
Towards Better Dental AI: A Multimodal Benchmark and Instruction Dataset for Panoramic X-ray Analysis

Jing Hao^{1*} Yuxuan Fan^{2*} Yanpeng Sun³ Kaixin Guo¹ Lizhuo Lin¹ Jinrong Yang^{4,5}
Qi Yong H. Ai⁶ Lun M. Wong⁷ Hao Tang^{8†} Kuo Feng Hung^{1†}

¹Faculty of Dentistry, The University of Hong Kong

²The Hong Kong University of Science and Technology (GZ)

³National University of Singapore ⁴CVTE ⁵Sun Yat-sen University

⁶Department of Diagnostic Radiology, The University of Hong Kong

⁷Imaging and Interventional Radiology, Faculty of Medicine, The Chinese University of Hong Kong

⁸School of Computer Science, Peking University



<https://github.com/isbrycee/OralGPT>



<https://huggingface.co/datasets/OralGPT/MMOral-OPG-Bench>

Abstract

Recent advances in large vision-language models (LVLMs) have demonstrated strong performance on general-purpose medical tasks. However, their effectiveness in specialized domains such as dentistry remains underexplored. In particular, panoramic X-rays, a widely used imaging modality in oral radiology, pose interpretative challenges due to dense anatomical structures and subtle pathological cues, which are not captured by existing medical benchmarks or instruction datasets. To this end, we introduce **MMOral**, the first large-scale multimodal instruction dataset and benchmark tailored for panoramic X-ray interpretation. MMOral consists of 20,563 annotated images paired with 1.3 million instruction-following instances across diverse task types, including attribute extraction, report generation, visual question answering, and image-grounded dialogue. In addition, we present **MMOral-Bench**, a comprehensive evaluation suite covering five key diagnostic dimensions in dentistry. We evaluate 64 LVLMs on MMOral-Bench and find that even the best-performing model, *i.e.*, GPT-4o, only achieves 43.31% accuracy, revealing significant limitations of current models in this domain. To promote the progress of this specific domain, we provide the supervised fine-tuning (SFT) process utilizing our meticulously curated MMOral instruction dataset. Remarkably, a single epoch of SFT yields substantial performance enhancements for LVLMs, *e.g.*, Qwen2.5-VL-7B demonstrates a 24.73% improvement. MMOral holds significant potential as a critical foundation for intelligent dentistry and enables more clinically impactful multimodal AI systems in the dental field.

1 Introduction

Recent advancements in large vision-language models (LVLMs) [33, 41, 12, 8, 9, 45, 62, 55, 60, 67, 68, 43, 70, 18] have driven significant progress in open-world medical image understanding, supported by benchmarks such as GMAI-MMBench [66], RadBench [64], and OmniMedVQA [30], as well as

*Equal contribution.

†Corresponding authors: hungkfg@hku.hk, haotang@pku.edu.cn

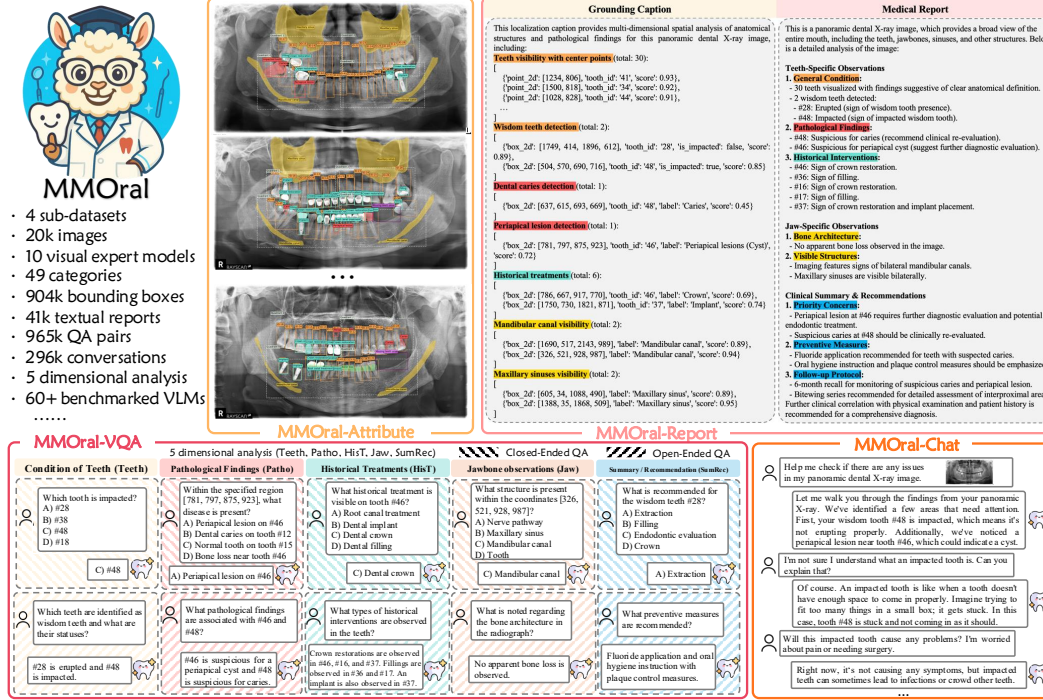


Figure 1: Overview of the MMOral. It consists of four sub-datasets: MMOral-Attribute, MMOral-Report, MMOral-VQA, and MMOral-Chat. MMOral-Attribute contains a total of 49 categories of anatomical structures within panoramic X-rays. MMOral-Report consists of two types of textual descriptions: the grounding caption and the medical report. MMOral-VQA includes closed-ended and open-ended QA pairs spanning five diagnostic dimensions. MMOral-Chat simulates the dialogue process between patients and radiology experts regarding the interpretation of panoramic X-rays.

models like LLaVA-Med [34], HuatuoGPT-Vision [15], MedDr [27], HealthGPT [38], and so on [47, 24, 46, 37]. These efforts focus on broad, general-purpose medical scenarios, aiming to evaluate and improve LVLMs across diverse modalities and tasks. However, these general-purpose benchmarks overlook the unique requirements of domain-specific medical fields. In particular, oral radiology—a critical specialty relying on dental imaging for diagnosis and treatment planning—remains largely absent from existing medical benchmarks. The panoramic X-ray is one of the most commonly used imaging modalities and has been widely accepted as a primary source of information for assessing oral health [58, 17]. It provides a comprehensive visualization of all teeth and surrounding structures in a single image, enabling a basic evaluation of dentition, periodontal bone loss, and lesions within the jawbones. The omission of this modality leaves a significant gap: the lack of tailored evaluation and instruction resources hampers the adaptation of LVLMs to dentistry-specific tasks.

Unlike other modalities, interpreting panoramic X-rays presents unique challenges, characterized by dense anatomical structures and fine-grained pathological cues. Addressing these challenges requires not only dental-specific instruction data but also a specialized benchmark aligned with the clinical knowledge of dental practitioners. To bridge this gap, we introduce **MMOral**, the first large-scale multimodal instruction dataset and benchmark tailored for panoramic X-ray understanding. MMOral comprises 20,563 annotated panoramic X-rays paired with 1.3 million instruction-following instances, spanning multiple task formats including attribute extraction, report generation, visual question answering, and image-grounded dialogue. Complementing the dataset, **MMOral-Bench** offers a curated evaluation suite covering five key diagnostic dimensions, including the condition of teeth, pathological findings, historical treatments, jawbone observations, and clinical summary & recommendations. This benchmark consists of 100 images, paired with 491 closed-ended and 578 open-ended questions. All cases in MMOral-Bench are manually chosen and checked from the MMOral to ensure their quality and reliability. Together, MMOral and MMOral-Bench lay a critical foundation for advancing intelligent dentistry and enabling clinically meaningful multimodal AI.

¹<https://platform.stepfun.com/>

²<https://www.volcengine.com/product/doubao/>

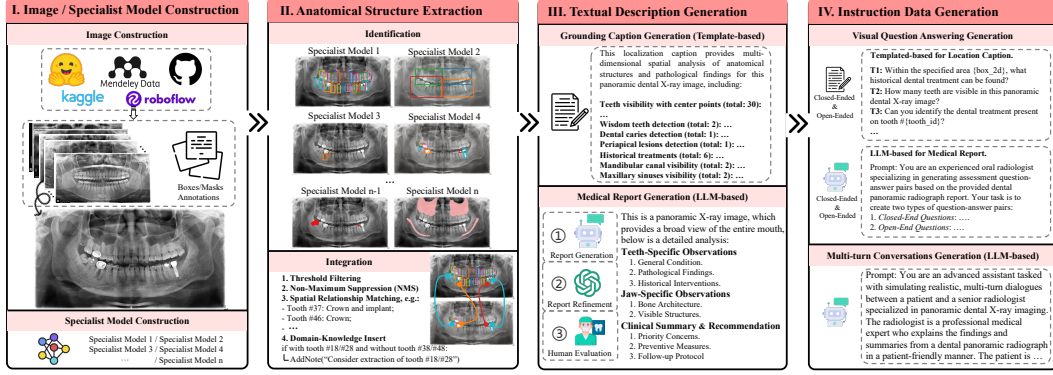


Figure 2: The MMOral dataset curation pipeline, which consists of four sequential steps.

We assess 53 publicly available LVLs (44 general-purpose and 9 medical-specific models) as well as 11 advanced proprietary LVLs such as GPT-4o [31], GPT-4V [31], Claude-3-7-Sonnet [11], Gemini-2.0-Flash [53], Gemini-2.5-Flash [53], Qwen-Max-VL [12], Step-1o series¹, and Doubao-1.5 series² on our Oral-Bench. We summarize five findings according to the evaluation outcomes: (1) MMOral-Bench is a significant challenge for current LVLs, even for GPT-4o, which achieves only a 43.31% average score. (2) The performance of existing universal medical LVLs is suboptimal in the field of dental applications. (3) Existing medical LVLs show no clear advantage over general-purpose LVLs for dentistry tasks. (4) Nearly all LVLs perform worse on open-ended questions compared to closed-ended questions in the MMOral-Bench. (5) LVLs exhibit a clear performance bias across five diagnostic dimensions (depending on the focus of anatomical structures), and they show relatively limited capability when it comes to fine-grained teeth-related questions.

To further validate the effectiveness of MMOral in enhancing the instruction-following capabilities of LVLs and promote the progress of intelligent dentistry, we conduct extensive supervised fine-tuning (SFT) experiments on the Qwen2.5-VL-7B model [12]. Experimental results show that the average score on MMOral-Bench could improve by 24.73% when conducting SFT utilizing MMOral instruction data (MMOral-Report, MMOral-VQA, MMOral-Chat) for one epoch. This obvious improvement highlights the value of the MMOral towards intelligent dentistry.

To summarize, our contributions are threefold: (1) We introduce MMOral, the first large-scale multimodal instruction dataset and benchmark tailored for panoramic X-ray understanding. MMOral-Bench offers a curated evaluation suite covering five key diagnostic dimensions, which could comprehensively reflect the capabilities of current LVLs in the dental field. (2) 64 existing representative LVLs are assessed on MMOral-Bench, including 11 proprietary models and 53 open-source models (44 general-purpose models and 9 medical-specific models). Evaluation results pave the way for the next optimization direction to enhance the interpretation of panoramic X-rays. (3) We implement supervised fine-tuning using our MMOral instruction data to verify its effectiveness. Experiments demonstrate the average score on MMOral-Bench could improve by 24.73% when fine-tuning the Qwen2.5-VL-7B model for only one epoch.

2 MMOral Dataset Curation

The MMOral dataset curation pipeline consists of four sequential steps, which are shown in Figure 2.

2.1 Images and Visual Specialists Construction

We collect panoramic X-ray images from two publicly available datasets: the TED3 dataset [26] and the dataset proposed by Hoang Viet Do [19]. After filtering out duplicate images, a total of 20,563 images is obtained. Subsequently, we build ten visual specialist models to simulate the interpretative process of oral radiology experts. These models are designed to recognize 49 categories of anatomical structures in radiographic images, encompassing visual elements such as tooth numbering (1 to 32 according to the FDI system [63]), four quadrants, pathological findings, historical treatments, bone loss, and visible bone structures. The category spaces of multiple visual specialist models overlap with each other, ensuring the reliability of anatomical structure extraction. Additionally, we utilize an

off-the-shelf OpenOCR model [20] to detect the acquisition time displayed within the images. More details of image source information, visual specialist model construction, and the specific category list of detected anatomical structures can be found in *Appendix B.2*.

2.2 Anatomical Structure Extraction

After obtaining the extracted anatomical structures from all visual specialist models, we further integrate these outcomes for two primary considerations: **(1) the necessity to deduplicate overlapping categories detected by multiple visual specialist models**, and **(2) the discrete nature of identified anatomical structures lacking interrelational information**. First, we meticulously design the anatomical structures post-processing pipeline to remove redundant information. Second, since the associations between dental pathological findings, historical treatment, and their corresponding tooth numbering systems remain undefined, we establish these correlations through their spatial relationships. Third, domain-specific clinical knowledge is inserted. For instance, extraction of maxillary third molars (teeth #18/28) is recommended when they lack opposing mandibular counterparts (teeth #48/38). The developed anatomical structure post-processing and relationship matching pipeline comprises eight systematic steps, as detailed in *Appendix B.2*.

2.3 Report Generation

We synthesize the discrete anatomical structure information into two coherent textual outputs: a **grounding caption** and a **medical report**. The grounding caption contains detailed coordinates, categories, and confidence scores of all anatomical structures, providing a multi-dimensional spatial observation and identification for panoramic X-rays. All anatomical information is systematically organized into structured textual descriptions following manually designed rules, with detailed explanations provided in *Appendix B.3*. An illustrative example of the grounding caption is shown in the left panel of the MMOral-Report part in Figure 1.

In contrast to grounding captions, medical reports focus on providing a condensed summary of key anatomical structures, abnormal findings, and corresponding diagnostic recommendations. Through extensive consultations with senior dental specialists, we structure the medical report into three principal sections: **Teeth-Specific Observations**, **Jaw-Specific Observations**, and **Clinical Summary & Recommendations**. The medical report is derived from two-stage LLM-based generation. First, we prompt DeepSeek-R1-Distill-Llama-70B³ to generate medical reports based on grounding captions. Following that, we manually check the generated medical reports and summarize several common errors, which are presented in detail in *Appendix B.3*. According to these errors, we carefully prepare the prompt for report correction and instruct GPT-4-turbo [10] to simultaneously output both revised reports and corresponding revision logs. By examining these revision logs, we can efficiently identify modified sections of the reports, thereby facilitating quality verification of the revised content. To comprehensively evaluate the quality of the revised reports, two professional dentists are engaged to conduct the human evaluation, which will be discussed in section 3.2. The details of prompts in two-stage generation are provided in *Appendix B.2*.

2.4 Instruction Data Generation

Based on the generated grounding captions and medical reports, we construct two types of single-turn instruction-following QA pairs—closed-ended and open-ended—using template-based and LLM-based approaches. For grounding captions, we generate both closed-ended multiple-choice QA pairs and open-ended QA pairs using manually designed question templates. The incorrect choices in the closed-ended QA are created by introducing random perturbations (*Appendix B.4*) to the ground truth for enhancing the model’s ability to recognize and understand subtle differences. For the medical report, we prompt GPT-4-turbo (see *Appendix B.3* for prompt designs) to simultaneously generate both closed-ended and open-ended QA pairs. To systematically evaluate panoramic X-ray analysis capabilities, we establish a taxonomy across five clinically grounded dimensions: **condition of teeth (Teeth)**, **pathological findings (Patho)**, **historical treatments (HisT)**, **jawbone observations (Jaw)**, and **clinical summary & recommendation (SumRec)**. Each QA pair is mapped to one or more of these diagnostic categories based on its clinical intent, forming a multi-dimensional analysis.

³<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>

Table 1: The brief description of four sub-datasets in MMOral and their corresponding data size.

Dataset	Sub-Dataset	Description	Size
MMOral	MMOral-Attribute	The category, position and correlation of anatomical structures shown in the panoramic X-ray image.	904k
	MMOral-Report	Two types of textual description for each panoramic X-ray image: grounding caption and medical report.	41k
	MMOral-VQA	Two types of visual question answering: closed-ended QA and open-ended QA.	965k
	MMOral-Chat	The multi-turn conversation between the assistant and a person asking questions about the panoramic X-ray.	296k

In addition to the single-turn QA pairs mentioned above, we also prompt GPT-4-turbo to generate a multi-turn conversation between the assistant and a person asking questions about the panoramic X-ray. The answers are in a tone as if the assistant is seeing the panoramic radiograph and explaining the findings and summaries in a patient-friendly manner. A diverse set of questions is asked about the visual content of the image and the assistant’s explanations, and only questions that have definite answers are considered. Please see *Appendix B.4* for the detailed prompt.

3 MMOral Dataset Analysis

3.1 Data Statistics

The MMOral comprises 20,563 images paired with 1.3 million instruction-following data instances, establishing a comprehensive multimodal resource for human-AI interaction research in the digital dentistry field. It consists of four distinct sub-datasets: MMOral-Attribute, MMOral-Report, MMOral-VQA, and MMOral-Chat. Each sub-dataset corresponds to specific tasks, including visual perception, report generation, visual question answering, and image-grounded dialogue. Table 1 provides detailed information on each component and its corresponding data size. Notably, a single panoramic X-ray contains an average of 44 bounding boxes, reflecting its structural diversity and dense representations, making its interpretation highly complex. Figure 3 (a) presents the category distribution of MMOral-Attribute, comprising 5 major categories and 49 subcategories of anatomical structures identified in panoramic X-rays, illustrating its diversity and comprehensive coverage. To the best of our knowledge, MMOral is the largest multimodal dataset for panoramic X-rays to date, forming a robust foundation for the development and evaluation of LVLs.

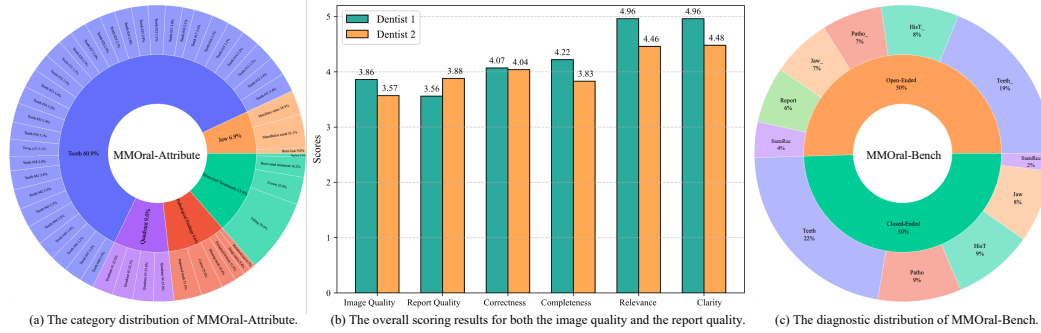


Figure 3: The data statistic distribution and human evaluation results.

3.2 Strategies for Ensuring Data Quality

We ensure the quality of the MMOral from the following three perspectives:

(1) Collaborative validation of anatomical structures by multiple specialist models. The anatomical structures present in panoramic X-rays serve as metadata for subsequent report generation and instruction data construction. If these visual structures contain errors, such inaccuracies will propagate throughout all datasets. To ensure precise detection of anatomical structures, we construct ten visual specialist models with overlapping category spaces. For instance, ten structures (e.g., caries, periapical lesion, impacted teeth, missing teeth area, filling, implant, root canal treatment, crown,

mandibular canal, and maxillary sinus) are validated by two or more visual specialists, with the final results obtained through post-processing. This approach ensures the reliability of the extraction of anatomical structures, providing trustworthy metadata for subsequent medical report generation and instruction data construction.

(2) Two-stage LLM-based report generation. Unlike most existing approaches that rely on a single LLM to generate image captions, we adopt a two-stage scheme of generation followed by correction to ensure report quality. Due to the complexity of generating structured medical reports based on discrete anatomical structures, LLMs inevitably introduce errors in areas such as structural organization, content description, and hallucinations. By manually reviewing the preliminary reports generated by DeepSeek-R1-Distill-Llama-70B, we identify common errors and summarize them into 8 rules for prompting GPT-4-turbo to revise the reports. Analysis of the revision logs revealed that **95.45%** of the reports are successfully corrected, significantly improving their overall quality.

(3) Professional dentist evaluation. We invite two professional dentists to evaluate the medical report and assess whether the associated panoramic X-ray image meets practical imaging quality standards. Both the image quality and the report quality are evaluated using a scoring system with five levels, ranging from 1 to 5, representing a progression from "very dissatisfied" to "very satisfied." For the report evaluation, in addition to assigning an overall score, the evaluators are required to provide further scores in four specific aspects: correctness, completeness, relevance, and clarity. Figure 3 (b) illustrates the overall scoring results for both the image quality and the report quality, showing that the average scores for all aspects exceeded 3.5, highlighting the high quality of MMOral. The evaluation guidelines are detailed in Appendix.

4 MMOral-Bench

4.1 Benchmark Construction

We construct MMOral-Bench by curating 491 closed-ended and 578 open-ended QA pairs with 100 images **through significant manual selection and validation**. To ensure image quality, we select images from the dataset proposed by Hoang Viet Do [19] because its acquisition process is clearer and more reliable. Moreover, we filter out QA pairs that could not be answered with the image, and incorrect answers are identified and re-annotated. MMOral-Bench covers five different clinically grounded dimensions (e.g., Teeth, Patho, HisT, Jaw, SumRec) and thus can comprehensively evaluate the ability of LVLMs to understand and interpret panoramic X-rays. Each QA pair is assigned to one or more diagnostic dimensions based on its clinical intent, enabling a multi-dimensional analysis. The distribution of the diagnostic dimension on MMOral-Bench can be seen in Figure 3(c), and some QA examples are present in the MMOral-VQA part of Figure 1.

4.2 Evaluation Metrics

We adopt two evaluation metrics tailored for closed-ended and open-ended questions, respectively. For closed-ended questions, we use accuracy as the evaluation metric. For open-ended questions, following previous works [71, 72], we construct a few-shot prompt and leverage GPT-4-turbo to assist with the evaluation. The few-shot prompt incorporates nine in-context examples with open-ended answers, covering fully correct cases, partially correct cases, and incorrect cases. GPT-4-turbo assigns a score ranging from 0 to 1 based on each sample’s input question, ground truth, and model output. We report the evaluation scores for each category as well as the overall performance. The full details of the few-shot prompt can be found in Appendix E.2. We integrate the evaluation of MMOral-Bench into the standard VLMEvalKit [21] framework, thereby facilitating subsequent capability assessments of newly developed LVLMs⁴.

5 Experiments

5.1 Experimental Setups

Benchmarked LVLMs. We conduct zero-shot evaluations across 64 LVLMs on our MMOral-bench, and we pick 36 out of 64 models for demonstration in the main text; additional results are provided in

⁴<https://anonymous.4open.science/r/MMOral-2459/README.md>

Table 2: Results on MMOral-Bench for existing various LVLMS across both closed-ended and open-ended VQA tasks. The best-performing model in each category is highlighted **in bold**, while the second-best is underlined. 36 out of 64 models for demonstration in the main text; additional results are provided in the *supplementary material*.

Model	Close-ended VQA						Open-ended VQA						Avg.	
	Teeth	Patho	His	Jaw	Summ	Overall	Teeth	Patho	His	Jaw	Summ	Report		Overall
Proprietary LVLMS														
GPT-4o-2024-11-20 [31]	36.16	41.14	36.48	47.85	57.89	39.92	41.99	27.20	41.96	66.00	41.79	60.40	46.70	43.31
GPT-4V [31]	38.21	35.76	43.42	52.79	50.00	42.57	31.46	23.79	39.51	69.81	34.29	43.70	39.38	40.98
Claude-3-7-Sonnet-20250219 [11]	39.61	34.78	42.76	34.66	57.89	39.10	41.07	37.88	51.84	62.69	37.86	42.70	45.31	42.21
Gemini-2.5-Flash-preview-04-17 [53]	16.40	12.99	23.94	29.43	20.00	19.14	28.04	24.77	31.90	47.81	12.98	16.70	27.84	23.49
Gemini-2.0-Flash [53]	42.42	37.01	51.41	41.89	53.33	42.97	36.19	27.73	47.06	62.88	33.57	45.50	41.82	42.40
Qwen-Max-VL-2025-04-08 [12]	24.42	27.27	33.10	34.34	62.22	29.94	2.10	4.47	7.06	11.62	7.98	5.50	5.29	17.62
Doubao-1-5-thinking-vision-pro-250428 ⁵	31.55	20.78	41.55	53.96	40.00	36.05	41.26	23.94	43.93	60.75	32.74	53.40	43.90	39.97
Open-Source LVLMS														
Deepseek-VL-7b-chat [43]	29.41	31.82	33.33	51.70	42.22	35.44	16.48	7.50	13.44	34.56	9.52	9.60	15.95	25.70
Emu3-chat [62]	39.75	31.17	38.73	49.81	42.22	40.12	17.03	6.36	12.88	36.06	14.76	19.60	18.13	29.13
Qwen2.5-VL-72B [13]	24.60	24.68	27.16	26.79	42.22	25.66	13.90	15.83	15.40	27.12	7.38	11.50	15.38	20.52
CogVLM2-19B [60]	29.77	29.87	43.83	41.13	55.56	35.23	29.29	18.03	27.79	52.19	15.83	27.10	29.81	32.52
GLM-4V-9B [22]	23.71	28.57	34.57	62.64	44.44	34.83	20.94	9.70	18.77	26.62	12.74	21.30	19.74	27.29
LLaVA-NeXT-13B-hf [41]	30.09	32.92	30.54	38.20	60.42	33.80	14.48	10.28	9.23	22.41	14.30	21.30	15.43	24.62
LLaVA-OneVision [33]	34.05	40.26	37.04	63.02	73.33	42.77	28.79	13.64	30.12	55.12	20.36	23.20	29.43	36.10
LLaMA-3.2-Vision-11B-Instruct [23]	31.37	27.27	35.19	52.83	60.00	36.46	27.42	21.89	26.63	41.06	22.14	21.40	27.02	31.74
Cambrian-34B [57]	36.90	37.01	45.06	60.75	55.56	44.60	30.22	23.11	32.57	47.90	11.68	17.03	27.09	35.85
Phi-4-multimodal-instruct [9]	33.51	27.92	40.74	46.79	42.22	37.07	28.52	21.21	27.85	44.06	12.98	14.60	26.19	31.63
InternVL3-38B [16]	26.56	22.08	22.22	33.58	28.89	26.48	34.77	19.17	30.18	47.69	17.74	40.10	34.05	30.27
Chameleon-7B [44]	28.34	33.12	33.95	31.70	46.67	31.98	5.21	3.18	10.86	14.00	5.12	8.80	7.65	19.82
PaliGemma-3B [14]	23.89	20.78	27.78	39.25	28.89	27.70	10.73	12.05	10.25	14.00	6.07	0.20	9.12	18.41
MiniCPM-O2.6 [65]	32.09	16.88	22.84	31.32	24.44	27.09	19.98	13.94	15.09	28.56	13.10	33.40	21.60	24.35
Kosmos-2 [50]	14.26	20.78	24.07	12.83	13.33	17.11	15.68	14.77	8.10	20.56	12.50	5.40	13.13	15.12
mPLUG-Owl3-7B [67]	29.77	27.27	30.86	51.32	53.33	34.83	17.17	13.94	11.23	35.06	8.81	11.20	16.83	25.83
Gemma3-12B [54]	22.28	12.99	25.93	33.58	35.56	23.83	28.38	22.73	23.56	34.44	23.93	31.70	28.22	26.03
XComposer2-VL-7B [75]	23.17	25.32	25.93	21.89	26.67	24.44	10.75	6.59	10.74	14.25	7.02	2.40	9.00	16.72
Molmo-72B-0924 [18]	35.65	31.17	27.78	23.02	31.11	31.16	9.82	4.47	6.75	12.81	3.21	11.00	9.03	20.09
Yi-Vision [70]	38.50	44.81	44.37	56.60	53.33	45.01	30.75	18.94	33.13	59.00	24.76	39.90	34.91	39.96
Qwen-QVQ-72B [56]	28.70	22.08	30.99	34.72	13.33	28.31	27.26	13.64	23.01	34.19	13.45	41.90	27.80	28.05
Ovis2-34B [45]	38.15	36.36	43.83	72.45	71.11	47.45	35.21	22.12	37.79	55.31	16.43	32.20	35.03	41.24
Kimi-VL-A3B-Thinking [55]	23.17	23.38	17.28	29.81	35.56	23.42	50.39	37.73	50.18	58.25	45.71	61.50	51.80	37.61
Medical Specific LVLMS														
LLaVA-Med [34]	27.09	29.22	27.46	18.11	24.44	25.66	0.91	1.52	0.00	0.00	0.00	24.50	4.76	15.21
HuatuoGPT-V-34B [15]	30.84	35.71	49.30	56.98	71.11	41.96	35.18	24.92	36.32	65.69	25.48	24.80	36.02	38.99
HealthGPT-XL32 [38]	41.71	42.21	59.15	69.43	66.67	51.53	30.64	25.83	27.98	51.12	17.02	8.00	27.80	39.67
MedVLM-R1 [47]	25.13	36.36	37.32	49.81	46.67	35.23	22.42	13.71	24.42	43.88	13.57	25.80	24.70	29.97
MedDr [27]	26.56	21.43	24.65	38.49	22.22	27.29	22.99	32.58	29.57	52.44	20.95	8.70	26.20	26.75

⁵ <https://www.volcengine.com/product/doubao/>

Appendix E.3. We evaluate 8 proprietary LVLMS via API: GPT-4o [31], GPT-4V [31], Claude-3-7-sonnet [11], Gemini-2.5-Flash [53], Gemini-2.0-Flash [53], Qwen-Max-VL [12], Step-1o-turbo, and Doubao-1-5-thinking-vision-pro. For medical-specific LVLMS, we test 5 powerful models, including LLaVA-Med [34], HuatuoGPT-V [15], MedVLM-R1 [47], MedDr [27], and HealthGPT [38]. We also evaluate 23 representative general-purpose LVLMS: Deepseek-VL-7b [43], Emu3 [62], Qwen2.5-VL-72B [13], CogVLM2-19B [60], GLM-4V-9B [22], LLaVA-NeXT-13B-hf [41], LLaVA-OneVision [33], LLaMA-3.2-Vision-11B [23], Cambrian-34B [57], Phi-4-multimodal-instruct [9], InternVL3-38B [16], Chameleon-7B [44], PaliGemma-3B [14], MiniCPM-O2.6 [65], Kosmos-2 [50], Kimi-VL-A3B-Thinking [55], Ovis2-34B [45], Qwen-QVQ-72B [56], Gemma3-12B [54], XComposer2-VL-7B [75], Molmo-72B [18], and Yi-VL-34B [70].

Supervised Fine-Tuning. We also implement supervised fine-tuning (SFT) on two popular LVLMS models with different scales (Qwen-2.5-VL-7B [12] and LLaVA-Next-13B [41]) using our MMOral instruction data to verify its effectiveness. The extensive experiments are implemented through the LLaMA-Factory framework [76] while maintaining default hyperparameters, with all models being trained for a single epoch. The results on Qwen-2.5-VL-7B will be discussed in Sec. 5.3, and the results on LLaVA-Next-13B can be found in *Appendix F*.

5.2 Evaluation Results

Following a comprehensive review of the evaluation outcomes, which are shown in Table 10, we have identified 5 key findings regarding the performance of existing LVLMS in the dental domain:

Finding 1. The MMOral-Bench poses significant challenges to ALL LVLMS. Even the most advanced model, GPT-4o, only achieves 43.31% overall performance, highlighting persistent challenges and fundamental limitations in current LVLMS’ capacity to interpret complex panoramic X-rays, which are characterized by anatomically dense structures and fine-grained pathological patterns. This critical performance gap reveals fundamental limitations of existing LVLMS’ capacity in dental-specific images, underscoring substantial room for improvement.

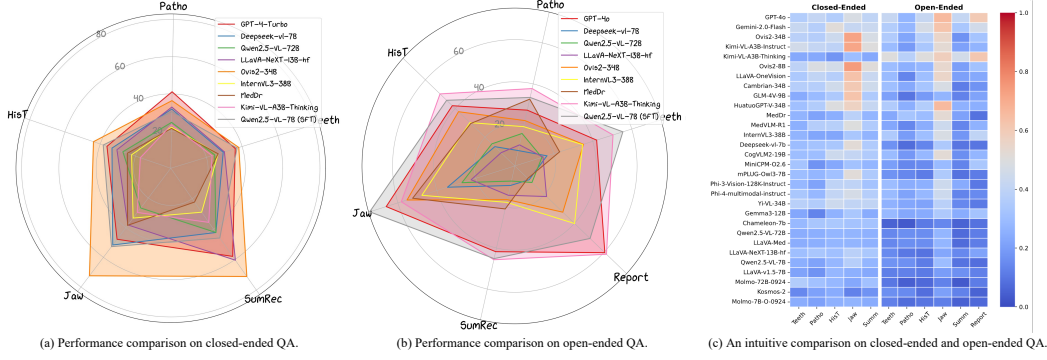


Figure 4: Performance comparison on both closed-ended and open-ended QA across multiple LVLMS.

Finding 2. The performance of existing universal medical LVLMS is suboptimal in the field of dental applications. Current universal medical LVLMS, which predominantly focus on enhancing capabilities for generalized clinical scenarios across diverse medical imaging modalities, have unsatisfactory performance when it comes to understanding panoramic X-rays—a specialized, fine-grained modality within dental imaging. The results reveal that general medical LVLMS achieve less than 40% average accuracy on MMOral-Bench, with HealthGPT-XL32 [38] attaining peak performance at 39.67%. This indicates that current universal medical LVLMS still require further exploration and improvement in their ability to interpret panoramic X-rays, which is an imaging modality characterized by complex and numerous anatomical structures.

Finding 3. Existing medical LVLMS show no significant advantage over general LVLMS in the field of dentistry. Existing medical-specific LVLMS, including the LLaVA-Med series, HuatuoGPT series, MedVLM-R1, MedDr, and HealthGPT, fail to outperform general-purpose models in our MMOral-bench. This indicates that current medical LVLMS lack adequate understanding and analytical capabilities specific to the oral region. Among all medical LVLMS evaluated, HealthGPT demonstrates the best performance, achieving an average score of 39.67%. However, this score remains lower than that of general-purpose open-source models such as the Ovis2 series and commercial models like GPT-4o and Claude-3-7-Sonnet. These results highlight the need for further improvements in medical-specific AI models to enhance their understanding of the oral region—an area intrinsically linked to essential human functions such as eating and speaking.

Finding 4. Nearly all LVLMS perform worse on open-ended questions compared to closed-ended questions in the MMOral-Bench. Figure 4 shows the performance comparison on both closed-ended and open-ended QA tasks. Although some LVLMS (e.g., LLaVA-NeXT-8B-hf, DeepSeek-VL-7B, Ovis2-34B) perform relatively well in closed-ended VQA tasks, they exhibit a significant performance drop in open-ended VQA tasks. Moreover, the proportion of open-sourced models with an overall score below 25% in open-ended VQA tasks is as high as 50.9% (33 out of 53 models). This highlights the current limitations of LVLMS in handling open-ended answer generation for dentistry-related questions and the urgent need for targeted optimization.

Finding 5. LVLMS demonstrate a clear bias across various question categories, depending on the focus of anatomical structures involved. The questions are categorized into five dimensions based on their focus as mentioned in Sec. 2.4: ‘Teeth’, ‘Patho’, ‘HisT’, ‘Jaw’, ‘SumRec’. We observe that current LVLMS perform relatively well on the ‘Jaw’ category—both in closed-ended and open-ended tasks—where the focus is on larger anatomical structures such as bone loss, mandibular canals, and maxillary sinuses. In contrast, the models generally exhibit poorer performance on categories that require fine-grained visual understanding, such as ‘Teeth’, ‘HisT’, and ‘Patho’. This suggests that current LVLMS still need significant improvement in their ability to perform fine-grained dental visual understanding and reasoning.

Table 3: The effectiveness verification of MMOral instruction data by supervised fine-tuning.

Model	SFT			Close-ended VQA						Open-ended VQA						Avg.	
	Report	VQA	Chat	Teeth	Patho	His	Jaw	Summ	Overall	Teeth	Patho	His	Jaw	Summ	Report		Overall
Qwen2.5-VL-7B [13]	✗	✗	✗	24.96	21.12	27.54	37.08	35.42	27.00	17.01	16.10	11.18	29.41	9.07	8.20	15.92	21.46
Qwen2.5-VL-7B [13]	✓	✗	✗	26.90	27.33	26.35	45.32	37.50	31.00	27.82	15.82	25.92	63.76	22.33	38.00	32.62	31.81
Qwen2.5-VL-7B [13]	✗	✗	✗	39.12	36.65	37.73	62.92	43.75	43.60	36.22	31.92	32.49	78.47	40.93	4.30	35.73	39.67
Qwen2.5-VL-7B [13]	✓	✓	✗	43.19	40.99	43.11	63.60	37.50	46.20	39.85	32.41	35.20	78.06	36.98	36.80	42.85	44.53
Qwen2.5-VL-7B [13]	✓	✓	✓	37.17	30.43	38.32	52.81	45.83	39.60	55.45	33.40	45.74	74.47	45.17	50.50	52.77	46.19

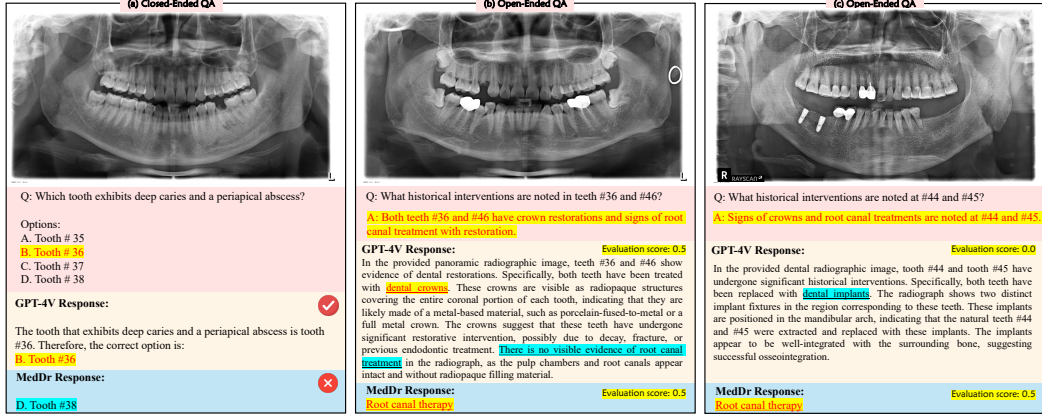


Figure 5: Three examples of case studies on closed-ended QA and open-ended QA, respectively. More examples can be found in the *Appendix*.

5.3 Efficacy Validation of MMOral Instruction Data

We implement supervised fine-tuning (SFT) on the Qwen2.5-VL-7B model using our MMOral instruction data, and the results are presented in Table 11. When using the MMOral-report or MMOral-VQA dataset individually for SFT, the average score on MMOral-Bench improved by 10.35% (from 21.46% to 31.81%) and 18.21% (from 21.46% to 39.67%), respectively. Furthermore, when both MMOral-report and MMOral-VQA are used together for SFT, the average score achieves a more significant improvement, rising from 21.46% to 44.53%. Based on this, incorporating the MMOral-Chat into the SFT process results in an additional 1.66% improvement in the average score. Notably, the model demonstrates significant improvements on open-ended QA tasks when MMOral-Chat is included in SFT, while exhibiting a slight decline in performance on closed-ended QA tasks. We hypothesize that image-grounded conversation data can significantly enhance the model’s instruction-following ability for open-ended questions and improve overall user experience.

5.4 Case Study

To provide a more intuitive demonstration of the current capabilities of LVLMs in understanding panoramic X-rays, we conduct a case study on a high-performing proprietary model (GPT-4-Turbo) and a medical-specific LVLm (MedDr [27]) by analyzing their performance on both closed-ended and open-ended question types. Figure 5 (a) illustrates a closed-ended QA case, where the question requires the LVLm to simultaneously understand tooth numbering according to the FDI standard, detect caries, and identify periapical abscesses. As shown, GPT-4V could successfully answer the question, while MedDr provides an incorrect response. For the open-ended case, both GPT-4V and MedDr demonstrate some level of understanding of panoramic X-rays but fall short of providing comprehensive outcomes. For instance, as depicted in Figure 5 (b), when examining teeth #36 and #46, which both exhibit crown restoration and root canal treatment, GPT-4-Turbo is able to detect the dental crown but explicitly states that no root canal treatment is identified. Conversely, MedDr detects the root canal treatment but overlooks the clearly visible crown restoration. Despite being among the top-performing models on the MMOral-Bench benchmark, GPT-4-Turbo and MedDr still exhibit significant limitations, highlighting substantial room for improvement in their ability to accurately interpret panoramic X-rays. Figure 5 (c) shows the case that GPT-4V misrecognises the tooth #44 and #45, leading to a completely incorrect answer.

Additionally, we observe that some proprietary models, such as Qwen-Max-VL and Qwen-QVQ, commonly refuse to provide answers due to safety concerns stemming from commercial policies. For example, these models often respond with statements like, “*Input data may contain inappropriate content.*” This strict adherence to safety protocols and ethical standards significantly limits their response capabilities in the dental field.

6 Conclusion

We introduce MMOral, the first large-scale multimodal instruction dataset and benchmark tailored for panoramic X-ray understanding. The evaluation outcomes on MMOral-Bench reveal that existing universal medical-specific LVLMs have unsatisfactory performance when it comes to understanding panoramic X-rays characterized by dense anatomical structures and fine-grained pathological cues. We hope that our instruction dataset and benchmark will serve as a pivotal resource for advancing digital dentistry towards more sophisticated and intelligent multimodal AI in oral healthcare.

Limitation and Future Work. MMOral exhibits limitations in imaging modality diversity compared to existing medical instruction datasets. However, panoramic X-rays hold significant clinical value owing to their comprehensive visualization of the entire oral anatomy. Automated interpretation of panoramic X-rays could substantially advance intelligent dental AI. Future efforts will aim to expand coverage to additional oral imaging modalities, including 2D modalities such as periapical X-rays and intraoral photographs, as well as 3D modalities like cone-beam computed tomography (CBCT).

References

- [1] Dentistry computer vision project, 2023. URL <https://universe.roboflow.com/nanyang-technological-university-kdgtt/dentistry-vibir>.
- [2] Dental caries detection computer vision project, 2023. URL <https://universe.roboflow.com/panoramic-xray-images/dental-caries-detection>.
- [3] 2023. URL <https://www.kaggle.com/datasets/reemsalahshehab/dental?select=data.yaml>.
- [4] Dental diseases, 2024. URL <https://www.kaggle.com/datasets/ayaalialnozahyy/dental-diseases/data>.
- [5] vzrad2 computer vision project, 2024. URL <https://universe.roboflow.com/arshs-workspace-radio/vzrad2>.
- [6] 2024. URL <https://universe.roboflow.com/arshs-workspace-radio/vzrad2>.
- [7] 2024. URL <https://www.kaggle.com/datasets/nadaaglan/dental-periapical-x-rayss>.
- [8] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [9] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- [10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [11] Anthropic. Claude 3.7 sonnet system card. URL <https://api.semanticscholar.org/CorpusID:276612236>.
- [12] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- [13] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [14] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschanen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.
- [15] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. H. Chen, X. Wang, R. Zhang, Z. Cai, K. Ji, et al. Huatuoogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.
- [16] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

- [17] T. S. de Oliveira Capote, M. de Almeida Gonçalves, A. Gonçalves, and M. Gonçalves. Panoramic radiography—diagnosis of relevant structures that might compromise oral and general health of the patient. In *Emerging Trends in Oral Health Sciences and Dentistry*. IntechOpen, 2015.
- [18] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.
- [19] H. V. Do, T. N. N. Vo, P. T. Nguyen, T. H. L. Luong, N. G. Cu, and H. S. Le. A dataset of apical periodontitis lesions in panoramic radiographs for deep-learning-based classification and detection. *Data in Brief*, 54:110486, 2024.
- [20] Y. Du, Z. Chen, H. Xie, C. Jia, and Y.-G. Jiang. Svtrv2: Ctc beats encoder-decoder models in scene text recognition. *CoRR*, abs/2411.15858, 2024. URL <https://arxiv.org/abs/2411.15858>.
- [21] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. 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.
- [22] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.
- [23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [24] Y. Guo and W. Huang. Llava-next-med: Medical multimodal large language model. In *2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC)*, pages 474–477. IEEE, 2025.
- [25] I. E. Hamamci, S. Er, E. Simsar, A. E. Yuksel, S. Gultekin, S. D. Ozdemir, K. Yang, H. B. Li, S. Pati, B. Stadlinger, et al. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. *arXiv preprint arXiv:2305.19112*, 2023.
- [26] J. Hao, Y. Zhu, L. He, M. Liu, J. K. H. Tsoi, and K. F. Hung. T-mamba: A unified framework with long-range dependency in dual-domain for 2d & 3d tooth segmentation. *arXiv preprint arXiv:2404.01065*, 2024.
- [27] S. He, Y. Nie, Z. Chen, Z. Cai, H. Wang, S. Yang, and H. Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints*, pages arXiv–2404, 2024.
- [28] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [29] Z. He, Y. Wang, A. Yan, Y. Liu, E. Y. Chang, A. Gentili, J. McAuley, and C.-N. Hsu. Medeval: a multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. *arXiv preprint arXiv:2310.14088*, 2023.
- [30] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao, and P. Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.
- [31] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [32] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- [33] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [34] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. 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.

- [35] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023.
- [36] S. Li, T. Lin, L. Lin, W. Zhang, J. Liu, X. Yang, J. Li, Y. He, X. Song, J. Xiao, et al. Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. *arXiv preprint arXiv:2504.13650*, 2025.
- [37] T. Li, Y. Su, W. Li, B. Fu, Z. Chen, Z. Huang, G. Wang, C. Ma, Y. Chen, M. 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.
- [38] T. Lin, W. Zhang, S. Li, Y. Yuan, B. Yu, H. Li, W. He, H. Jiang, M. Li, X. 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.
- [39] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
- [40] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [41] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [42] M. Liu, W. Hu, J. Ding, J. Xu, X. Li, L. Zhu, Z. Bai, X. Shi, B. Wang, H. Song, et al. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*, 7(4):1116–1128, 2024.
- [43] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [44] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023.
- [45] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.
- [46] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
- [47] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.
- [48] K. Panetta, R. Rajendran, A. Ramesh, S. P. Rao, and S. Agaian. Tufts dental database: a multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE journal of biomedical and health informatics*, 26(4):1650–1659, 2021.
- [49] J. Park, S. Kim, B. Yoon, J. Hyun, and K. Choi. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. *arXiv preprint arXiv:2408.16213*, 2024.
- [50] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [51] Z. Qin, Y. Yin, D. Campbell, X. Wu, K. Zou, Y.-C. Tham, N. Liu, X. Zhang, and Q. Chen. Lmod: A large multimodal ophthalmology dataset and benchmark for large vision-language models. *arXiv preprint arXiv:2410.01620*, 2024.
- [52] Y. Sun, H. Wu, C. Zhu, S. Zheng, Q. Chen, K. Zhang, Y. Zhang, D. Wan, X. Lan, M. Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer, 2024.
- [53] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

- [54] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [55] K. Team, A. Du, B. Yin, B. Xing, B. Qu, B. Wang, C. Chen, C. Zhang, C. Du, C. Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [56] Q. Team. Qvq: To see the world with wisdom, December 2024. URL <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- [57] P. Tong, E. Brown, P. Wu, S. Woo, A. J. V. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.
- [58] N. Turosz, K. Chęcińska, M. Chęciński, A. Brzozowska, Z. Nowak, and M. Sikora. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. *Dentomaxillofacial Radiology*, 52(7):20230284, 2023.
- [59] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [60] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, S. XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.
- [61] X. Wang, F. Wang, Y. Li, Q. Ma, S. Wang, B. Jiang, C. Li, and J. Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. *arXiv preprint arXiv:2410.00379*, 2024.
- [62] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [63] T. Wong and J.-L. Eiselé. Fdi world dental federation: Responding to new realities of oral health, 2015.
- [64] C. Wright and P. Reeves. Radbench: benchmarking image interpretation skills. *Radiography*, 22(2): e131–e136, 2016.
- [65] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [66] J. Ye, G. Wang, Y. Li, Z. Deng, W. Li, T. Li, H. Duan, Z. Huang, Y. Su, B. Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.
- [67] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.
- [68] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13040–13051, 2024.
- [69] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.
- [70] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, G. Wang, H. Li, J. Zhu, J. Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [71] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [72] W. Yu, Z. Yang, L. Ren, L. Li, J. Wang, K. Lin, C.-C. Lin, Z. Liu, L. Wang, and X. Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.
- [73] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

- [74] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [75] P. Zhang, X. Dong, Y. Zang, Y. Cao, R. Qian, L. Chen, Q. Guo, H. Duan, B. Wang, L. Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.
- [76] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our motivation, contribution, and scope are clearly described in both the abstract and the introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We describe our limitations and future work in Sec. 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: Our paper aims to construct a multimodal instruction dataset and benchmark, which are not related to the theoretical assumption/proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We ensure that all experimental results are fully reproducible. For all benchmarked LVLMS evaluated on MMOral-Bench, we provide standardized evaluation results based on the EvalVLMKit framework. Additionally, we offer the complete evaluation code. For the SFT experiments, we provide the training details and hyperparameter settings in Sec. 5.1, ensuring the reproducibility of our experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided **ALL** the code for dataset instructions. In addition, we have also provided the designed prompts used for prompting large language models. The evaluation suite for our MMOral-Bench is built upon the standardized library EvalVLMKit, which is also open-sourced, ensuring the reproducibility of the experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: **All** training and test details are explicitly specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Given that our benchmark involves heavy use of costly proprietary models (GPT-4-turbo) for evaluations, we did not repeat the experiments. We provide all our prompts and evaluation code at <https://github.com/isbrycee/OralGPT/tree/main/MMOral-Omni-Bench-Eval>

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have provided the specific experimental requirements in Appendix H, including the financial cost of the API we used and the computational resources for experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: All the research presented in this paper complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The proposed MMOral holds significant potential as a critical foundation for intelligent dentistry and enables more clinically impactful multimodal AI systems in the dental field. It is a valuable asset for training the dental-specific large vision-language models for assisting the panoramic X-ray interpretation. The proposed MMOral-Bench suite could reflect the capabilities of existing LVLMs and reveal limitations of current models in this domain, paving the way for the next optimization direction to enhance the interpretation of panoramic X-rays.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use two public datasets and strictly follow their usage license. Besides, we cite their original paper to express our gratitude for their work. The details of the dataset source used in this paper are described in Appendix B.1.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The well-prepared documentation of our proposed MMOral instruction dataset and MMOral-Bench can be found on our project page at <https://github.com/isbrycee/OralGPT>

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We utilize LLMs as one of the key components for generating the instruction data, and we have provided **ALL** the prompts used in our experiments.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Towards Better Dental AI: A Multimodal Benchmark and Instruction Dataset for Panoramic X-ray Analysis

Appendix

Contents

A	Related works	23
B	MMOral Curation Details	23
B.1	Image and Visual specialists construction	23
B.2	Anatomical structure Extraction	24
B.3	Report generation	25
B.4	Instruction Data Generation	26
C	MMOral Visualization	27
D	MMOral-Bench Analysis	28
E	Evaluation	30
E.1	Evaluation Metrics	30
E.2	LLM as the Evaluator for Open-ended Questions: A Feasibility Analysis	33
E.3	Evaluation results	34
F	Efficacy Validation of MMOral Instruction Data	34
G	Limitations	36
H	Experiments Compute Resources	36
I	Case Study	36

A Related works

Over the past few years, the evaluation landscape for large vision-language models (LVLMs) has evolved significantly. Benchmarking plays a crucial role in assessing model capabilities, identifying model deficiencies, and guiding future optimization directions. Within the medical domain, existing benchmarks can be classified into two primary categories based on their alignment with the imaging modality coverage: general-purpose benchmarks for broad applicability and specialized benchmarks for discipline-specific evaluation.

There have been numerous efforts toward advancing general medical AI, such as LLaVA-Med [34], GMAI-VL [37], MedDr [27], HealthGPT [38], and HuatuoGPT [15]. Alongside these advancements, several general-purpose medical benchmarks spanning diverse imaging modalities and medical domains have been proposed, such as MMMU [73], OminimedVQA [30], MedEval [29], MedBench [42], MMT-Bench [69], and GMAI-Bench [66]. While these medical general-purpose benchmarks enable broader assessments across multiple medical fields, they inevitably fall short in their coverage of imaging modalities and specific medical domains. Therefore, they risk overlooking advancements in specific medical domains not encompassed by the general-purpose benchmarks.

Conversely, specialized benchmarks are concentrated on a particular imaging modality or medical discipline. For instance, PathVQA [28], PathMMU [52], and PathBench³ focus on pathology analysis, while SLAKE [39], VQA-RAD [32], and RadBench [64] target radiology understanding. In addition, Eyecare-Bench [36] and LMOD [51] assess the overall performance of LVLMs on intelligent ophthalmic diagnosis tasks. CXPMRG-Bench [61] and M4CXR [49] provide an evaluation suite for chest X-ray interpretation tasks. These specialized benchmarks facilitate in-depth evaluations within their respective disciplines, which are better suited to advancing the development of specific medical fields.

Notably, both current medical general-purpose benchmarks and specialized benchmarks overlook the evaluation of MLLMs within oral radiology, particularly in panoramic X-ray understanding—a widely used imaging modality that serves as a primary diagnostic source for assessing oral health. Consequently, there is an urgent demand for more comprehensive and robust benchmarks to address this gap and advance the development of LVLMs in oral healthcare.

B MMOral Curation Details

B.1 Image and Visual specialists construction

The method used to identify eligible panoramic X-ray datasets for image curation and visual specialist model construction was adapted from previous studies published in prestigious international peer-reviewed journals [66, 30, 51, 42] and globally recognized preprint platforms [36, 29]. Specifically, we collect panoramic X-ray images from two publicly available datasets: the TED3 dataset [26] and the dataset proposed by Hoang Viet Do [19]. The TED3 dataset is a large-scale semantic segmentation dataset constructed by aggregating 18 publicly available datasets from various public platforms, including Kaggle, Grand Challenge, and Tianchi. We filter out duplicate images according to the naming rules and ultimately obtain 16,639 unique images. The dataset proposed by Hoang Viet Do [19] is designed for detecting apical periodontitis lesions in panoramic radiographs. This dataset is obtained from the high-quality Dental Treatment Centre, School of Dentistry, Hanoi Medical University, and consists of a total of 3,924 images. Therefore, the final curated dataset comprised 20,563 images, and it exhibits significant diversity across various dental conditions, such as dentate and edentulous dentitions, tooth misalignment, impacted teeth, dental caries, root canal treatment, apical lesions, periodontal bone loss, dental implants, and various types of metallic and non-metallic dental restorations. Besides, it includes panoramic x-rays sourced from at least six countries, including Iran, the Democratic Republic of the Congo, Switzerland, the United States, China, and Vietnam, demonstrating the diversity of its demographic coverage.

The TED3 dataset [26] and the dataset proposed by Hoang Viet Do [19] are licensed under the Apache License 2.0 and CC BY 4.0, respectively. Both licenses allow for the reproduction and distribution of copies of the original datasets with modifications. Therefore, we utilize these two publicly available datasets as the image sources for MMOral.

³<https://smartlab.cse.ust.hk/showcase/PathBench/>

Subsequently, we leverage visual specialist models to simulate the interpretative process of oral radiology experts. This process aims to recognize as many anatomical structures as possible from the image, covering attributes from teeth to bone structures, historical treatments, and potential existing diseases. Owing to the inherent complexity of anatomical variations and fine-grained pathological cues observed in panoramic X-rays, we build a total of ten specialized visual models that are capable of detecting a total of 49 different anatomical structures. These models are fine-tuned on public datasets specifically related to the panoramic radiograph to extract various visual attributes from the images, and object detection and instance segmentation models are selected for training following the annotation protocols of these public datasets [25, 48, 1, 2, 4, 5, 3, 6, 7, 20]. The details of developed visual specialists and the corresponding category list of detected anatomical structures are listed in Table 4.

The specialist models were trained using the DINO [74] or MaskDINO [35] framework depending on the annotation protocols of public datasets. All hyperparameters followed the default configurations in the official implementations. During the process of training specialist models, we evaluated their performance by randomly splitting each annotated public dataset into training and validation sets at a 9:1 ratio. These ten specialist models achieved an average mAP50 of 81.25%, ensuring the initial quality of anatomical structure extraction. To further enhance the reliability of attribute extraction, only anatomical structures detected by at least two visual specialists were retained. All checkpoints and configurations of the specialist models are available on [Hugging Face](#).

Table 4: The details of visual specialists and the corresponding category list of detected anatomical and pathological structures.

Dataset Source	Task Type	Category Space	# Categories	# Samples
[25, 48, 1]	Object Detection	1 to 32 tooth numbering following the FDI tooth numbering system	32	2798
[48]	Object Detection	4 Quadrants	4	634
[48]	Object Detection	Caries, Deep Caries, Periapical lesions, Impacted tooth	4	705
[2]	Object Detection	Caries, Filling	2	448
[4]	Object Detection	Caries, Crown, Filling, Implant, Malaligned, Mandibular canal, Missing teeth area, Periapical lesion, Retained root, Root canal treatment, Impacted tooth	11	9206
[19]	Object Detection	Granuloma, Cyst, Abscess	3	3924
[5]	Instance Segmentation	Caries, Filling	2	448
[3]	Instance Segmentation	Bone loss	1	7986
[6]	Instance Segmentation	Mandibular canal, Maxillary sinus	2	327
[7]	Instance Segmentation	Caries, Crown, Root canal treatment, Badly Decayed, Restoration, Normal	6	1899
[20]	Object Detection	Optical character recognition (OCR)	N/A	N/A

B.2 Anatomical structure Extraction

We construct ten visual specialist models with **overlapping category spaces** to ensure precise detection of anatomical and pathological structures. For instance, ten structures (e.g., caries, periapical lesion, impacted teeth, missing teeth area, filling, implant, root canal treatment, crown, mandibular canal, and maxillary sinus) are validated by two or more visual specialists. To process the detected redundant visual elements, we meticulously design a post-processing pipeline that integrates anatomical structures and establishes associations between dental pathological findings, historical treatments, and their corresponding tooth notations based on their spatial relationships. The designed anatomical structure integration and relationship generation pipeline comprises eight systematic steps, as detailed in Algorithm 1.

Algorithm 1: Anatomical Structure Integration and Relationship Generation Pipeline.

Input: Panoramic X-ray images dataset $I = \{I_1, I_2, \dots, I_n\}$, Expert models $\{M_1, M_2, \dots, M_{10}\}$

Output: Visual attributes and relationships $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n\}$ for all images

for each panoramic X-ray image $I_i \in I$ do

- // Step 1: Detect imaging timestamp*
Detect imaging timestamp in I_i (if present) and save as t_i ;
- // Step 2: Detect teeth locations and tooth notations*
Detect all teeth positions and their corresponding notations following FDI tooth numbering system $\mathcal{T}_i = (\mathcal{P}_i, \mathcal{N}_i)$;
- // Step 3: Divide the image into four quadrants*
Divide the panoramic X-ray image into four quadrants: $\mathcal{Q}_i = \{Q_{UR}, Q_{UL}, Q_{LR}, Q_{LL}\}$;
- // Step 4: Anatomical identification using specialist models*
Initialize bounding boxes $\mathcal{B}_i = \emptyset$, $\mathcal{L}_i = \emptyset$, and $\mathcal{S}_i = \emptyset$;
for each $M_j \in \{M_3, \dots, M_{10}\}$ do
 - $(\mathcal{B}_j, \mathcal{C}_j, \mathcal{S}_j) \leftarrow M_j(I_i)$;
 - $\mathcal{B}_i \leftarrow \mathcal{B}_i \cup \mathcal{B}_j$;
 - $\mathcal{L}_i \leftarrow \mathcal{L}_i \cup \mathcal{L}_j$;
 - $\mathcal{S}_i \leftarrow \mathcal{S}_i \cup \mathcal{S}_j$;
- // Step 5: Post-process bounding boxes*
 $\mathcal{B}_i \leftarrow \text{Filter}(\mathcal{B}_i, \tau)$, retaining boxes with $s_k \geq \tau$;
 $\mathcal{B}_i \leftarrow \text{CategoryIntegration}(\mathcal{B}_i, \mathcal{L}_i)$;
 $\mathcal{B}_i \leftarrow \text{NMS}(\mathcal{B}_i, \mathcal{S}_i)$;
- // Step 6: Assign tooth-related observations to the specific tooth*
 $\mathcal{R}_i \leftarrow \text{Assign}(\mathcal{T}_i, \mathcal{B}_i, \mathcal{L}_i)$
- // Step 7: Insert domain knowledge rules*
if #18 $\in \mathcal{N}_i$ **and** #48 $\notin \mathcal{N}_i$ **or** #28 $\in \mathcal{N}_i$ **and** #38 $\notin \mathcal{N}_i$ **then**
 - $\mathcal{O}_i \leftarrow \text{Comment}(\text{"Consider extraction of tooth \#18/\#28"});$
- // Step 8: Generate final visual structures and relationships*
 $\mathcal{A}_i \leftarrow \{t_i, \mathcal{T}_i, \mathcal{Q}_i, \mathcal{B}_i, \mathcal{L}_i, \mathcal{S}_i, \mathcal{R}_i, \mathcal{O}_i\}$;

return $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n\}$;

B.3 Report generation

We designed a manual rule to organize discrete anatomical structures into coherent grounding captions. Specifically, each grounding caption begins with: "This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including." The caption is then composed of several parts: teeth visibility with center points, wisdom teeth detection, dental caries detection, periapical lesion detection, historical treatments, mandibular canal visibility, and maxillary sinuses visibility. Each part contains detailed coordinates of the anatomical structure, the corresponding tooth ID, and confidence scores, formatted as a JSON list string.

Through extensive consultations with senior dental specialists, we structure the medical report of a panoramic x-ray into three principal sections: Teeth-Specific Observations, Jaw-Specific Observations, and Clinical Summary & Recommendations. Each section is further subdivided into some subsections, and the specific content covered within each part is outlined in Table 5.

Thanks to the robust text comprehension and instruction-following capabilities of LLMs, we prompt LLMs to automatically generate medical reports for panoramic X-ray images based on the grounding caption generated by human-designed templates. To ensure precise and structured output, we adopt a two-stage LLM-based generation method to generate the medical report. First, the DeepSeek-R1-Distill-Llama-70B model is selected for medical report generation due to its exceptional performance on text understanding, logical reasoning, and instruction-following abilities. The entire generation process requires approximately 48 hours utilizing 4xNVIDIA A100 80G GPUs. We meticulously craft a system prompt and manually prepare an example for in-context learning to query the LLM.

Table 5: The detailed construction of the medical report within three sections and their corresponding content.

Section	Subsection	Content
Teeth-Specific Observations	General Condition	Number of teeth, presence and number of wisdom teeth, and cases of impaction.
	Pathological Findings	Presence of cavities (caries, deep caries) and periapical lesions (e.g., granuloma, cyst, abscess).
	Historical Interventions	Past dental treatments, including fillings, crowns, root canal treatments, and implants.
Jaw-Specific Observations	Bone Architecture	Assessment of periodontal bone loss.
	Visible Structures	Observation of key anatomical features such as mandibular canals and maxillary sinuses.
Clinical Summary & Recommendations	-	Summary of priority concerns, proposed preventive measures, and recommended follow-up protocols.

The details of the system prompt and manually prepared example are shown in Figure 6 and Figure 7, respectively.

During the inspection of generated reports, we identify and summarize several common issues even when grounding captions were provided:

- **Hallucination:** The LLM occasionally introduced dental conditions or diseases that were not present in the grounding captions.
- **Terminology Inaccuracy:** The model sometimes misrepresented attributes, such as stating "4 missing teeth" instead of the more accurate "4 missing teeth regions," since our specialist models detected contiguous areas of tooth loss.
- **Structural Errors:** There were instances where the LLM mis-organized content, such as placing pathological findings in historical intervention subsections.

We hypothesize that these issues are inevitable due to the inherent complexity of the task of generating medical reports based on location captions. This specified task presents significant complexity as it requires LLMs to simultaneously perform multiple cognitive sub-tasks, including text comprehension, organization, classification, structuring, summarization, and extrapolation. As a result, relying solely on the DeepSeek-R1-Distill-Llama-70B model cannot produce high-quality and entirely accurate reports in a single step, despite the model’s impressive natural language processing capabilities. To further enhance the quality of reports, we employ GPT-4-Turbo to refine reports generated by DeepSeek. Based on several common issues in the reports, we carefully design rules for validation and corrections and instruct GPT-4-Turbo to simultaneously output both revised reports and corresponding revision logs. By examining these revision logs, we can efficiently identify modified sections of the reports, thereby facilitating quality verification of the revised content. The details of the designed prompt for GPT-4o-Turbo are provided in Figure 8.

B.4 Instruction Data Generation

For the closed-ended VQA pairs, we introduced perturbations to the ground truth in four aspects: bounding box coordinates, tooth count, anatomical structure categories, and tooth IDs, generating incorrect choices for closed-ended questions. This method aims to improve the model’s ability to detect and interpret subtle differences. The details of the designed perturbations are summarized in Table 6. We believe this perturbation strategy enables a more robust evaluation, accurately assessing model performance across multiple dimensions.

Based on the refined medical report, we prompt GPT-4-turbo to generate the instruction data, including the visual question-answering data (MMOral-VQA) and the image-grounded conversation data (MMOral-Chat). The designed prompts for these two sub-datasets are shown in Figure 9 and Figure 10, respectively.

You are a professional oral radiologist assistant tasked with generating precise and clinically accurate oral panoramic X-ray examination reports based on structured localization data.

The structured localization data contains all teeth and potential dental conditions as well as diseases detected by multiple visual expert models, along with their corresponding visual absolute position coordinates. Each condition/disease is associated with specific tooth numbers. For those conditions/diseases that do not correspond to specific teeth (where the tooth_id is labeled as "unknown"), they must include the side (e.g., upper left, lower right, etc.) visible in the panoramic X-ray.

Generate a formal and comprehensive oral examination report **ONLY** containing three mandatory sections:

1. Teeth-Specific Observations
2. Jaw-Specific Observations
3. Clinical Summary & Recommendations

The Teeth-Specific Observations section comprises three subsections: General Condition, Pathological Findings, and Historical Interventions.

The General Condition outlines overall dental status, including tooth count and wisdom teeth status (e.g., presence or impaction).

Pathological Findings document dental diseases such as caries or periapical periodontitis.

Historical Interventions detail prior treatments like fillings, crowns, root canal treatments, or implants.

The Jaw-Specific Observations section evaluates bone status and visible anatomical structures (e.g., Bone loss, Mandibular Canal, Maxillary Sinus).

Besides, each condition/disease is associated with a confidence score. Apply the following processing rules on the pathological finding subsection:

- For confidence scores <0.80 : Include terms like "suspicious for..." or "suggest clinical re-evaluation" in the description;
- For confidence scores ≥ 0.80 : Use definitive descriptors such as "sign of..." or "shows evidence of...", etc.

Note that the confidence scores are **ONLY** used to express the degree of certainty regarding the condition/disease from visual expert models and **MUST NOT** appear in the report. Only the pathological findings subsection needs to include specific certainty terms based on the confidence scores; other sections do not require this rules. **Do not include or reference confidence scores in any form in the output.**

Please strictly follow the following requirements:

- Strict adherence to FDI numbering system
- Use professional medical terminology while maintaining clarity whenever possible
- Don't include ANY confidence score in the provided structured localization data
- Don't generate any form of formatted content in a standard report, such as 'Patient Name', 'Date', 'Designation', etc.
- Exclude any speculative content beyond the provided structured localization data

Example Format:

Input:

```
{input_example}
```

Output:

```
{output_example}
```

Now generate a new report for the following input:

Figure 6: The prompt for DeepSeek-R1-Distill-Llama-70B to generate the medical report of panoramic X-ray images.

C MMOral Visualization

To provide an intuitive demonstration of the information conveyed in the textual description, we visualize the most frequently occurring words in the MMOral-Report, MMOral-VQA, and MMOral-Chat through word cloud maps, as demonstrated in Figure 11. Besides, Figures 14 - 17 show six examples in the MMOral-Attribute and MMOral-Report sub-datasets.

input_example = "" This localization caption provides multi-dimensional spatial analysis of anatomical structures and pathological findings for this panoramic dental X-ray image, including: Teeth visibility with center points (total: 26): <pre>[{'point_2d': [1290, 550], 'tooth_id': '21', 'score': 0.9}, {'point_2d': [1272, 812], 'tooth_id': '31', 'score': 0.89}, ...]</pre> Wisdom teeth detection (total: 3): <pre>[{'box_2d': [672, 332, 787, 561], 'tooth_id': '18', 'is_impacted': false, 'score': 0.65}, {'box_2d': [1680, 565, 1883, 731], 'tooth_id': '38', 'is_impacted': false, 'score': 0.59},]</pre> Dental caries detection (total: 2): <pre>[{'box_2d': [1695, 615, 1767, 679], 'tooth_id': '38', 'label': 'Caries', 'score': 0.56}, {'box_2d': [1667, 556, 1912, 759], 'tooth_id': '38', 'label': 'Deep caries', 'score': 0.46}]</pre> Periapical lesions detection (total: 1): <pre>[{'box_2d': [1473, 854, 1521, 901], 'tooth_id': '35', 'label': 'Periapical Lesions (Granuloma)', 'score': 0.84}]</pre> Historical treatments (total: 9): <pre>[{'box_2d': [1732, 578, 1771, 615], 'tooth_id': '38', 'label': 'Filling', 'score': 0.81}, ...]</pre> Mandibular canal visibility (total: 2): <pre>[{'box_2d': [1601, 531, 2066, 1062], 'label': 'Mandibular canal', 'score': 0.94}, {'box_2d': [443, 413, 855, 975], 'label': 'Mandibular canal', 'score': 0.9}]</pre> Maxillary sinuses visibility (total: 2): <pre>[{'box_2d': [1380, 16, 1801, 317], 'label': 'Maxillary sinus', 'score': 0.81}, {'box_2d': [699, 15, 1084, 302], 'label': 'Maxillary sinus', 'score': 0.82}]</pre> ""	output_example = "" This is a panoramic dental X-ray image, which provides a broad view of the entire mouth, including the teeth, jawbones, sinuses, and other structures. Below is a detailed analysis of the image: Teeth-Specific Observations - General Condition: - 26 teeth visualized with findings suggestive of clear anatomical definition - 3 wisdom teeth detected: - #18: Erupted (suspected, recommend clinical review) - #28: Erupted (suspected, recommend clinical review) - #38: Erupted (suspected, recommend clinical review) - Pathological Findings: - #18: sign of caries - #38: Suspected deep caries - #35: Imaging features sign of periapical granuloma - Historical Interventions: - #15,16,23,24,25,38: sign of fillings - #36: Imaging features sign of dental implant with crown restoration - #25: signs of root canal treatment with post-core restoration Jaw-Specific Observations - Bone Architecture: - No apparent bone loss in the image - Visible Structures: - Imaging features signs of bilateral mandibular canals Clinical Summary & Recommendations - Priority Concerns: - Periapical lesion at #35 requires endodontic evaluation - Deep caries in #38 needs immediate intervention - Preventive Measures: - Monitor suspected caries at #18 with radiographic follow-up - Follow-up Protocol: - 6-month recall for caries monitoring (particularly suspected areas) - Bitewing series recommended for interproximal caries detection Further clinical correlation with physical examination and patient history is recommended for a comprehensive diagnosis. ""
---	---

Figure 7: The manually designed in-context examples for medical report generation. Given a grounding caption of panoramic X-rays (left side), the ideal medical report exemplar is shown on the right side.

Table 6: Transformation perturbations and constraints for closed-ended VQA pairs.

Category	Transformation Perturbations	Additional Constraints
Bounding Box	Translation, scaling, or mirroring	IoU with ground truth < 0.3
Tooth Count	Randomly adding large differences (± 3 -5) or small differences (± 1 -2)	Range constrained between 0 and 32
Anatomical Structure Categories	Generating related but incorrect options (e.g., for dental caries: periapical lesions, normal teeth, dental fillings)	–
Tooth IDs	Selecting tooth IDs in the same quadrant or adjacent to the correct ID	Incorrect options randomly shuffled to avoid fixed correct answer positions

Table 7: Summary of dental conditions in MMOral-Bench.

Dental Condition	Count
Crown	285
Filling	228
Caries	147
Periapical lesion	138
Root canal treatment	131
Impacted teeth	74
Implant	30
Bone loss case	26
Retained root	21

D MMOral-Bench Analysis

MMOral-Bench comprises 491 closed-ended and 578 open-ended QA pairs with 100 images through significant manual selection and validation, and it offers sufficient representativeness and generaliz-

You are a professional oral radiology report auditor. Strictly validate AI-generated reports against original structured location captions and correct errors according to the following protocols.

The AI-generated report follows three mandatory sections below:

1. Teeth-Specific Observations
 - This section comprises three subsections: General Condition, Pathological Findings, and Historical Interventions.
 - The General Condition outlines overall dental status, including tooth count and wisdom teeth status (e.g., presence or impaction).
 - Pathological Findings document dental diseases such as caries or periapical periodontitis.
 - Historical Interventions detail prior treatments like fillings, crowns, root canal treatments, or implants.
2. Jaw-Specific Observations
 - This section evaluates bone status and visible anatomical structures (e.g., Bone loss, Mandibular Canal, Maxillary Sinus).
3. Clinical Summary & Recommendations

Rules for Validation and Corrections

1. Authenticity Verification
 - Ensure that all pathological findings (caries, periapical lesions) and historical interventions (filling, crown, root canal treatment, or implant) mentioned in the Teeth/Jaw-Specific Observations section of the medical report are included in the structured location caption.
 - Error Example: Reporting "#15: sign of filling" when location captions show no such finding
2. Historical Interventions Protocol
 - When describing the absence of historical interventions, avoid specifying individual teeth and ensure this characterization pertains to the entire image as a whole.
3. Bone Loss Placement
 - Bone loss must only appear in: Jaw-Specific Observations → Bone Architecture
 - Incorrect Example: Listing "bone loss" under Teeth-Specific Observations → Pathological Findings
 - Move or rephrase details about bone loss as needed.
4. Missing Teeth Documentation
 - The phrase "**missing teeth detected**" should refer to **regions** rather than the exact count of missing teeth.
 - Incorrect: 4 missing teeth are detected in the upper jaw
 - Correct: 4 missing teeth regions are detected in the upper jaw
5. Pathological Findings check
 - The Pathological Findings subsection must NOT include any details about historical interventions (fillings, crowns, root canal treatments, or implants)
6. Locational Descriptions
 - Ensure all positional terms (e.g., lower, upper, left, right) match the phrasing and descriptions provided in the **Structured Location Captions**.
 - Do not introduce new locational terms or modify the original phrasing.
7. Do NOT modify any descriptive terms in the original medical report, such as 'suspicious, suspected or others'.
8. Do NOT remove any subsections directly or modify the original structure of the report.

Input Format

You will receive the following:

1. **Structured Location Captions:** A list of findings and descriptions of specific areas in the panoramic dental X-ray image.
2. **LLM-Generated Report:** A report generated based on the location captions.

Output Format

If no revision is needed, output:

```
{
  "need revision": false
}
```

If revision is required, output:

```
{
  "need revision": true,
  "Revised med report": {
    "Revised Report": "...",
    "Revision Log": [
      "1. Change description (Rule X)",
      "2. Change description (Rule Y)"
    ]
  }
}
```

Now generate a new report for the following input:

Figure 8: The prompt for GPT-4-Turbo to revise the generated report. We manually check the generated medical reports from the DeepSeek-R1-Distill-Llama-70B and summarize several rules for validation and correction. We ask the GPT-4-Turbo simultaneously output both revised reports and corresponding revision logs for convenient human verification.

ability to real-world clinical scenarios. The inclusion of categories and their respective counts for dental conditions are presented in Table 7. All conditions are precisely mapped to tooth positions using the standard FDI numbering system, significantly enhancing clinical relevance through accurate localization.

You are an experienced oral radiologist specializing in generating assessment question-answer pairs based on the provided dental panoramic radiograph report. Your task is to create two types of question-answer pairs:

1. Closed-End Questions:

These are multiple-choice questions with 4 options (A, B, C, D), where only one is correct.

The incorrect options should be plausible and relevant, using adjacent tooth numbers or similar pathologies/interventions to test comprehension and critical thinking.

2. Open-End Questions:

These are free-response questions targeting specific details from the report.

Answers should be concise (1-2 sentences) and directly reference the report terminology. Avoid vague or overly broad questions.

Ensure that all critical information from the following sections is included:

- Teeth general condition & wisdom teeth status
- Pathological findings (caries, periapical lesion)
- Historical interventions (filling, implant, crown, root canal treatment)
- Bone/jaw observations
- Clinical recommendations (excluding follow-up protocol-related content)

Please strictly follow the following requirements:

- Strict adherence to FDI numbering system in the provided report
- Distractors must be logical (e.g., incorrect options use adjacent tooth numbers or related pathologies)
- Answers must strictly align with the report and avoid any speculation beyond the stated findings
- Each Q-A pair must address unique information points without overlapping content or redundant queries
- Creating sufficient Q-A pairs to cover all reported findings while maintaining clinical relevance
- Strictly exclude any questions about follow-up recommendations/protocols

Output format:

```
```json
{
 "Closed-End Questions":[
 {
 "Question": "...",
 "Options": "A) ... B) ... C) ... D) ..."
 "Answer": "...",
 },
 ...
],
 "Open-End Questions":[
 {
 "Question": "...",
 "Answer": "..."
 },
 ...
]
}
```

Now generate two types of question-answer pairs for the following report:

Figure 9: The prompt for GPT-4-Turbo to generate both closed-ended and open-ended question-answering pairs based on the medical report.

## E Evaluation

### E.1 Evaluation Metrics

In this section, we describe the evaluation process in detail. For closed-ended questions, we adopt the assessment pipeline designed by the CMMU benchmark. Specifically, we use options as keywords to extract model responses through robust regular expressions, selecting the response with the highest number of options as the final answer. If there is no valid answer in the model’s response, random selection is performed for multiple-choice questions. We adopt accuracy as the evaluation metric.

For open-ended questions, following MM-Vet [71] and MM-Vetv2 [72], we construct a few-shot prompt and leverage GPT-4-turbo to assist with the evaluation. The few-shot prompt is demonstrated in Figure 12. Specifically, our designed prompt includes nine in-context examples with open-ended answers. These examples encompass fully correct responses (i.e., 1.0), entirely incorrect responses (i.e., 0.0), and cases illustrating various types of "partially correct" answers. The LLM-based evaluator enables the evaluation of any style of model’s prediction using a unified and consistent metric.

You are an advanced assistant tasked with simulating realistic, multi-turn dialogues between a **patient** and a senior **radiologist** specialized in panoramic dental X-ray imaging. The radiologist is a professional medical expert who explains the findings and summaries from a dental panoramic radiograph in a patient-friendly manner. The patient is a layperson with limited medical knowledge, asking questions based on the radiologist's explanations. The patient may express concerns, request clarifications, or ask follow-up questions about treatment options.

Your task is to:

1. Interpret and utilize the input data, which includes:
  - A **structured location caption** from the dental panoramic image (e.g., positions and labels of teeth, caries, periapical periodontitis, filling, crown, root canal treatment, implant, bone conditions or other observations).
  - A **textual examination report** (findings and summary) written by the radiologist.
2. Generate a realistic multi-turn conversation between the patient and the radiologist:
  - The **radiologist** should explain the findings and summary in simple terms, avoiding overly technical jargon.
  - The **patient** should respond naturally, asking questions or confirming their understanding.
3. Ensure the conversation is coherent, informative, and empathetic, addressing the patient's potential concerns.

Please strictly obey the following **rules and constraints**:

1. Dialogue Tone and Style:
  - The radiologist must maintain a professional, calm, and empathetic tone.
  - The patient should sound natural and relatable, expressing curiosity, concern, or a need for clarification, depending on the context.
2. Medical Accuracy:
  - Ensure that explanations provided by the radiologist are factually correct and must align with the input structured location caption and examination report.
  - Avoid making medical recommendations unless explicitly stated in the input data.
3. Patient Understanding:
  - Ensure explanations are simple and clear, using analogies or examples if necessary.
  - Address the patient's concerns with empathy and patience, ensuring they feel reassured and informed.
4. Dialogue Flow:
  - The conversation should alternate between the patient and the radiologist.
  - Each response should naturally follow from the previous turn.
  - The number of turns should be between 8-12 exchanges to allow for sufficient detail and interaction.
5. Output Format:
  - The output format must follow the standard JSON format with the structure as follows:

```
```json
{
  "conversation": [
    {
      "round": 1,
      "role": "Patient",
      "content": "..."
    },
    {
      "round": 1,
      "role": "Radiologist",
      "content": "..."
    },
    ...
  ]
}
```

Figure 10: The prompt for GPT-4-Turbo to generate a multi-turn conversation between the assistant and a person asking questions about the panoramic X-ray.

By inputting the prompt, GPT-4 automatically generates scores for each sample based on the input question, ground truth, and model's prediction. Each sample is assigned a score ranging from 0 to 1.

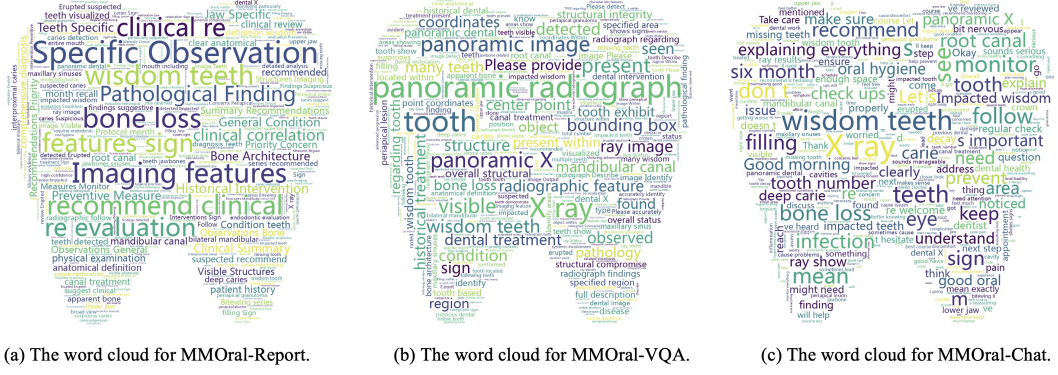


Figure 11: The word cloud maps for MMOral-Report, MMOral-VQA, and MMOral-Chat sub-datasets.

Given the question, compare the ground truth and prediction from AI models, to generate a correctness score for the prediction. The correctness score is 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). Just complete the last space of the correctness score.			
Question	Ground truth	Prediction	Correctness
---	---	---	---
How many teeth are visualized in the radiograph?	30 teeth are visualized with clear anatomical definition.	30	1.0
How many teeth are visualized in the radiograph?	30 teeth are visualized with clear anatomical definition.	29 teeth are visualized with clear anatomical definition.	0.0
What is the status of the wisdom teeth in the radiograph?	Three wisdom teeth are detected, all of which are impacted: #18, #28, and #48.	#18: impacted, #28: impacted, #48: erupted	0.7
What is the condition of the teeth #26 and #14?	Teeth #26 and #14 show signs of periapical abscesses.	Teeth #26 and #23 show signs of periapical abscesses.	0.5
What is the condition of the bone architecture and visible structures in the jaw?	No apparent bone loss is observed. Bilateral mandibular canals and maxillary sinuses are clearly visible.	Bilateral mandibular canals and maxillary sinuses are clearly visible.	0.5
What is the clinical priority concern regarding the periapical lesions?	Periapical cysts at #11 and #12, and granuloma at #46 require endodontic evaluation.	Periapical lesions at #11, #12, and #46 require endodontic evaluation.	0.8
What radiographic features are visible in tooth #31 on the panoramic X-ray?	[{"Teeth position": {"point_2d": [1242, 726]}}, {"Crown": {"box_2d": [1220, 637, 1266, 741]}}]	Crown	0.8
What radiographic features are visible in tooth #31 on the panoramic X-ray?	[{"Teeth position": {"point_2d": [1242, 726]}}, {"Crown": {"box_2d": [1220, 637, 1266, 741]}}]	Crown at position: [1230, 627, 1276, 750]	0.9
What radiographic features are visible in tooth #31 on the panoramic X-ray?	[{"Teeth position": {"point_2d": [1242, 726]}}, {"Crown": {"box_2d": [1220, 637, 1266, 741]}}]	Teeth at position: {"point_2d": [1242, 726]}], {"Crown at position: {"box_2d": [1230, 627, 1276, 750]}}	1.0
Q G P			

Figure 12: Few-shot prompt for evaluating model predictions using GPT-4-Turbo, where Q is the question, G is the ground truth and P is the model's prediction for the question. The prompt demonstrates exemplar scoring criteria for diverse open-ended responses. Taking the prompt filled with Q , G , and P , GPT-4-Turbo will generate a soft grading score from 0 to 1.

The total scores are calculated by

$$S = \frac{\sum_{i=1}^N s_i}{N} \times 100\% \quad (1)$$

where s_i is the score of sample i , and N is the number of samples. The score regarding each sub-category can derived by

$$S = \frac{\sum s_i}{N_c} \times 100\%, i \in C \quad (2)$$

where C is the set of samples belonging to a specific sub-category (e.g., Teeth, Patho, HisT, Jaw, SumRec, Report), and N_c is the number of samples in this set.

The evaluation on MMOral-Bench is conducted using the standard VLMEvalKit framework. We have prepared the automatic evaluation script for MMOral-Bench⁴, and it will be supported by the VLMEvalKit framework in the future.

E.2 LLM as the Evaluator for Open-ended Questions: A Feasibility Analysis

Effectiveness. To verify the effectiveness of LLM-based evaluation for open-ended tasks, 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 determine the evaluation criteria. The two dentists then independently scored the predictions of GPT-4o and HealthGPT-XL32 on 600 cases from the MMOral-Bench open-ended QA task based on these criteria. The absolute differences between human scores and evaluators’ scores are shown in Table 8, represented as Δ .

Overall, the absolute differences of the “Overall” metric given by dentists are approximately 1 point lower than those of the LLM-based evaluation for the predictions of both LVLMS (GPT-4o and HealthGPT-XL32). This indicates that human scoring preferences align with the trends of LLM-based evaluation. However, it also suggests that the dentists’ scoring is more stringent compared to the LLM-based evaluation, potentially due to subjective differences in their interpretation of the evaluation criteria provided 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 Teeth, Patho, HisT, and SumRec categories, whereas the differences are larger for the Jaw and Report categories. Although Dentist B exhibits slightly larger differences with LLM-based scoring across all subcategories, their “overall” score difference is only 0.37 points. This indicates that LLM-based scoring aligns well with human preferences in reflecting the overall performance of LVLMS on MMOral-Bench open-ended tasks. At the same time, we speculate that the score fluctuations in each subcategory are strongly associated with the subjective perceptions of human evaluators.

Table 8: Average absolute differences (Δ) between the evaluation scores of the LLM-based evaluator and the dentist-annotated scores on the open-ended QA task in MMOral-Bench.

Model	Evaluators	Teeth	Patho	HisT	Jaw	SumRec	Report	Overall
GPT-4o [31]	Dentist A	30.16	27.65	40.03	67.21	33.30	33.35	35.43
	GPT-4-Turbo	31.48	26.05	37.56	57.42	30.37	42.50	37.50
	$\Delta(\downarrow)$	-1.32	+1.6	+2.47	+9.79	+2.93	-9.15	-2.07
HealthGPT-XL32 [38]	Dentist B	26.51	29.26	21.66	43.68	28.20	14.5	26.80
	GPT-4-Turbo	29.80	22.16	24.11	47.82	24.77	10.00	27.17
	$\Delta(\downarrow)$	-3.29	+7.1	-2.45	-4.14	+3.43	+4.50	-0.37

Stability. Since using LLMs as evaluators inevitably introduces randomness, even with the temperature hyperparameter set to 0, we conduct multiple repeated experiments to verify the stability of LLMs as evaluators. Specifically, we evaluate the prediction results of GPT-4o [31], HealthGPT-XL32 [38], Qwen2.5-VL-7B [12], and Ovis2-34B [45] on open-ended questions using GPT-4-Turbo [10] with the same prompt five times. The obtained mean, standard deviation, and coefficient of variation (CV) of the metric “overall” are shown in Table 9. For proprietary models, medical-specific models, and general-purpose LVLMS, the standard deviation of the metric “overall” is within 0.45 when evaluated 5 times using GPT-4-Turbo with our designed few-shot prompt. Specifically, for the prediction results of Ovis2-34B, the standard deviation of the scores is 0.434, while for Qwen2.5-VL-7B, the standard deviation is as low as 0.096. 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 1%, which demonstrates the evaluation stability of using LLMs as evaluators. The detailed results across each specific category are demonstrated in Figure 13.

⁴<https://github.com/isbrycee/OralGPT>

Table 9: Stability verification of using LLMs as evaluators: Standard deviation and coefficient of variation (CV) are reported across four LVLMs from five repeated evaluations.

Model	Mean	StdDev	CV (Coefficient of Variation, %)
GPT-4o [31]	37.567	0.330	0.879
HealthGPT-XL32 [38]	27.284	0.172	0.631
Qwen2.5-VL-7B [12]	15.894	0.096	0.607
Ovis2-34B [45]	32.671	0.434	1.329

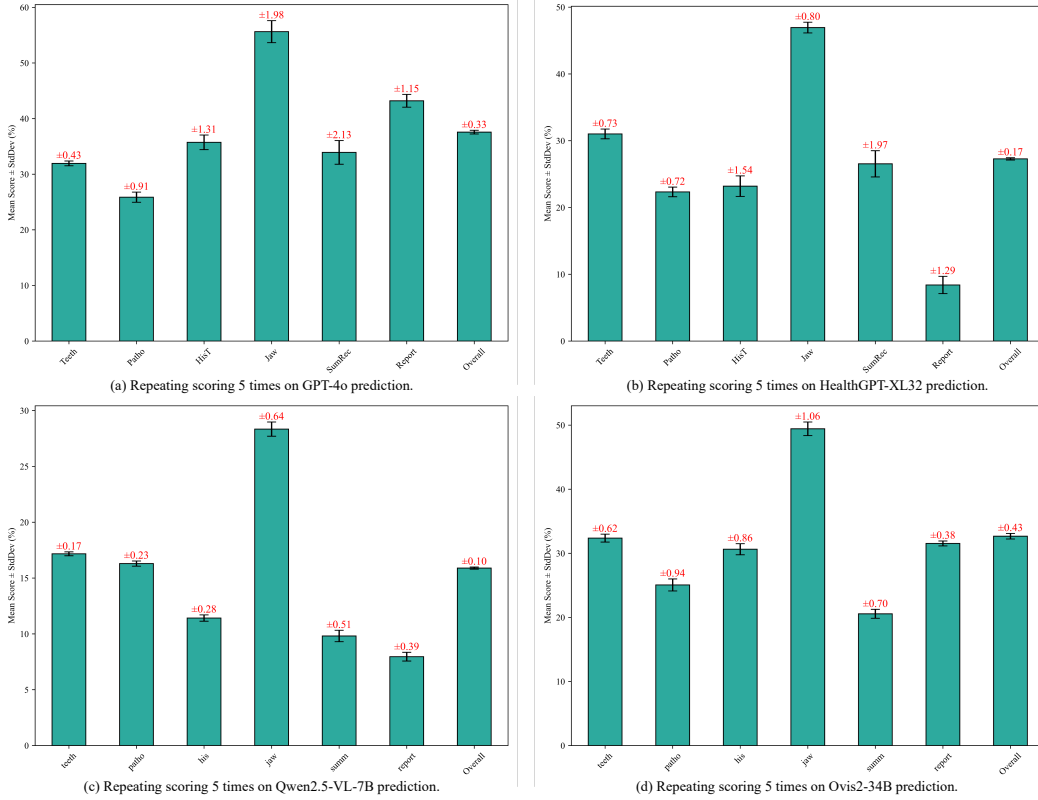


Figure 13: The means and standard deviations of each category on 5 repeated evaluations across four LVLMs' predictions.

E.3 Evaluation results

We conduct zero-shot evaluations across 64 LVLMs on our MMOral-Bench, the results are provided in Table 10.

F Efficacy Validation of MMOral Instruction Data

We implement supervised fine-tuning (SFT) on two LVLMs of different scales, Qwen2.5-VL-7B [12] and LLaVA-Next-13B-hf [41], using our MMOral instruction data, and the results are presented in Table 11. We use the LLaMA-Factory [76] framework to perform SFT for one epoch while maintaining other default hyperparameters. When fine-tuning Qwen2.5-VL-7B and LLaVA-Next-13B-hf with MMOral-Report, MMOral-VQA, and MMOral-Chat together, the average scores improve by 24.73% and 18.42%, respectively. This clearly demonstrates the effectiveness of the MMOral instruction data and its potential value in advancing digital AI applications in the field of oral healthcare.

Table 10: Results on MMOral-Bench for existing various LVLMS across both closed-ended and open-ended VQA tasks. The best-performing model in each category is highlighted **in-bold**, while the second-best is underlined.

Model	Close-ended VQA						Open-ended VQA						Avg.	
	Teeth	Patho	His	Jaw	Summ	Overall	Teeth	Patho	His	Jaw	Summ	Report		Overall
Proprietary LVLMS														
GPT-4o-2024-11-20 [31]	36.16	41.14	36.18	47.85	57.89	39.92	41.99	27.20	41.96	66.00	41.79	60.40	46.70	43.31
GPT-4V [31]	38.21	35.76	43.42	52.79	50.00	42.57	31.46	23.79	39.51	69.81	34.29	43.70	39.38	40.98
Claude-3-7-Sonnet-20250219 [111]	39.61	34.78	42.76	34.66	57.89	39.10	41.07	37.88	51.84	62.69	37.86	42.70	45.31	42.21
Gemini-2.5-Flash-preview-04-17 [53]	16.40	12.99	23.94	29.43	20.00	19.14	28.04	24.77	31.90	47.81	12.98	16.70	27.84	23.49
Gemini-2.0-Flash [53]	42.42	37.01	51.41	41.89	53.33	42.97	36.19	27.73	47.06	62.88	33.57	45.50	41.82	42.40
Qwen-Max-VL-2025-04-08 [12]	24.42	27.27	33.10	34.34	62.22	29.94	2.10	4.47	7.06	11.62	7.98	5.50	5.29	17.62
Step-1o-vision-32k ³	28.12	26.87	25.34	26.62	28.07	27.00	28.54	17.92	26.11	46.72	30.40	25.18	29.38	28.19
Step-1o-turbo ³	32.71	25.03	37.84	56.45	40.72	38.00	33.81	22.34	29.44	52.02	32.26	44.12	37.00	37.50
Doubao-1.5-vision-lite-250315 ⁴	21.98	13.62	27.45	26.40	32.12	22.30	31.24	13.35	30.72	61.43	27.85	47.15	35.80	29.05
Doubao-1.5-vision-pro-250328 ⁴	29.42	28.90	27.47	28.90	49.12	30.60	32.18	16.32	37.82	58.77	31.42	48.36	37.15	33.88
Doubao-1.5-thinking-vision-pro-250428 ⁴	31.55	20.78	41.55	53.96	40.00	36.05	41.26	23.94	43.93	60.75	32.74	53.40	43.90	39.97
Open-Source LVLMS														
Deepseek-VL-7b-chat [43]	29.41	31.82	33.33	51.70	42.22	35.44	16.48	7.50	13.44	34.56	9.52	9.60	15.95	25.70
Emu3-chat [62]	39.75	31.17	38.73	49.81	42.22	40.12	17.03	6.36	12.88	36.06	14.76	19.60	18.13	29.13
Qwen2-VL-7B [59]	26.42	25.83	22.18	21.35	15.73	24.20	17.82	11.75	12.42	24.55	10.27	12.78	16.00	20.10
Qwen2.5-VL-7B [13]	24.96	21.12	27.54	37.08	35.42	27.00	17.01	16.10	11.18	29.41	9.07	8.20	15.92	21.46
Qwen2.5-VL-32B [13]	26.40	25.85	22.48	21.32	15.82	24.30	17.66	11.74	12.28	24.47	10.18	12.74	15.90	20.10
Qwen2.5-VL-72B [13]	24.60	24.68	27.16	26.79	42.22	25.66	13.90	15.83	15.40	27.12	7.38	11.50	15.38	20.52
CogVLM-9B [60]	27.05	22.23	33.42	31.47	63.32	29.40	24.36	15.72	26.85	41.18	14.80	13.72	23.80	26.60
CogVLM2-19B [60]	29.77	29.87	43.83	41.13	55.56	35.23	29.29	18.03	27.79	52.19	15.83	27.10	29.81	32.52
GLM-4V-9B [22]	23.71	28.57	34.57	62.64	44.44	34.83	20.94	9.70	18.77	26.62	12.74	21.30	19.74	27.29
LLaVA-vl-5.7B [40]	21.32	16.85	27.45	25.10	46.72	23.40	11.84	12.37	10.92	19.63	16.25	12.32	13.00	18.20
LLaVA-vl-5.13B [40]	20.58	21.35	29.42	30.18	40.82	25.00	10.42	11.32	14.87	18.12	12.72	12.20	12.85	18.93
LLaVA-NeXT-8B-hf [41]	35.72	34.82	40.05	50.12	63.42	41.90	16.42	6.28	10.05	26.45	13.65	23.72	16.90	29.40
LLaVA-NeXT-13B-hf [41]	30.09	32.92	30.54	38.20	60.42	33.80	14.48	10.28	9.23	22.41	14.30	21.30	15.43	24.62
LLaVA-OneVision [33]	34.05	40.26	37.04	63.02	73.33	42.77	28.79	13.64	30.12	55.12	20.36	23.20	29.43	36.10
LLaMA-3.2-Vision-11B-Instruct [23]	31.37	27.27	35.19	52.83	60.00	36.46	27.42	21.89	26.63	41.06	22.14	21.40	27.02	31.74
Cambrian-8B [57]	26.42	24.82	43.12	63.25	49.18	37.20	26.80	11.64	23.05	38.22	21.42	20.38	24.90	31.05
Cambrian-13B [57]	33.42	26.40	44.25	69.22	40.82	41.60	25.05	13.47	21.42	42.35	16.34	15.58	23.40	32.50
Cambrian-34B [57]	36.90	37.01	45.06	60.75	55.56	44.60	30.22	23.11	32.57	47.90	11.68	17.03	27.09	35.85
Phi-3-Vision-128K-Instruct [8]	31.45	31.84	41.82	43.25	63.42	37.70	19.42	18.65	17.40	45.28	21.12	9.52	20.80	28.75
Phi-3.5-Vision-Instruct [8]	28.12	21.42	39.80	39.82	55.12	31.90	29.65	24.32	25.12	43.85	23.64	10.35	26.15	29.03
Phi-4-multimodal-instruct [9]	33.51	27.92	40.74	46.79	42.22	37.07	28.52	21.21	27.85	44.06	12.98	14.60	26.19	31.63
InternVL2.5-8B [16]	33.10	28.72	25.42	35.68	36.32	30.60	28.65	21.42	27.52	39.12	21.45	25.82	27.90	29.25
InternVL3-8B [16]	22.85	23.35	22.48	36.82	30.12	25.60	28.70	15.25	26.02	36.15	17.25	24.62	26.00	25.80
InternVL3-14B [16]	32.35	28.92	27.42	36.40	26.15	31.40	25.45	18.42	28.05	43.12	21.85	35.47	29.60	30.50
InternVL3-38B [16]	26.56	22.08	22.22	33.58	28.89	26.48	34.77	19.17	30.18	47.69	17.74	40.10	34.05	30.27
Chameleon-7B [44]	28.34	33.12	33.95	31.70	46.67	31.98	5.21	3.18	10.86	14.00	5.12	8.80	7.65	19.82
PaliGemma-3B [14]	23.89	20.78	27.78	39.25	28.89	27.70	10.73	12.05	10.25	14.00	6.07	0.20	9.12	18.41
MiniCPM-V2.6 [65]	27.65	26.30	30.85	33.25	15.72	28.30	26.55	16.52	25.21	48.52	16.85	28.82	28.70	28.50
MiniCPM-O2.6 [65]	32.09	16.88	22.84	31.32	24.44	27.09	19.98	13.94	15.09	28.56	13.10	33.40	21.60	24.35
Kosmos-2 [50]	14.26	20.78	24.07	12.83	13.33	17.11	15.68	14.77	8.10	20.56	12.50	5.40	13.13	15.12
Ovis1.5-8B [45]	26.05	28.22	38.00	69.42	40.25	37.60	32.02	27.45	29.62	49.35	18.20	18.92	30.60	34.10
Ovis2-8B [45]	39.25	48.65	45.80	74.42	71.92	51.10	29.45	25.35	27.42	54.12	19.95	31.55	31.10	41.10
mPLUG-Owl2-7B [68]	24.42	19.65	42.25	43.32	55.30	32.10	12.65	17.62	10.82	25.80	17.82	15.32	15.10	23.60
mPLUG-Owl3-7B [67]	29.77	27.27	30.86	51.32	53.33	34.83	17.17	13.94	11.23	35.06	8.81	11.20	16.83	25.83
Gemma3-12B [54]	22.28	12.99	25.93	33.58	35.56	23.83	28.38	22.73	23.56	34.44	23.93	31.70	28.22	26.03
Gemma3-12B [54]	22.28	12.99	25.93	33.58	35.56	23.83	28.38	22.73	23.56	34.44	23.93	31.70	28.22	26.03
Molmo-7B-O-0924 [18]	24.45	20.95	25.42	26.22	13.65	24.10	12.35	7.58	9.47	11.82	4.80	6.85	9.50	16.80
Molmo-72B-0924 [18]	35.65	31.17	27.78	23.02	31.11	31.16	9.82	4.47	6.75	12.81	3.21	11.00	9.03	20.09
Yi-VL-34B [70]	37.65	35.82	44.05	40.32	69.52	40.50	25.80	22.52	21.42	38.22	14.42	10.75	23.40	31.95
Yi-Vision [70]	38.50	44.81	44.37	56.60	53.33	45.01	30.75	18.94	33.13	59.00	24.76	39.90	34.91	39.96
Qwen-QVQ-72B [56]	28.70	22.08	30.99	34.72	13.33	28.31	27.26	13.64	23.01	34.19	13.45	41.90	27.80	28.05
Ovis2-34B [45]	38.15	36.36	43.83	72.45	71.11	47.45	35.21	22.12	37.79	55.31	16.43	32.20	35.03	41.24
Kimi-VL-A3B-Instruct [55]	45.42	41.35	39.18	71.85	65.32	50.10	29.12	18.85	33.25	52.42	20.82	26.45	30.50	40.30
Kimi-VL-A3B-Thinking [55]	23.17	23.38	17.28	29.81	35.56	23.42	50.39	37.73	50.18	58.25	45.71	61.50	51.80	37.61
Medical Specific LVLMS														
LLaVA-Med [34]	27.09	29.22	27.46	18.11	24.44	25.66	0.91	1.52	0.00	0.00	0.00	24.50	4.76	15.21
LLaVA-NeXT-Med [24]	18.42	23.15	27.82	30.20	42.32	25.10	3.25	1.85	1.02	1.72	1.45	0.82	1.70	13.40
HuatuoGPT-V-7B [115]	26.05	22.62	25.42	26.32	13.65	24.70	0.75	0.82	1.10	1.20	0.55	1.42	0.90	12.80
HuatuoGPT-V-34B [115]	30.84	35.71	49.30	56.98	71.11	41.96	35.18	24.92	36.32	65.69	25.48	24.80	36.02	38.99
HealthGPT-M3 [38]	46.65	43.32	47.52	73.85	73.62	53.10	31.22	19.70	27.75	42.22	20.22	18.65	25.90	39.50
HealthGPT-XL32 [38]	41.71	42.21	59.15	69.43	66.67	51.53	30.64	25.83	27.98	51.12	17.02	8.00	27.80	39.67
MedGemma-4B [54]	33.70	30.62	42.24	60.42	40.82	40.30	28.90	12.18	27.25	40.75	21.78	38.52	30.10	35.20
MedVLM-R1 [47]	25.13	36.36	37.32	49.81	46.67	35.23	22.42	13.71	24.42	43.88	13.57	25.80	24.70	29.97
MedDr [27]	26.56	21.43	24.65	38.49	22.22	27.29	22.99	32.58	29.57	52.44	20.95	8.70	26.20	26.75

³ <https://platform.stepfun.com/>

⁴ <https://www.volcengine.com/product/doubao/>

Table 11: The effectiveness verification of MMOral instruction data by supervised fine-tuning.

Model	SFT			Close-ended VQA						Open-ended VQA							Avg.
	Report	VQA	Chat	Teeth	Patho	HisT	Jaw	SumRec	Overall	Teeth	Patho	HisT	Jaw	SumRec	Report	Overall	
Qwen2.5-VL-7B [13]	X	X	X	24.96	21.12	27.54	37.08	35.42	27.00	17.01	16.10	11.18	29.41	9.07	8.20	15.92	21.46
Qwen2.5-VL-7B [13]	✓	X	X	26.90	27.33	26.35	45.32	37.50	31.00	27.82	15.82	25.92	63.76	22.33	38.00	32.62	31.81
Qwen2.5-VL-7B [13]	✓	✓	X	39.12	36.65	37.73	62.92	43.75	43.60	36.22	31.92	32.49	78.47	40.93	4.30	35.73	39.67

G Limitations

The ground truth reports generated in this project were based on the ground truth labels provided by previous studies published in esteemed international peer-reviewed journals [48, 19] and globally recognized preprint and dataset platforms [1, 2, 4, 5, 3, 6, 7, 25]. However, the potential inaccuracies in the provided ground truth labels cannot be entirely neglected, as their accuracy has not been validated by independent third-party organizations. Given the considerably large volume of annotated data used to construct the visual specialist model (e.g., 10 datasets comprising 28,375 images), it is not practical for a single centre to manually verify the accuracy of these ground truth labels in a short period of time.

Nevertheless, we have utilized multiple visual specialist models with overlapping category spaces to identify the same anatomical and pathological structures, thereby minimizing the risk of potential inaccuracies in the generated ground truth reports. For instance, ten structures (e.g., caries, periapical lesion, impacted teeth, missing teeth area, filling, implant, root canal treatment, crown, mandibular canal, and maxillary sinus) are validated by two or more visual specialist models, with the final results obtained through post-processing. In addition, we have adopted a two-stage LLM-based scheme of generation followed by correction to ensure the report quality. First, we utilize the DeepSeek-R1-Distill-Llama-70B to generate preliminary reports. Subsequently, through manual review of these preliminary reports, we identify common errors and summarize them into eight key rules (see in Figure 8) for prompting GPT-4-turbo to revise reports. An analysis of the revision logs shows that 95.45% of the reports are successfully corrected, leading to a significant improvement in their overall quality. Future efforts should focus on third-party validation of ground truth accuracy in these public datasets to further ensure their reliability.

H Experiments Compute Resources

The experimental section of this paper, involving the construction of the MMOral dataset and the evaluation of MMOral-Bench, requires the use of a proprietary LLMs API. The total cost of the experiments is approximately 1000 USD, with around 600 USD spent on building the MMOral dataset and about 400 USD on evaluating existing LVLMS. Furthermore, the SFT experiments conducted in this paper are performed on 4×NVIDIA A100 80G GPUs.

I Case Study

In this section, we provide additional examples of the performance of various models on closed-ended and open-ended QA tasks. Figures 18 - 23 show examples of closed-ended QA, while Figures 24 - 26 show examples of open-ended QA.

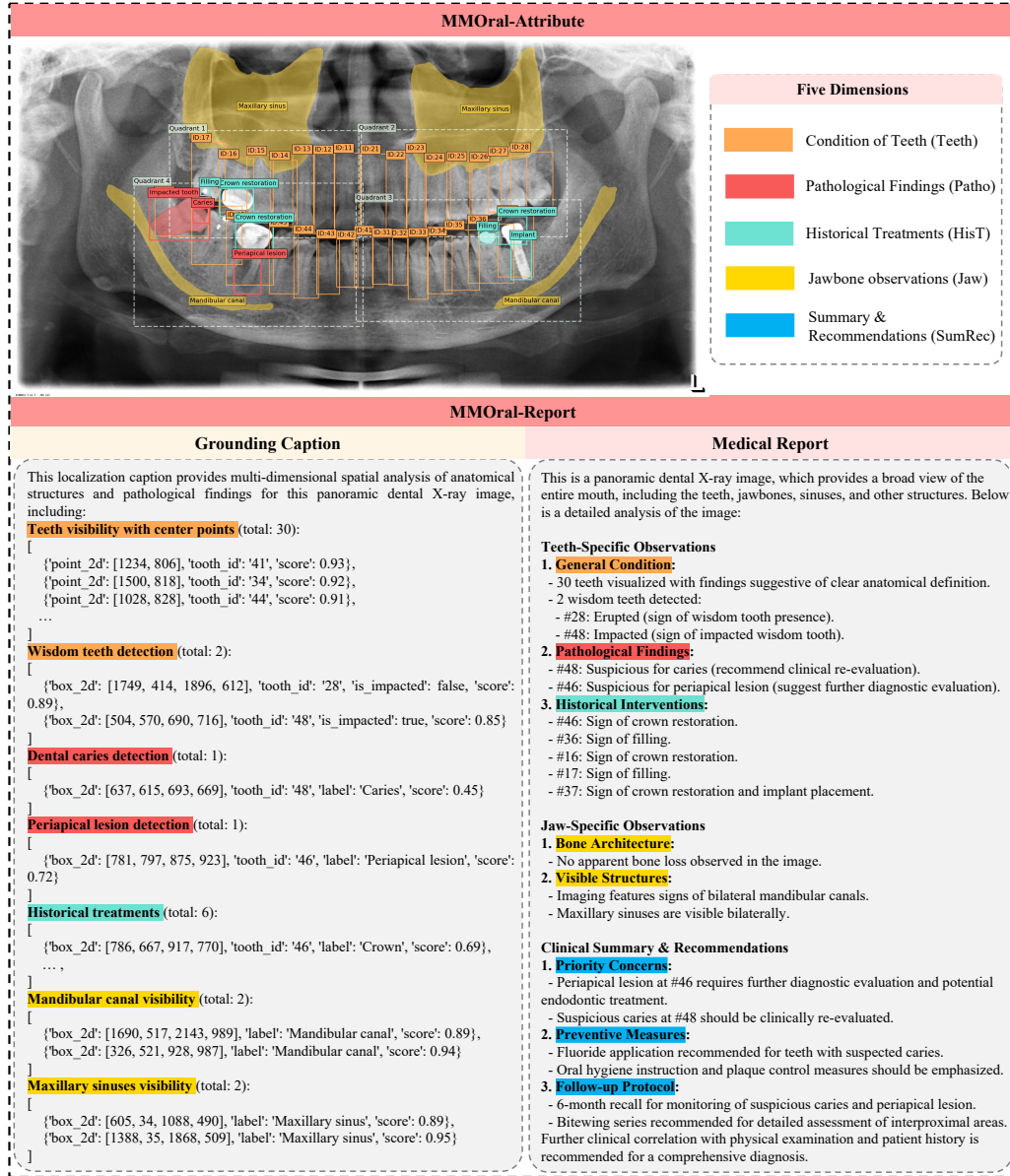


Figure 14: An example of MMOral-Attribute and MMOral-Report.

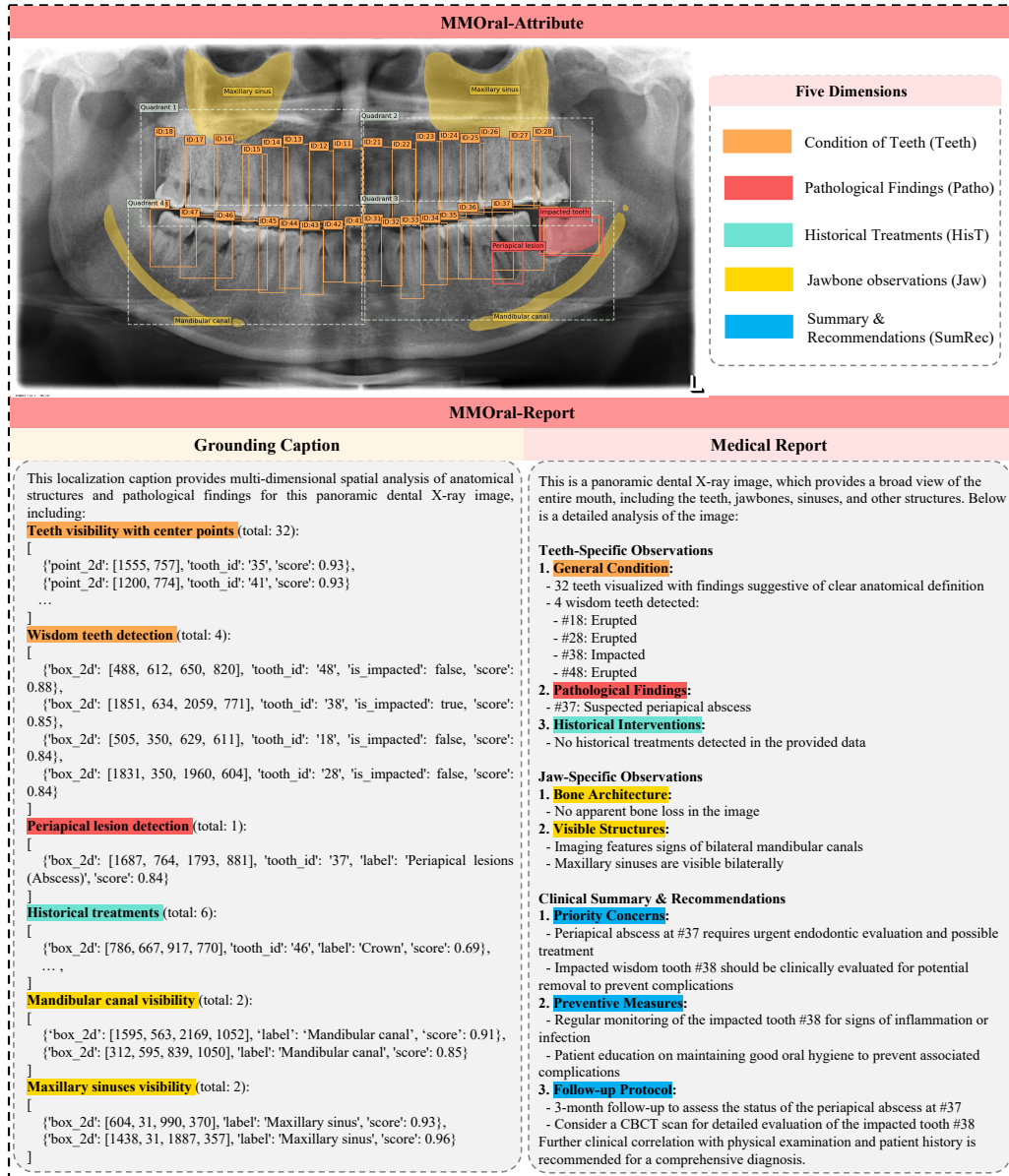


Figure 15: An example of MMOral-Attribute and MMOral-Report.

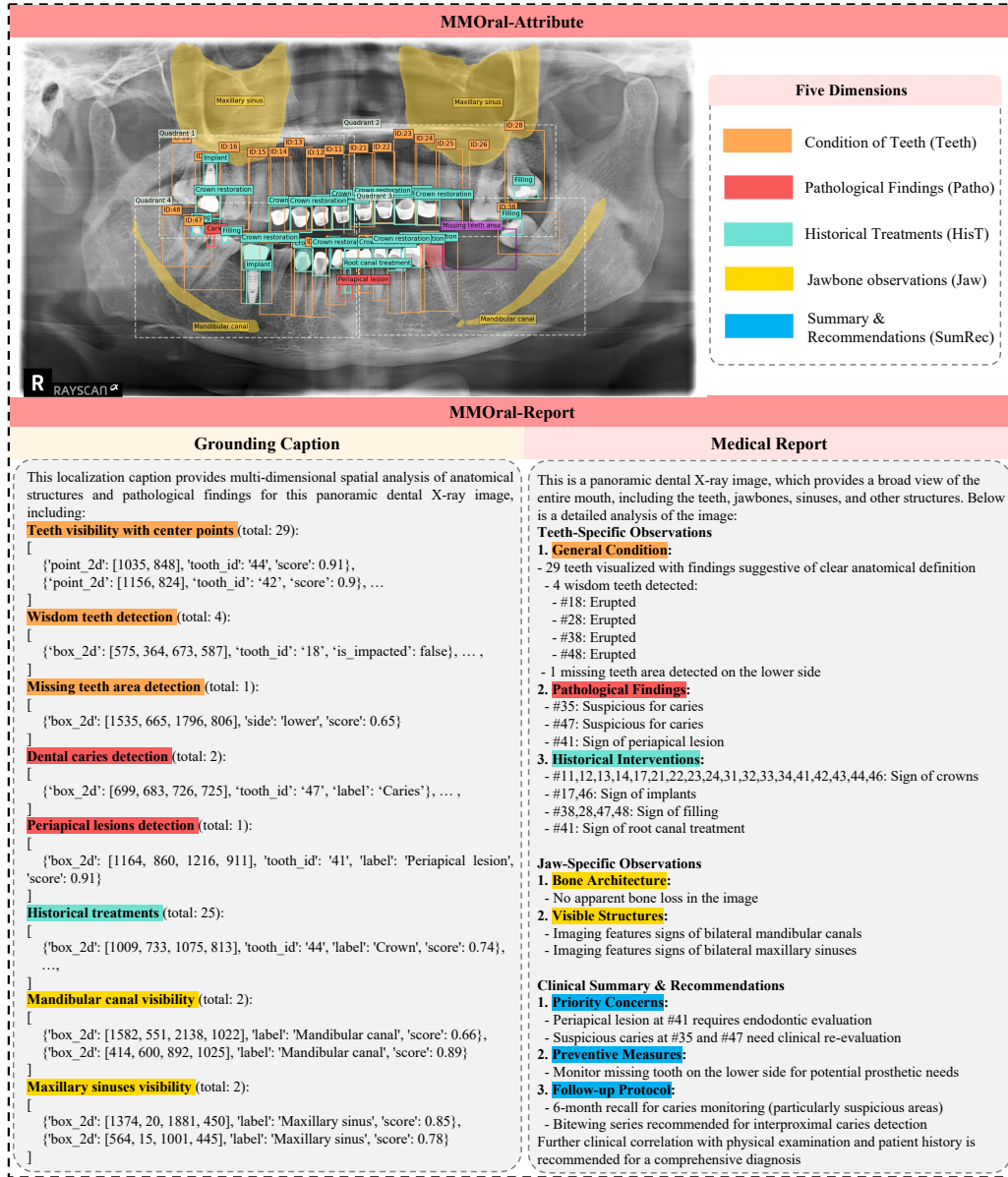


Figure 16: An example of MMOral-Attribute and MMOral-Report.

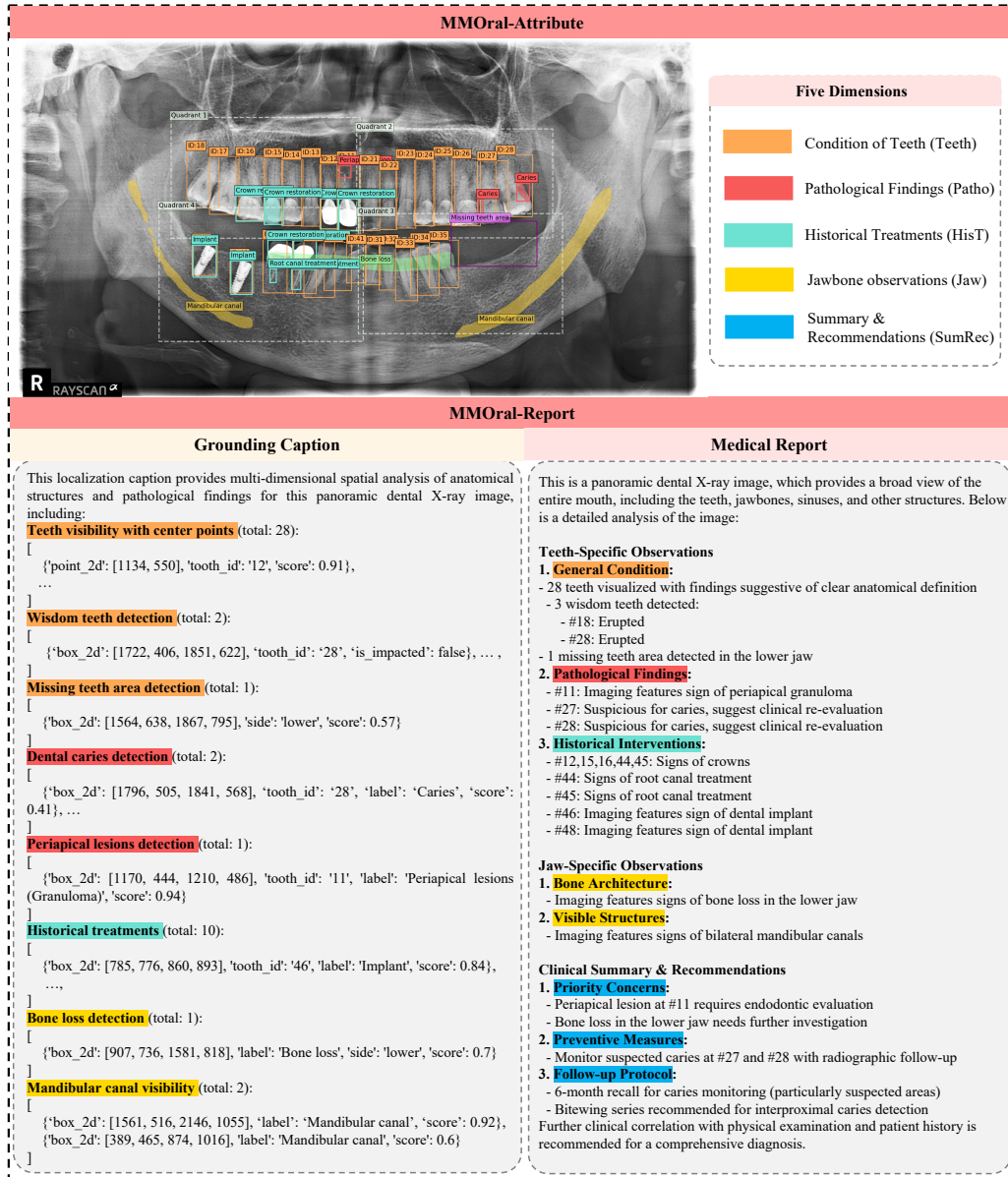


Figure 17: An example of MMOral-Attribute and MMOral-Report.

Closed-Ended QA



Q: Which teeth are suspected to have deep caries?

Options:

A. #15, #25, #35, #45

B. #18, #28, #38, #48

C. #11, #21, #31, #41

D. #52, #62, #72, #82

(Category: Teeth, Patho)

GPT-4V Response:



The image provided is a panoramic dental X-ray. Upon examining the X-ray, the third molars (wisdom teeth) appear to have significant radiolucency, which is indicative of deep caries. These teeth are numbered as #18, #28, #38, and #48.

Therefore, the correct option is:

B. #18, #28, #38, #48

Figure 18: A closed-ended QA example. **Red** highlights the right answer.

Closed-Ended QA



Q: Which of the following is TRUE regarding the wisdom teeth in this patient?

Options:

- A. All wisdom teeth are impacted
- B. Wisdom teeth are partially erupted
- C. No wisdom teeth are present

D. All wisdom teeth are erupted

(Category: Teeth)

GPT-4V Response:

The panoramic X-ray image shows that there are no visible third molars (wisdom teeth) present in the patient's mouth. This indicates that the wisdom teeth may have been extracted or are congenitally missing.

Therefore, the correct answer is:

C. No wisdom teeth are present

Figure 19: A closed-ended QA example. **Red** highlights the right answer. **Blue** highlights the wrong answer.

Closed-Ended QA



Q: Which tooth has a suspected deep caries?

Options:

A. #37

B. #38

C. #39

D. #40

(Category: Teeth, Patho)

Doubao-1.5-Vision-Pro Response:



In many cases, the third molar (#38) is more likely to have deep caries due to its position (partially or fully impacted, difficult to access for cleaning).

Correct option: \boxed{B}

Figure 20: A closed-ended QA example. **Red** highlights the right answer.

Closed-Ended QA



Q: Which tooth has a periapical lesion?

Options:

A. #36

B. #46

C. #26

D. #16

(Category: Teeth, Patho)

Doubao-1.5-Vision-Pro Response:



To answer this question, we need to recall the tooth numbering system. In the FDI (Federation Dentaire Internationale) system:

- #16 is the maxillary right first molar
- #26 is the maxillary left first molar
- #36 is the mandibular left first molar
- #46 is the mandibular right first molar

Looking at the radiograph, the periapical granuloma (a radiolucent area at the root apex) is visible on the mandibular left first molar.

So the answer is boxed{A}.

Figure 21: A closed-ended QA example. **Red** highlights the right answer. **Blue** highlights the wrong answer.

Closed-Ended QA



Q: Which of the following is true regarding historical interventions?

Options:

- A. None detected
- B. Root canal treatment found
- C. Implants present
- D. Crowns visible

(Category: HisT)

MedDr Response:



A. None detected

Figure 22: A closed-ended QA example. Red highlights the right answer.

Closed-Ended QA



Q: What object is located within the coordinates [496, 548, 693, 696]?

Options:

- A. Wisdom tooth #48
- B. Missing tooth on tooth #48
- C. Normal tooth on tooth #27
- D. Impacted tooth on tooth #24

(Category: Teeth)

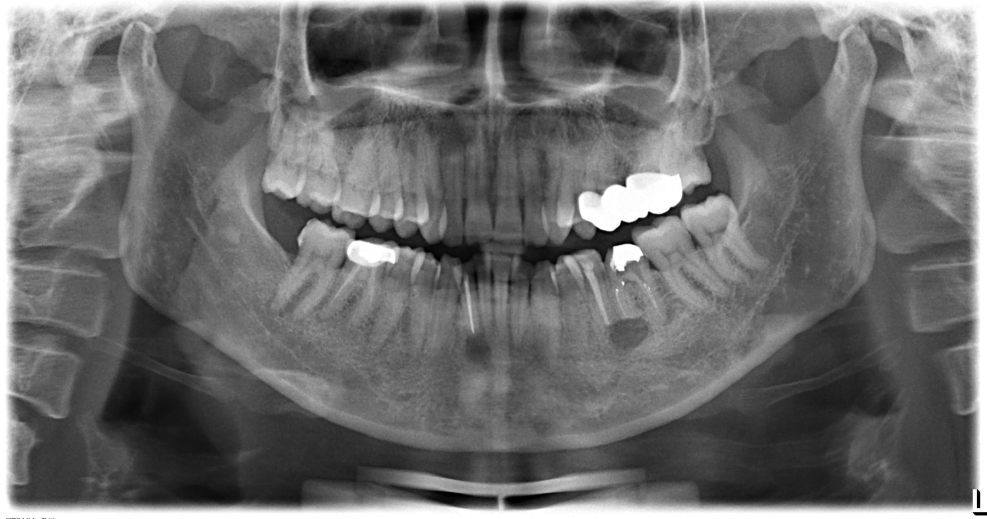
MedDr Response:



D. Impacted tooth on tooth #24

Figure 23: A closed-ended QA example. Red highlights the right answer. Blue highlights the wrong answer.

Open-Ended QA



Q: What interventions are noted on tooth #46?

A: Tooth #46 shows signs of a crown and root canal treatment.

(Category: Teeth, HisT)

GPT-4V Response:

Evaluation score: 1.0

Tooth #46, which is the lower right first molar, exhibits several interventions based on the provided panoramic radiographic image:

1. **Dental Crown**: The tooth has a radiopaque structure covering its occlusal surface, indicating the presence of a dental crown. This suggests that the tooth has been restored, likely due to extensive decay or structural damage.
2. **Root Canal Treatment**: The presence of radiopaque material within the root canals of tooth #46 indicates that it has undergone endodontic treatment. This is typically done to remove infected pulp tissue and seal the canals to prevent further infection.

These interventions suggest that tooth #46 has been treated for significant dental issues, such as decay or infection, requiring both structural restoration and endodontic therapy.

Figure 24: An open-ended QA example. Red highlights the right description.

Open-Ended QA



Q: Please detect the mandibular canal in the panoramic image.

A: [{"box_2d": [1577, 656, 2099, 1060], "label": "Mandibular canal"},
{"box_2d": [370, 691, 924, 1090], "label": "Mandibular canal"}]

(Category: Jaw)

GPT-4V Response:

Evaluation score: 0.5

In the provided panoramic radiographic image, the mandibular canal can be identified as a radiolucent (darker) linear structure within the mandible. It typically runs bilaterally from the mandibular foramen, located near the angle of the mandible, to the mental foramen, which is situated near the premolar region.

In this image, the mandibular canal appears as a dark band below the roots of the lower teeth. It is visible on both the left and right sides of the mandible. Starting from the posterior aspect near the angle of the mandible, the canal extends anteriorly, maintaining a relatively consistent path parallel to the lower border of the mandible. As it approaches the premolar region, it becomes closer to the alveolar crest, and its anterior termination is near the mental foramen, which is typically located between the roots of the first and second premolars.

The visibility of the mandibular canal can vary depending on the quality of the image and the anatomical variations of the patient. In this image, the canal is reasonably well-defined, allowing for clear identification.

Figure 25: An open-ended QA example. Red highlights the right description.

Open-Ended QA



Q: Which tooth shows signs of a periapical abscess?

A: Tooth #46 shows signs of a periapical abscess.

(Category: Teeth, Patho)

GPT-4o Response:

Evaluation score: 1.0

In the provided panoramic radiograph, a periapical abscess is typically identified by the presence of a radiolucent (dark) area around the apex (tip) of the root of a tooth, indicating bone loss or infection in the periapical region.

Upon examining the image:

- The **lower right first molar (tooth #46)** shows a radiolucent area around its root apex. This is indicative of a **periapical abscess**. The tooth appears to have a large restoration (filling), which could have led to pulpal necrosis and subsequent infection at the root apex.

Key Observations:

1. **Radiolucency at the apex:** The dark area around the root tip of the lower right first molar is a hallmark sign of a periapical abscess.
2. **Restoration:** The tooth has a large filling, which may have compromised the pulp, leading to infection.
3. **Bone loss:** The radiolucency suggests localized bone destruction due to the abscess.

Conclusion:

The **lower right first molar (tooth #46)** is the tooth showing signs of a periapical abscess. This condition requires prompt dental intervention, such as root canal therapy or extraction, to address the infection.

Figure 26: An open-ended QA example. Red highlights the right description.