OMNI-Dent: Towards an Accessible and Explainable AI Framework for Automated Dental Diagnosis

Leeje Jang[1] Yao-Yi Chiang[1] Angela M. Hastings[1] Patimaporn Pungchanchaikul[2]
Martha B. Lucas[1] Emily C. Schultz[3] Jeffrey P. Louie[1] Mohamed Estai[4]
Wen-Chen Wang[5] Ryan H.L. Ip[6] Boyen Huang[1]
[1]University of Minnesota, United States [2]Khon Kaen University, Thailand
[3]Minnesota State University, Mankato, United States [4]The University of Western Australia, Australia
[5]Kaohsiung Medical University, Taiwan [6]Auckland University of Technology, New Zealand
jang0124@umn.edu yaoyi@umn.edu hasti070@umn.edu patpun@kku.ac.th
lucas288@umn.edu emily.schultz.2@mnsu.edu louie003@umn.edu mohamed.estai@uwa.edu.au
wcwang@kmu.edu.tw ryan.ip@aut.ac.nz huan2321@umn.edu

Abstract

Accurate dental diagnosis is essential for oral healthcare, yet many individuals lack access to timely professional evaluation. Existing AI-based methods primarily treat diagnosis as a visual pattern recognition task and do not reflect the structured clinical reasoning used by dental professionals. These approaches also require large amounts of expert-annotated data and often struggle to generalize across diverse real-world imaging conditions. To address these limitations, we present OMNI-Dent, a data-efficient and explainable diagnostic framework that incorporates clinical reasoning principles into a Vision-Language Model (VLM)-based pipeline. The framework operates on multi-view smartphone photographs, embeds diagnostic heuristics from dental experts, and guides a general-purpose VLM to perform tooth-level evaluation without dental-specific fine-tuning of the VLM. By utilizing the VLM's existing visual-linguistic capabilities, OMNI-Dent aims to support diagnostic assessment in settings where curated clinical imaging is unavailable. We design OMNI-Dent as an early-stage assistive tool to help users identify potential abnormalities and determine when professional evaluation may be needed, thereby offering a practical option for individuals with limited access to in-person care.

1. Introduction

Oral health plays an important role in maintaining quality of life across the human lifespan, yet many individuals, particularly those in underserved or rural communities [19, 20], struggle to access timely and reliable dental diagnosis. Limited availability of dental professionals often leads to delayed treatment and preventable deterioration in oral health. Prior work on teledentistry [4, 11–14, 16, 22] provides remote diagnostic support to improve accessibility; however, limited expert availability and the resources required for remote evaluation still remain ongoing challenges.

Recent advances in artificial intelligence (AI) motivate the development of automated dental diagnostic systems. Existing approaches [1, 7, 21, 24, 27] typically rely on deep learning models trained to detect or classify dental conditions from clinical imaging modalities such as radiographs or intraoral photographs taken with professional devices. Although effective in controlled settings, these methods require large expert-labeled datasets in which clinicians manually annotate tooth locations and conditions, making the annotation process labor-intensive and difficult to scale. Model performance is also sensitive to visual similarity between the training data and deployment environments, limiting applicability across the wide range of real-world dental imaging conditions. A fundamental limitation of current AI-based dental systems is that they treat diagnosis largely as a visual pattern recognition task. In practice, however, dental professionals use structured clinical reasoning that involves comparing neighboring and contralateral teeth, considering multi-surface morphology, and interpreting structural changes such as wear, erosion, or cavities in the context of the full dentition. Existing AI-based methods do not encode these diagnostic heuristics, so they fail to reflect how clinicians reason through a diagnosis and result in a lack of interpretability.

The emergence of Generative AI (GenAI), particularly Vision-Language Models (VLMs) [2, 5, 6, 18, 25] trained on large-scale image-text pairs, introduce new possibili-

ties for advancing automated dental diagnosis. These models learn from various image-understanding tasks (e.g., image captioning, visual question answering, object detection) across different domains. However, applying VLMs to dentistry [10, 26] typically require additional domain-specific datasets annotated by experts, leading to substantial labeling costs and systems that often overfit to narrow clinical imaging conditions. Recent work also introduces VLMs [23] and benchmark datasets tailored for medical imaging [15], but these efforts continue to focus on professionally captured modalities such as radiographs or intraoral photographs obtained with specialized devices rather than accessible smartphone-based images. These limitations highlight the need for an approach that minimizes reliance on dental-specific annotations and controlled clinical imaging environments while effectively leveraging the existing capabilities of modern VLMs.

To address these challenges, we propose OMNI-Dent, an explainable and data-efficient diagnostic framework designed to emulate the reasoning process of dental clinicians and to serve as an early, first-line assistive tool for identifying subtle abnormalities. OMNI-Dent operates on multi-view smartphone photographs consisting of frontal, upper occlusal, and lower occlusal views, making it suitable for use outside specialized clinical imaging environments. A key component of the framework is a clinical reasoning module that directs a general-purpose VLM to follow expert-defined diagnostic steps, enabling tooth-level evaluation without any fine-tuning of the VLM. The goal is to provide accessible, initial screening support that helps individuals detect potential dental issues at an early stage and determine when professional care is warranted. We examine how a state-of-the-art VLM pretrained on broad image-text corpora can be guided to follow clinician-like diagnostic reasoning. By leveraging relational cues within the dentition and embedding expert heuristics, OMNI-Dent performs tooth-level diagnosis while retaining the VLM's general visual capabilities. Operating directly on smartphone photographs rather than specialized clinical images enables early at-home assessment for individuals who face barriers to in-person evaluation, including those in underserved or geographically isolated communities.

We summarize the main contributions of this work as follows:

- We present OMNI-Dent, an explainable and data-efficient diagnostic framework that performs tooth-level diagnosis from smartphone photographs using a general-purpose VLM, without any fine-tuning.
- We propose a clinical reasoning module that guides the VLM through structured diagnostic steps inspired by how dental professionals evaluate teeth, using explicit visual cues to support reliable tooth-level interpretation.
- We demonstrate how state-of-the-art VLMs perform dental imaging tasks in both zero-shot and few-shot in-context learning (ICL) settings, achieving strong quantitative performance across multiple diagnostic categories.
- We highlight the potential of OMNI-Dent as an early at-home screening tool that operates on smartphone photographs, offering accessible diagnostic support for individuals who face barriers to in-person dental care.

2. Related Work

2.1. Accessible dental diagnosis

Oral health represents an essential component of overall health and quality of life, yet access to dental care remains uneven across populations. Prior work shows that individuals in underserved or geographically isolated communities [20] continue to face substantial barriers in obtaining timely professional evaluation. Existing studies [4, 11–14, 16, 22] therefore explore teledentistry as a remote diagnosis framework to facilitate online diagnosis and consultation for individuals who cannot easily reach a clinic. Specifically, recent work [16] evaluates the diagnostic accuracy of smartphone photographs for traumatic dental injuries (TDI), demonstrating strong potential for expanding remote dental care through smartphone-based assessment. Although existing studies report encouraging diagnostic performance for teledentistry-based assessments, these services still depend on human experts to interpret remotely submitted information. This dependence introduces practical constraints because expert availability and the resources required for remote evaluation are not reliably accessible in a timely manner.

2.2. AI for Dental Diagnosis

AI-based methods increasingly serve dental diagnosis, with prior work [1, 7, 21, 24, 27] primarily adopting convolutional neural networks and other image-driven architectures to identify conditions in radiographs [7] and clinical photographs [27]. Training these models typically relies on supervised learning with dental images, which requires large amounts of task-specific data that share similar visual appearances and diagnosis codes, along with expert-annotated labels that demand substantial human effort. Although these approaches learn visual patterns associated with disease, they still treat diagnosis primarily as a pattern recognition task, which makes it difficult to understand how the model reaches a particular decision and results in limited interpretability.

2.3. VLM for Dental Diagnosis

Vision-language models, particularly Large-VLMs that combine a visual encoder with an LLM-style decoder to generate natural language descriptions of images (e.g., GPT4 [2], OVIS [18], Qwen [5], and the InternVL se-

ries [6, 25]), demonstrate strong capabilities across general visual-linguistic tasks and support the integration of visual information with textual interpretation and reasoning. VLMs learn from large amounts of image-text pairs that include tasks such as image captioning and visual question answering (VQA). In addition, in-context learning (ICL) enables the models to adapt from general-purpose training to domain-specific tasks without additional training by using relevant question-answer pairs as references during inference.

To leverage these general-purpose VLMs for dentistry, recent studies [10, 26] apply them to dental diagnosis and report promising performance in both diagnostic reasoning and explainability. DentalVLM [10] uses a large-scale bilingual dataset of oral images paired with visual question-answer annotations, covering multiple 2D dental imaging modalities and a broad range of diagnostic tasks. However, these approaches rely on extensive image-text annotations that require domain experts to manually inspect and describe each case, making it difficult to apply these methods across the wide variety of specific dental cases encountered in practice. Furthermore, interest continues to grow in vision–language foundation models for medical applications. MedGemma [23], for example, introduces a VLM that leverages medical training data, and MMOral [15] provides a benchmark dataset for general medical analysis with VLMs. However, these models and datasets still primarily rely on specialized clinical imaging such as X-rays or 2D/3D CT/MRI slices, which limits their direct applicability to accessible, smartphone-based dental photographs.

3. OMNI-Dent

In this section, we describe the design and implementation of OMNI-Dent. Sec. 3.1 presents an overview of the full diagnostic pipeline and its multi-view input setting. Sec. 3.2 details the tooth-level identification module, which localizes and indexes each tooth from the input views. Sec. 3.3 explains the clinical reasoning module, which replicates expert diagnostic workflows through structured, expert-defined instructions applied to a pretrained VLM. Sec. 3.4 describes the diagnosis integration module, which consolidates per-view predictions into an individual-level diagnostic summary.

3.1. Framework Overview

Figure 1 shows the overview of OMNI-Dent. OMNI-Dent takes a set of multi-view smartphone photographs as input, captured from each individual. The inputs consist of three perspectives: (1) a frontal view, (2) an upper occlusal view, and (3) a lower occlusal view, all acquired using commonly available and widely accessible smartphones. These complementary views reveal tooth surfaces that are not fully visible from a single angle and provide sufficient visual diversity to support reliable tooth-level analysis under real-world imaging conditions.

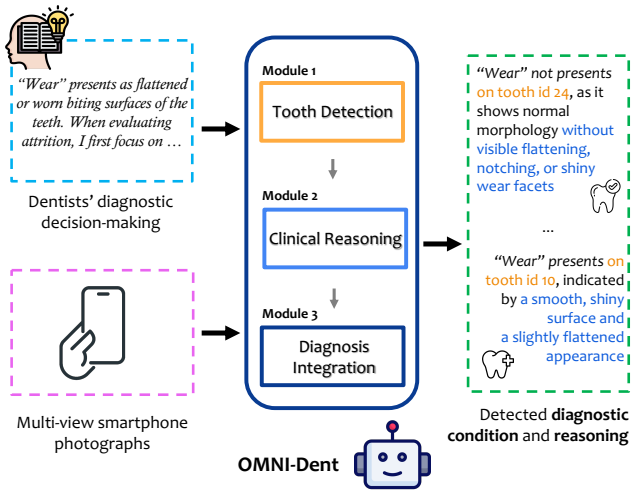

Figure 1. Overview of OMNI-Dent. Replicating dentists' clinical diagnostic reasoning processes, the framework processes multi-view smartphone photographs through tooth detection, clinical reasoning, and diagnosis integration modules. The output of OMNI-Dent provides tooth-level diagnostic conditions with corresponding reasoning.

OMNI-Dent comprises three main components: the *tooth identification module* (Sec. 3.2), the *clinical reasoning module* (Sec. 3.3) and the *diagnosis integration module* (Sec. 3.4). The tooth-level identification module directly localizes each visible tooth across the multi-view inputs and assigns its identifier according to the universal numbering system, thereby establishing precise and standardized indexing. The clinical reasoning module predicts diagnostic conditions by replicating expert-provided natural-language diagnostic steps. The diagnosis integration module consolidates the overall diagnosis for each tooth. The framework outputs a tooth-level diagnostic condition along with a corresponding reasoning description.

3.2. Tooth Detection Module

The tooth detection module aims to localize each tooth in the input images and assign its corresponding universal tooth-numbering identifier. This step is essential for the clinical diagnosis stage (Sec. 3.3) because it provides explicit tooth-level visual guidance, ensuring that the VLM focuses on clinically relevant evidence at the tooth level rather than relying on broader or ambiguous image regions. We formulate this task as an object detection problem and adopt a YOLOv11-based [17] model to perform joint localization and tooth-ID classification. We train the detector on tooth-level annotations and predicts bounding boxes together with their universal numbering labels, enabling con-

sistent and standardized indexing across images.

3.3. Clinical Reasoning Module

Visual Input Setting. After the tooth detection module, OMNI-Dent assigns a bounding box and the corresponding universal tooth number to each detected tooth and overlays this information on the image, following the general paradigm of VQA-style queries used in LVLMs (e.g., asking a model to identify the object highlighted by a blue bounding box). This setup specifies the individual tooth region and provides a consistent reference for subsequent processing. By retaining the bounding box while preserving the full image context, the framework enables the model to consider the target tooth in relation to adjacent and contralateral teeth, which is essential for capturing the comparative patterns and structural relationships used in clinical reasoning.

Clinical Reasoning guidance. The clinical reasoning module aims to replicate the diagnostic workflow of human dental experts within a VLM pretrained on large-scale, diverse data. To achieve this goal, OMNI-Dent introduces expert-defined guidance that encodes the procedural steps, visual criteria, and heuristic assessments used in clinical practice.

Figure 2 illustrates the three-step diagnostic reasoning process. First, OMNI-Dent begins with role assignment and primary assessment, a step that localizes early indicators by specifying the anatomical regions to examine and the morphological or chromatic cues associated with each condition category. Second, OMNI-Dent performs structural and pattern verification. This step incorporates heuristic patterns routinely applied by clinicians, including contralateral comparison for symmetry assessment, evaluation of adjacent-tooth relationships to identify cross-tooth patterns, and consistency checks across multiple views of the same tooth. Finally, OMNI-Dent generates the final diagnosis by integrating the findings from the previous steps. The VLM processes these stages sequentially while retaining intermediate reasoning states, enabling it to emulate clinical diagnostic logic without requiring large-scale dental fine-tuning. For the VLM, we employ the InternVL3 [25] model, a state-of-the-art vision-language framework pretrained on large-scale image-text data.

3.4. Diagnosis Integration Module

After generating diagnoses for each image, the diagnosis integration module aggregates the per-tooth predictions across all views to produce a consolidated condition summary for each individual. In addition to assigning condition labels, the module also retains the accompanying reasoning outputs, enabling the framework to provide explanations that remain aligned with the underlying diagnostic logic.

4. Experiments

4.1. Datasets

We evaluate OMNI-Dent using a multi-view dental image dataset, collected at clinical and research settings in the USA and Thailand, with appropriate IRB approval. This paper focuses exclusively on adult participants from the dataset. Each participant contributes three smartphone photographs consisting of frontal, upper occlusal, and lower occlusal views. Licensed dental professionals provide tooth-level diagnostic labels. Given the lower diagnostic accuracy reported for assessing posterior teeth using smartphone-captured images [14], we focus on the upper and lower anterior teeth, including the central incisors, lateral incisors, and canines on both sides, for a total of twelve teeth. Under the Universal Numbering System, the corresponding teeth are 6–8, 9–11, 22–24, and 25–27. This design choice ensures consistent visibility across participants and supports reliable evaluation under real-world smartphone imaging conditions. We fully de-identify all images prior to analysis, and data collection and use follow appropriate ethics approval and informed consent. Due to privacy and consent constraints, the dataset remains under restricted access and is not publicly available.

Tooth Detection. To identify the exact location of each tooth for the tooth detection module, we manually annotate bounding boxes for every individual tooth along with its FDI tooth number, resulting in 313 annotated images in the dataset. We train the object detection model using a two-step supervised learning strategy: we first pretrain it on 5K open-source tooth-detection images from Roboflow [8, 9], and then fine-tune it on 229 manually annotated smartphone images to align the model with real-world imaging conditions. We evaluate the model on a separate set of 84 manually annotated images from 26 participants, which we do not include during fine-tuning.

Tooth diagnosis. To ensure fair evaluation of the full end-to-end framework, we consistently use the same 84 images from the 26 participants for assessing the performance of OMNI-Dent and its components. We evaluate OMNI-Dent on overall abnormality, which refers to whether each tooth shows any abnormal condition, including tooth wear, uncomplicated crown fracture, or dental caries. We also report the performance for each of the three diagnostic categories individually.

4.2. Evaluation Metrics

To evaluate tooth localization in the *tooth detection module*, we measure multi-class object detection performance, covering both bounding-box localization and tooth ID classification. For the *overall diagnosis* evaluation of OMNI-

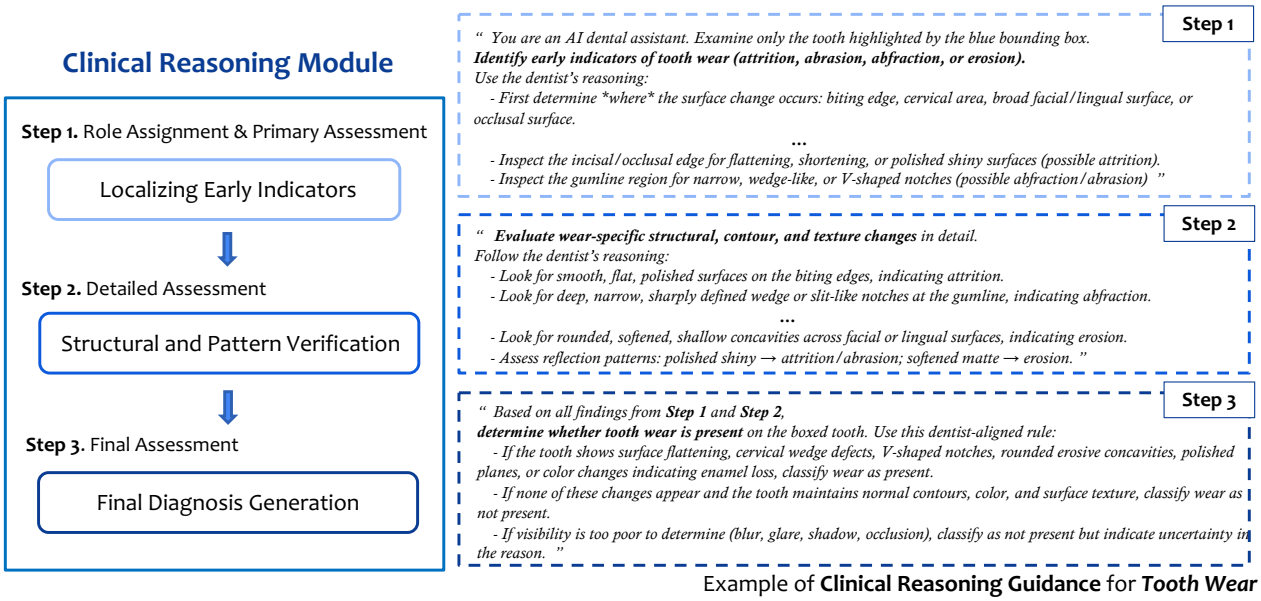

Figure 2. Three-step diagnostic reasoning in the clinical reasoning module (left). For example, the module assesses tooth wear in three stages (right), replicating a dentist's diagnostic process: Step 1 localizes early surface changes; Step 2 examines structural and textural patterns using clinician-guided criteria (e.g., attrition, abfraction/abrasion, erosion); and Step 3 integrates these findings to produce the final diagnosis.

Dent, we report tooth-level precision, recall, and F1-score to quantify how well the model's predictions align with expert diagnoses across the assessed condition categories.

4.3. Baselines

To assess the contribution of each component in OMNI-Dent, we compare the framework with three baseline settings while keeping the same VLM across all experiments. In the first baseline, the VLM analyzes cropped single-tooth images without any clinical reasoning guidance (Exp-1). In the second baseline, the model receives full images without tooth-level localization cues or reasoning guidance (Exp-2). In the third baseline, the model processes full images with a single-tooth bounding box but still does not receive any clinical reasoning prompts (Exp-3). These baselines allow us to isolate the effects of visual input configuration and clinical reasoning guidance, enabling a controlled comparison with the full OMNI-Dent setup. For all baseline conditions, we use a general VQA-style question that omits clinical reasoning cues and asks the model to determine whether specific diagnostic categories (e.g., dental caries or tooth wear) are present based solely on its overall visual impression of the individual tooth.

4.4. Implementation Details

We conduct all experiments for OMNI-Dent on a GPU server equipped with NVIDIA A100-SXM4-40GB GPUs, using the same hardware configuration for both tooth detection inference and clinical reasoning. For the clinical reasoning module, we employ the InternVL3 model with approximately 14 billion parameters as the VLM backbone. We use the model in its pretrained form without any task-specific fine-tuning. For the tooth detection module, we train a YOLOv11-based [17] detector using supervised learning. We train the model using the Adam optimizer with an initial learning rate of 1e-3 for pretraining and 1e-4 for fine-tuning, a batch size of 48, and a total of 40 training epochs. The loss function follows the standard YOLOv11 formulation [17], incorporating bounding-box regression, objectness prediction, and multi-class classification terms. We pretrain the detector on open-source datasets and fine-tune it on manually annotated images from the Minnesota State Fair dataset to better reflect real-world smartphone imaging conditions.

5. Results

5.1. Framework Evaluation

Table 1 reports the evaluation results for OMNI-Dent. We first examine overall abnormality, which indicates whether each tooth shows any condition such as tooth wear, uncomplicated crown fracture, or dental caries. We then present observations for each diagnostic category individually.

Overall abnormality. The overall abnormality category includes tooth wear, uncomplicated crown fracture, and

Diagnosis Category	Experiment	Visual Input Setting	Clinical Reasoning Guidance	Actual Positive	TP	FP	FN	Precision	Recall	F1-score
Overall Abnormality	Exp-1	Cropped single-tooth image	No	271	89	25	182	0.78	0.32	0.46
	Exp-2	Full image	No		155	22	116	**0.87**	0.57	0.69
	Exp-3	Full image + single tooth bounding box	No		141	33	130	0.81	0.52	0.63
	OMNI-Dent	Full image + single tooth bounding box	Yes		264	62	7	0.80	**0.97**	**0.88**
Wear	Exp-1	Cropped single-tooth image	No	215	38	29	177	0.56	0.17	0.26
	Exp-2	Full image	No		86	38	129	0.69	0.40	0.50
	Exp-3	Full image + single tooth bounding box	No		68	29	147	**0.70**	0.31	0.43
	OMNI-Dent	Full image + single tooth bounding box	Yes		206	115	9	0.64	**0.95**	**0.76**
Uncomplicated Crown Fracture	Exp-1	Cropped single-tooth image	No	43	0	3	43	0.00	0.00	0.00
	Exp-2	Full image	No		1	1	42	**0.50**	0.02	0.04
	Exp-3	Full image + single tooth bounding box	No		0	3	43	0.00	0.00	0.00
	OMNI-Dent	Full image + single tooth bounding box	Yes		27	159	16	0.14	**0.62**	**0.23**
Dental Caries	Exp-1	Cropped single-tooth image	No	16	5	41	11	0.10	0.31	0.16
	Exp-2	Full image	No		7	84	9	0.07	0.43	0.13
	Exp-3	Full image + single tooth bounding box	No		11	82	5	0.11	**0.68**	0.20
	OMNI-Dent	Full image + single tooth bounding box	Yes		11	59	5	**0.15**	**0.68**	**0.25**

Table 1. Evaluation results across four experimental settings for overall abnormality detection and three diagnostic categories (tooth wear, uncomplicated crown fracture, and dental caries). Bold values indicate the best performance, and underlined values indicate the second-best.

dental caries, totaling 271 positive cases. This setting best reflects OMNI-Dent's role as an early-stage screening tool.

Across the baselines, Exp-1 produces the lowest performance because the cropped single-tooth image removes surrounding anatomical context and prevents the VLM from leveraging cross-tooth relationships. Exp-2 and Exp-3 achieve similar performance levels: Exp-2 gains broader visual cues from the full image, and Exp-3 stabilizes attention by highlighting the target tooth, but both settings still lack the diagnostic reasoning needed to detect subtle abnormalities. As a result, the best baseline score appears in Exp-2 with an F1-score of 0.69. In contrast, OMNI-Dent reaches an F1-score of 0.88, outperforming all baselines by a substantial margin. Notably, OMNI-Dent achieves this improvement without any additional training that relies on large-scale annotations, highlighting the effectiveness of explicit visual guidance and clinical reasoning in enabling a general-purpose VLM to perform reliable abnormality detection.

Tooth Wear. Tooth wear includes 215 positive cases, the largest among the diagnostic categories. Exp-1 produces the lowest performance because the cropped single-tooth view removes adjacent-tooth context and prevents the VLM from comparing surface patterns across neighboring teeth. This limitation results in an F1-score far below that of OMNI-Dent, which improves performance by 0.5. Exp-2 and Exp-3 achieve higher scores than Exp-1 by providing full-image context or explicit localization, yet both settings still fall short of OMNI-Dent by a substantial margin, with F1-score differences of 0.26–0.29. These gaps highlight that, even under the same VLM, reliable tooth-wear detection requires both explicit visual guidance and clinical reasoning cues.

Uncomplicated Crown Fracture. Uncomplicated crown fracture includes 43 positive cases and typically appears as subtle enamel fractures or faint craze lines involving dentin [3]. These features challenge visual interpretation in smartphone photographs, especially when images are blurry or cues are minimal. These characteristics make uncomplicated crown fractures particularly difficult for a general-purpose VLM to recognize in smartphone photographs, and all baseline settings perform extremely poorly as a result. Exp-1, Exp-2, and Exp-3 all yield near-zero F1-scores (with the best baseline reaching only 0.04). In contrast, OMNI-Dent attains an F1-score of 0.23, representing a substantial improvement over all baselines and highlighting the essential role of clinical reasoning guidance in detecting fine-grained structural defects. Identifying uncomplicated crown fractures requires deliberate evaluation of enamel continuity across multiple views, a process that general-purpose VLMs cannot perform without explicit reasoning guidance.

Dental Caries. Dental caries includes 16 positive cases, which makes the evaluation sensitive to small prediction errors. Exp-1 relies solely on a cropped view and often confuses staining or food debris with dental caries, which lowers precision. Exp-2 adds global context but provides no localization, causing the model to attend to irrelevant regions and further degrade precision. Exp-3 localizes the target tooth and improves recall, yet the VLM still struggles to differentiate discoloration from true lesions without diagnostic reasoning. While both OMNI-Dent and the baselines show lower precision than recall for dental caries detection, OMNI-Dent improves recall to 0.68 and attains an F1-score of 0.25, outperforming the baseline methods, which remain substantially lower in both recall and F1-score.

Category	Experiment	Precision	Recall	F1-score
Uncomplicated Crown Fracture	OMNI-Dent	0.14	**0.62**	0.23
	OMNI-Dent+ ICL	**0.30**	0.37	**0.33**
Dental Caries	OMNI-Dent	0.15	**0.68**	**0.25**
	OMNI-Dent+ ICL	**0.25**	0.06	0.10

Table 2. Comparison of OMNI-Dent and OMNI-Dent + ICL on two diagnostic categories. Bold indicates the best score.

Evaluation Summary. Across all diagnostic categories, OMNI-Dent consistently outperforms the baseline configurations by a large margin. Exp-1, Exp-2, and Exp-3 each reveal different limitations of a general-purpose VLM, such as restricted context, ambiguous localization, or the absence of diagnostic reasoning, which lead to low recall and unstable performance. In contrast, OMNI-Dent combines explicit visual guidance with clinical reasoning and achieves substantial gains without any additional annotation-heavy training. These results demonstrate that clinical reasoning guidance is essential for enabling a VLM to reliably interpret subtle dental cues in real-world smartphone images. Furthermore, we provide an analysis of common failure patterns in the Section 6.1.

5.2. In-Context Learning (ICL) Capabilities

In-context learning (ICL) allows VLMs to adapt to domain-specific tasks without additional training by referencing a few example question–answer pairs during inference. In this work, we further examine whether OMNI-Dent benefits from ICL when diagnosing subtle dental conditions. For each patient, the dataset includes paired views of the same tooth (e.g., frontal and occlusal). We construct an ICL prompt using two paired images that include the tooth bounding box and the corresponding expert-provided diagnosis. At inference time, we provide two such reference pairs (four images in total) and evaluate the model's predictions on a separate image of the same tooth.

We evaluate ICL on uncomplicated crown fracture and dental caries, the two categories with the most subtle and variable visual appearances. Table 2 reports the quantitative results. ICL yields a clear improvement for uncomplicated crown fracture, raising the F1-score from 0.23 to 0.33. Because fracture lines exhibit relatively consistent structural cues across views, the reference examples help the VLM better distinguish these defects from normal surface texture. In contrast, ICL does not improve performance for dental caries. Caries exhibits highly variable visual patterns and often resembles staining or food debris. With only two paired ICL examples, the model lacks sufficient information to resolve these ambiguities, and the additional examples may even narrow or distort the model's internal decision patterns.

Pretrained	fine-tuned	Precision	Recall	F1-score
Yes	No	0.80	0.79	0.80
Yes	Yes	**0.96**	**0.89**	**0.92**

Table 3. Tooth detection performance before and after fine-tuning on smartphone images.Bold indicates the best score.

5.3. Tooth detection evaluation

We evaluate the performance of the tooth detection module by examining the impact of the two-step training strategy. We first pretrain the detector on open-source tooth-detection images, each containing bounding-box annotations for individual teeth. We then fine-tune the model on manually annotated images from the dataset to adapt the detector to real-world smartphone imaging characteristics. Table 3 summarizes the detection results. The pretrained model achieves a precision of 0.80, a recall of 0.79, and an F1-score of 0.80, indicating reasonable transferability from general-purpose tooth images. After fine-tuning on smartphone-specific images, performance improves substantially, reaching a precision of 0.96, a recall of 0.89, and an F1-score of 0.92. These results demonstrate that domain adaptation through fine-tuning is crucial for reliable tooth localization and ID classification in smartphone photographs.

6. Discussion

6.1. Failure Cases Analysis

We examine the failure cases of OMNI-Dent to understand the diagnostic errors that arise in smartphone images. Because OMNI-Dent produces explicit tooth-level reasoning for each prediction, we can directly inspect the visual cues and decision steps that lead to misclassification. This interpretability enables a more detailed and clinically meaningful failure analysis than prediction-only models.

For tooth wear, errors often occur when the occlusal surface is not clearly visible. In these situations, the model relies on secondary cues such as the incisal edge, where mild flattening or reflections can resemble attrition and lead to overdiagnosis. For uncomplicated crown fracture, false positives frequently arise from normal surface irregularities or reflections that mimic faint enamel fractures. Illumination and viewing angles in smartphone photographs can exaggerate these cues, making trauma detection particularly difficult. For dental caries, staining or food debris often creates dark regions that resemble early carious lesions. Limited contextual information and the small number of caries cases increase the likelihood of misinterpreting these regions as pathological.

6.2. Clinical Implications

OMNI-Dent suggests that a general-purpose VLM, when guided by expert-defined diagnostic reasoning, can approximate key aspects of clinical decision-making without requiring labor-intensive fine-tuning on large, expert-annotated dental datasets. The strong recall achieved across all diagnostic categories is particularly meaningful for an assistive tool intended to encourage timely dental care, as it reduces the likelihood of missed abnormalities that could otherwise delay treatment.

In addition, the stepwise reasoning outputs generated by OMNI-Dent provide interpretable justifications aligned with clinical logic, which may support early-career clinicians and offer patients clearer explanations of why a finding warrants professional evaluation. These explicit reasoning traces also contribute to explainable AI by revealing how the model arrives at its conclusions, enabling more targeted feedback and facilitating iterative refinement of the system's diagnostic behavior. Consequently, OMNI-Dent has the potential to connect accessible smartphone-based imaging with clinically informed diagnostic guidance, improving access to preliminary oral health assessment for individuals in underserved or marginalized communities who face barriers to timely professional care.

6.3. Deployment Considerations

Considering that OMNI-Dent operates on widely accessible smartphone photographs, it supports deployment in settings that lack specialized dental imaging equipment. However, reliable and effective deployment depends on several factors. First, the system guides users to capture images of sufficient quality to ensure reliable tooth localization and diagnosis. Second, privacy protections play a critical role, as smartphone images may include identifiable facial features or metadata, and the system handles this information securely. Finally, although the system demonstrates strong recall, it does not replace professional diagnosis. Appropriate human oversight remains necessary, and deployment emphasizes that the system provides preliminary assessments to assist, rather than substitute, clinical judgment.

6.4. Limitations

Despite the promising findings, OMNI-Dent exhibits several limitations. First, the evaluation dataset includes limited size and demographic diversity, which may constrain generalizability across broader populations and imaging conditions. Certain conditions, such as decay, appear infrequently in the dataset (e.g., 15 instances), reducing the reliability of condition-specific performance estimates. Second, improving diagnostic performance (Sec. 3.3) remains challenging for conditions whose visual presentation is subtle or easily confused with normal variations (e.g., uncomplicated crown fracture). These cases often lack distinct visual boundaries and may require additional contextual or non-visual information for reliable differentiation. Third, the current framework performs limited reasoning during the aggregation of multi-view predictions. The diagnosis integration module (Sec. 3.4) does not yet incorporate view-specific contextual descriptions or confidence-based weighting, which may lead to less reliable combined diagnoses across views.

6.5. Future Work

Future work will expand the diagnostic scope of OMNI-Dent and improve its robustness under real-world smartphone imaging conditions. Extending the set of diagnostic categories, including restorations and other clinically relevant conditions, and improving the reliability of the tooth detection module (Sec. 3.2) across diverse visual presentations remain important directions. Future work will also strengthen the clinical reasoning module (Sec. 3.3). Exploring medical-domain VLMs, integrating external knowledge sources, and enhancing reasoning through in-context learning (ICL) may improve performance for conditions with subtle or ambiguous cues. In addition, refining multi-view aggregation with view-specific context or confidence-based weighting in the diagnosis integration module (Sec. 3.4) may further improve diagnostic consistency. Finally, prospective evaluations in clinical and community settings will be essential for assessing usability, user trust, and the broader impact of OMNI-Dent, particularly for populations with limited access to professional care.

7. Conclusion

This paper presents OMNI-Dent, an explainable and data-efficient framework for automated dental diagnosis using multi-view smartphone photographs. By combining tooth-level detection with expert-defined clinical reasoning instructions, the framework enables a general-purpose VLM to emulate structured diagnostic workflows without large-scale dental fine-tuning. Experimental results show that OMNI-Dent consistently outperforms baseline configurations across all evaluated conditions, which is essential for an assistive system intended to encourage timely professional evaluation. Through its integration of clinically grounded reasoning, transparent outputs, and accessible imaging requirements, OMNI-Dent contributes toward more interpretable, reliable, and widely accessible dental assessment that may benefit individuals with limited access to direct dental care.

References

[1] Lyndon P Abbott, Ankita Saikia, and Robert P Anthonappa. Artificial intelligence platforms in dental caries detection: A

systematic review and meta-analysis. *Journal of Evidence-Based Dental Practice*, 25(1):102077, 2025. 1, 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[3] Austė Antipovienė, Julija Narbutaitė, and Jorma I Virtanen. Traumatic dental injuries, treatment, and complications in children and adolescents: a register-based study. *European journal of dentistry*, 15(03):557–562, 2021. 6

[4] Somayyeh Azimi, Basheer Bennamoun, Maryam Mehdizadeh, Janardhan Vignarajan, Di Xiao, Boyen Huang, Heiko Spallek, Michelle Irving, Estie Kruger, Marc Tennant, et al. Teledentistry improves access to oral care: A cluster randomised controlled trial. In *Healthcare*, page 2282. MDPI, 2025. 1, 2

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1, 3

[7] Omri Dan, Samuel Lilek, Ariel Hirschhorn, Lazar Kats, Nahum Kiryati, and Arnaldo Mayer. Artifact correction in panoramic radiographs using deep de-shadowing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1008–1016, 2025. 1, 2

[8] dental. dent dataset, 2023. Open source dataset, visited on 2025-12-02. 4

[9] DentalMate6v. Intraoral tooth numbering fdi dataset, 2024. Open source dataset, visited on 2025-12-02. 4

[10] Chenlin Du, Xiaoxuan Chen, Jingyi Wang, Junjie Wang, Zhongsen Li, Zongjiu Zhang, and Qicheng Lao. Prompting vision-language models for dental notation aware abnormality detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 687–697. Springer, 2024. 2, 3

[11] Mohamed Estai, Yogesan Kanagasingam, Boyen Huang, Hellen Checker, Lesley Steele, Estie Kruger, and Marc Tennant. The efficacy of remote screening for dental caries by mid-level dental providers using a mobile teledentistry model. *Community Dentistry and Oral Epidemiology*, 44(5): 435–441, 2016. 1, 2

[12] Mohamed Estai, Yogesan Kanagasingam, Di Xiao, Janardhan Vignarajan, Boyan Huang, Estie Kruger, and Marc Tennant. A proof-of-concept evaluation of a cloud-based store-and-forward telemedicine app for screening for oral diseases. *Journal of telemedicine and telecare*, 22(6):319–325, 2016.

[13] Mohamed Estai, Yogesan Kanagasingam, Boyen Huang, Julia Shiikha, Estie Kruger, Stuart Bunt, and Marc Tennant. Comparison of a smartphone-based photographic method with face-to-face caries assessment: a mobile teledentistry model. *Telemedicine and e-Health*, 23(5):435–440, 2017.

[14] Mohamed Estai, Yogesan Kanagasingam, Maryam Mehdizadeh, Janardhan Vignarajan, Richard Norman, Boyen Huang, Heiko Spallek, Michelle Irving, Amit Arora, Estie Kruger, et al. Mobile photographic screening for dental caries in children: diagnostic performance compared to unaided visual dental examination. *Journal of Public Health Dentistry*, 82(2):166–175, 2022. 1, 2, 4

[15] Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lin Lizhuo, Jinrong Yang, Qiyong Hemis Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2, 3

[16] Boyen Huang, Mohamed Estai, Patimaporn Pungchanchaikul, Karin Quick, Sarbin Ranjitkar, Emily Fashingbauer, Abdirahim Askar, Josiah Wang, Fatma Diefalla, Margaret Shenouda, et al. Mobile health assessment of traumatic dental injuries using smartphone-acquired photographs: a multicenter diagnostic accuracy study. *Telemedicine and e-Health*, 30(10):2592–2600, 2024. 1, 2

[17] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 3, 5

[18] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 1, 2

[19] Marco A Peres, Lorna MD Macpherson, Robert J Weyant, Blánaid Daly, Renato Venturelli, Manu R Mathur, Stefan Listl, Roger Keller Celeste, Carol C Guarnizo-Herreño, Cristin Kearns, et al. Oral diseases: a global public health challenge. *The Lancet*, 394(10194):249–260, 2019. 1

[20] Md Shahinoor Rahman, Jeffrey C Blossom, Ichiro Kawachi, Renuka Tipirneni, and Hawazin W Elani. Dental clinic deserts in the us: spatial accessibility analysis. *JAMA Network Open*, 7(12):e2451625–e2451625, 2024. 1, 2

[21] Airton Oliveira Santos-Junior, Rocharles Cavalcante Fontenele, Frederico Sampaio Neves, Saleem Ali, Reinhilde Jacobs, and Mário Tanomaru-Filho. A unique ai-based tool for automated segmentation of pulp cavity structures in maxillary premolars on cbct. *Scientific Reports*, 15(1):5509, 2025. 1, 2

[22] Emily C Schultz, Boyen Huang, Margaret Shenouda, Mohamed Estai, Sarbin Ranjitkar, Jeffrey P Louie, and Patimaporn Pungchanchaikul. In-review: Perspectives of front-line clinicians and remote reviewers on smartphone-based photography for assessing traumatic dental injuries: A qualitative study. 2025. 1, 2

[23] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 2, 3

[24] Lin Wang, Yanyan Xu, Weiqian Wang, and Yuanyuan Lu. Application of machine learning in dentistry: insights, prospects and challenges. *Acta Odontologica Scandinavica*, 84:43345, 2025. 1, 2

[25] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 3, 4

[26] Yongjia Wu, Yun Zhang, Yange Wu, Qianhan Zheng, Xiaojun Li, and Xuepeng Chen. Chatios: Improving automatic 3-dimensional tooth segmentation via gpt-4v and multimodal pre-training. *Journal of Dentistry*, 157:105755, 2025. 2, 3

[27] Xuan Zhang, Yuan Liang, Wen Li, Chao Liu, Deao Gu, Weibin Sun, and Leiying Miao. Development and evaluation of deep learning for screening dental caries from oral photographs. *Oral diseases*, 28(1):173–181, 2022. 1, 2