

# DentalGPT: Incentivizing Multimodal Reasoning in Dentistry

Anonymous ACL submission

## Abstract

Reliable interpretation of multimodal dental data is essential for automated oral healthcare, yet existing multimodal large language models (MLLMs) exhibit limited understanding of dental images. Although complex reasoning improves performance, its gains in dentistry are substantially smaller than in other medical domains, indicating that complex reasoning is not yet sufficiently incentivized for dental diagnosis, likely due to insufficient domain knowledge and limited reinforcement learning on dental questions. We present **DentalGPT**, a dentistry-specialized MLLM trained via staged multimodal alignment and reinforcement learning. By constructing the largest annotated multimodal dental dataset to date with over 120k images, multimodal alignment provides the necessary domain knowledge foundation to support and incentivize complex reasoning, which is further strengthened through reinforcement learning. Experiments on expert-annotated benchmarks and dental subsets of medical VQA benchmarks show that DentalGPT achieves superior performance on disease classification and dental VQA tasks, outperforming many state-of-the-art MLLMs despite its compact 7B parameter scale.

## 1 Introduction

Dental healthcare is an essential area of public health, yet the workload of dental professionals continues to increase each year (Centers for Disease Control and Prevention, 2024; Jiang et al., 2024; Zheng et al., 2025; Lilford et al., 2025). To support both clinicians and patients, multimodal large language models (MLLMs) (Liu et al., 2023; Bai et al., 2025b; Team et al., 2023; Chen et al., 2024d) capable of interactive communication through dialogue have recently attracted significant attention, offering new possibilities for intelligent dental care. However, despite their promising performance in general medical applications (Awadalla et al., 2023; Li et al., 2023; Wu

et al., 2023, 2024a; Chen et al., 2024b; Su et al., 2025; Xu et al., 2025), current MLLMs still face notable limitations when dealing with more specialized medical imaging problems, such as images in dentistry.

In recent studies (Su et al., 2025; Bai et al., 2025a; Team et al., 2025b; Pan et al., 2025), enabling complex reasoning has yielded substantial gains on challenging multimodal benchmarks. Motivated by this, we evaluate mainstream MLLMs on a dentistry benchmark (Hao et al., 2025a) with and without complex reasoning mode, and observe consistent improvements when reasoning is enabled; however, these gains are markedly smaller than those reported in broader medical settings (Su et al., 2025; Zhang et al., 2025; Pan et al., 2025), suggesting that complex reasoning in dentistry is not yet sufficiently incentivized. Based on this observation, we hypothesize two challenges: 1) insufficient visual understanding of dental images, likely due to scarce dental VQA data, which narrows the model’s reasoning perspectives and limits its exploration space for dental questions; and 2) limited reinforcement learning on dentistry-related questions, which constrains the acquisition of domain-specific reasoning patterns.

To address these challenges, we inject dentistry-related knowledge into MLLMs and incentivize complex reasoning. Specifically, we curate a large-scale dental dataset by combining online dental images with textual descriptions or labels and professionally annotated images from dental hospitals, resulting in over 120k images with detailed descriptions and task-specific QA pairs. The descriptions emphasize diagnostically relevant visual cues to improve multimodal alignment, while the QA data supports downstream dental tasks. After that, we apply a reinforcement learning stage using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to encourage more explanatory reasoning for dental questions. This process yields **DentalGPT**,

085 a dentistry-specific MLLM equipped with complex  
086 reasoning capabilities.

087 A comprehensive evaluation was conducted to  
088 assess DentalGPT’s dental image analysis capa-  
089 bility. We first evaluated question answering using  
090 existing dentistry-focused VQA benchmarks,  
091 and then assessed disease identification using pro-  
092 fessionally annotated benchmarks. Despite hav-  
093 ing only 7B parameters, DentalGPT outperforms  
094 existing MLLMs in dental image understanding  
095 and question answering, demonstrating strong effi-  
096 ciency and domain specialization. We further ana-  
097 lyze the impact of the two training stages through  
098 detailed ablations. The multimodal alignment stage  
099 substantially enriches dental knowledge and im-  
100 proves performance across image analysis tasks.  
101 These gains are further amplified by reinforcement  
102 learning, leading to higher disease classification  
103 accuracy and more professional identification of  
104 diagnostically relevant visual cues.

105 In summary, our contributions are: (1) We intro-  
106 duce **DentalGPT**, a dentistry-specialized MLLM  
107 capable of fine-grained dental image understand-  
108 ing and multimodal complex reasoning. (2) We  
109 curate the largest multimodal dental dataset to date,  
110 comprising over 120k dental images with detailed  
111 annotations of diagnostically relevant visual fea-  
112 tures. (3) We show that multimodal alignment,  
113 combined with targeted reinforcement learning,  
114 is crucial for incentivizing complex reasoning in  
115 domain-specific MLLMs.

## 116 2 On Incentivizing Multimodal 117 Reasoning in Dentistry

### 118 2.1 Observations

119 Dentistry relies on identifying lesion-related vi-  
120 sual cues in dental images. Although multimodal  
121 complex reasoning has shown substantial gains  
122 across medical tasks (Zhang et al., 2025; Pan et al.,  
123 2025; Su et al., 2025), its effectiveness in dentistry  
124 remains unclear. We therefore evaluate leading  
125 MLLMs<sup>1</sup> with and without complex reasoning  
126 on MMOral-OPG-Bench (Hao et al., 2025a), the  
127 only publicly available multimodal dentistry bench-  
128 mark. As shown in Figure 1, current MLLMs  
129 exhibit limited performance on dentistry-related

<sup>0</sup>Complex reasoning mode: GPT5-2025-08-07, Gemini-  
2.5-Pro-Thinking, Claude-Sonnet-4-5-20250929-Thinking,  
Qwen3-VL-235B-A22B-Thinking. **Without complex rea-  
soning mode:** GPT5-chat-2025-08-07, Gemini-2.5-Pro-  
NoThinking, Claude-Sonnet-4-5-20250929, Qwen3-VL-  
235B-A22B-Instruct.

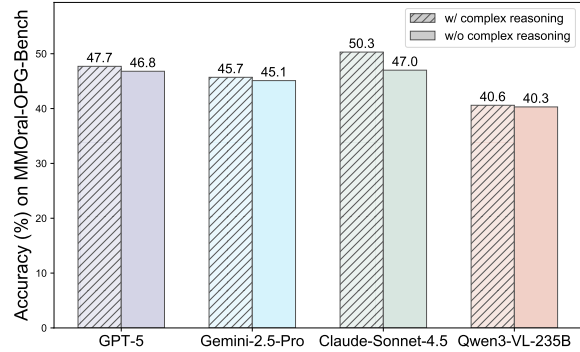


Figure 1: Accuracy (%) of MLLMs *with and without the complex reasoning mode* on the MMOral-OPG-Bench.

multimodal tasks, indicating insufficient ability to  
perceive lesion-related cues in dental images. This  
is further supported by Appendix A, leading to the  
following observation.

**Observation 1:** *Current MLLMs struggle to reliably perceive lesion-related visual cues in dental images.*

Moreover, by comparing multiple MLLMs with  
and without the complex reasoning mode enabled,  
we further observe that:

**Observation 2:** *Complex reasoning consistently improves performance, yet yields limited gains in dentistry.*

### 2.2 Analysis

Based on the above observations, we analyze why  
complex reasoning remains insufficiently incen-  
tivated in dentistry-related multimodal tasks. We  
group the potential reasons into two complemen-  
tary levels: *perception* and *cognition*.

**Perception-Level Limitation** *Observation 1* sug-  
gests that insufficient multimodal alignment in den-  
tistry limits visual grounding, narrows reasoning  
perspectives, and constrains the exploration space  
needed for incentivizing complex reasoning.

**Cognition-Level Limitation** *Observation 2* in-  
dicates that limited gains from complex reason-  
ing stem from insufficient dental-specific optimiza-  
tion, as task-specific reinforcement learning for  
dentistry remains underexplored (Su et al., 2025;  
Zhang et al., 2025; Pan et al., 2025).

## 3 DentalGPT: From Multimodal Alignment to Complex Reasoning

To address perception- and cognition-level limita-  
tions in dental imaging, we propose a dentistry-  
specific framework to incentivize multimodal com-  
plex reasoning (Figure 2), with Stage I focusing on

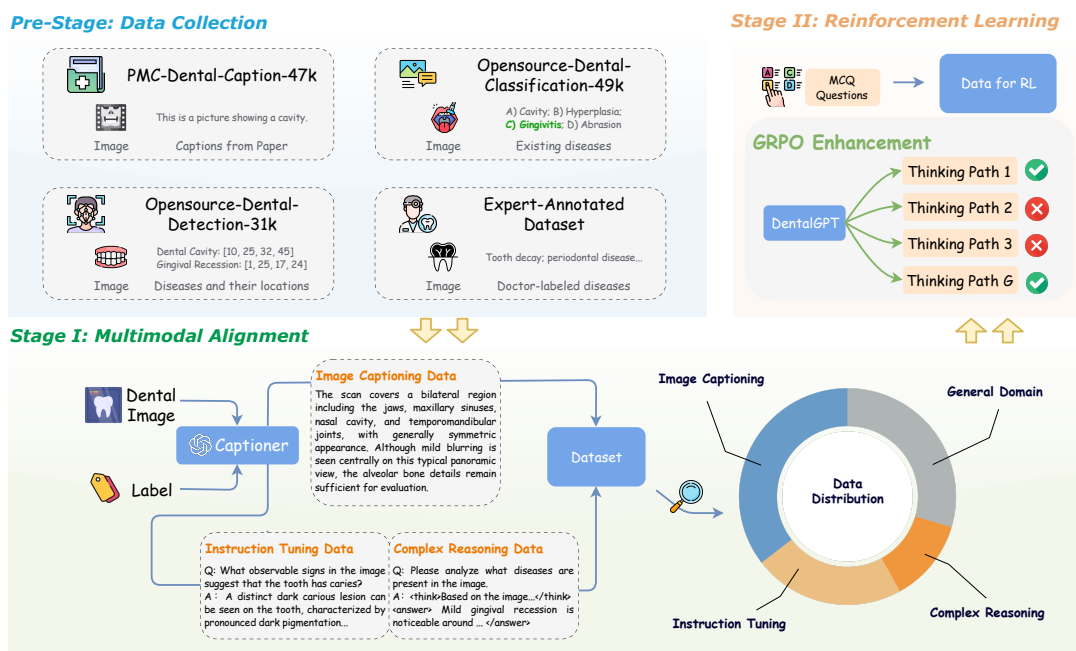


Figure 2: The process of building DentalGPT. Dentistry-related datasets from multiple sources are collected at the *Pre-stage: Data Collection*; *Stage I: Multimodal Alignment* uses the dataset to align the model’s medical knowledge with its multimodal understanding; *Stage II: Reinforcement Learning* then strengthens complex reasoning ability.

Dataset	Annotation Type	Learning Signal
PMC-Dental-Caption-47k	Image–text pairs from PubMed Central	Visual understanding with dental domain knowledge grounding
Opensource-Dental-Classification-49k	Expert-defined disease labels	Disease classification and key visual feature recognition
Opensource-Dental-Detection-31k	Bounding boxes for teeth and lesions	Spatial reasoning, localization, and lesion counting

Table 1: Overview of dental image datasets used for multimodal training, covering captioning, classification, and detection supervision.

perception via multimodal alignment and Stage II on cognition via dentistry-oriented reinforcement learning.

### 3.1 Pre-stage: Data Collection

To support a comprehensive understanding of dental knowledge, diverse dental images were collected from multiple sources, including intraoral photographs, panoramic radiographs, and partial panoramics. These datasets provide complementary information that enhances the model’s ability in dental knowledge learning, disease recognition, and spatial perception of lesions. Details of collection and filtering are provided in Appendix B.

Table 1 summarizes the dental image datasets used for multimodal training. PMC-Dental-Caption-47k provides rich image–text supervision from biomedical literature, enabling visual grounding and integration of dental domain knowledge. Opensource-Dental-Classification-49k supplies expert-labeled disease categories, supporting robust dental disease recognition. Opensource-

Dental-Detection-31k further introduces spatial supervision through bounding box annotations, allowing the model to localize lesions, reason about spatial relationships, and quantify abnormalities. Together, these datasets offer complementary supervision signals for learning visual understanding, clinical semantics, and spatial reasoning in dentistry.

**Expert-Annotated Dataset** A subset of dental images was selected from internet sources, hospital imaging databases, and open-source datasets. After removing duplicates with the test set, professional dentists annotated these images with disease categories. Labels with high cross-validation agreement and sufficient sample sizes were retained for model evaluation.

### 3.2 Stage I: Multimodal Alignment

To address the perception-level limitations of existing MLLMs in dentistry, this stage enhances multimodal alignment by incorporating high-quality,

domain-specific dental knowledge. We construct a large-scale, professionally curated image–text dataset to improve visual grounding and broaden perceptual coverage. GPT-5 is used to generate detailed descriptions, strengthening fine-grained visual understanding and supporting basic downstream dental tasks.

**Data Construction** A large and comprehensive dataset was constructed, consisting of several components (Details can be found at Appendix C):

- **Image Captioning** This component enhances the model’s ability to capture diagnostically relevant visual details in dental images. GPT-5 is instructed to describe all observable features that may aid diagnosis, while referencing original descriptions and labels and avoiding diagnostic assumptions. These observation-based captions are paired with predefined questions and answers to form a caption-based VQA subset, improving dental image interpretation and reducing visual information gaps.
- **Instruction Tuning** To improve performance on downstream tasks, GPT-5 is prompted to generate questions from the collected images and descriptions, simulating real diagnostic scenarios. In addition, multiple vision models annotate a subset of the data, from which only high-confidence labels are retained. GPT-5 then refines these annotations and converts them into structured question–answer pairs, further strengthening the model’s dental question-answering capability.
- **Complex Reasoning** To support the subsequent post-training stage, complex chain-of-thought data were constructed by GPT-5 following the HuatuoGPT-o1 (Chen et al., 2024a) methodology. Fixed reasoning templates were added to activate the model’s basic multi-step reasoning ability, enabling it to analyze before answering.
- **General Domain** General-domain and general medical data (Liu et al., 2023; An et al., 2025; Chen et al., 2024c) were included to preserve the model’s broad visual understanding and prevent degradation of language capabilities during domain specialization.

After these construction steps were completed, GPT-5-mini was used to perform a secondary verification of all data. Entries that diverged from the original image descriptions or labels were removed, further ensuring data accuracy.

**Training Settings** The model was then trained on this dataset for two epochs with a batch size of 256 and a learning rate of  $2 \times 10^{-5}$ . All parameters were fully updated during training, with the first 5% of steps allocated for learning-rate warmup.

### 3.3 Stage II: Reinforcement Learning

After gaining new knowledge, the MLLM must learn to apply it for improved complex reasoning in multimodal diagnosis. Recent works such as DeepSeek-R1 (DeepSeek-AI, 2025) and GPT-o1 (Jaech et al., 2024) show that reinforcement learning can encourage long chain-of-thought generation and enhance reasoning quality. Following this paradigm, we adopt GRPO to optimize the reasoning process of *DentalGPT*.

**Data Composition** To achieve this goal, we selected a set of dental images that were not used during the Stage I to construct a new dataset. Based on the original labels and their label sets, we generated multiple-choice questions with correct answers, enabling rule-based correctness checking, which is crucial for reward computation in GRPO.

**Training Strategy** During this stage, we adopt the GRPO algorithm to improve the model’s reasoning ability on dental multiple-choice tasks. GRPO optimizes relative advantages among sampled responses, enabling efficient training without a value network. We use answer accuracy as the primary reward and combine it with a format reward to enforce a fixed response template, weighted 90% and 10%, respectively.

**Training Settings** During GRPO optimization, the model was optimized using grouped rollouts with a sampling size of 10 responses per prompt. Training was conducted with a rollout batch size of 256 and a learning rate of  $1 \times 10^{-6}$ . The optimization ran for 5 epochs, and the maximum response length was capped at 8192 tokens to accommodate long CoT reasoning. This configuration ensured stable exploration within each action group while maintaining sufficient capacity for detailed reasoning outputs.

## 4 Expert-annotated Benchmark

To comprehensively evaluate the model, a large set of dental images was collected and annotated with disease labels by professional dentists, ensuring clinical validity and allowing further assessment aligned with expert consensus.

## 4.1 Expert Annotation Workflow

**Label Definition** To maximize data diversity and ensure clinically reliable model outputs, we collected dental images from multiple sources and invited professional dentists to perform expert annotations. Guided by clinical practice, we defined a set of commonly observed dental diseases and diagnostically relevant signs that either directly indicate pathology or serve as auxiliary evidence in clinical reasoning. These labels are used to assess whether models can correctly identify key visual cues in dental images.

**Cross-Validation** To ensure annotation reliability, we implemented a rigorous cross-validation workflow. Each image was independently annotated by at least 2 dentists, who selected all clinically relevant labels present in the image. Annotators could mark images as *uncertain* when visual quality was insufficient, reducing forced or ambiguous judgments. Labels with any disagreement in cross-validation were removed, and only fully consistent labels were retained.

**Benchmark Curation** After expert annotation and quality filtering, only high-confidence and clinically reliable labels were retained to form the final dataset. This curated dataset serves as the foundation for constructing expert-annotated benchmarks and enables rigorous evaluation of multimodal diagnostic performance.

## 4.2 Benchmark Composition

Based on the above annotation workflow, we construct three benchmarks to evaluate the multimodal diagnostic performance of DentalGPT. These benchmarks cover both clinical and in-the-wild dental images across multiple modalities, enabling systematic assessment of reliability, generalization, and robustness in diverse dental scenarios. Each benchmark supports multi-label classification across both clinical and in-the-wild scenarios, enabling comprehensive evaluation of multimodal diagnostic ability. To avoid biased evaluation, we further applied strict data balancing strategies: the label distributions and the ratios of positive and negative samples for each category were aligned across subsets. This ensures that accuracy reliably reflect model performance rather than being influenced by label frequency or overrepresented diseases.

Table 2 summarizes three dental image classification benchmarks with complementary evaluation

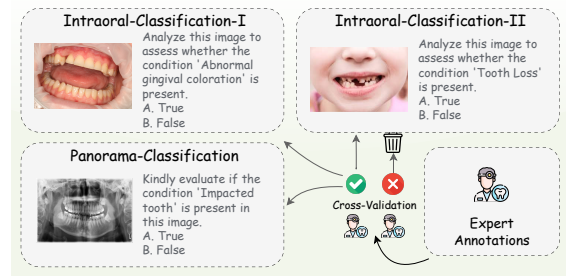


Figure 3: Examples of expert-annotated benchmarks. Only samples that pass cross-validation are retained.

settings. **Intraoral-Classification-I** uses dentist-acquired intraoral photographs collected under standardized clinical conditions and evaluates performance on high-quality, in-distribution images. **Intraoral-Classification-II** focuses on robustness and generalization, using patient-captured, in-the-wild images with diverse lighting and viewpoints. **Panorama-Classification** evaluates panoramic radiographs that reveal internal anatomical and pathological information beyond surface appearance. Together, these benchmarks assess clinical accuracy, robustness to real-world data, and cross-modality diagnostic capability.

## 5 Experiment

### 5.1 Evaluation of Dentistry Expertise

**Training Settings** DentalGPT is built upon Qwen2.5-VL-7B-Instruct. In both training stages, all model parameters are fully updated to enable domain adaptation for dentistry and complex reasoning. Detailed hyperparameters and implementation details are provided in Section 3. All experiments are conducted on a cluster with 8 NVIDIA H200 GPUs.

**Evaluation Settings** We conduct evaluations on the curated dentistry-specific benchmark (Section 4), which assesses models across dental disease classification, lesion recognition, and common dental consultation scenarios. All models are required to provide responses to the given tasks. For models marked with an asterisk (\*), complex reasoning mode is enabled, while other models are instructed to directly output the correct option without additional reasoning steps.

**Benchmarks** We evaluate DentalGPT using both existing and newly constructed benchmarks to assess dental multimodal diagnostic ability. Existing benchmarks measure general dental image understanding, including *MMOral-OPG-Bench* (Hao

Benchmark	Data Source and Key Characteristics	Label Set
<b>Intraoral-Classification-I</b>	Dentist-acquired intraoral photographs from AlphaDent (Sosnin et al., 2025), captured under standardized clinical conditions with consistent lighting and viewpoints, representing in-distribution clinical images.	Tooth discoloration; abnormal gingival coloration; gingival recession; dental caries; tooth pigmentation; tooth defect or loss; tooth loss; dental calculus; abnormal tooth morphology; abnormal gingival morphology.
<b>Intraoral-Classification-II</b>	Patient-captured intraoral images collected from the internet, exhibiting diverse lighting, resolution, and shooting angles, reflecting non-clinical, in-the-wild conditions and domain shift.	Tooth pigmentation; abnormal gingival coloration; dental calculus; tooth loss; dental caries; abnormal gingival morphology; gingival recession.
<b>Panorama-Classification</b>	Hospital-collected panoramic radiographs (X-ray), providing internal anatomical and pathological cues beyond surface appearance and requiring radiographic interpretation.	Periodontal disease; root canal treatment; tooth defect or loss; jawbone lesion; periapical lesion; impacted tooth.

Table 2: Three dental image classification benchmarks highlighting differences in acquisition setting (clinical vs. in-the-wild), image modality (photograph vs. radiograph), and clinical complexity.

Model	MMOral OPG-Bench	DentalBench Mixed	Intraoral Classification-I	Intraoral Classification-II	Panorama Classification	Avg.
Open-source MLLMs						
Deepseek-VL2 (Wu et al., 2024b)	39.1	22.6	51.1	59.4	55.1	45.5
Mistral-Large-2512 (AI, 2024)	41.9	48.2	50.7	58.0	44.2	48.6
Phi-4-Multimodal-Instruct (Abouelenin et al., 2025)	38.5	44.4	52.2	63.3	61.5	52.0
Ernie-4.5-VL-424B-A47B* (Baidu-ERNIE-Team, 2025)	45.0	51.4	58.1	65.1	44.9	52.9
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025a)	40.3	51.6	50.7	58.0	55.8	51.3
Gemma-3-27B-it (Team, 2025)	42.2	43.0	51.5	61.4	59.6	51.5
GLM-4.5v* (Team et al., 2025b)	45.7	51.4	54.8	64.7	54.5	54.2
Qwen3-VL-235B-A22B-Thinking* (Bai et al., 2025a)	40.6	51.6	56.7	65.7	60.3	55.0
LLaMA-4-Maverick (AI, 2025)	<u>51.4</u>	53.9	<u>61.1</u>	67.1	59.0	58.5
Proprietary MLLMs						
Claude-Sonnet-4.5 (Anthropic, 2025)	47.0	50.4	51.9	59.4	50.0	51.7
Claude-Sonnet-4.5-Thinking* (Anthropic, 2025)	50.3	53.9	55.2	66.7	55.8	56.4
Grok-4.1-Fast	47.1	52.2	57.0	65.2	62.2	56.7
Gemini-2.5-Pro-Thinking* (Comanici et al., 2025)	45.7	<b>57.4</b>	57.0	65.2	<u>64.1</u>	57.9
GPT-4.1 (OpenAI, 2024)	47.2	51.7	60.4	70.5	61.5	58.3
GPT-5* (OpenAI, 2025)	47.7	54.3	59.3	<u>71.0</u>	63.5	<u>59.2</u>
DentalGPT and Its Backbone						
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	27.0	46.1	48.8	61.8	50.0	46.7
<b>DentalGPT*</b>	<b>60.0</b>	<u>54.4</u>	<b>64.1</b>	<b>72.9</b>	<b>84.0</b>	<b>67.1</b>

Table 3: Accuracy (%) of MLLMs on Dental-related VQA Benchmarks. **Bold** indicates the best score; underlines marks the second-best. \* indicates that the model has activated complex reasoning.

et al., 2025a), which contains panoramic dental X-ray images with expert annotations and uses the open-ended test split, and **DentalBench-Mixed**, built by filtering tooth-related samples from medical VQA benchmarks such as PMC-VQA (Zhang et al., 2023), OmniMedVQA (Hu et al., 2024), and MedXpertQA-MM (Zuo et al., 2025). We further introduce dentistry-specific benchmarks that emphasize common dental diseases and signs, enabling more comprehensive evaluation in diagnostically critical scenarios.

**Results and Analysis** As shown in Table 3, DentalGPT consistently outperforms both comparable and substantially larger MLLMs across all expert-annotated benchmarks, achieving significant gains over its backbone and highlighting the effectiveness of the proposed data pipeline and domain-aligned

training. Despite its compact 7B scale, DentalGPT matches or surpasses many general-purpose models exceeding 100B parameters. It also maintains strong performance on *MMOral-OPG-Bench* and *DentalBench-Mixed*, demonstrating robust generalization across diverse dental tasks. Overall, these results position DentalGPT as a strong multimodal foundation model for dental image understanding.

## 5.2 Human Evaluation

To further assess DentalGPT’s capabilities, four licensed dentists were invited to evaluate the model outputs based on predefined criteria (Details in Appendix D), focusing on obvious errors, the soundness of diagnoses, and omissions of critical information.

Specifically, 80 dental images were curated from

medical textbooks and licensed dentist examinations, and each model was tasked with analyzing the images and identifying all potential dental conditions. DentalGPT was compared against several widely used general-purpose models through pairwise comparisons, and win/tie/loss statistics were reported. As shown in Figure 4, DentalGPT is preferred in most cases, with its outputs judged as more clinically accurate and better aligned with dentists. Notably, despite its smaller parameter scale, DentalGPT achieves performance comparable to leading models, offering a favorable trade-off between model size and clinical quality.

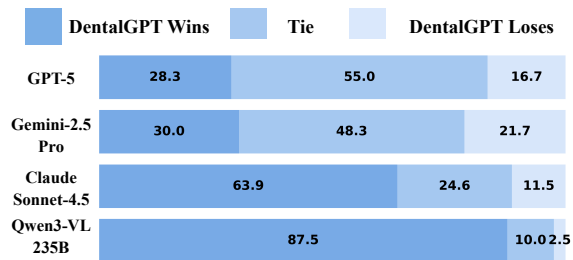


Figure 4: Results of the human evaluation. *Win/Tie/Loss* indicates the ratios of expert preferences on responses.

### 5.3 Ablation Study of Multimodal Alignment (on Stage I)

Multimodal alignment equips MLLMs with a richer understanding of dentistry-related images. In this subsection, we examine whether such alignment facilitates subsequent complex reasoning.

To characterize changes in complex reasoning, we analyze the evolution of the accuracy-based reward during reinforcement learning after alignment. In this process, the model repeatedly relies on its internalized knowledge to explore different solution paths for improving downstream task accuracy. Consequently, the accuracy reward serves as an indicator of whether the injected alignment knowledge supports more diverse exploration and more elaborate reasoning behaviors, thereby facilitating complex multimodal reasoning.

**Experimental Setup** Specifically, three controlled settings using 0%, 30%, and 100% of the Stage-I dataset were evaluated. The impact of alignment strength was assessed by tracking accuracy-based reward improvements during the subsequent RL stage. To ensure fairness, duplicate images between alignment and RL data were removed, and all *complex reasoning* samples were excluded from Stage I. Following the DentalGPT setup,

Qwen2.5-VL-7B was adopted as the backbone, and each model was trained for 30 RL steps, with reward trajectories on the validation set used to assess multimodal reasoning performance.

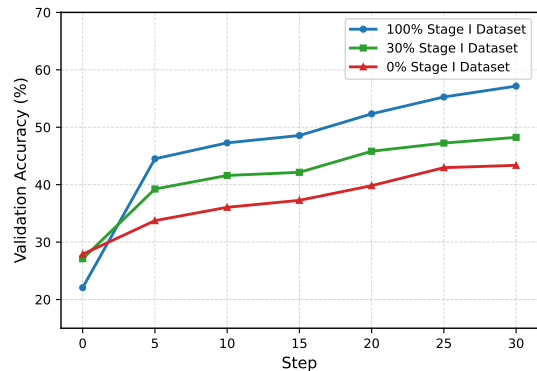


Figure 5: Accuracy reward (%) of the MLLM during RL training under different Stage I dataset scales, showing that stronger multimodal alignment facilitates more effective reward optimization.

**Results Analysis** As shown in Figure 5, the model trained with 0% Stage-I data exhibits only marginal reward gains during RL, indicating limited reasoning improvement without dental-domain alignment. In contrast, increasing the amount of Stage-I data consistently elevates the reward ceiling throughout RL training. These results suggest that *multimodal alignment provides the model with richer perspectives for exploration during the RL stage, thereby further incentivizing complex multimodal reasoning in dentistry.*

### 5.4 Ablation Study of RL (on Stage II)

This section investigates how the reinforcement learning (RL) stage further shapes the model’s capabilities. After applying RL training on 10k multiple-choice dental questions, the performance of the model was reevaluated across the same set of benchmarks. As shown in Table 4, reinforcement learning brings consistent improvements across all tasks, demonstrating that it further enhances the model’s ability to execute downstream dental tasks. These gains confirm that *RL effectively strengthens both the accuracy and reliability of the model’s reasoning in dental image understanding.*

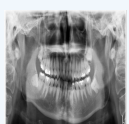
### 5.5 Case Study

We further analyze the outputs of **DentalGPT** across different training stages together with the original backbone model (Figure 6). The backbone model struggles with this example, describing some relevant visual features but failing to identify any

Benchmarks	Qwen2.5-VL Backbone	Qwen2.5-VL + Stage I w/o Stage II	DentalGPT w/ Stage I & Stage II
MMOral-OPG-Bench	27.0	56.8	<b>60.0</b>
DentalBench-Mixed	46.1	51.7	<b>54.4</b>
Intraoral-Classification-I	48.8	61.5	<b>64.1</b>
Intraoral-Classification-II	61.8	67.6	<b>72.9</b>
Panorama-Classification	50.0	78.4	<b>84.0</b>
Total	46.7	63.2	<b>67.1</b>

Table 4: Accuracy(%) comparison among Qwen2.5-VL, Qwen2.5-VL with Stage I training, and DentalGPT (Stages I + II) on dentistry-specific benchmarks.

teeth with fillings. After Stage-I multimodal alignment, the model can perform basic dental image interpretation and detect most filled teeth, missing only a less salient case, highlighting the role of multimodal alignment in enabling visual grounding and establishing a foundation for reasoning. After completing all training stages, **DentalGPT** with complex reasoning conducts a more systematic multimodal analysis and, despite intermediate inaccuracies, iteratively verifies and refines its predictions to arrive at the correct answer. This case study underscores the importance of reinforcement learning in supporting self-correction and accurate complex multimodal reasoning in dentistry.



**Input Question:**  
How many visible teeth in the image appear to have dental fillings based on their radiopaque characteristics?

**Ground Truth:**  
10

**Qwen2.5-VL-7B-Instruct (Backbone of DentalGPT)**  
Determine how many ..., we need to look for areas of increased density (radiopacity) within the tooth structure that do not correspond with natural anatomy such as enamel or dentin. ...  
Final Answer : 9

**DentalGPT w/o Complex Reasoning**  
Dental fillings are typically made from materials such as amalgam or composite resin, ...

Upper Jaw (Maxilla): Careful examination reveals three teeth with clear radiopaque spots indicative of dental fillings. These are distributed across both the left and right sides of the upper jaw.

Lower Jaw (Mandible): The lower jaw shows six teeth with similar radiopaque features, suggesting the presence of fillings. These are primarily located in the molar and premolar regions, where restorations are most common due to higher risk of decay, ...

In summary, the panoramic X-ray demonstrates 9 teeth with visible dental fillings.

**DentalGPT w/ Complex Reasoning**  
◀think> Examining the panoramic X-ray image for signs of radiopaque materials that might represent dental fillings. Fillings typically appear as bright white areas within the crown of the teeth.

Starting from the front teeth (incisors) and moving towards the molars, ... After identifying radiopaque regions, I noted that a few molars and premolars in both the upper and lower jaw displayed these features. The exact number of teeth with fillings is confirmed to be nine ...

However, after reviewing the image and reasoning, I observed discrepancies in the count. The conclusion is false, and further refinement is necessary. ... I systematically inspected each visible tooth in the panoramic image, focusing on molars and premolars where fillings are more common. ...

Starting from the right upper quadrant to the left upper quadrant, then proceeding from the left lower quadrant to the right lower quadrant, I recounted the teeth exhibiting radiopaque (bright white) regions indicative of fillings on the X-ray image. I identified ten teeth with these characteristics. ... </think>

<answer> ... In summary, the panoramic X-ray demonstrates 10 teeth with visible dental fillings, easily identified by their radiopaque appearance. </answer>

Figure 6: Case study of **DentalGPT** and its backbone on a dental multimodal task. Red and green indicate incorrect and correct analyses, while orange highlights reflective reasoning.

## 6 Related Work

**Medical MLLMs** Medical imaging is a key application area for MLLMs due to its reliance on visual data. General medical MLLMs (Chen et al., 2024b; Lin et al., 2025; Team et al., 2025a; Chao et al., 2024; Yu et al., 2025) demonstrate basic medical QA and preliminary visual analysis across diverse clinical scenarios, emphasizing broad generalization through large-scale multimodal medical data and benchmarks (Zhong et al., 2023; Liu et al., 2021; Xie et al., 2024; Cai et al., 2025).

**MLLMs for Specific Medical Domains** To address complex clinical settings, recent work adapts MLLMs to specific medical modalities, including 3D medical imaging (Wu et al., 2023, 2025b,a), biomedical signal understanding (Zhao et al., 2024; Liu et al., 2024; Chen et al., 2025b; Zeng et al., 2025), and fine-grained perception of high-resolution clinical images via specialized pipelines or training strategies (Chen et al., 2025a; Wang et al., 2025).

**MLLMs for Dentistry** Dentistry is an important and emerging multimodal domain for medical AI. In this context, DentVLM (Meng et al., 2025) leverages large-scale dental reports to support dental tasks, while OralGPT (Hao et al., 2025a,b) introduces dentistry-focused multimodal benchmarks and enables diverse dental modalities. We hope that our work and theirs can jointly contribute to the advancement of intelligent AI in dentistry.<sup>2</sup>

## 7 Conclusion

This work presents **DentalGPT**, a specialized MLLM for multimodal dental diagnosis. By constructing a large-scale multimodal dataset in dentistry and adopting a staged training strategy that combines multimodal alignment and reinforcement learning, DentalGPT effectively captures fine-grained visual cues and supports reliable disease-related reasoning. Extensive evaluations across different dental VQA benchmarks show that DentalGPT achieves strong performance with only 7B parameters, outperforming many state-of-the-art general-purpose MLLMs.

## Limitations

Despite extensive data collection efforts, the availability of dental imaging data remains limited, es-

<sup>2</sup>These models are currently not publicly released and thus cannot be evaluated in our experiments.

554	pecially for high-quality annotated datasets. As	Ding, Chang Gao, Chunjiang Ge, and 1 others.	605
555	a result, our data construction relies on expert an-	2025a. Qwen3-vl technical report. <i>arXiv preprint</i>	606
556	notation and human-in-the-loop curation, which	<i>arXiv:2511.21631</i> .	607
557	is costly and difficult to scale. Future work may	Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-	608
558	explore more efficient human–AI collaboration and	bin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie	609
559	improved data validation to expand dataset scale	Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl	610
560	while preserving quality.	technical report. <i>arXiv preprint arXiv:2502.13923</i> .	611
561	<b>Potential Risks</b>	Baidu-ERNIE-Team. 2025. Ernie 4.5 techni-	612
562	Although DentalGPT shows strong performance	cal report. <a href="https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf">https://ernie.baidu.com/blog/</a>	613
563	and high preference among licensed dentists, it is	<a href="https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf">publication/ERNIE_Technical_Report.pdf</a> .	614
564	not intended for direct clinical use. Real-world	Zhenyang Cai, Junying Chen, Rongsheng Wang, Wei-	615
565	deployment would require further adaptation, rig-	hong Wang, Yonglin Deng, Dingjie Song, Yize Chen,	616
566	orous validation, and integration into clinical work-	Zixu Zhang, and Benyou Wang. 2025. Exploring	617
567	flows to ensure safety and reliability. Clinical deci-	compositional generalization of multimodal llms for	618
568	sions should remain under the supervision of quali-	medical imaging. In <i>Proceedings of the 63rd An-</i>	619
569	fied professionals.	<i>Annual Meeting of the Association for Computational</i>	620
570	<b>References</b>	<i>Linguistics (Volume 1: Long Papers)</i> , pages 13057–	621
571	Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkin-	13079.	622
572	son, Hany Awadalla, Nguyen Bach, Jianmin Bao,	Centers for Disease Control and Prevention. 2024. 2024	623
573	Alon Benhaim, Martin Cai, Vishrav Chaudhary, Cong-	oral health surveillance report: Selected findings. Ac-	624
574	cong Chen, and 1 others. 2025. Phi-4-mini techni-	cessed: 2025-04-28.	625
575	cal report: Compact yet powerful multimodal lan-	Chieh-Ju Chao, Imon Banerjee, Reza Arsanjani, Chadi	626
576	guage models via mixture-of-loras. <i>arXiv preprint</i>	Ayoub, Andrew Tseng, Garvan C. Kane, Jae K Oh,	627
577	<i>arXiv:2503.01743</i> .	Li Fei-Fei, Ehsan Adeli, and Curtis Langlotz. 2024.	628
578	Meta AI. 2025. Llama 4 maverick. <a href="https://www.llama.com/models/llama-4/">https://www.</a>	<i>Echogpt: A large language model for echocardiogra-</i>	629
579	<a href="https://www.llama.com/models/llama-4/">llama.com/models/llama-4/</a> . Accessed: 2025-	<i>phy report summarization. medRxiv</i> .	630
580	12-06.	Jingyun Chen, Linghan Cai, Zhikang Wang, Yi Huang,	631
581	Mistral AI. 2024. Mistral large instruct 2407.	Songhan Jiang, Shenjin Huang, Hongpeng Wang, and	632
582	<a href="https://huggingface.co/mistralai/Mistral-Large-Instruct-2407">https://huggingface.co/mistralai/</a>	Yongbing Zhang. 2025a. Pathagent: Toward inter-	633
583	<a href="https://huggingface.co/mistralai/Mistral-Large-Instruct-2407">Mistral-Large-Instruct-2407</a> . Accessed:	pretable analysis of whole-slide pathology images	634
584	2025-12-06.	via large language model-based agentic reasoning.	635
585	Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang,	<i>arXiv preprint arXiv:2511.17052</i> .	636
586	Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen	Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang,	637
587	Xu, Changrui Chen, Chunsheng Wu, Huajie Tan,	Wanlong Liu, Rongsheng Wang, Jianye Hou, and	638
588	Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin	Benyou Wang. 2024a. Huatuogpt-o1, towards med-	639
589	Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3	ical complex reasoning with llms. <i>arXiv preprint</i>	640
590	others. 2025. Llava-onevision-1.5: Fully open frame-	<i>arXiv:2412.18925</i> .	641
591	work for democratized multimodal training. In <i>arXiv</i> .	Junying Chen, Zhenyang Cai, Zhiheng Liu, Yunjin	642
592	Anthropic. 2025. Claude opus 4-	Yang, Rongsheng Wang, Qingying Xiao, Xiangyi	643
593	5 system card. <a href="https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf">https://assets.</a>	Feng, Zhan Su, Jing Guo, Xiang Wan, and 1 oth-	644
594	<a href="https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf">anthropic.com/m/64823ba7485345a7/</a>	ers. 2025b. Shizhengpt: Towards multimodal llms	645
595	<a href="https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf">Claude-Opus-4-5-System-Card.pdf</a> . Accessed:	for traditional chinese medicine. <i>arXiv preprint</i>	646
596	2025-11-30.	<i>arXiv:2508.14706</i> .	647
597	Anas Awadalla, Irena Gao, Josh Gardner, Jack Hes-	Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe	648
598	sel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe,	Gao, Shunian Chen, Guiming Hardy Chen, Xi-	649
599	Yonatan Bitton, Samir Gadre, Shiori Sagawa, and	dong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji,	650
600	1 others. 2023. Openflamingo: An open-source	Guangjun Yu, Xiang Wan, and Benyou Wang. 2024b.	651
601	framework for training large autoregressive vision-	<i>Huatuogpt-vision, towards injecting medical visual</i>	652
602	language models. <i>arXiv preprint arXiv:2308.01390</i> .	<i>knowledge into multimodal llms at scale. Preprint,</i>	653
603	Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen,	<i>arXiv:2406.19280</i> .	654
604	Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei	Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Cong-	655
		hui He, Jiaqi Wang, Feng Zhao, and Dahua Lin.	656
		2024c. Sharegpt4v: Improving large multi-modal	657
		models with better captions. In <i>European Confer-</i>	658
		<i>ence on Computer Vision</i> , pages 370–387. Springer.	659

660	Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo	Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan,	718
661	Chen, Sen Xing, Muyan Zhong, Qinglong Zhang,	Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang,	719
662	Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl:	Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao,	720
663	Scaling up vision foundation models and aligning	Hui Lin, Yueting Zhuang, and Beng Chin Ooi.	721
664	for generic visual-linguistic tasks. In <i>Proceedings of</i>	2025. <a href="#">Healthgpt: A medical large vision-language</a>	722
665	<i>the IEEE/CVF conference on computer vision and</i>	<a href="#">model for unifying comprehension and generation</a>	723
666	<i>pattern recognition</i> , pages 24185–24198.	<a href="#">via heterogeneous knowledge adaptation</a> . <i>Preprint</i> ,	724
667	Gheorghe Comanici, Eric Bieber, Mike Schaekermann,	arXiv:2502.09838.	725
668	Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-	Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and	726
669	cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and	Xiao-Ming Wu. 2021. Slake: A semantically-labeled	727
670	1 others. 2025. Gemini 2.5: Pushing the frontier with	knowledge-enhanced dataset for medical visual ques-	728
671	advanced reasoning, multimodality, long context, and	tion answering. In <i>2021 IEEE 18th international sym-</i>	729
672	next generation agentic capabilities. <i>arXiv preprint</i>	<i>posium on biomedical imaging (ISBI)</i> , pages 1650–	730
673	<i>arXiv:2507.06261</i> .	1654. IEEE.	731
674	DeepSeek-AI. 2025. <a href="#">Deepseek-r1: Incentivizing rea-</a>	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae	732
675	<a href="#">soning capability in llms via reinforcement learning</a> .	Lee. 2023. Visual instruction tuning. <i>Advances in</i>	733
676	<i>Preprint</i> , arXiv:2501.12948.	<i>neural information processing systems</i> , 36:34892–	734
677	Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo,	34916.	735
678	Lizhuo Lin, Jinrong Yang, Qi Yong H Ai, Lun M	Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping	736
679	Wong, Hao Tang, and Kuo Feng Hung. 2025a. To-	Zhang. 2024. Teach multimodal llms to compre-	737
680	wards better dental ai: A multimodal benchmark	hend electrocardiographic images. <i>arXiv preprint</i>	738
681	and instruction dataset for panoramic x-ray analy-	<i>arXiv:2410.19008</i> .	739
682	sis. <i>arXiv preprint arXiv:2509.09254</i> .	Zijie Meng, Jin Hao, Xiwei Dai, Yang Feng, Jiayang	740
683	Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai	Liu, Bin Feng, Huikai Wu, Xiaotang Gai, Hengchuan	741
684	Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu	Zhu, Tianxiang Hu, and 1 others. 2025. Dentvlm: A	742
685	Zhang, Yanqi Yang, and 1 others. 2025b. Oralgpt-	multimodal vision-language model for comprehen-	743
686	omni: A versatile dental multimodal large language	sive dental diagnosis and enhanced clinical practice.	744
687	model. <i>arXiv preprint arXiv:2511.22055</i> .	<i>arXiv preprint arXiv:2509.23344</i> .	745
688	Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao,	OpenAI. 2024. Gpt-4.1. <a href="https://openai.com/index/gpt-4-1/">https://openai.com/</a>	746
689	Junjun He, Yu Qiao, and Ping Luo. 2024. Omn-	<a href="https://openai.com/index/gpt-4-1/">index/gpt-4-1/</a> . Accessed: 2024-11-30.	747
690	imedvqa: A new large-scale comprehensive evalu-	OpenAI. 2025. Gpt-5 system card. Available at <a href="https://cdn.openai.com/gpt-5-system-card.pdf">https:</a>	748
691	ation benchmark for medical lvlm. <i>arXiv preprint</i>	<a href="https://cdn.openai.com/gpt-5-system-card.pdf">/cdn.openai.com/gpt-5-system-card.pdf</a> .	749
692	<i>arXiv:2402.09181</i> .	Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan	750
693	Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richard-	Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang,	751
694	son, Ahmed El-Kishky, Aiden Low, Alec Helyar,	and Daniel Rueckert. 2025. Medvlm-r1: Incentiviz-	752
695	Aleksander Madry, Alex Beutel, Alex Carney, and 1	ing medical reasoning capability of vision-language	753
696	others. 2024. Openai o1 system card. <i>arXiv preprint</i>	models (vlms) via reinforcement learning. In <i>Inter-</i>	754
697	<i>arXiv:2412.16720</i> .	<i>national Conference on Medical Image Computing</i>	755
698	Xiaochen Jiang, Zhiguo Ding, Yanlei Su, Fei Wang,	<i>and Computer-Assisted Intervention</i> , pages 337–347.	756
699	Weifeng Wang, Ziyang Wang, Xueling Qiu, Chenxi	Springer.	757
700	Sun, Fan Sun, Lu Tang, and 1 others. 2024. Dentists’	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu,	758
701	views on the role orientation of dental hygienists in	Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu,	759
702	china: A qualitative content analysis. <i>BMC Oral</i>	and Daya Guo. 2024. <a href="#">Deepseekmath: Pushing the</a>	760
703	<i>Health</i> , 24(1):1563.	<a href="#">limits of mathematical reasoning in open language</a>	761
704	Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto	<a href="#">models</a> . <i>Preprint</i> , arXiv:2402.03300.	762
705	Usuyama, Haotian Liu, Jianwei Yang, Tristan Nau-	Evgeniy I. Sosnin, Yuriy L. Vasilev, Roman A. Solovyev,	763
706	mann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-	Aleksandr L. Stempkovskiy, Dmitry V. Telpukhov,	764
707	med: Training a large language-and-vision assis-	Artem A. Vasilev, Aleksandr A. Amerikanov, and	765
708	tant for biomedicine in one day. <i>arXiv preprint</i>	Aleksandr Y. Romanov. 2025. <a href="#">Alphadent: A dataset</a>	766
709	<i>arXiv:2306.00890</i> .	<a href="#">for automated tooth pathology detection</a> . <i>Preprint</i> ,	767
710	Richard J Lilford, Benjamin Daniels, Barbara McPake,	arXiv:2507.22512.	768
711	Zulfiqar A Bhutta, Robert Mash, Frances Griffiths,	Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Jun-	769
712	Akinyinka Omigbodun, Elzo Pereira Pinto, Radhika	zhi Ning, Cheng Tang, Siboj Ju, Jin Ye, Pengcheng	770
713	Jain, Gershim Asiki, and 1 others. 2025. Supply-	Chen, Ming Hu, and 1 others. 2025. Gmai-vl-r1: Har-	771
714	side and demand-side factors affecting allopathic pri-	nessing reinforcement learning for multimodal medi-	772
715	mary care service delivery in low-income and middle-	cal reasoning. <i>arXiv preprint arXiv:2504.01886</i> .	773
716	income country cities. <i>The Lancet Global Health</i> ,		
717	13(5):e942–e953.		

774	Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. <i>arXiv preprint arXiv:2312.11805</i> .	
775		
776		
777		
778		
779		
780	Gemma Team. 2025. <b>Gemma 3</b> . <i>arXiv preprint arXiv:2503.19786</i> .	
781		
782	LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025a. <b>Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning</b> . <i>Preprint</i> , arXiv:2506.07044.	
783		
784		
785		
786		
787		
788		
789		
790	V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025b. <b>Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning</b> . <i>Preprint</i> , arXiv:2507.01006.	
791		
792		
793		
794		
795		
796		
797		
798	Sheng Wang, Ruiming Wu, Charles Herndon, Yihang Liu, Shunsuke Koga, Jeanne Shen, and Zhi Huang. 2025. <b>Pathology-cot: Learning visual chain-of-thought agent from expert whole slide image diagnosis behavior</b> . <i>arXiv preprint arXiv:2510.04587</i> .	
799		
800		
801		
802		
803	Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. <b>Pmc-llama: toward building open-source language models for medicine</b> . <i>Journal of the American Medical Informatics Association</i> , page ocae045.	
804		
805		
806		
807		
808	Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025a. <b>Towards generalist foundation model for radiology by leveraging web-scale 2d&amp;3d medical data</b> . <i>Nature Communications</i> , 16(1):7866.	
809		
810		
811		
812		
813	Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. <b>Towards generalist foundation model for radiology</b> . <i>arXiv preprint arXiv:2308.02463</i> .	
814		
815		
816		
817	Jing Wu, Yuli Wang, Zhushi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X Bai. 2025b. <b>Vision-language foundation model for 3d medical imaging</b> . <i>npj Artificial Intelligence</i> , 1(1):17.	
818		
819		
820		
821	Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024b. <b>Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding</b> . <i>arXiv preprint arXiv:2412.10302</i> .	
822		
823		
824		
825		
826		
827	Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2024. <b>Medtrinity-25m: A large-scale multimodal dataset with multi-granular annotations for medicine</b> . <i>arXiv preprint arXiv:2408.02900</i> .	829
828		830
		831
		832
	Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. <b>Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning</b> . <i>arXiv preprint arXiv:2506.07044</i> .	833
		834
		835
		836
		837
		838
	Yuechun Yu, Han Ying, Haoan Jin, Wenjian Jiang, Dong Xian, Binghao Wang, Zhou Yang, and Mengyue Wu. 2025. <b>Medkgeval: A knowledge graph-based multi-turn evaluation framework for open-ended patient interactions with clinical llms</b> . <i>Preprint</i> , arXiv:2510.12224.	839
		840
		841
		842
		843
		844
	Ziyi Zeng, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Yipeng Liu, Siqi Cai, Benyou Wang, Zhiguo Zhang, and 1 others. 2025. <b>Wavemind: Towards a conversational eeg foundation model aligned to textual and visual modalities</b> . <i>arXiv preprint arXiv:2510.00032</i> .	845
		846
		847
		848
		849
		850
	Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. <b>Medrlvr: Emerging medical reasoning from a 3b base model via reinforcement learning</b> . <i>arXiv preprint arXiv:2502.19655</i> .	851
		852
		853
		854
		855
	Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. <b>Pmc-vqa: Visual instruction tuning for medical visual question answering</b> . <i>arXiv preprint arXiv:2305.10415</i> .	856
		857
		858
		859
		860
	Yubao Zhao, Tian Zhang, Xu Wang, Puyu Han, Tong Chen, Linlin Huang, Youzhu Jin, and Jiaju Kang. 2024. <b>Ecg-chat: A large ecg-language model for cardiac disease diagnosis</b> . <i>arXiv preprint arXiv:2408.08849</i> .	861
		862
		863
		864
		865
	Peixin Zheng, Xiaoting Qiu, Lingxiao Zhang, Peizhang Liu, Zeyi Peng, and Zhijian Huang. 2025. <b>Comparative analysis of oral disorder burden in china and globally from 1990 to 2021 based on gbd data</b> . <i>Scientific Reports</i> , 15(1):10061.	866
		867
		868
		869
		870
	Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. <b>Medxpertqa: Benchmarking expert-level medical reasoning and understanding</b> . <i>arXiv preprint arXiv:2501.18362</i> .	871
		872
		873
		874
		875
	<b>A Case Study</b>	876
	Dentistry is a key medical field that relies on clinicians analyzing patients' imaging data and communicating with them for diagnosis, yet even leading MLLMs with strong general multimodal abilities still fall far behind professional dentists in multimodal diagnosis.	877
		878
		879
		880
		881
		882

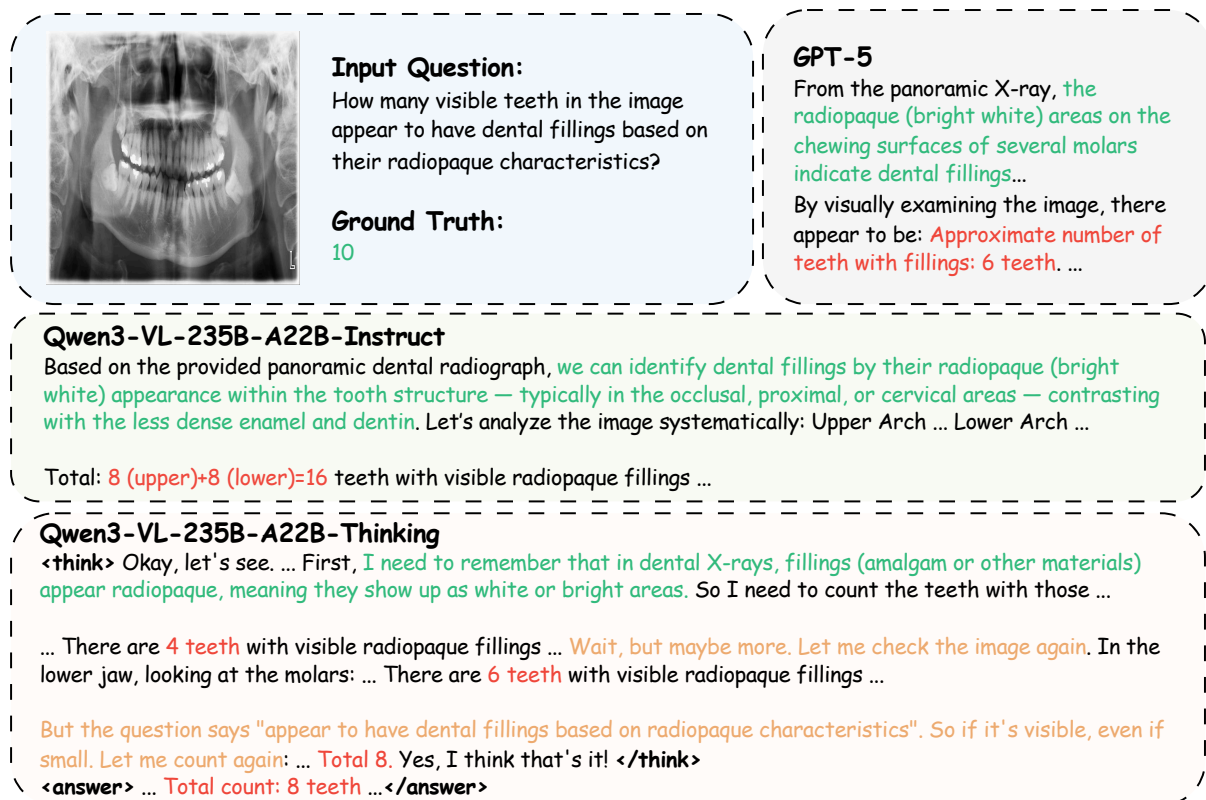


Figure 7: Examples of top-tier general-purpose MLLMs analyzing a dental image task. Red indicates incorrect analysis, green indicates correct analysis, and orange highlights reflective turns in the complex reasoning process.

To investigate this phenomenon, we conducted a case study on a dental multimodal task using one leading commercial MLLM as well as both the *complex reasoning* and *non-complex reasoning* modes of a leading MLLM (Figure 7)<sup>3</sup>. After analyzing the model outputs, we found that although they could identify the relevant visual features to be counted, they still produced incorrect answers. Notably, by examining the reasoning trajectory of Qwen3-VL-235B-A22B-Thinking, we observed that it repeatedly reflected on and revised its own intermediate conclusions. Through this type of complex reasoning pattern, the model gradually approached the correct answer step by step. Although it still failed to produce the correct final prediction, this behavior highlights the potential of applying complex reasoning strategies to achieve more accurate multimodal diagnosis in dentistry.

<sup>3</sup>Qwen3-VL-235B-A22B is a fully open-source model family with both Thinking and Instruct models, corresponding to the with and without complex reasoning modes. Its fully accessible reasoning paths and diverse model variants make it the primary model used in this study.

## B Data Collection

Before improving an MLLM’s understanding and reasoning on dental images, it is necessary to collect a sufficiently large set of training samples, which provides the foundation for the model to effectively use its knowledge during image interpretation (Liu et al., 2023).

### B.1 Existing Annotated Data

To efficiently obtain a sufficiently large number of dentistry-specific multimodal datasets, we first sourced datasets from a variety of open-source platforms. By leveraging certified, high-quality dental image datasets and existing literature, we aim to enrich the model’s understanding of relevant radiographic data.

**PMC-Dental-Caption-47k** PubMed Central (PMC) is a publicly accessible biomedical repository that hosts a vast collection of peer-reviewed medical publications. It is considered a reliable and widely used data source in previous research. From PMC, we filtered a large number of dental images and retained the associated captions and labels provided within the original papers. This rich textual context is expected to provide valuable information

925	for enhancing visual understanding and facilitating	974
926	the integration of dental domain knowledge into	975
927	our model.	976
928	 <b>Open-source-Dental-Classification-49k</b> To	977
929	further leverage image datasets that have previ-	978
930	ously been used to train classification models, we	979
931	collected a wide range of dental-related classifica-	980
932	tion datasets and consolidated them into a larger	981
933	corpus of dental images with corresponding labels.	982
934	Specifically, each image is associated with one or	983
935	more disease labels; for multi-class or multi-label	984
936	datasets, we also retain the negative labels so that	985
937	all available, clinically validated annotations can	
938	be fully exploited. This unified resource supports	<b>C Data Curation</b> 986
939	MLLMs in better aligning common dental disease	<b>C.1 Curation Strategies</b> 987
940	categories with key feature identification.	To enrich the model’s multimodal alignment,
941	 <b>Open-source-Dental-Detection-31k</b> We also	instruction-following ability, and complex reason-
942	collected a number of datasets previously utilized	ing in dentistry, we further curate and refine all
943	for dental lesion localization tasks, in which each	collected data using GPT-5. This data curation pro-
944	image is annotated with one or more lesion in-	cess is designed to provide high-quality supervision
945	stances together with their spatial coordinates. Al-	for both fine-grained dental image understanding
946	though our model is not explicitly trained to predict	and the elicitation of multi-step reasoning, serv-
947	bounding boxes, such annotations provide MLLMs	ing as a foundation for subsequent reinforcement
948	with implicit spatial cues and lesion counts, thereby	learning. The overall process consists of image cap-
949	supporting the model’s capability to understand	tioning, instruction tuning, and complex reasoning
950	spatial relationships and quantify dental abnormali-	data construction, as outlined below. 998
951	ties.	<b>Image Captioning</b> This component focuses on 999
952		enhancing the model’s ability to capture diagnosti-
953	<b>B.2 Newly Annotated Data</b>	cally relevant visual details in dental images. Given 1000
954	Building upon the open-source datasets described	a dental image $\mathcal{I}$ , a reference label $l$ , and a caption-
955	above, we observed that although they contain a	ing instruction $q_{\text{cap}}$ , the LLM is prompted to gener-
956	considerable number of annotations, the diagnos-	ate an observation-based description $d$ that enumer-
957	tic focus is predominantly centered on a few com-	ates all visible cues related to the label while
958	mon dental conditions. From a clinical perspective,	avoiding diagnostic assumptions (Detailed prompts
959	however, there exist additional critical abnormali-	are in Appendix E): 1007
960	ties and visual manifestations that warrant greater	$d = \text{LLM}(\mathcal{I}, q_{\text{cap}}, l). \quad 1008$
961	attention yet are underrepresented in existing data	These captions are then paired with predefined den-
962	sources. To address this gap, we expanded the diag-	tal questions and answers to form caption-based 1009
963	nostic label set and further curated a new subset of	dental questions and answers to form caption-based 1010
964	dental images, which were annotated by certified	VQA samples, allowing the model to better asso-
965	dental experts with an emphasis on clinically sig-	ciate fine-grained visual evidence with correspond-
966	nificant conditions and indicative visual features.	ing textual concepts and reducing visual informa-
967	 <b>Expert-Annotated Dataset</b> We collected	tion gaps. 1014
968	some dental images from internet sources, hospital	<b>Instruction Tuning</b> To improve performance 1015
969	imaging archives, and publicly available datasets.	on downstream dental tasks, we further construct 1016
970	After removing duplicates and low-resolution im-	instruction-following supervision using a two-step 1017
971	ages, we curated a candidate dataset that was sub-	generation process. In the first step, given the den-
972	sequently annotated by professional dental clini-	tal image $\mathcal{I}$ , an instruction $q_{\text{inst}}$ , and the reference la-
973	cians. To ensure annotation quality, a strict cross-	bel $l$ , the LLM generates an intermediate question-
	validation mechanism was applied with different	answer pair $(q^*, a^*)$ that reflects a plausible dental 1021

1022 diagnostic query:

$$1023 (q^*, a^*) = \text{LLM}(\mathcal{I}, q_{\text{inst}}, l).$$

1024 In the second step, the LLM produces a final an-  
1025 swer  $a$  conditioned on the image, the instruction,  
1026 and the generated intermediate pair:

$$1027 a = \text{LLM}(\mathcal{I}, q_{\text{inst}}, q^*, a^*).$$

1028 In addition, several vision models are used to anno-  
1029 tate a subset of images, and only high-confidence  
1030 predictions are retained. The LLM further refines  
1031 these annotations and converts them into structured  
1032 question–answer pairs, which are incorporated into  
1033 the instruction tuning data to improve dental ques-  
1034 tion answering under diverse task formulations.

1035 **Complex Reasoning** To support the subsequent  
1036 reinforcement learning stage, we construct complex  
1037 chain-of-thought (CoT) data following the method-  
1038 ology of HuatuoGPT-01 (Chen et al., 2024a). This  
1039 process builds upon the instruction tuning data and  
1040 is designed to explicitly elicit and refine complex  
1041 reasoning.

1042 Given a dental image  $\mathcal{I}$  and an question–answer  
1043 pair  $(q^*, a^*)$  in instruction tuning data, the LLM  
1044 is first prompted to analyze the question  $q^*$  and  
1045 produce an initial reasoning path  $r_1$  together with  
1046 a tentative answer summary  $a_1$ :

$$1047 (r_1, a_1) = \text{LLM}(q^*).$$

1048 The generated answer  $a_1$  is then compared with the  
1049 reference answer  $a^*$  using the LLM as an evaluator.  
1050 If  $a_1$  is inconsistent with  $a^*$ , the LLM is prompted  
1051 again with all previously generated information  
1052 to continue the reasoning process and produce a  
1053 refined reasoning step  $r_2$  and answer  $a_2$ :

$$1054 (r_t, a_t) = \text{LLM}(q^*, r_{1:t-1}, a_{1:t-1}), \text{ until } a_t = a^*.$$

1055 This iterative process continues until the gen-  
1056 erated answer matches the reference answer.  
1057 All intermediate reasoning steps and answers  
1058  $\{(r_1, a_1), (r_2, a_2), \dots\}$  are then concatenated to  
1059 form a complete reasoning trajectory  $R$ .

1060 Finally, conditioned on the dental image  $\mathcal{I}$ , the  
1061 question  $q^*$ , and the aggregated reasoning trajec-  
1062 tory  $R$ , the LLM generates a final summarized an-  
1063 swer  $a$ :

$$1064 a = \text{LLM}(\mathcal{I}, q^*, R).$$

1065 The resulting data are formatted into a stan-  
1066 dardized CoT template of the form `<think> R`

`</think><answer> a </answer>`, which trains a  
1067 consistent response template and encourages reflec-  
1068 tive analysis, providing a cold start for the subse-  
1069 quent reinforcement learning stage. 1070

## 1071 C.2 Data Filtering

1072 After completing all data construction steps, we  
1073 apply an additional quality control stage using  
1074 GPT-5-mini to perform secondary verification.  
1075 Samples that deviate from the original image con-  
1076 tent or reference labels are filtered out, yielding  
1077 a finalized training dataset with improved consis-  
1078 tency and reliability.

## 1079 C.3 Quality Assessment

1080 As described in the methodology, the training  
1081 dataset of DentalGPT was generated by GPT-5  
1082 while referencing existing image labels or descrip-  
1083 tions to minimize hallucinations and ensure pro-  
1084 fessional domain knowledge injection. To validate  
1085 the effectiveness of this approach, five evaluation  
1086 dimensions were defined, and a randomly sampled  
1087 set of 3,000 entries was assessed and compared  
1088 against data obtained through direct GPT-5 distil-  
1089 lation. To ensure fairness, all comparisons were  
1090 evaluated using Gemini-2.5-Pro as the judge.

1091 **Evaluation Setup** Specifically, the dataset was  
1092 evaluated across the following five dimensions:  
1093 1) *Description Completeness*: Whether all ob-  
1094 servable visual details in the image are thor-  
1095 oughly described, with particular attention to  
1096 features that may contribute to dental diagno-  
1097 sis. 2) *Terminology Consistency*: Whether pro-  
1098 fessional dental terminology is used correctly  
1099 and consistently throughout the description. 3)  
1100 *Content Safety*: Whether the content adheres to  
1101 medical ethics and safety standards, avoiding sen-  
1102 sitive, discriminatory, misleading, or inappropriate  
1103 statements. 4) *Text–Image Consistency*: Whether  
1104 the textual description is well written and accu-  
1105 rately aligned with the corresponding image con-  
1106 tent. 5) *Knowledge Depth*: Whether the descrip-  
1107 tion demonstrates an appropriate level of dental  
1108 knowledge.

1109 Gemini-2.5-Pro was asked to score each di-  
1110 mension on a scale of 1 to 5, and the final dataset  
1111 quality was reported using the average score across  
1112 all evaluated samples.

1113 **Results Analysis** Results are shown in Figure 8,  
1114 our training dataset demonstrates clear advantages  
1115 over the directly distilled version across multiple

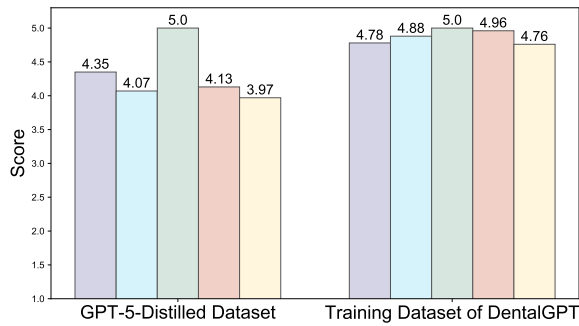


Figure 8: Gemini-2.5-Pro’s multi-dimensional evaluation of GPT-5–distilled data and the training dataset of DentalGPT (scores range from 0 to 5). For each dataset, the color blocks from left to right represent *Description Completeness*, *Terminology Consistency*, *Content Safety*, *Text-Image Consistency*, and *Knowledge Depth*.

evaluation metrics. It can be observed that the data generated with label references shows the most significant improvements in *terminology consistency* and *knowledge depth*, and also achieves notable gains in the completeness of visual detail descriptions. Interestingly, Gemini-2.5-Pro assigns a perfect score for safety to both datasets, indicating that GPT-5–generated data perform exceptionally well in medical safety, avoiding harmful diagnostic suggestions and providing timely guidance to reduce potential risks.

In conclusion, the results indicate that our dataset, by leveraging annotations from public datasets and descriptions from academic literature, provides more comprehensive and more professional knowledge injection for the model. Such a foundation ensures substantial improvement in the performance of DentalGPT.

## D Human Evaluation

To further assess the clinical value of DentalGPT, we conduct a human evaluation study to measure the degree to which licensed dentists prefer the model’s outputs in realistic diagnostic scenarios. Unlike automatic metrics, this evaluation focuses on whether the model’s responses are aligned with professional clinical judgment and would be considered acceptable in practice.

**Participants** Four licensed dentists participated in the evaluation. All evaluators are professionally certified, with both the mean and median years of clinical practice being three years. As the evaluators are members of the same collaborative re-

search project, compensation for participation was covered by the project’s existing research funding.

**Evaluation Data** Evaluation images were randomly sampled from ten standard medical textbooks and three publicly available dental examination papers, resulting in a total of 80 dental images. These images cover a range of common dental conditions and realistic clinical scenarios. Each model was prompted to provide a detailed analysis of the image and a corresponding diagnostic conclusion.

**Evaluation Protocol** For each image, the response generated by DentalGPT was paired with that of a competing model. Dentists were asked to compare the two responses and indicate which one they preferred based on predefined criteria, including clinical accuracy, diagnostic soundness, and completeness of information. The prompt presented to the evaluators is shown in Figure 9. Each dentist evaluated all comparison pairs independently.

**Result Aggregation** For each model comparison, we aggregate the dentists’ judgments and report the average proportions of *win*, *tie*, and *loss* outcomes. These statistics provide a quantitative measure of dentists’ preferences and offer insight into the relative clinical acceptability of the model outputs.

### The prompt for human evaluation

image

As a dental professional, please perform a rigorous clinical analysis of the provided image. Your response must prioritize **diagnostic accuracy**, the **absence of clinical errors**, and **comprehensive coverage** of all visible dental pathologies.

Please evaluate the image based on the following key dimensions: 1. **Clinical Correctness**: Ensure there are no theoretical or factual errors in the assessment. 2. **Diagnostic Accuracy**: Provide a precise diagnosis of the conditions shown (e.g., caries, periodontal status, restorations). 3. **Completeness**: Include all relevant visual details, such as tooth numbering (if applicable), location, and severity of the condition.

Use professional dentistry terminology while maintaining clinical relevance. Focus solely on the practical value of the diagnosis.

<caption>  
{caption}  
</caption>

```
<image context> (<image> represents the location of the image)
{image_context}
</image context>
```

Please output a detailed diagnostic analysis of the image only. Do not generate any content unrelated to the clinical task.

Figure 9: Prompt for human evaluation.

## E Data Curation Prompts

All the prompts for data curation are listed below:

### The prompt for Image Captioning

image

Please generate a professional, detailed, and high-quality description for the medical image I provide. The description should include as many dentistry-related visual details as possible to ensure clinical readability, professionalism, and comprehensiveness, while also being understandable to the general public. Aim to make the description as rich and detailed as possible, providing extensive visual information. You may refer to the image's context to improve the accuracy and completeness of your description, but do not reveal that you referenced the context.

```
<caption>
{caption}
</caption>
```

```
<image context> (<image> represents the location of the image)
{image_context}
</image context>
```

Please output a detailed description of the image only. Do not generate any content unrelated to the task.

Figure 10: Prompt for generating observation-based dental image captions.

### The prompt for Instruction Tuning: QA Generation

image

Please generate a dentistry-related question and a short corresponding ground truth answer about the medical image I provide. The question should assess the model's visual capabilities. Avoid being too specific—design the question so that it requires the model to look at the image to answer. The question should demand strong visual understanding as well as some knowledge of dentistry. You may refer to the provided image caption and contextual information to improve the quality of the question. However, \*\*do not mention or reference the

caption or context in the question itself—assume they are not available\*\*.

```
<caption>
{caption}
</caption>
```

```
<image context> (<image> represents the location of the image)
{image_context}
</image context>
```

Please generate the question and the ground truth answer directly. Do not include any content unrelated to the task.

Figure 11: Prompt for generating dentistry-related question–answer pairs for instruction tuning.

### The prompt for Instruction Tuning: Answer Generation

image

```
<question>
{caption}
</question>
```

```
<answer>
{ground_truth_answer}
</answer>
```

You are now required to look at the image I provide and answer the user's question about dentistry. Make sure your response directly addresses the user's query, follows instructions well, and is as detailed and rich as possible, with the style and quality characteristic of GPT-5.

You may refer to the image caption and contextual information I secretly provide to you in order to improve the accuracy and completeness of your answer. However, \*\*do not mention or reference the caption or context in your response—assume they are not available\*\*.

```
<caption>
{caption}
</caption>
```

```
<image context> (<image> represents the location of the image)
{image_context}
</image context>
```

Please generate the answer directly. Do not include any unrelated content.

Figure 12: Prompt for generating final answers during instruction tuning.

### The prompt for answer verification

```
<Model Response>
{model_response}
</Model Response>
```

```
<Reference Answer>
{reference_answer}
</Reference Answer>
```

You are provided with a model-generated response (<Model Response>) and a reference answer (<Reference Answer>). Compare the model response with the reference answer and determine its correctness.

**Task:** Output True if the model response is correct, and False otherwise.

Please output only True or False, without any additional explanation.

Figure 13: Prompt for verifying the correctness of generated answers during iterative reasoning.

### The prompt for Initial CoT Generation

```
<question>
{question}
</question>
```

Please respond to the above question <question> according to the provided image using the Chain of Thought (CoT) reasoning method. Your response should consist of multiple steps, each of which includes three types of actions: **Inner Thinking**, **Final Conclusion**, and **Verification**:

- **Inner Thinking**: This is the step where thinking is done. Note that multiple 'Inner Thinking' steps are required to describe thorough reasoning. Each step should first generate a brief title.
- **Final Conclusion**: At this stage, you summarize the correct reasoning from previous 'Inner Thinking' steps and provide the final answer. No title is required here.
- **Verification**: At this stage, you verify the conclusion from the "Final Conclusion" step. If the conclusion holds, end the process. If not, return to "Inner Thinking" for further reasoning. No title is required here.

The output format must strictly follow the JSON structure below:

```
“json "CoT": [ "action": "Inner Thinking", "title": "...",
"content": "...", ..., "action": "Final Conclusion", "content":
"...", "action": "Verification", "content": "..."]
```

Figure 14: Prompt for initial construction of chain-of-thought reasoning.

### The prompt for CoT Refinement with New Thinking

```
<question>
{question}
</question>
```

```
<previous reasoning>
{previous_reasoning}
</previous reasoning>
```

```
<response requirements>
```

Your response must include the following steps, each composed of three types of actions: **Inner Thinking**, **Final Conclusion**, and **Verification**:

1. **Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose.
2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section.
3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement.

```
</response requirements>
```

<question> represents the image-based question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning by exploring new approaches to solve this problem and construct a new Final Conclusion.

**Output Format**

Strictly follow the JSON structure below. You do not need to repeat your previous reasoning. Begin directly from the next 'Verification' stage.

```
“json "CoT": [ "action": "Verification", "content": "...",
"action": "Inner Thinking", "title": "...", "content": "...",
..., "action": "Final Conclusion", "content": "...", "action":
"Verification", "content": "..."]
```

Figure 15: Prompt for refining CoT by exploring new reasoning strategies.

### The prompt for CoT Refinement via Backtracking

```
<question>
{question}
</question>
```

```
<previous reasoning>
{previous_reasoning}
</previous reasoning>
```

```
<response requirements>
```

Your response must include the following steps, each composed of three types of actions: **Inner Thinking**, **Final Conclusion**, and **Verification**:

1. **Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose.
2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section.
3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement.

</response requirements>

<question> represents the image-based question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning using **backtracking** to revisit earlier points of reasoning and construct a new Final Conclusion.

#### Output Format

Strictly follow the JSON structure below. You do not need to repeat your previous reasoning. Begin directly from the next 'Verification' stage.

```
““json "CoT": [ "action": "Verification", "content": "...",  
"action": "Inner Thinking", "title": "...", "content": "...",  
..., "action": "Final Conclusion", "content": "...", "action":  
"Verification", "content": "..."]
```

```
““json "CoT": [ "action": "Verification", "content": "...",  
"action": "Inner Thinking", "title": "...", "content": "...",  
..., "action": "Final Conclusion", "content": "...", "action":  
"Verification", "content": "..."]
```

1189

Figure 17: Prompt for refining CoT by exploring new visual regions.

Figure 16: Prompt for refining CoT via backtracking.

### The prompt for CoT Refinement by Exploring New Image Regions

<question>  
{question}  
</question>

<previous reasoning>  
{previous\_reasoning}  
</previous reasoning>

<response requirements>

Your response must include the following steps, each composed of three types of actions: **Inner Thinking**, **Final Conclusion**, and **Verification**:

- Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose.
- Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section.
- Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement.

</response requirements>

<question> represents the image-based question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning by focusing on new areas of the picture to solve this problem and construct a new Final Conclusion.

#### Output Format

Strictly follow the JSON structure below. You do not need to repeat your previous reasoning. Begin directly from the next 'Verification' stage.

1187

1188