring aligns well with human preferences in reflecting the overall performance of LVLMs on the MMOral-Omni benchmark. At the same time, we speculate that the score fluctuations in each subcategory are strongly associated with the subjective perceptions of human evaluators.

**Stability.** Since using LLMs as judges inevitably introduces randomness, even with the temperature hyperparameter set to 0, we conduct multiple repeated experiments to verify the stability of LLMs as judges. Specifically, we evaluate the prediction results of GPT-5 [37], Qwen3-VL-8B [53], Lingshu-7B [52], and HuatuoGPT-Vision-7B [7] on the MMOral-Omni benchmark using GPT-5-mini [37] with the same prompt five times. The obtained mean, standard deviation, and coefficient of variation (CV) of the metric "overall" are shown in Table 6. For proprietary models, medical-specific models, and general-purpose LVLMs, the standard deviation of the metric "overall" is no more than 0.212 when evaluated 5 times using GPT-5-mini with our designed few-shot prompt. Specifically, for the prediction results of GPT-5, the standard deviation of the scores is 0.179, while for Qwen3-VL-8B, the standard deviation is as low as 0.039. Meanwhile, CV (Coefficient of Variation), as a standardized measure of dispersion of a probability distribution, can be used to assess the stability of scores across multiple experiments. The CV values for the prediction results of these four models, after being scored 5 times, are all around 0.5%, which demonstrates the evaluation stability of using LLMs as evaluators. The detailed results across each specific category are demonstrated in Figure 7.




(a) Repeating scoring 5 times on GPT-5 prediction.

(b) Repeating scoring 5 times on Qwen3-VL-8B prediction.

(c) Repeating scoring 5 times on Lingshu-7B prediction.

(d) Repeating scoring 5 times on HuatuoGPT-Vision-7B prediction.

Figure 7. The means and standard deviations of each category on 5 repeated evaluations across four LVLMs' predictions.

## 6. Case Study and Clinical Validity of OralGPT-Omni

We provide more case studies to demonstrate the superiority of our OralGPT-Omni, as illustrated in Figure 30, Figure 31, and Figure 32. From a clinical perspective, the ability to accurately identify common dental diseases and conditions is a fundamental prerequisite for diagnosis and treatment planning. To explore the clinical validity of the OralGPT-Omni, we invite a professoriate faculty member in oral-maxillofacial radiology with over ten years of experience in diagnostic imaging studies to conduct in-depth clinical assessments on the outputs of the top-performing models (GPT-5, Lingshu-7B, and our OralGPT-Omni). The clinical validation of the model outputs on four representative modalities is demonstrated in Figure 33, Figure 34, Figure 35, and Figure 36. The cases and clinical feedback effectively demonstrate the superior performance and potential clinical value of our OralGPT-Omni.

Table 6. Stability verification of using LLMs as judges: Standard deviation and coefficient of variation (CV) are reported across four LVLMs from five repeated evaluations.

| Model | Mean | StdDev | CV % |
|---|---|---|---|
| GPT-5 [37] | 36.446 | 0.179 | 0.490 |
| Qwen3-VL-8B [53] | 23.436 | 0.039 | 0.164 |
| Lingshu-7B [52] | 26.910 | 0.127 | 0.472 |
| HuatuoGPT-Vision-7B [7] | 25.258 | 0.212 | 0.838 |

Table 7. The category, visual appearance, and relevant dental knowledge utilized in prompts for the TRACE-CoT data construction pipeline across multiple datasets.

| Dataset | Category | Appearence | Dental Knowledge |
|---|---|---|---|
| AlphaDent URL | Abrasion | Tooth with mechanical wear of hard tissues | Abrasion typically appears as flat, smooth, and well-defined areas of hard tissue loss, often wedge-shaped at the cervical margins, with a shiny polished surface. |
| | Filling | Filling | A filling is visually recognized as a distinct restorative material embedded in the tooth, usually with a regular geometric outline and a surface texture or color different from natural enamel or dentin. |
| | Crown | Installed crown | A crown presents as a full-coverage artificial cap over the tooth, with smooth and uniform surfaces, clear margins near the gingiva, and contours that may look slightly different from natural tooth anatomy depending on the material. |
| | Caries 1 class | Caries in fissures and blind pits of teeth (occlusal surfaces of molars and premolars, buccal surfaces of molars, lingual surfaces of upper incisors) | Caries are seen as irregular areas of demineralization or cavitation, with rough surfaces and discoloration ranging from chalky white to brown or black, often with undermined enamel edges. |
| | Caries 2 class | Caries of the contact surfaces of molars and premolars. | |
| | Caries 3 class | Caries of the contact surfaces of incisors and canines without damage to the cutting edge. | |
| | Caries 4 class | Caries of the contact surfaces of incisors and canines with damage to the cutting edge. | |
| | Caries 5 class | Cervical caries of the vestibular and lingual surfaces. | |
| | Caries 6 class | Caries of the cutting edges of the front teeth and the cusps of the chewing teeth. | |
| Dental Caries Detection URL | Primary tooth decay | Occlusal surface of a primary molar shows a small, dark brown spot in the central pit, with surrounding enamel appearing slightly chalky. Minor enamel surface irregularities are visible along the grooves. | The affected primary tooth shows a localized brownish or dark spot on the occlusal or smooth surfaces, with enamel appearing softened or chalky. The child may exhibit mild sensitivity to sweet or cold foods, occasional discomfort when chewing, and sometimes visible plaque accumulation near the lesion. Gingival tissue surrounding the tooth may appear slightly inflamed. |
| | Permanent tooth decay | Occlusal surface of a permanent molar shows a large, dark brown cavitated lesion extending into the central pit and fissures. Enamel margins appear undermined, and surface texture is roughened. | Permanent teeth with decay display larger cavitated lesions, often dark brown or black in color, with enamel margins appearing undermined. Patients may complain of prolonged sensitivity to cold, sweet, or acidic foods, intermittent pain while chewing, and visible plaque or calculus accumulation at the affected site. |

| Dataset | Category | Appearence | Dental Knowledge |
|---|---|---|---|
| FDTooth URL | Anterior teeth with fenestration and dehiscence | Localized gingival recession with partially exposed root surfaces, a thin labial mucosa allowing clear visualization of root contours, cementoenamel junction, and subtle convexities, and small fenestration windows or linear dehiscence defects present in the alveolar bone beneath. | Anterior teeth with fenestration and dehiscence exhibit localized gingival recession with exposed root surfaces. The thin labial mucosa allows visualization of the underlying root contour, and alveolar bone may be partially or completely deficient in the affected regions. Clinically, patients are often asymptomatic or may report mild sensitivity at the exposed sites. |
| GINGIVITIS URL | Gingivitis | The region shows marked redness, edema, and/or marginal gingival hypertrophy of the unit or spontaneous bleeding, papillary, congestion, or ulceration. | Severe inflammation of gingivitis presents with intense redness and deep erythema of the gingival margins and interdental papillae, pronounced swelling, a shiny and moist surface, spontaneous bleeding in some areas, hypertrophy and deformation of papillae, and occasional ulcerations. |
| Dental Plaque URL | Dental plaque | A soft, adherent film on the tooth surface, typically pale yellow to light brown, often along the gingival margin or in pits and fissures. | Dental plaque appears as a soft, adherent film on tooth surfaces, typically pale yellow to light brown, often along the gingival margin or in pits and fissures. Clinically, patients are usually asymptomatic or may report mild sensitivity to sweet foods, with slight gingival redness and minimal inflammation. |
| OMNI_COCO URL | Tooth torsion | A condition where a tooth is rotated or twisted around its longitudinal axis, resulting in an abnormal orientation within the dental arch. | Tooth torsion results in a tooth noticeably rotated around its longitudinal axis, causing misalignment within the dental arch; this rotation creates uneven spacing with adjacent teeth, sometimes overlapping or crowding neighboring teeth, which can lead to difficulty in chewing, localized food impaction, accumulation of plaque, and mild inflammation of the surrounding gingival tissue. |
| | Deep overjet | A dental malocclusion where the upper front teeth (maxillary incisors) horizontally overlap the lower front teeth (mandibular incisors) to an excessive degree. | Deep overjet occurs when the upper front teeth are prominently protruded over the lower front teeth, producing a pronounced horizontal overlap; this may impair lip closure, cause difficulty in biting or chewing, increase the risk of trauma to the anterior teeth or lips, and create aesthetic concerns due to facial profile changes. |
| | Invisible orthodontic attachment | These attachments, part of clear aligner or aesthetic orthodontic treatments, are tooth-colored or transparent devices bonded to teeth to aid in controlled tooth movement without the visibility of traditional braces. | Invisible orthodontic attachment involves small, tooth-colored or transparent attachments bonded to the teeth to guide clear aligner therapy, blending almost seamlessly with the natural tooth surface; patients may experience minor gum irritation, localized plaque accumulation, or slight sensitivity during initial treatment stages. |
| | Tooth emergence | The process by which a tooth moves through the alveolar bone and soft tissue to appear in the oral cavity. This is a natural phase of dental development. | Tooth emergence occurs when a tooth is erupting through the gingival tissue, causing visible swelling and redness around the emerging crown; this process may result in mild tenderness, discomfort during chewing, and localized inflammation, sometimes accompanied by increased salivation or irritability. |

| Dataset | Category | Appearence | Dental Knowledge |
|---|---|---|---|
| | Cast fixed orthodontic appliance | Custom-made orthodontic devices, such as fixed space maintainers or lingual braces, which are fabricated using dental impressions (casts) of the patient's teeth and are permanently attached to the teeth to correct or maintain dental alignment. | Cast fixed orthodontic appliance consists of custom-made fixed devices bonded to teeth, precisely matching the dental arch; these appliances may be metallic or tooth-colored, producing initial soreness, localized gum irritation, and requiring meticulous oral hygiene to prevent plaque accumulation. |
| | Tooth misalignment | Improper positioning of teeth within the dental arch, leading to malocclusion. This includes various forms of malpositions such as crowding, spacing, and deviations from the ideal arch form. | Tooth misalignment occurs when teeth are irregularly positioned in the dental arch, showing crowding, spacing, or tilting; this misalignment can interfere with normal bite function, create difficulty in chewing, increase plaque accumulation, and elevate the risk of gingival inflammation or caries. |
| | Mandibular retrusion | A skeletal condition characterized by the posterior positioning of the mandible relative to the maxilla, often contributing to the class II malocclusion. | Mandibular retrusion presents with the lower jaw positioned posteriorly relative to the upper jaw, visible in profile view; this skeletal discrepancy can cause Class II malocclusion, impaired occlusion, difficulty chewing, and altered facial aesthetics with a receding chin. |
| | Orthodontic brace | A comprehensive orthodontic appliance consisting of brackets, wires, and sometimes bands, which is fixed to the teeth. This device applies continuous pressure to move teeth into the desired position. | Orthodontic brace consists of fixed braces with metal or ceramic brackets bonded to teeth with connecting archwires, sometimes including colored elastics; these appliances apply continuous force to correct alignment, causing mild soreness, gum irritation, and requiring careful oral hygiene. |
| | Fixed orthodontic device | Any orthodontic apparatus that is bonded to the teeth and remains in place throughout the treatment period. These devices, which include traditional braces and certain types of retainers, are used to correct dental and skeletal discrepancies by exerting controlled forces over time. | Fixed orthodontic device remains permanently bonded to teeth throughout treatment, including traditional braces or fixed retainers; they may cause mild soreness, gum irritation, difficulty in maintaining oral hygiene, and localized plaque accumulation. |

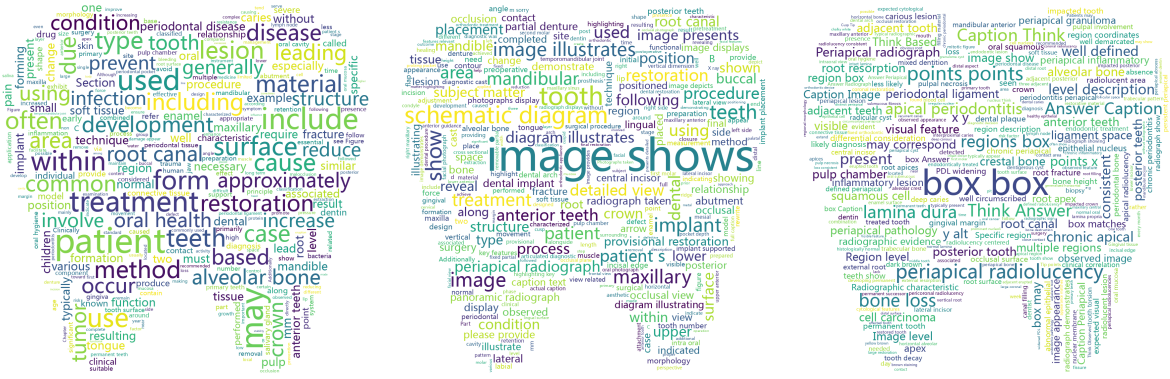



Figure 8. The word cloud maps for training datasets used in the DKI, DCA, and SFT stages, respectively.

**System Prompt:**

You are a senior dentist. Your task is to generate a realistic and medically accurate clinical reasoning summary, explaining the observable findings in the image and how they support the diagnosis. You MUST strictly follow the output format, including all opening and closing tags exactly as written. Never omit or misspell any tag.

**User Prompt: <IMAGE>**

This is an intraoral image of the oral cavity. What dental diseases or conditions are observed in this image?

Output format:

<Caption> Describe the intraoral image in detail, including the tooth and/or arch location, tooth loss condition, any noticeable color changes, and the morphology, texture, and distribution of any deposits. Incorporate visual features strongly associated with {***CATEGORY***}, or with other diseases/conditions you are highly confident are present.
</Caption>

<Think>
1. Explain the visible features in the image that match the known characteristics of {***CATEGORY***}.
2. Recall the typical visual appearance and symptoms of {***CATEGORY***} from professional knowledge.
3. Confirm that the observed image appearance matches the knowledge-based visual representation of {***CATEGORY***}, and rule out other inconsistent conditions.
</Think>

<Answer>Summarize the matched disease(s)/condition(s) from step 3 of the above <Think> pipeline into one concise, fluent, professional medical diagnosis statement that clearly specifies the oral disease(s)/condition(s) present in the image, without including any treatment recommendations or management advice.
</Answer>

Figure 9. The prompt for GPT-5-mini to generate the TRACE-CoT reasoning chains for image-level diagnosis of intraoral images.

Table 8. Categories of analysis for abnormal conditions and diseases.

| **Categories of analysis for abnormal conditions and diseases** | | | |
|---|---|---|---|
| Caries | Gingivitis | Ulcer | Tooth discoloration |
| Defective dentition | Cancer | Normality | Orthodontics |
| Abrasion | Filling | Crown | Caries |
| Fenestration and dehiscence | Tooth torsion | Deep overjet | Invisible orthodontic attachment |
| Tooth emergence | Case fixed orthodontic appliances | Tooth misalignment | Mandibular retrusion |
| Orthodontic brace | Fixed orthodontic device | Dental plaque | Impacted tooth |
| Pulpitis | Periodontitis | Apical periodontitis | Bone loss |
| Root canal treatment | Restoration | Mixed dentition | Leukoplakia without dysplasia |
| Leukoplakia with dysplasia | Oral squamous cell carcinoma | Oral submucous fibrosis | healthy epithelial nucleus |
| Abnormal epithelial nucleus | Blood cell nucleus | Reactive cell nucleus | Dividing nucleus |

**System Prompt:**

You are a senior dentist. Your task is to generate a realistic and medically accurate description, focusing only on the observable findings in the image without providing any diagnostic reasoning. You must strictly follow the required output format, including all opening and closing tags exactly as written. Never omit or misspell any tag.

**User Prompt: <IMAGE>**

Output format:

<Caption>
1. Image-level inspection.
Provide a clear and concise description of the intraoral image, capturing its overall visual appearance, including color changes, cavitations, enamel defects, or localized lesions. Describe only what is visibly present.

2. Region-level inspection.

```
region_descriptions= " "
for box in boxes:
    region_description = f"In region <box>[x1, y1, x2, y2]</box>, the image shows {APPEARANCE}."
    region_descriptions = " ".join([region_descriptions, region_description])
```

</Caption>

<Think>
1. These regions {<box>[x1, y1, x2, y2]</box>} may correspond to the {***CATEGORY***}.
2. {***DENTAL KNOWLEDGE REFERENCE FROM ACADEMIC PUBILICTIONS OR WIKIPEDIA***}.
3. In multiple regions {<box>[x1, y1, x2, y2]</box>}, the observed image appearance matches the expected visual features of {***CATEGORY***}.
</Think>

<Answer>
The {***CATEGORY***} is present in regions {<box>[x1, y1, x2, y2]</box>}.
</Answer>

Figure 10. The prompt for GPT-5-mini to generate the TRACE-CoT reasoning chains for region-level diagnosis of intraoral images.

Table 9. Average absolute differences (Δ) between the evaluation scores of the LLM-based evaluator and the dentist-annotated scores on the MMOral-Omni benchmark.

| Model | Evaluators | II-Loc | II-Dx-I | II-Dx-R | PA | CE | PI | TP | IV | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| GPT-5 [37] | GPT-5-mini | 44.60 | 45.24 | 25.16 | 31.43 | 41.27 | 40.52 | 80.67 | 56.00 | 36.42 |
| | Dentist A | 45.82 | 43.24 | 27.58 | 35.30 | 39.50 | 38.17 | 60.56 | 49.80 | 34.94 |
| | Δ(↓) | +1.22 | -2.00 | +2.42 | +3.87 | -1.77 | -2.35 | -20.11 | -6.20 | -1.48 |
| Lingshu-7B [52] | GPT-5-mini | 12.00 | 30.58 | 25.77 | 27.48 | 20.50 | 30.94 | 48.00 | 20.00 | 27.08 |
| | Dentist B | 18.20 | 32.96 | 28.95 | 23.34 | 26.83 | 28.26 | 42.24 | 14.50 | 29.51 |
| | Δ(↓) | +6.20 | +2.38 | +3.18 | -4.14 | +6.33 | -2.68 | -5.76 | -5.50 | +2.43 |

**System Prompt:**

You are a senior dentist. You will be given a periapical X-ray image and a preliminary assessment indicating possible diseases or conditions in the image. This preliminary assessment is for reference only and may be inaccurate. It may be provided at the image level or the region level. Your task is to generate a realistic and medically accurate clinical reasoning summary, explaining the observable findings in the image and how they support the diagnosis. Note that the preliminary assessment may be incorrect and may also miss additional diseases/conditions present in the image; carefully examine the visual features to reach the final conclusion. Do not attempt to predict whether a tooth is in the upper or lower jaw, and do not attempt to predict its specific type (e.g., second premolar) unless you are 100% certain. You must strictly follow this output format, including all opening and closing tags exactly as written. Never omit or misspell any tag. Do not add any extra sections.

**User Prompt: <IMAGE>**

<Caption> Provide a detailed description of the periapical X-ray. The preliminary assessment for this periapical X-ray is {***CATEGORY***}. </Caption>

<Think>
1. Analyze the visual features listed in <Caption> and determine which diseases or conditions they may correspond to.
2. Recall the typical radiographic characteristics of the suspected diseases or conditions, including common variations and differential considerations.
3. Compare the observed features with the typical characteristics, weigh alternative explanations, and decide on the most likely final diagnosis.

Do not restate the instruction text above; instead, directly write your own reasoning content under each numbered step.
</Think>

<Answer>
Based on the conclusion from Step 3 in <Think>, summarize the diagnosis in a single concise, fluent, and professional radiology report-style sentence. Avoid repeating the reasoning process here.
</Answer>

Figure 11. The prompt for GPT-5-mini to generate the TRACE-CoT reasoning chains for diagnosis of periapical radiographs.

Table 10. Detailed compositions of tasks and modalities in the MMOral-Omni benchmark.

| **Task** | **Modality** | **Number** |
|---|---|---|
| Diagnosis of Abnormality | Intraoral Image | 1462 |
| | Periapical X-ray | 539 |
| | Pathological Image | 383 |
| CVM Stage Prediction | Cephalometric X-ray | 300 |
| Tooth Location & Counting | Intraoral Image | 100 |
| Treatment Planning | Interleaved Image-Text Data | 15 |
| Dental Treatment Video Comprehension | Intraoral Video | 10 |

**System Prompt:**

You are an experienced oral pathologist. Your task is to generate a realistic and medically accurate description, focusing only on the observable findings in the image without providing any diagnostic reasoning or interpretation. You must strictly follow this output format, including all opening and closing tags exactly as written.

**User Prompt: <IMAGE>**

Output format:

<Caption> # *for histopathological images*
Provide a clear and detailed description of this {***CATEGORY***} pathological image, focusing on its morphological and visual characteristics.
</Caption>

<Caption> # *for cytopathology image*
Provide a clear and concise description of the cytopathology image. Focus specifically on the visible cytological features, including cellular morphology, nuclear and cytoplasmic characteristics, chromatin texture, cell arrangement, and background elements such as necrosis, inflammation, or debris. Describe only what is visibly present in the field. Do not mention the absence of atypia, inflammation, or other features.
</Caption>

<Think>
1. Based on the histologic field description above, the observed patch demonstrates features consistent with {***CATEGORY***}.
2. {***DENTAL KNOWLEDGE REFERENCE FROM ACADEMIC PUBILICTIONS OR WIKIPEDIA***}.
3. The observed patch appearance matches the expected histopathological/cytological features of {***CATEGORY***}.
</Think>

<Answer>
The patch is consistent with {***CATEGORY***}.
</Answer>

Figure 12. The prompt for GPT-5-mini to generate the TRACE-CoT reasoning chains for diagnosis of pathological images.

**System Prompt:**

You are a clinical dental expert. You are analyzing a real intra-oral surgical video recorded during conservative dental treatment. The video captures complex clinical conditions inside the oral cavity, where the scene is crowded with multiple dental instruments and artifacts. It may include conditions such as occlusions, frequent appearance variations, tool–tooth interactions, bleeding, water spray, splashing fluids, motion blur, strong light reflections, and occasional camera fouling. The footage could show non-standard tools, intra-oral mirrors, and other interfering objects that partially obstruct the view.

**User Prompt: <VIDEO>**

Generate a detailed and structured description of the intra-oral surgical procedure shown in the video. Focus on:

(1) clinical environment and imaging conditions,
(2) visible tools, tissues, and anatomy,
(3) dynamic interactions during treatment,
(4) procedural sequence,
(5) notable clinical findings and challenges,
(6) possible procedural intents.

Conclude with a final summary that integrates descriptions of the above six aspects. Do not add any instructions, suggestions, or optional tasks.

Figure 13. The prompt for GPT-5-mini to generate the TRACE-CoT reasoning chains for comprehension of intraoral videos.

## Training Data Example

**Question**: What is the count of teeth observed in this intraoral image, what kinds are they, and where exactly are their centers located?

**Ground Truth**: This is an intraoral photography image containing 14 teeth. There are 7 tooth types identified, including 2 2nd Molars located at [<points>(456, 136)</points>, <points>(158, 152)</points>]; 2 1st Premolars located at [<points>(207, 290)</points>, <points>(412, 274)</points>]; 2 Canines located at [<points>(206, 337)</points>, <points>(409, 321)</points>]; 2 2nd Premolars located at [<points>(423, 233)</points>, <points>(194, 247)</points>]; 2 1st Molars located at [<points>(442, 186)</points>, <points>(171, 202)</points>]; 2 Lateral Incisors located at [<points>(249, 343)</points>, <points>(366, 317)</points>]; 2 Central Incisors located at [<points>(289, 366)</points>, <points>(341, 361)</points>].

**Question**: This is an oral photograph. What condition is visible?

**Ground Truth**:
**<Caption>**
The photo shows the inner aspect of the lower lip (lower labial mucosa) of an individual, displaying multiple small, well-demarcated, round to oval shallow mucosal breaks. Each lesion has a shallow yellow-white base with clear erythematous borders; lesions are a few millimetres in diameter and located on non-keratinized mucosa. Adjacent mucosa appears otherwise normal without diffuse swelling.
**</Caption>**
**<Think>** 1. Visible features matching an oral ulcer: there are localized mucosal breaches with round/oval shallow depressions, a central yellow-white fibrinous base and an erythematous margin on the non-keratinized lower labial mucosa; multiple small discrete lesions are present.
2. Typical clinical appearance and symptoms of oral aphthous ulcer: minor aphthous ulcers are painful, small (<1 cm), shallow, with a yellow/white fibrinous center and red halo, commonly occur on non-keratinized mucosa (labial/buccal mucosa, ventral tongue) and are frequent in children and young adults.
3. Correlation and differential: the size, number, shallow depth, fibrinous center and erythematous border on the inner lower lip match minor aphthous (canker) ulcers. Alternative entities are less likely: primary herpetic lesions often begin as vesicles and commonly involve gingiva/palate with systemic symptoms; traumatic ulcers tend to be solitary, irregular and associated with an obvious source; candidiasis appears as removable white plaques rather than discrete ulcers. Thus the visual pattern is most consistent with minor aphthous ulceration.
**</Think>**
**<Answer>**
Minor aphthous ulcers (minor recurrent aphthous stomatitis) involving the lower labial mucosa.
**</Answer>**

Figure 14. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

## Training Data Example

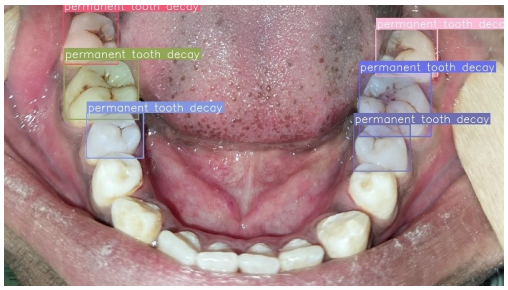

**Question**: This is an intraoral photograph. Can you identify the regions where caries is observed? Please provide the regions as bounding boxes in <box></box> format.

**Ground Truth**:

**<Caption>**
1. Image-level description: Multiple left teeth show chalky white opacities present on the buccal surfaces. Localized brown staining is present on interproximal and cervical aspects of the posterior teeth. A carious lesion is observed along the buccal fissure of the left mandibular first molar. 2. Region-level description: In region <box>[719,363,1091,588]</box>, buccal surface of permanent teeth display multiple brown-black discoloration and clear enamel demineralization. Surface texture seems uneven.
**</Caption>**
**<Think>**
1. Based on the above region description, this region <box>[719,363,1091,588]</box> may correspond to permanent tooth decay.
2. Decay in permanent teeth often appears as deep pits or fissures, with enamel softening or discoloration ranging from white chalky spots to dark brown areas. Clinically, patients may report sharp pain upon biting, sensitivity to temperature changes, and slight gingival inflammation around the lesion.
3. The observed image appearance of this region <box>[719,363,1091,588]</box> matches the expected visual features of permanent tooth decay.
**</Think>**
**<Answer>**
The permanent tooth decay is present in regions <box>[719,363,1091,588]</box>.
**</Answer>**

Figure 15. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

**Training Data Example**




**Question**: Locate the areas of caries in this intraoral photograph. Please provide the coordinates as bounding boxes in <box></box> format.

**Ground Truth**:

**<Caption>**

1. Image-level description:
Multiple posterior mandibular teeth show dark brown to black staining concentrated in the occlusal pits and fissures. Several molar occlusal surfaces exhibit focal enamel breakdown with small, rounded cavitated areas in the central grooves. Premolar occlusal fossae show smaller areas of brown discoloration and shallow, pit-like defects. The discoloration and cavitation are most pronounced on the most posterior molars bilaterally.
2. Region-level description:
In multiple regions, permanent premolar exhibits deep pits with black discoloration, surrounded by a chalky white zone of demineralized enamel. Surface enamel is irregular and partially eroded.
Specific region coordinates are:\n<box>[272,11,584,313]</box>, <box>[269,247,694,588]</box>, <box>[1817,102,2168,369]</box>, <box>[1730,309,2141,674]</box>, <box>[1709,573,2032,838]</box>, <box>[388,516,711,774]</box>.

**</Caption>**

**<Think>**

1. Based on the above region description, these regions <box>[272,11,584,313]</box>, <box>[269,247,694,588]</box>, <box>[1817,102,2168,369]</box>, <box>[1730,309,2141,674]</box>, <box>[1709,573,2032,838]</box>, <box>[388,516,711,774]</box> may correspond to permanent tooth decay.
2. Permanent tooth decay often presents with enamel breakdown, brownish-black lesions on occlusal or interproximal surfaces, and surrounding enamel demineralization. Clinical symptoms include prolonged sensitivity to temperature or sugary foods, intermittent discomfort while chewing, and visible plaque accumulation near the lesion, sometimes with mild gingival swelling.
3. The observed image appearance of these regions <box>[272,11,584,313]</box>, <box>[269,247,694,588]</box>, <box>[1817,102,2168,369]</box>, <box>[1730,309,2141,674]</box>, <box>[1709,573,2032,838]</box>, <box>[388,516,711,774]</box> matches the expected visual features of permanent tooth decay.

**</Think>**

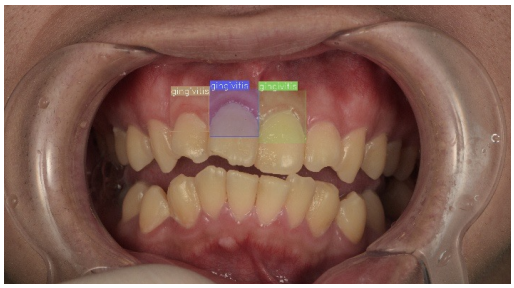
**<Answer>**

The permanent tooth decay is found in regions <box>[272,11,584,313]</box>, <box>[269,247,694,588]</box>, <box>[1817,102,2168,369]</box>, <box>[1730,309,2141,674]</box>, <box>[1709,573,2032,838]</box>, <box>[388,516,711,774]</box>.

**</Answer>**

Figure 16. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

**Training Data Example**



**Question**: This is an intraoral photograph. Which disease is visible?

**Ground Truth**:

**<Caption>**

1. Image-level description:

Intraoral frontal view with cheek retractors in place showing the maxillary and mandibular anterior segments. The maxillary central and lateral incisors exhibit irregular, chipped incisal edges and appear slightly discolored (yellowish). There is a separation between the maxillary and mandibular teeth with no occlusal contact present. The maxillary lateral incisors and canines are visible on both sides. The mandibular incisors display crowding with rotations and mild overlapping, particularly of the central and lateral incisors. Incisal wear is apparent on several anterior teeth. The tooth surfaces overall show a yellowish extrinsic discoloration. The marginal gingiva in the anterior sextants is visible, generally pink with areas of slightly increased redness along the gingival margin; interdental papillae are present. No restorative materials, crowns, or obvious large calculus deposits are seen in the exposed anterior areas. The posterior dentition is partially visible bilaterally but not fully captured in this view.

2. Region-level description:

In multiple regions, the image shows slightly increased redness along the gingival margin. Specific region coordinates are: <box>[2259,1651,2957,2313]</box>,<box>[2789,1550,3682,2404]</box>, <box>[3474,1550,4341,2492]</box>

**</Caption>**

**<Think>**

1. Based on the above region description, these regions <box>[2259,1651,2957,2313]</box>, <box>[2789,1550,3682,2404]</box>, <box>[3474,1550,4341,2492]</box> may correspond to inflammation associated with gingivitis.

2. Inflammation associated with gingivitis manifests as reddened gingival margins and papillae, edema causing puffiness and distortion of normal gingival contour, and a smooth and shiny surface.

3. So, the observed image appearance of these regions <box>[2259,1651,2957,2313]</box>, <box>[2789,1550,3682,2404]</box>, <box>[3474,1550,4341,2492]</box> matches the expected visual features of inflammation associated with gingivitis.

**</Think>**

**<Answer>**

There are three teeth in the upper jaw showing possible signs of gingivitis. They are detected in regions <box>[2259,1651,2957,2313]</box>,<box>[2789,1550,3682,2404]</box>, <box>[3474,1550,4341,2492]</box>.

**</Answer>**

Figure 17. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

Training Data Example

**Question**: This is a periapical X-ray image. What disease(s) or condition(s) does it suggest

**Ground Truth**:

**<Caption>** Periapical radiograph of a posterior quadrant showing four posterior teeth. The two most posterior teeth (on the right side of the image) demonstrate radiolucent defects in the occlusal and interproximal regions: the second-from-right tooth has an obvious radiolucency in the occlusal and interproximal regions extending into the dentin toward the pulp chamber., and The most posterior tooth shows a subtle mesio-occlusal radiolucency. Root contours and periodontal ligament spaces appear preserved without a clearly defined periapical radiolucency; alveolar bone height between teeth is within expected limits for this region.

**</Caption>**

**<Think>**

1. Visual feature analysis: A radiolucent lesion is evident at the occlusal and distal surfaces of the second most posterior tooth, with contours consistent with carious destruction of enamel and underlying dentin. A subtle radiolucent lesion is noted at the mesio-occlusal surface of the most posterior tooth. The radiolucency of the second most posterior tooth is relatively large and extends close to the pulpal chamber space.\n- No clear periapical radiolucency or loss of lamina dura around root apices is visible; periodontal ligament spaces appear within normal limits and alveolar crest height is maintained.
2. Radiographic characteristics and differentials: Dental caries appear radiolucent on radiographs; interproximal caries typically present as triangular radiolucencies pointing toward the DEJ/pulp, while occlusal caries can appear as less well-defined radiolucencies in the occlusal fossa that may be underestimated on periapical films. When caries extend into deep dentin and approach or involve the pulp chamber, radiographs may show loss of the normal radiopaque dentin margin and proximity to the pulp horn; if pulpal necrosis had occurred, a periapical radiolucency would be expected, which is not seen here.\n- Other possibilities (restorative radiolucency, artifact) are less likely given the anatomic shape and location of the defects and absence of restorative margins.
3. Synthesis and most likely diagnosis: The image is most consistent with occlusal and interproximal dental caries affecting the two most posterior teeth, with the largest lesion on the second most posterior tooth extending deep into dentin and likely approaching or involving the pulp; there is no radiographic evidence of periapical pathology at the time of image acquisition.

**</Think>**

**<Answer>**

The periapical radiograph demonstrates a deep occlusal and distal carious lesion of the second most posterior tooth, extending into dentin and approaching the pulp, without clear radiographic evidence of periapical pathology. A possible mesio-occlusal carious lesion is also noted on the most posterior tooth.

**</Answer>**

Figure 18. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

## Training Data Example

**Question**: Analysis of the lateral cephalometric radiograph requires:

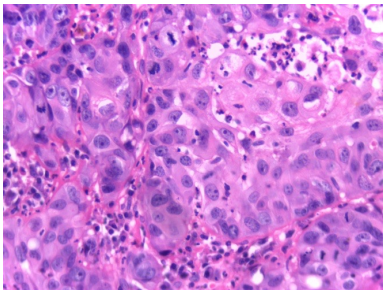
1. Landmark identification: Provide all anatomical landmarks with coordinates in the format: <points>(x='…', y='…', alt='…')</points>
2. CVM staging: What is the estimated Cervical Vertebral Maturation (CVM) stage based on this analysis?

**Ground Truth**:
In this lateral cephalometric radiograph, 29 landmarks were detected:
<points x='1626' y='1196' alt='A-point'></points>,
<points x='1652' y='1151' alt='Anterior Nasal Spine'></points>,
<points x='1578' y='1494' alt='B-point'></points>,
<points x='1551' y='1630' alt='Menton'></points>,
<points x='1645' y='794' alt='Nasion'></points>,
<points x='1543' y='1011' alt='Orbitale'></points>,
<points x='1588' y='1578' alt='Pogonion'></points>,
<points x='1285' y='1172' alt='Posterior Nasal Spine'></points>,
<points x='1857' y='1103' alt='Pronasale'></points>,
<points x='1102' y='1207' alt='Ramus'></points>,
<points x='1134' y='867' alt='Sella'></points>,
<points x='1066' y='1084' alt='Articulare'></points>,
<points x='1059' y='1004' alt='Condylion'></points>,
<points x='1575' y='1610' alt='Gnathion'></points>,
<points x='1123' y='1364' alt='Gonion'></points>,
<points x='998' y='1007' alt='Porion'></points>,
<points x='1488' y='1346' alt='Lower 2nd PM Cusp Tip'></points>,
<points x='1620' y='1349' alt='Lower Incisor Tip'></points>,
<points x='1440' y='1337' alt='Lower Molar Cusp Tip'></points>,
<points x='1471' y='1339' alt='Upper 2nd PM Cusp Tip'></points>,
<points x='1601' y='1214' alt='Upper Incisor Apex'></points>,
<points x='1649' y='1388' alt='Upper Incisor Tip'></points>,
<points x='1435' y='1334' alt='Upper Molar Cusp Tip'></points>,
<points x='1540' y='1505' alt='Lower Incisor Apex'></points>,
<points x='1733' y='1403' alt='Labrale inferius'></points>,
<points x='1734' y='1306' alt='Labrale superius'></points>,
<points x='1673' y='825' alt='Soft Tissue Nasion'></points>,
<points x='1679' y='1544' alt='Soft Tissue Pogonion'></points>,
<points x='1729' y='1195' alt='Subnasale'></points>.

Based on the morphological characteristics of the cervical vertebrae (C2–C4) derived from these landmarks, the cervical vertebral maturation (CVM) stage was estimated. The analysis considered vertebral body shape, height-to-width ratios, and the degree of concavity at the lower borders.
The prediction indicates that the patient is at CVM Stage S6.

Figure 19. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

Training Data Example



**Question**: A male patient over 60 years old, a former smoker and drinker, who reports no sun exposure. The pathological image from the gingiva demonstrates a lesion measuring 3.0 cm. Evaluate this pathological patch's morphology and indicate if it corresponds to leukoplakia without dysplasia, leukoplakia with dysplasia, or oral squamous cell carcinoma.

**Ground Truth**:

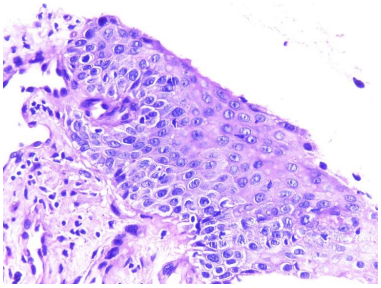
**<Caption>**
H&E-stained high-power field showing sheets and nests of polygonal epithelial cells. Cells display marked nuclear pleomorphism with variation in nuclear size and shape, coarse irregular chromatin, and one or more prominent nucleoli in many cells. Many cells have a high nuclear-to-cytoplasmic ratio; cytoplasm ranges from scant to moderately eosinophilic. Occasional cells show dense eosinophilic cytoplasm and faint intercellular bridges. Frequent mitotic figures are present, including several that appear atypical. The tumor cell nests are separated by a fibrovascular stroma containing a mixed inflammatory infiltrate composed predominantly of small lymphocytes and scattered neutrophils. Foci of individual cell keratinization and scattered keratinized debris are visible within some cell groups. Overall tissue architecture is disrupted with irregular, tightly packed cellular aggregates and intervening stromal reaction.
**</Caption>**

**<Think>**
1. Based on the histologic field description above, the observed patch demonstrates features consistent with oral squamous cell carcinoma.
2. Oral squamous cell carcinoma is characterized by malignant squamous proliferation with loss of normal epithelial stratification. Tumor nests infiltrate the lamina propria, forming keratin pearls and dyskeratotic cells. Stroma shows dense fibroblastic reaction, chronic inflammatory infiltrates, and frequent abnormal mitoses, sometimes with focal necrosis.
3. The overall histologic features support a diagnosis of oral squamous cell carcinoma.
**</Think>**

**<Answer>**
Histopathological findings indicate oral squamous cell carcinoma.
**</Answer>**

Figure 20. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

Training Data Example



**Question**: This is a patch of a pathological image from the oral cavity. Can you determine whether it shows histologically normal oral mucosa, or oral squamous cell carcinoma?

**Ground Truth**:

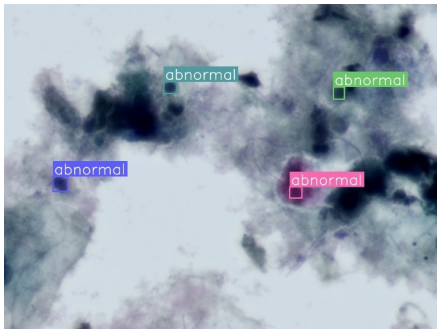
**<Caption>**
Light-microscopic view of stratified squamous epithelium showing an expansile, thickened epithelial plate with irregular surface contour and loss of orderly maturation. The epithelial cells are markedly pleomorphic with variation in cell and nuclear size and shape, prominent hyperchromatic nuclei, and frequent large eosinophilic nucleoli. There is an increased nuclear-to-cytoplasmic ratio and nuclear crowding with overlap of adjacent cells. Numerous mitotic figures are visible, some appearing atypical. Individual cell keratinization/dyskeratotic cells and small aggregates of more eosinophilic, dense cytoplasm are present within the epithelial layer. At the interface, epithelial strands and nests extend irregularly into the underlying connective tissue, which shows a reactive fibroblastic stroma with a mixed chronic inflammatory infiltrate. The overall architecture is disordered with loss of normal polarity of the epithelial layers.
**</Caption>**

**<Think>**
1. From the histologic appearance, the patch exhibits features indicative of oral squamous cell carcinoma.
2. Oral squamous cell carcinoma generally exhibits malignant squamous proliferation with prominent nuclear irregularity, hyperchromatic nuclei, and loss of normal epithelial stratification. Tumor nests infiltrate the lamina propria, often forming keratin pearls or dyskeratotic cells. The surrounding stroma shows marked desmoplastic response, chronic inflammatory infiltrates, and occasional microvascular proliferation. Focal necrosis, ulceration, and abnormal mitoses are frequently present, indicating aggressive malignancy.
3. The overall histologic features support a diagnosis of oral squamous cell carcinoma.
**</Think>**

**<Answer>**
The patch is consistent with oral squamous cell carcinoma.
**</Answer>**

Figure 21. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

Training Data Example




**Question**: This is a patch of a cellular pathological image from the oral cavity. Please identify all types of nuclei present in this region based on their characteristics. Please provide the bounding box coordinates in <box></box> format.

**Ground Truth**:

**<Caption>**

1. Image-level description:
Low-to-moderate cellularity smear with abundant pale basophilic, fibrillar/mucoid background and scattered granular debris and dense basophilic clumps. Cells are present singly and in loose small groups. Many cells are small to intermediate in size, round to oval, with scant to moderate pale bluish cytoplasm and centrally placed round nuclei showing finely to slightly coarsely granular chromatin. A single larger polygonal cell near the right-center has abundant deeply eosinophilic/magenta cytoplasm with a well-defined cell border and a single round, dark nucleus. Scattered small, densely stained round nuclei or nuclear fragments are seen throughout. Thin filamentous strands and proteinaceous material are interspersed in the background.
2. Region-level description:
In multiple regions, the image shows the nuclei vary in size, are irregular in shape, with dense or uneven chromatin, irregular nuclear membranes, and have prominent or asymmetrically distributed nucleoli.
Specific region coordinates are:
<box>[504,322,530,345]</box>, <box>[83,302,115,335]</box>, <box>[581,144,605,170]</box>, <box>[280,135,308,160]</box>.

**</Caption>**

**<Think>**

1. Based on the regional morphology, these regions <box>[504,322,530,345]</box>, <box>[83,302,115,335]</box>, <box>[581,144,605,170]</box>, <box>[280,135,308,160]</box> may correspond to abnormal epithelial nucleus.
2. Abnormal nuclei often show size variation, irregular shapes, dense or clumped chromatin, irregular nuclear membranes, and enlarged or displaced nucleoli, commonly seen in malignancy or severe pathology.
3. The observed appearance of these regions <box>[504,322,530,345]</box>, <box>[83,302,115,335]</box>, <box>[581,144,605,170]</box>, <box>[280,135,308,160]</box> matches the expected cytological features of abnormal epithelial nucleus.

**</Think>**

**<Answer>**

The abnormal epithelial nucleus is detected in regions <box>[504,322,530,345]</box>, <box>[83,302,115,335]</box>, <box>[581,144,605,170]</box>, <box>[280,135,308,160]</box>.

**</Answer>**

Figure 22. Data example used in the training stage. It contains multimodal information, including the image, question, and ground truth.

**System Prompt:**

You are an expert in textual diagnostic reasoning in the field of oral healthcare. You will compare a model's diagnostic thinking output (**Candidate**) with the reference gold-standard report (**Reference**) . Your task is to evaluate the **Candidate** based on three dimensions:

1. **Logical Coherence**
   - Evaluate whether the reasoning in <Think> (steps 1, 2, and 3) is logically consistent with the information in <Caption>. Look for any jumps, contradictions, or illogical conclusions.
   - Score: 0 or 1. 0 indicates a clear logical inconsistency or contradiction, 1 indicates fully coherent reasoning.
2. **Factual Knowledge Accuracy**
   - Assess whether the professional knowledge presented in step 2 of <Think> is correct, complete, and consistent with established oral medicine or pathology facts.
   - Score: 0–2. 0 = incorrect or seriously inaccurate, 1 = partially correct, 2 = fully correct.
3. **Answer Consistency**
   - Evaluate whether the reference (gold-standard answer) can be directly derived from the description in Step 3 of <Think>. Ensure that the reasoning trace fully justifies the reference (gold-standard answer) by checking for contradictions, reliance on unstated information, or unsupported claims.
   - Score: 0–2. 0 = inaccurate, 1 = partially accurate, 2 = fully accurate.
4. **Overall Thinking Reward**
   - Calculate as:
   overall_thinking_reward = (logical_coherence + factual_knowledge_accuracy + answer_consistency) / 5
   - Output the result as a decimal value.

**User Prompt:**
**Reference**: {***SOLUTION***}
**Candidate**: {***COMPLETION***}

Please strictly follow this JSON format for your output:

```
{
    “logical coherence": X,
    “factual knowledge accuracy": Y,
    “answer consistency": Z,
    “overall thinking reward": R
}
```

Figure 23. The prompt for computing the TRACE-based reward using GPT-5-nano.

**System Prompt:**

You are an expert in diagnostic reasoning for oral healthcare. Your task is to assign a **numerical score** (0–1) that measures how clinically equivalent a model's answer (**Candidate**) is to the gold-standard (**Reference**). The Reference may include:

- Bounding boxes (<box>[x1,y1,x2,y2]</box>)
- Tooth numbers (e.g., "Teeth 11, 12, 21, 22")
- Diagnostic description

### Scoring Rules

1. **Content Score** (0–1)
   - Measures semantic/diagnostic equivalence between Candidate and Reference text.
   - Minor omissions → small penalty; contradictions or false positives → large penalty.
2. **Box/Tooth Score** (0–1)
   - Only if Reference contains boxes or teeth.
   - Box Scoring: For each Reference box, find Candidate box with highest IoU:
     - ≥0.9 → 1.0
     - 0.5–0.9 → 0.6–0.8
     - 0.1–0.5 → 0.3–0.5
     - <0.1 → 0
   - Tooth Scoring: For each Reference tooth, assign:
     - 100% match → 1.0
     - >50% match → 0.6–0.8
     - 10–50% match → 0.3–0.5
     - <10% match → 0
   - If both exist, average → Box/Tooth Score
3. **Format Bonus**
   - Only if Reference has boxes or teeth. Candidate correctly uses `<box>` or tooth notation → +1, else 0

### Weighting & Final Score

- **If Reference contains boxes/teeth**:
  - w_content = 0.4, w_box_tooth = 0.4, w_format = 0.2
  - Final Score = Content × w_content + Box/Tooth × w_box_tooth + Format Bonus × w_format
  - If Candidate misses required format → Final Score = 0
- **If Reference has only diagnosis text**:
  - Only Content Score counts; Box/Tooth Score = 0, Format Bonus = 0
  - Final Score = Content Score
  - Weight sum remains 1 automatically

**User Prompt:**
**Reference**: {***SOLUTION***}
**Candidate**: {***COMPLETION***}

Please output the score **only** in this format: Score: <float between 0.0 and 1.0>

Figure 24. The prompt for computing the answer reward using GPT-5-nano.

MMOral-Uni Benchmark

**Question**: What is the count of teeth observed in this intraoral image, what kinds are they, and where exactly are their centers located?

**Category**: II_Loc

**Ground Truth**: This is an intraoral photography image containing 14 teeth. There are 7 tooth types identified, including 2 Central Incisors located at [<points>(349, 394)</points>, <points>(297, 391)</points>]; 2 Lateral Incisors located at [<points>(399, 379)</points>, <points>(247, 382)</points>]; 2 Canines located at [<points>(205, 349)</points>, <points>(446, 352)</points>]; 2 1st Premolars located at [<points>(489, 291)</points>, <points>(168, 285)</points>]; 2 2nd Premolars located at [<points>(510, 217)</points>, <points>(150, 216)</points>]; 2 1st Molars located at [<points>(538, 132)</points>, <points>(120, 129)</points>]; 2 2nd Molars located at [<points>(104, 38)</points>, <points>(563, 39)</points>].

**Question**: This is an intraoral photograph of the oral cavity. What is the likely diagnosis?

**Category**: II_Dx-I

**Ground Truth**: The image is most consistent with a minor aphthous (recurrent aphthous) ulcer on the maxillary labial/vestibular mucosa.

**Question**: This is an intraoral photograph. Which oral disease is depicted here?

**Category**: II_Dx-I

**Ground Truth**: Advanced occlusal dental caries (large cavitated carious lesion with dentinal involvement) in a mandibular posterior molar (third molar; wisdom tooth).

**Question**: This is an intraoral photograph. Can you identify the regions where caries is observed? Please provide the regions as bounding boxes in <box></box> format.

**Category**: II_Dx-R

**Ground Truth**: Permanent tooth decay is present in the regions <box>[108,443,674,820]</box>, <box>[2190,447,2808,1013]</box>.

Figure 25. Some examples in our MMOral-Uni benchmark. Each example contains the image, question, category, and corresponding ground truth validated by experienced dentists.

MMOral-Uni Benchmark

**Question**: This is a periapical X-ray image. What disease(s) or condition(s) does it suggest?

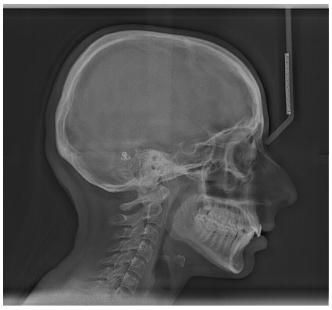
**Category**: PA

**Ground Truth**: The radiograph demonstrates carious lesions affecting multiple posterior multi-rooted teeth. Periapical radiolucencies and furcation involvement are displayed in some roots, suggesting chronic apical periodontitis, periodontal bone loss, and combined endodontic-periodontal lesions.

**Question**: This is a periapical X-ray image. Please identify the disease(s) or condition(s).

**Category**: PA

**Ground Truth**: The periapical radiograph demonstrates occlusal and interproximal caries on adjacent, the canine and first premolar. The first premolar shows a deep carious lesion extending into the dentin and approaching/involving the pulp, with associated radiographic evidence of periapical pathology.



**Question**: Analysis of the lateral cephalometric radiograph requires:

1. Landmark identification: Provide all anatomical landmarks with coordinates in the format: <points>(x='…', y='…', alt='…')</points>

2. CVM staging: What is the estimated Cervical Vertebral Maturation (CVM) stage based on this analysis?

**Category**: CE

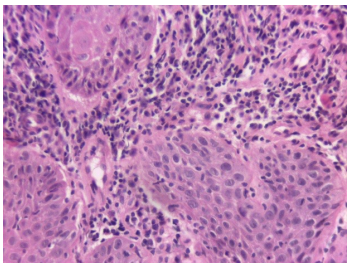
**Ground Truth**:

In this lateral cephalometric radiograph, 29 landmarks were detected:

<points x='1446' y='1145' alt='A-point'></points>,
<points x='1497' y='1086' alt='Anterior Nasal Spine'></points>,
<points x='1355' y='1606' alt='B-point'></points>,
……
Based on the morphological characteristics of the cervical vertebrae (C2–C4) derived from these landmarks, the cervical vertebral maturation (CVM) stage was estimated. The analysis considered vertebral body shape, height-to-width ratios, and the degree of concavity at the lower borders.
The prediction indicates that the patient is at CVM Stage S5.

Figure 26. Some examples in our MMOral-Uni benchmark. Each example contains the image, question, category, and corresponding ground truth validated by experienced dentists.

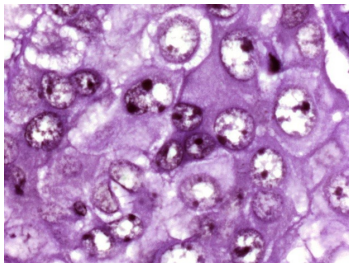
MMOral-Uni Benchmark



**Question**: The patient is a 40- to 60-year-old female who uses tobacco, does not consume alcohol, and has sun exposure. The pathological image from the buccal mucosa demonstrates a lesion measuring 2.0 cm. Evaluate this pathological patch's morphology and indicate if it corresponds to leukoplakia without dysplasia, leukoplakia with dysplasia, or oral squamous cell carcinoma.

**Category**: PI

**Ground Truth**: Histopathological findings indicate leukoplakia with dysplasia.



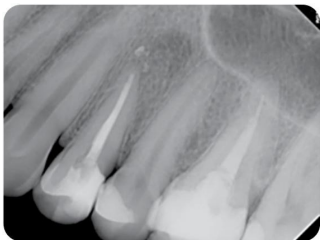
**Question**: This is a patch of a pathological image from the oral cavity. Can you determine whether it shows histologically normal oral mucosa, oral submucous fibrosis, or oral squamous cell carcinoma?

**Category**: PI

**Ground Truth**: The patch is consistent with oral squamous cell carcinoma.

Figure 27. Some examples in our MMOral-Uni benchmark. Each example contains the image, question, category, and corresponding ground truth validated by experienced dentists.

MMOral-Uni Benchmark

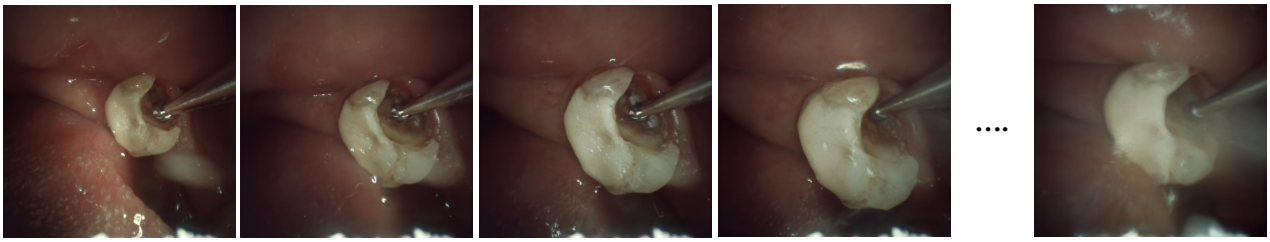
**Category**: TP

**Question**: Based on the information provided by the patient, including Chief Complaint, Medical History, Dental History, Clinical Evaluation, Extra-oral Examination, Intra-oral Examination, and Radiographic Findings, together with the corresponding oral-related images, formulate an appropriate Pretreatment Diagnosis and a comprehensive Treatment Plan.
## Chief Complaint
"I had a root canal re-done on this tooth and I am still having pain when I chew with it."
## Medical History
The patient (Pt) was a 26-year-old Caucasian male. He had no known drug allergies (NKDA). Vital signs were: blood pressure (BP) 122/84 mmHg right arm seated (RAS), respiratory rate (RR) 18 breaths per minute and regular, pulse 72 beats per minute (BPM) and regular.
The Pt was American Society of Anesthesiologists Physical Status Scale (ASA) Class I.
## Dental History
The Pt had tooth #12 treated with root canal therapy (RCT) and restorative composite 5-7 years prior. The tooth became symptomatic and it was re-treated. After re-treatment, the Pt experienced postoperative pain for a few weeks and it was decided that apical microsurgery should be performed.
## Clinical Evaluation
### Extra-oral Examination (EOE)
Clinical examination revealed no lymphadenopathy of the submandibular and neck areas. Perioral and intra-oral soft tissue appeared normal. Extra-oral soft tissues appeared satisfactory in color and texture. The temporomandibular joint showed no popping/clicking or deviation on opening and was otherwise asymptomatic.
### Intra-oral Examination (IOE)
The intra-oral soft tissue examination had normal appearance, but the root apex had tenderness upon palpation and percussion tenderness was noted.
### Radiographic Findings
Tooth #12 showed dense root canal obturation with possible extrusion of gutta-percha (GP)/sealer into the periapical (PA) tissues. No final crown was present, but adequate composite restoration and glass ionomer material were placed within access and cavosurface. Tooth #11 was clinically intact without any restorations. Tooth #13 had a mesial-occlusal composite. Tooth #14 had previous RCT with incomplete root filling, …

**Ground Truth**:
# Pretreatment Diagnosis:
## Pulpal: Previously treated, tooth #12
## Apical: Symptomatic Apical Periodontitis, tooth #12
# Treatment Plan
## Recommended
Emergency: None
Definitive: PA surgery to tooth #12 with Root End Filling.
## Alternative
Extraction and dental implant, fixed partial denture, or no treatment
## Restorative
Composite or amalgam build up with either onlay or full crown coverage

Figure 28. Some examples in our MMOral-Uni benchmark. Each example contains the image, question, category, and corresponding ground truth validated by experienced dentists.

MMOral-Uni Benchmark



**Question**: This is an intraoral video recording of the dental treatment. Help me to understand it accurately.

**Category**: IV

**Ground Truth**:

1) Clinical environment and imaging conditions\n- The footage is intra-oral, recorded with a close-focus intraoral camera or loupe-mounted video system. The camera provides a tight field of view focused on a single posterior tooth and adjacent soft tissues.\n- Lighting is bright and produces specular highlights on wet tooth and mucosal surfaces; reflections and glare intermittently obscure fine detail.\n- The scene shows variable focus and some motion blur as instruments and the camera move; intermittent water spray or irrigation mist and saliva create transient haze and small reflective droplets.\n- The operative field is crowded by the working instrument and surrounding soft tissues (buccal mucosa and tongue), with only a limited visual window to the occlusal surface. Occasional camera fouling/obscuration and framing shifts occur during the sequence.
2) Visible tools, tissues, and anatomy\n- Tools: a metallic rotary instrument tip is consistently visible. It appears as a slender cylindrical/rounded metal cutting or excavation tip (consistent with a rotary bur or a contra-angle slow/low-speed handpiece instrument). There is evidence of irrigation/aerosol but no direct view of a saliva ejector or mirror in most frames.\n- Hard tissues: a posterior tooth with a large occlusal cavitation.\n- Soft tissues: adjacent buccal mucosa, tongue and saliva are visualized; gingival margins near the cavitated area are visible but not the focus. There is little overt bleeding in the frames provided.\n- Internal anatomy: the excavation reaches a dark central area which may represent deep carious dentin and/or the roof of the pulp chamber in a large lesion.
3) Dynamic interactions during treatment\n- The rotary tip is engaged repeatedly in the central occlusal defect and along the cavosurface margin; motion includes axial penetration into the lesion and circumferential shaping of the cavity.\n- Cutting action produces removal of soft/brown carious material and detachment of undermined enamel; particulate debris and irrigant spray are intermittently seen leaving a transient fog over the field.\n- Small fragments of tooth structure are dislodged and displaced while the remaining coronal portion is preserved and manipulated; the instrument occasionally contacts and abrades the inner margins of the cavity.\n- The operator repositions the tip multiple times to access different aspects of the defect; the soft tissues shift slightly with instrument movement and irrigation, requiring continuous reorientation of the camera/instrument.
4) Procedural sequence (as observed)\n- Initial close inspection of a large occlusal lesion with visible dark (carious) center.\n- Engagement of a metallic rotary instrument into the lesion center, with progressive excavation of the dark carious material.\n- Enlargement and shaping of the cavity: removal of undermined enamel and excavation of decayed dentin, creating a deeper, more defined occlusal preparation.\n- Repeated irrigation/air-water spray cycles observed during and after cutting, with transient mist and removal of debris.\n- Continued contouring of cavity walls and apparent detachment of a coronal fragment or cusp remnant as the procedure progresses.\n- Final frames show a shaped occlusal defect with smoother internal surfaces and less visible loose debris, consistent with completion of the initial caries excavation/cavity preparation stage.
5) Notable clinical findings and challenges\n- ….

Figure 29. Some examples in our MMOral-Uni benchmark. Each example contains the image, question, category, and corresponding ground truth validated by experienced dentists.

Intraoral Image

**Question**: This is an intraoral photograph of the oral cavity. Which oral disease is depicted here?

**Category**: II_Dx-I, Caries

**Ground Truth**: Advanced occlusal dental caries (large cavitated carious lesion with dentinal involvement) of a mandibular posterior molar.

**GPT-5 Response**: Evaluation score: 0.8

The image shows signs of dental caries, particularly on the molars. The visible decay and discoloration suggest advanced tooth decay. Additionally, there is evidence of gingivitis or periodontal disease, as indicated by the redness and inflammation of the gums. These findings are consistent with poor oral hygiene and potential neglect in dental care.

**Lingshu-7B Response:** Evaluation score: 0.8

Dental caries (a cavity) on a molar.

**OralGPT-Omni Response:** Evaluation score: 1.0

**<Caption>** The image shows the right posterior mandibular quadrant with the tongue retracted to expose the buccal and occlusal surfaces of the lower molars. There is a large, dark brown–black cavitated lesion occupying much of the occlusal surface of the most posterior molar, with loss of normal enamel contour and a central area of exposed dentin. Adjacent posterior teeth show yellow-brown staining on occlusal surfaces. The concavities are irregular in shape, with rough, undermined margins and discoloration; there is no obvious metallic restoration visible in the affected tooth.
**</Caption>**
**<Think>**
1. Visible features matching caries: The photo demonstrates classic signs of dental caries — focal enamel breakdown with cavitation, dark brown/black discoloration of the lesion, and exposure of softened dentin within the occlusal pit/fissure system. The irregular, undermined enamel margins and the central darkened area indicate progressive demineralization and tissue loss rather than a simple stain or abrasion.
2. Typical visual appearance and symptoms of caries from professional knowledge: Carious lesions commonly present as white spot lesions in early stages, progressing to brown/black staining and eventual cavitation when enamel and dentin are destroyed. Advanced occlusal caries often appear as deep, dark, soft-appearing defects with loss of normal occlusal anatomy and may be associated with sensitivity or pain if dentin/pulp are involved.
3. Confirmation that observed appearance matches caries: The combination of a large cavitated occlusal defect, dark discoloration, and surrounding enamel breakdown on the posterior molar corresponds well with advanced dental caries (deep occlusal caries) affecting the posterior mandibular tooth.
**</Think>**
**<Answer>**
Advanced dental caries (deep occlusal caries) involving the posterior mandibular molar, with extensive cavitation and darkly discolored dentinal involvement.
**</Answer>**

Figure 30. Case study on the intraoral image modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

Intraoral Image

**Question**: This is an intraoral photograph of the oral cavity. What is the disease present in this image?

**Category**: II_Dx-I, Gingivitis

**Ground Truth**: The image shows clinical features consistent with plaque-induced gingivitis (marginal/papillary gingival inflammation) affecting the anterior maxillary and mandibular gingiva, characterized by erythema, edema, and focal bleeding.

**GPT-5 Response**: Evaluation score: 0.8

Gingivitis — plaque-induced inflammation of the gums (erythematous, swollen margins with bleeding).

**Lingshu-7B Response:** Evaluation score: 0.2

Based on the clinical presentation shown in the intraoral photograph, the patient appears to have a condition called gingival hyperplasia or overgrowth of the gums. The gums are significantly enlarged and swollen, covering a large portion of the teeth and creating a thickened, irregular appearance.

**OralGPT-Omni Response:** Evaluation score: 1.0

**<Caption>**

Frontal intraoral view showing both maxillary and mandibular anterior teeth (canine to canine). The marginal gingiva appears erythematous (reddened) and edematous with loss of normal scalloped contour; interdental papillae are blunted and there is visible bleeding at the gingival margin. Supragingival plaque and calculus deposits are evident on several lower anterior teeth. No obvious large ulceration or purulent exudate is seen in this image.

**</Caption>**

**<Think>**

1. Visible features matching Gingivitis: The gingiva demonstrate diffuse erythema and swelling of the marginal and papillary tissues, blunting of interdental papillae, and active bleeding at the gingival margin — all classic signs of inflammation. The presence of supragingival plaque/calculus adjacent to the inflamed tissue supports a local irritant-driven inflammatory process.
2. Typical appearance/symptoms of Gingivitis from professional knowledge: Gingivitis commonly presents as red, swollen, tender gingiva that bleeds easily on provocation (brushing/probing), often associated with plaque accumulation; changes are usually confined to the soft tissues without radiographic bone loss.
3. Confirmation: The observed combination of marginal/papillary erythema, edema, bleeding, and adjacent plaque/calculus corresponds well with the expected clinical presentation of plaque-induced gingivitis.</Think>

**<Answer>**

Clinical diagnosis: Generalized plaque-induced gingivitis affecting the anterior maxillary and mandibular gingiva, characterized by marginal and papillary erythema, edema, and bleeding with visible supragingival plaque/calculus.

**</Answer>**

Figure 31. Case study on the intraoral image modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

## Periapical X-ray

**Question**: This is a periapical X-ray image. Please identify the disease(s) or condition(s).

**Category**: PA, Pulpitis

**Ground Truth**: The periapical radiograph demonstrates occlusal and interproximal caries on adjacent, the canine and first premolar. The first premolar shows a deep carious lesion extending into the dentin and approaching/involving the pulp, with associated radiographic evidence of periapical pathology.

**GPT-5 Response**: Evaluation score: 0.0

Multiple anterior teeth with very wide, open apices and large pulp chambers.\n- Lamina dura and periodontal ligament space appear normal; This radiograph is most consistent with immature permanent teeth exhibiting incomplete root development (open apices). This is a normal developmental condition in adolescents, not disease.

**Lingshu-7B Response:** Evaluation score: 0.3

1. Dental Caries: There is evidence of dental caries in the molars, particularly visible as radiolucent areas near the crown and root regions of the teeth. 2. Periapical Lesions: The presence of radiolucent areas at the apices of some teeth suggests periapical lesions, which could be indicative of periapical abscesses or granulomas. 3. Root Resorption: There appears to be some degree of root resorption, especially noticeable in the lower left molar where the root structure seems compromised. 4. Bone Loss: The alveolar bone surrounding the teeth shows signs of reduced density and height, indicating possible periodontal disease or chronic inflammation.

**OralGPT-Omni Response:** Evaluation score: 0.8

**<Caption>**
The periapical radiograph of several adjacent teeth showing. Well-defined are present on the proximal surfaces of at least one tooth consistent with approximal caries. Radiographic evidence of periapical pathology is observed at the root apex of a tooth with a cavity. Lamina dura around the apices is relatively preserved.
**</Caption>**
**<Think>**
1. Visual feature analysis: Focal triangular/rounded radiolucencies on proximal crown/root surfaces consistent with approximal carious lesions.
2. Radiographic characteristics recall and differentials: Approximal caries appear as radiolucent areas at the contact/proximal enamel-dentin region and can be distinguished from bone defects by their location relative to the tooth structure.
3. Synthesis and diagnostic decision: The focal proximal radiolucencies indicate concomitant approximal caries on at least one tooth. Therefore the most likely diagnosis is interproximal caries; recommend clinical correlation (pulp testing) and treatment planning restorative care.
**</Think>**
**<Answer>** Periapical radiograph demonstrates proximal carious lesions on one or more teeth with associated radiographic evidence of periapical pathology.
**</Answer>**

Figure 32. Case study on the periapical X-ray modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

Intraoral Image

**Question**: What dental diseases or conditions are observed in this image?

**Category**: II_Dx-I, Calculus

**Ground Truth**: Supragingival dental calculus and plaque are accumulated mainly at the cervical and interproximal areas of the upper and lower anterior teeth, accompanied by extrinsic staining and localized gingival inflammation, suggesting gingivitis.

**GPT-5 Response**: Evaluation score: 0.0

Dental fluorosis (mottled enamel) due to excessive fluoride exposure during tooth development.

**Lingshu-7B Response:** Evaluation score: 0.1

The patient appears to have gingival hyperplasia, which is an overgrowth or enlargement of the gums. The gums are enlarged and swollen, covering a large portion of the teeth.

**OralGPT-Omni Response:** Evaluation score: 0.9

**<Caption>**The frontal intraoral view shows the maxillary and mandibular anterior teeth with supragingival deposits. …
**</Caption>**
**<Think>**
1. Visible features matching Gingivitis: The image shows marginal gingival erythema and edema localized to …
2. Typical appearance of Gingivitis from professional knowledge: Gingivitis commonly presents as redness …
3. Confirmation that observed appearance matches Gingivitis: The combination of marginal erythema, edema…
**</Think>**
**<Answer>**Localized plaque-induced gingivitis involving the anterior maxillary and mandibular gingiva, characterized by marginal erythema, edema, and the presence of supragingival plaque and calculus deposits.
**</Answer>**

**Clinical validity assessment from a radiologist with over ten years of experience:**

| | | |
|---|---|---|
| The output from GPT-5 is not accurate. No dental fluorosis is observed in this image. GPT-5 could have mistakenly identified dental plaque, calculus, and light reflections as enamel abnormalities. Therefore, the overall accuracy level is considered low. | The output from Lingshu-7B is not entirely accurate. While mild gingival swelling is observed in this image, there is no obvious overgrowth or enlargement covering a large portion of the teeth, which suggests that the gingival hyperplasia indicated in the output may be overstated. Therefore, the overall accuracy level is considered moderate. | The output from OralGPT-Omni is comprehensive and accurate. The response shows correct clinical reasoning and considers multiple relevant factors. Therefore, the overall accuracy level is considered high. |

Figure 33. Case study and clinical validity on the intraoral image modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

**Periapical X-ray**

**Question**: Please specify the disease(s) or condition(s) demonstrated in this image.

**Category**: PA, Caries

**Ground Truth**: The radiograph demonstrates mesial interproximal carious lesions on the posterior teeth, extending into the pulp (severe approximal caries) and through the enamel into the dentin (moderate approximal caries).

**GPT-5 Response**: Evaluation score: 0.0

Internal root resorption (inflammatory resorption) of the affected tooth. The film shows a well-defined, round/ovoid radiolucent "ballooning" within the mid-root that is continuous with the root canal, typical of internal resorption. Clinical and endodontic evaluation is indicated.

**Lingshu-7B Response:** Evaluation score: 0.1

The periapical X-ray image shows a tooth with an abnormal root canal morphology, specifically a bifurcation in the root structure. The bifurcation appears to be located at the mid-root level, which is an uncommon anatomical variation. There is no evidence of periapical radiolucency, indicating that there is no significant periapical pathology such as infection or cyst formation.

**OralGPT-Omni Response:** Evaluation score: 1.0

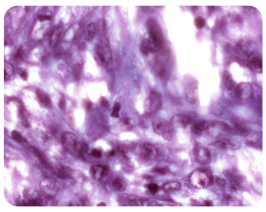
**<Caption>**The radiograph shows four maxillary posterior teeth. Radiolucent cavities are present at the proximal contact …
**</Caption>**
**<Think>** 1. The image shows localized radiolucent cavities at the interproximal contact point between two …
2. Radiographically, approximal (interproximal) caries typically presents as a radiolucent notch or triangle at …
3. The observed interproximal radiolucency most closely matches approximal caries extending into dentin and …
**</Think>**
**<Answer>**Interproximal (approximal) carious lesion involving the contacting surfaces of adjacent posterior teeth with radiographic extension into dentin and pulp, and no clear evidence of periapical pathology.
**</Answer>**

**Clinical validity assessment from a radiologist with over ten years of experience:**

| | | |
|---|---|---|
| The output from GPT-5 is not accurate. There is no evidence of internal root resorption in this image. It appears that GPT-5 could have mistakenly identified a large cavity as internal root resorption. Therefore, the overall accuracy level is considered low. | The output from Lingshu-7B is not accurate. It states that there is a bifurcation in the root structure but does not specify the tooth site, and no bifurcation is visible in any of the teeth in this image. It appears that Lingshu-7B could have mistakenly identified overlapping roots of posterior teeth as a bifurcation in the root structure. Therefore, the overall accuracy level is considered low. | The output from OralGPT-Omni is comprehensive and accurate. The response shows correct clinical reasoning and appropriately emphasizes all teeth with visible cavities in the image. Therefore, the overall accuracy level is considered high. |

Figure 34. Case study and clinical validity on the periapical X-ray modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

## Histopathological Image



**Question**: Does it represent histologically normal oral mucosa, oral submucous fibrosis, or oral squamous cell carcinoma?

**Category**: PI, Oral Squamous Cell Carcinoma

**Ground Truth**: The patch is consistent with oral squamous cell carcinoma.

**GPT-5 Response**: Evaluation score: 0.0

Oral submucous fibrosis. Rationale: The image shows dense, hyalinized collagenous stroma with scattered fibroblasts or inflammatory cells and no invasive epithelial nests or keratin pearls.

**Lingshu-7B Response:** Evaluation score: 0.0

Based on the provided image and description, this appears to be a case of oral submucous fibrosis (OSF). The key features that support this diagnosis include the presence of dense collagen bundles in the lamina propria.

**OralGPT-Omni Response:** Evaluation score: 1.0

**<Caption>** High-power photomicrograph showing densely cellular tissue with loss of normal orderly epithelial stratification ...**</Caption>**

**<Think>**

1. From the histologic appearance, the patch exhibits features indicative of oral squamous cell carcinoma.
2. Oral squamous cell carcinoma generally demonstrates invasive nests and sheets of atypical squamous cells penetrating into the connective tissue. Cells show marked nuclear pleomorphism, hyperchromasia, prominent …
3. The observed morphological features are consistent with oral squamous cell carcinoma.

**</Think>**

**<Answer>**

This patch represents oral squamous cell carcinoma.

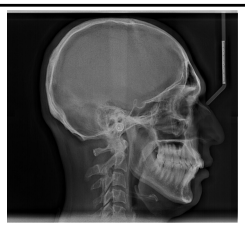
**</Answer>**

**Clinical validity assessment from a radiologist with over ten years of experience:**

| | | |
|---|---|---|
| The output from GPT-5 is not accurate. It could have mistakenly identified histopathological features characteristic of oral submucous fibrosis. Although the reasoning appears relatively reliable, the final diagnosis was determined as oral submucous fibrosis rather than the ground truth, which is oral squamous cell carcinoma. Therefore, the overall accuracy is considered low. | The output from Lingshu-7B is not accurate. It could have mistakenly identified histopathological features characteristic of oral submucous fibrosis. Although the reasoning appears relatively reliable, the final diagnosis was determined as oral submucous fibrosis rather than the ground truth, which is oral squamous cell carcinoma. Therefore, the overall accuracy is considered low. | The output from OralGPT-Omni is comprehensive and accurate. It correctly identified histopathological features characteristic of oral squamous cell carcinoma and demonstrated reliable reasoning. Therefore, the overall accuracy is considered high. |

Figure 35. Case study and clinical validity on the histopathological image modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.

## Cephalometric Radiograph



**Question**: Analysis of the lateral cephalometric radiograph requires:
1. Landmark identification: Provide all anatomical landmarks with coordinates in the format: <points>(x='…', y='…', alt='…')</points>
2. CVM staging: What is the estimated Cervical Vertebral Maturation (CVM) stage based on this analysis?",
**Category**: CE

**Ground Truth**: In this lateral cephalometric radiograph, 29 landmarks were detected.
<points x='590' y='475' alt='A-point'></points>,
<points x='573' y='588' alt='B-point'></points>, …
Based on the morphological characteristics of the cervical vertebrae (C2–C4) derived from these landmarks, the cervical vertebral maturation (CVM) stage was estimated. The analysis considered vertebral body shape, height-to-width ratios, and the degree of concavity at the lower borders.
The prediction indicates that the patient is at CVM Stage S5.

**GPT-5 Response**: Evaluation score: 0.4
<points>(x='470', y='305', alt='Sella (S): center of hypophyseal fossa')</points>
<points>(x='640', y='280', alt='Nasion (N): frontonasal suture')</points> …
Estimated CVM stage: CS6 (post-pubertal, growth complete).
Reasoning: Clear concavities are present on the inferior borders of C2, C3, and C4, and the bodies of C3–C4 appear more rectangular/vertical than wide, consistent with CS6 rather than CS5.

**Lingshu-7B Response:** Evaluation score: 0.3
CVM staging: Based on the analysis of the cervical vertebrae, the estimated Cervical Vertebral Maturation (CVM) stage appears to be advanced, indicating a more mature skeletal development.

**OralGPT-Omni Response:** Evaluation score: 1.0
In this lateral cephalometric radiograph, 29 landmarks were detected:
<points x='578' y='460' alt='A-point'></points>,
<points x='562' y='594' alt='B-point'></points>, …
Based on the morphological characteristics of the cervical vertebrae (C2–C4) derived from these landmarks, the cervical vertebral maturation (CVM) stage was estimated. The analysis considered vertebral body shape, height-to-width ratios, and the degree of concavity at the lower borders.
The prediction indicates that the patient is at CVM Stage S5.

**Clinical validity assessment from a radiologist with over ten years of experience:**

| | | |
|---|---|---|
| The output from GPT-5 is not entirely accurate. While it provided estimated x, y, and z coordinates for more than 20 anatomical landmarks and showed relatively reliable reasoning, the final CVM staging was determined as stage S6. This is close but not correct, as the ground truth indicates stage S5. Therefore, the overall accuracy is considered moderate. | The output from Lingshu-7B is not entirely accurate. It provided a general estimation of the CVM stage as advanced but did not specify the exact staging classification. Therefore, the overall accuracy is considered moderate. | The output from OralGPT-Omni is comprehensive and accurate. It provided estimated x, y, and z coordinates for nearly 30 anatomical landmarks and showed reliable reasoning. The final CVM staging was determined as stage S5, which aligns with the ground truth. Therefore, the overall accuracy is considered high. |

Figure 36. Case study and clinical validity on the cephalometric radiograph modality, with correct responses highlighted in green and obvious incorrect responses highlighted in red.