OPGAgent: An Agent for Auditable Dental Panoramic X-ray Interpretation

Zhaolin Yu[1,2], Litao Yang[1,2](✉), Ben Babicka[3], Ming Hu[1,2], Jing Hao[4], Anthony Huang[5], James Huang[5], Yueming Jin[6], Jiasong Wu[7], and Zongyuan Ge[1,2,5](✉)

[1] AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia
[2] Faculty of Information Technology, Monash University, Melbourne, Australia
[3] Curae Health, Melbourne, Australia
[4] Faculty of Dentistry, The University of Hong Kong, Hong Kong
[5] Airdoc-Monash Research, Monash University, VIC 3800, Australia
[6] National University of Singapore, Singapore City, Singapore
[7] Southeast University, Nanjing, China
litao.yang@monash.edu, zongyuan.ge@monash.edu

Abstract. Orthopantomograms (OPGs) are the standard panoramic radiograph in dentistry, used for full-arch screening across multiple diagnostic tasks. While Vision Language Models (VLMs) now allow multi-task OPG analysis through natural language, they underperform task-specific models on most individual tasks. Agentic systems that orchestrate specialized tools offer a path to both versatility and accuracy, yet no agent has been designed for dental imaging. To address this gap, we propose OPGAgent, a multi-tool agentic system for auditable OPG interpretation. OPGAgent coordinates specialized perception modules with a consensus mechanism through three components: (1) a Hierarchical Evidence Gathering module that decomposes OPG analysis into global, quadrant, and tooth-level phases while dynamically invoking tools, (2) a Specialized Toolbox encapsulating spatial, detection, utility, and expert zoos, and (3) a Consensus Subagent that resolves conflicts through anatomical constraints. We further propose OPG-Bench, a structured-report benchmark based on (Location, Field, Value) triples derived from real clinical reports, which enables a comprehensive review of findings and hallucinations that goes beyond the scope of VQA indicators. On our OPG-Bench and the public MMOral-OPG benchmark, OPGAgent outperforms current dental VLMs and medical agent frameworks across both structured-report and VQA evaluation. Code and the structured-report schema are available at `https://github.com/Zhaolin-Yu/OPGAgent`.

Keywords: Radiography, Panoramic · Artificial Intelligence · Diagnosis, Computer-Assisted · Tooth Diseases

1 Introduction

The Orthopantomogram (OPG) captures the maxilla, mandible, and full dentition in a single panoramic exposure and is routinely used for screening, diagnosis,

and treatment planning [4]. Deep learning models now exist for individual steps of OPG analysis, including caries detection [1], alveolar bone loss assessment [14], tooth segmentation [22,24], and tooth enumeration [8]. However, as recent reviews note [7,19], since these separate models cannot connect with each other, clinicians must use multiple tools one by one, making their workflow inefficient. This creates a need for a unified dental model that can handle multiple tasks at once.

To address this need for multi-tasking, Vision Language Models (VLMs) take a different approach by framing diverse visual tasks as a unified text generation problem. Dental VLMs such as DentalGPT [2] and OralGPT-Omni [10], together with general medical VLMs like LLaVA-Med [15] and MedGemma [6], can perform pathology identification, anatomical description, and report generation through natural language. Their accuracy on individual tasks, however, lags behind that of specialized models, where local spatial cues dominate [5]. Therefore, overcoming this performance limitation while sustaining multi-task adaptability remains a challenge.

To bridge this gap, AI agents [26] coordinate specialized models and external tools, integrating high-precision outputs into a unified language framework. MedAgents [20] and MDAgents [13] use collaborative reasoning for clinical decision-making; MedAgent-Pro [25] adds external perception tools to improve diagnostic accuracy. While these frameworks excel in specific modalities or broad clinical applications, they are not designed to capture the specialized nuances of dentistry. Effective processing of dental imaging relies heavily on domain knowledge like FDI notation [8,12] and specialized mechanisms such as dynamic Region of Interest (ROI) cropping. Our experiments reveal a significant performance gap when applying medical agent designs (Table 1) to specialized dental tasks, underscoring the necessity for a domain-specific agent.

To reliably test these models, we need a complete benchmark. However, existing VQA benchmarks, whether closed-ended or open-ended, measure precision only on the questions asked; recall depends entirely on whether those questions cover all clinically relevant findings. As a result, (1) any finding type absent from the question set is invisible to evaluation, and (2) hallucination severity cannot be quantified, since fabricated findings in unprompted regions go undetected [3,28]. Furthermore, this question-and-answer paradigm fundamentally diverges from real-world clinical practice. Dentists do not analyze an OPG by answering isolated prompts; rather, they conduct comprehensive visual screenings to generate structured, holistic reports. Consequently, current benchmarks fail to reflect a model's true diagnostic utility in an actual clinical workflow.

To address these issues, we propose OPGAgent, a multi-tool agentic system for auditable OPG interpretation that combines specialized perception modules with a consensus mechanism. The system has three components: (1) a **Hierarchical Evidence Gathering** module that decomposes analysis into global, quadrant, and tooth-level phases; (2) a **Specialized Toolbox** encapsulating four categories; and (3) a **Consensus Subagent** that aggregates votes from detection tools and expert zoos and resolves conflicts through anatomical con-

straints. Furthermore, to align our evaluation with real clinical workflows rather than isolated prompts, we construct OPG-Bench, a structured-report benchmark utilizing (Location, Field, Value) triples to overcome the failure of VQA in reflecting true diagnostic capabilities. **In summary, our contributions are threefold:** (1) **OPGAgent**, the first agentic system specifically designed for OPG interpretation, featuring Hierarchical Evidence Gathering, a Specialized Toolbox, and a Consensus Subagent; (2) **OPG-Bench**, a novel evaluation protocol derived from real clinical reports that explicitly audits both pathological findings and hallucinations; and (3) **State-of-the-art performance**, demonstrating that OPGAgent outperforms all baselines on both OPG-Bench and MMOral-OPG [9], achieving higher F1 scores with fewer false positives.

2 OPGAgent

Overview. Fig. 1 illustrates the workflow of OPGAgent. Given an input OPG image I, OPGAgent is a training-free planner powered by GPT-5.2 following the ReAct paradigm [26] that orchestrates three modules: Hierarchical Evidence Gathering, Specialized Toolbox, and Consensus Subagent. In the **Hierarchical Evidence Gathering** module, OPGAgent decomposes the analysis into three phases: global dentition, quadrant-level screening, and tooth-level screening, with findings stored in Memory. The **Specialized Toolbox** wraps perception models and VLM experts as agent-callable tools in four categories: spatial, detection, utility, and expert zoos. The **Consensus Subagent** aggregates evidence from detection tools and expert zoos, resolves conflicts using anatomical constraints, and commits only well-supported findings to the final report.

2.1 Hierarchical Evidence Gathering

To generate refined, verifiable findings, we design a Hierarchical Evidence Gathering module that progressively refines results through three phases, from full images to the tooth level, and stores all findings in Memory.

Phase 1: Global Analysis. First, the agent queries VLM experts on the full image to generate an initial reading. Simultaneously, it invokes detection tools to establish the anatomical baseline: total tooth count, missing teeth, and FDI notation mapping [12] for each tooth. This phase creates a structured coordinate system in Memory for all subsequent phases.

Phase 2: Quadrant-Level Screening. Using the global context in Memory, OPGAgent scans each quadrant (Q1–Q4). It uses coordinates from Phase 1 to generate dynamic quadrant crops, which are sent to VLM experts to screen for gross pathologies such as alveolar bone loss or large lesions. The marked regions will be sent to Phase 3 for tooth-level investigation.

Phase 3: Tooth-Level Screening. For detailed pathologies (e.g., caries, impaction status), OPGAgent crops dynamic ROIs using coordinates in Memory. Because bounding boxes preserve local context, relatively higher resolution crops are sent to VLM experts for assessment. If Phase 2 flags a potential issue in

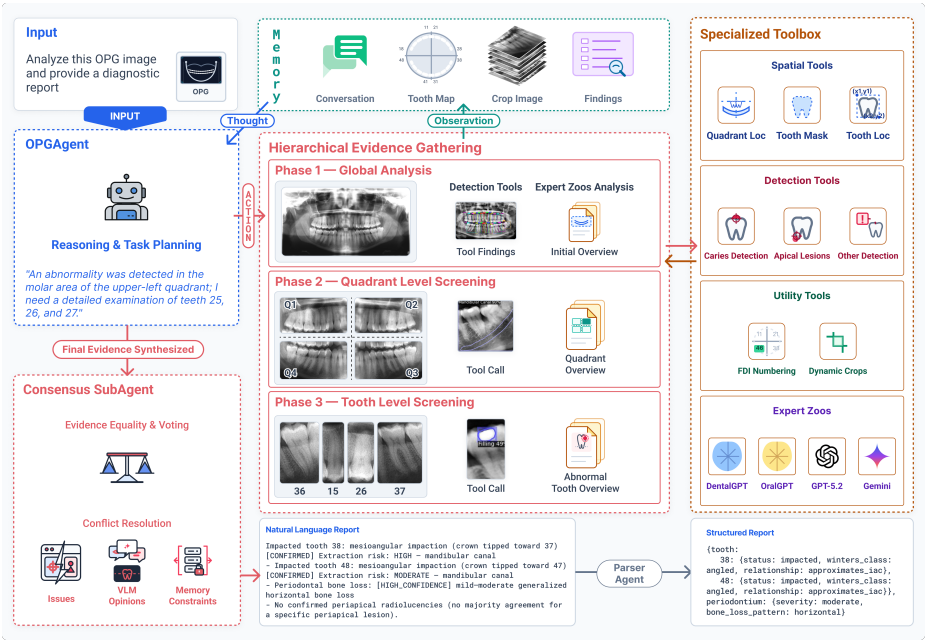

Fig. 1. Overview of OPGAgent. The Agent orchestrates three modules: Hierarchical Evidence Gathering, Specialized Toolbox, and Consensus Subagent.

a region where Phase 1 detected no tooth (e.g., a root remnant), OPGAgent requests a quadrant crop from independent anatomical detection (not tooth coordinates), allowing VLMs to recover false negatives missed by the tooth-level scan.

2.2 Specialized Toolbox

To provide structured visual perception for OPGAgent, we implement perception modules and expose them as agent-friendly tools. These tools are powered by four underlying model types: (1) YOLO [23] trained on public dental datasets [8] for quadrant and tooth detection; (2) open-source dental vision experts from OralGPT-Omni [10], based on MaskDINO [16] and DINO [27], for maxillary sinus, mandibular canal, and pathology detection; (3) MedSAM [17] for tooth mask segmentation; and (4) multiple VLM experts for whole-image or ROI-level analysis.

Spatial Tools. These tools return masks or bounding boxes with coordinates for teeth, quadrants, and detected findings. These coordinates form the spatial reference for all subsequent phases.

Detection Tools. These tools detect specific or all abnormal conditions using various pathological models (caries, periapical lesions, apical fillings, etc.) and determine which teeth are associated with these abnormalities.

Utility Tools. This suite manages FDI numbering, ROI extraction, and spatial reasoning. To assess clinical risks such as the proximity of a tooth root to the mandibular canal, the agent computes the minimum contour distance $d(m_i, m_j)$ between two polygon masks m_i and m_j:

$$d(m_i, m_j) = \min_{p \in \partial m_i,\, q \in \partial m_j} \|p - q\|_2 \tag{1}$$

where $d(m_i, m_j) = 0$ if the polygons intersect. For matching diseases to teeth, the agent prioritizes Intersection over Union (IoU), using the Euclidean distance between box centers as a fallback when IoU is below a threshold.

Expert Zoos. These tools call DentalGPT, OralGPT-Omni, GPT-5.2, and Gemini-3-Flash to collect multiple expert opinions on whole images, whole images with bounding box, or specific ROIs. These opinions form the evidence base for the Consensus Subagent.

2.3 Consensus Subagent

To mitigate hallucination risks in generative models, we design a Consensus Subagent that aggregates evidence from detection tools and expert zoos.

Evidence Aggregation. The Consensus Subagent collects opinions from N sources and confirms a finding when ≥ 3 sources agree or ≥ 2 sources report the same finding. All sources receive equal voting weight so that rare conditions missed by rule-based detectors can still pass through VLM votes.

Conflict Resolution. Disagreements often concern exact attributes rather than the presence of a finding. When a finding is confirmed by majority vote but sources disagree on its attributes (e.g., tooth number or severity grade), the subagent checks the conflicting claims against hard constraints provided by the detection tools. For instance, if VLMs agree on a "lower left implant" but assign different tooth numbers, the subagent consults the detection tool's FDI coordinate map to assign the finding to the spatially correct tooth, correcting the VLM output with deterministic coordinates.

2.4 Structured Reporting and Benchmark Design

Hierarchical Clinical Ontology. To align the agent's output with real-world clinical triage, we formalize the OPG interpretation into a structured, anatomy-centric schema. We define a comprehensive dental report as a set of pathological findings $\mathcal{S} = \{t_1, t_2, \ldots, t_N\}$, where each finding is abstracted as an information triple $t_i = (l_i, f_i, v_i) \in \mathcal{L} \times \mathcal{F} \times \mathcal{V}$. Here, $\mathcal{L}$ denotes the anatomical location space strictly adhering to the FDI World Dental Federation notation (ISO 3950) for teeth, alongside global regions (e.g., TMJ, maxillary sinus). $\mathcal{F}$ represents the clinical field, and $\mathcal{V}$ is the categorical value space. To ensure clinical completeness, $\mathcal{V}$ is constructed upon standardized dental guidelines, mapping complex numerical scales into semantic severity levels (e.g., ICDAS [11] for caries, AAP/EFP 2017 [21] for periodontitis, and PAI [18] for periapical health). Following the sparse representation principle of clinical reporting, $\mathcal{S}$ only encapsulates anomalous findings; normal anatomical structures are intentionally omitted.

Constrained Parser Agent. A dedicated Parser Agent transforms the system's Memory output into the formal $\mathcal{S}$ space. To eliminate hallucinations and enforce schema adherence, it utilizes constrained decoding—via JSON mode, Pydantic validation, and few-shot learning. This deterministically maps findings into valid (l_i, f_i, v_i) triples, strictly preventing unsupported free-text diagnoses.

Triple-based Hierarchical Evaluation (OPG-Bench). Overcoming VQA limitations, we evaluate findings as $(location, field, value)$ triples via two complementary metrics. The **Exact Match** requires perfect alignment of all elements. To reflect clinical workflows, we also compute a step-wise partial match: **Detection** of pathology presence (Precision/Recall/F1), **Localization** of the anatomical target l_i (Precision/Recall/F1), and **Classification** of the grade/type v_i (Accuracy). Each step conditions on the previous one. This hierarchical protocol cleanly disentangles visual perception from clinical reasoning.

3 Experiments

Dataset and Implementation. The OPG-Bench dataset comprises 1,006 anonymized OPGs with paired clinical and structured reports, and 5,219 unguided VQA pairs; evaluation follows the triple-based protocol defined in §2.4. Sourced from multiple clinics, the dataset is restricted to patients $\geq$ 16 years to focus on permanent dentition. Ground truths derive from real clinical reports and are validated via manual spot-checks. Since real-world reports typically prioritize chief complaints over incidental findings, our reported False Positive rates may be slightly inflated when the agent detects undocumented yet valid pathologies.

Evaluation Protocol. We evaluate OPGAgent against state-of-the-art general VLMs, dental-specific VLMs, and medical agents. To isolate our framework's architectural contribution, MedAgent-Pro [25] is configured with the same LLM and tools as OPGAgent. To decouple diagnostic accuracy from formatting, we employ a unified generation-evaluation pipeline. All models receive identical prompts (temperature 0.3) to generate natural-language reports. A shared Parser Agent then converts these into structured formats—ensuring parser bias applies equally across methods—followed by manual filtering to retain only valid clinical findings.

Results on OPG-Bench. We evaluate OPGAgent against general VLMs, and dental VLMs [2,10] on the OPG-Bench dataset. As shown in Table 1, OPGAgent reaches an exact-match F1 of 42.3% and an aggregate score of 49.7%, surpassing Gemini-3-Flash and DentalGPT. While Gemini-3-Flash obtains higher recall (45.1%), its precision drops to 27.6% with 10.58 false positives per case; OPGAgent maintains precision at 43.1% while keeping false positives at 4.89. Domain-specific models like OralGPT-Omni minimize false positives (4.02) but at the cost of low coverage (F1 6.2%). OPGAgent thus balances sensitivity and reliability better than single-pass approaches. A paired bootstrap (B=1000, n=1006) confirms this margin is significant: OPGAgent's Score 95% CI $[48.2, 51.3]$ does not overlap that of the strongest baseline Gemini-3-Flash $[41.5, 44.4]$ (bootstrap standard error $\leq 0.84\%$ for all methods; Holm-corrected

Table 1. Quantitative comparison on OPG-Bench dataset.

Methods	Score	Exact Match				Det.			Loc.			Cls.
		P	R	F1	FP↓	P	R	F1	P	R	F1	Acc
Gemini-3-Flash	0.428	0.276	**0.451**	0.343	10.58	0.440	**0.596**	0.506	0.311	**0.488**	0.380	0.796
Kimi-k2.5	0.399	0.252	0.345	0.291	9.20	0.481	0.500	0.490	0.348	0.463	0.397	0.764
Gemini-3-Pro	0.399	0.323	0.294	0.308	5.60	0.556	0.431	0.485	0.458	0.319	0.376	0.732
Qwen3.6-27B	0.368	0.216	0.335	0.263	10.85	0.557	0.464	0.506	0.247	0.390	0.303	0.748
GPT-5.2	0.357	0.283	0.273	0.278	6.17	0.624	0.359	0.456	0.392	0.193	0.258	0.754
Qwen3-VL-235B	0.230	0.106	0.191	0.137	14.50	0.316	0.436	0.366	0.107	0.175	0.133	0.614
DentalGPT	0.295	0.128	0.200	0.156	12.22	0.464	0.483	0.473	0.241	0.273	0.256	0.715
OralGPT-Omni	0.157	0.094	0.046	0.062	**4.02**	0.457	0.151	0.227	0.145	0.075	0.099	0.603
MedAgent-Pro	0.278	0.171	0.183	0.177	8.00	0.412	0.438	0.425	0.206	0.204	0.205	0.632
$\text{OPGAgent}_{\text{GPT-5.2}}$	**0.497**	**0.431**	0.415	**0.423**	4.89	**0.715**	0.456	**0.557**	**0.625**	0.376	**0.469**	**0.801**
$\text{OPGAgent}_{\text{Qwen3.6-27B}}$	0.452	0.402	0.347	0.372	4.61	0.659	0.407	0.503	0.570	0.353	0.436	0.779

$Score = 0.5\times Exact\text{-}Match\_F1 + 0.2\times Det\_F1 + 0.2\times Loc\_F1 + 0.1\times Cls\_Acc$ (balancing exact and partial matches). Det. (Presence), Loc. (Location), Cls. (Grade/Type), FP (False Positives/Case). Subscripts on OPGAgent denote the planner LLM. ***Best*** *and second-best per column.*

Table 2. Performance comparison on VQA benchmarks.

Methods	MMOral-OPG					OPG-Bench (VQA)					
	Acc.	Tee.	Pat.	Jaw	HisT	Acc.	Mis.	Muc.	Bone	Anat.	Api.
Gemini-3-Flash	48.47%	54.1%	42.6%	59.7%	66.2%	59.9%	48.4%	47.9%	60.0%	65.1%	**42.2%**
GPT-5.2	43.58%	37.2%	32.4%	65.9%	52.1%	51.5%	48.8%	44.7%	50.5%	65.1%	27.7%
OralGPT-Omni	59.06%	53.5%	44.6%	**77.5%**	69.7%	36.1%	13.6%	15.7%	23.2%	30.8%	17.0%
DentalGPT	48.27%	42.3%	40.5%	65.1%	47.2%	27.8%	11.4%	22.2%	4.2%	13.0%	17.5%
OPGAgent	**62.53%**	**59.7%**	**52.7%**	72.1%	**71.8%**	**61.2%**	**67.0%**	**65.2%**	**76.8%**	**80.8%**	35.0%

Tee. (Teeth-related), Pat. (Pathology), HisT (History & Treatment). Mis. (Missing Teeth), Muc. (Mucosal Change), Bone (Bone Variant), Anat. (Anatomical Variants), Api. (Apical Status).

$p < 0.001$ against every baseline). To assess dependence on the planner, we replace the GPT-5.2 planner with the open-source Qwen3.6-27B: the score drops only 4.5 points (49.7%→45.2%) yet still leads all baselines, while the same model used directly as a VLM reaches only 36.8%. This confirms that the gains stem from the agentic framework rather than the raw capability of the planner.

Results on VQA Benchmarks. As shown in Table 2, OPGAgent leads on both benchmarks (62.53% on MMOral-OPG, 61.2% on OPG-Bench). On MMOral-OPG, OralGPT-Omni scores highest on Jaw questions (77.5%) but falls behind on Teeth and Pathology categories, where OPGAgent's multi-tool pipeline provides clearer gains. On OPG-Bench, domain-specific VLMs (OralGPT-Omni 36.1%, DentalGPT 27.8%) score much lower than general VLMs, despite OralGPT-Omni performing well on its own MMOral-OPG benchmark (59.06%). This gap suggests that models trained on generated QA pairs may not fully capture the distribution of real clinical reports.

Results on Ablation Studies. Table 3 adds components cumulatively. By design, the agent LLM serves only as a planner. Without external tools (Row 1), it is forced to act as the sole diagnostician in a ReAct loop, yielding a 27.78% F1. Adding **Expert Zoos** (Row 2) improves precision (38.62%) and minimizes

Table 3. Ablation study on OPG-Bench with cumulative component additions.

Components			Exact Match				Det.	Loc.	Cls.
Expert Zoos	Spatial Tools	Detection Tools	P	R	F1	FP/Case↓	F1	F1	Acc
✗	✗	✗	28.32%	27.25%	27.78%	6.17	45.57%	25.84%	75.40%
✓	✗	✗	38.62%	16.64%	23.25%	2.37	33.20%	24.19%	66.41%
✓	✓	✗	37.13%	35.98%	36.55%	5.47	49.07%	37.05%	76.23%
✓	✓	✓	**43.15%**	**41.48%**	**42.30%**	**4.89**	**55.67%**	**46.94%**	**80.10%**

Expert Zoos (GPT-5.2, Gemini-3-Flash, DentalGPT, OralGPT-Omni), Detection Tools (Pathology Detection), Spatial Tools (Tooth/Quadrant Localization). Det. and Loc. report F1; Cls. reports accuracy.

false positives (6.17 → 2.37), but plummets recall (16.64%) because findings lack exact FDI coordinates and fail strict triple matching. **Spatial Tools** (Row 3) resolve this grounding bottleneck, restoring recall (35.98%) and F1 (36.55%). Finally, **Detection Tools** (Row 4) provide complementary localized evidence, achieving the peak 42.30% F1. The partial metrics follow the same trend, with detection F1, localization F1, and classification accuracy all peaking for the full system (55.67%, 46.94%, 80.10%).

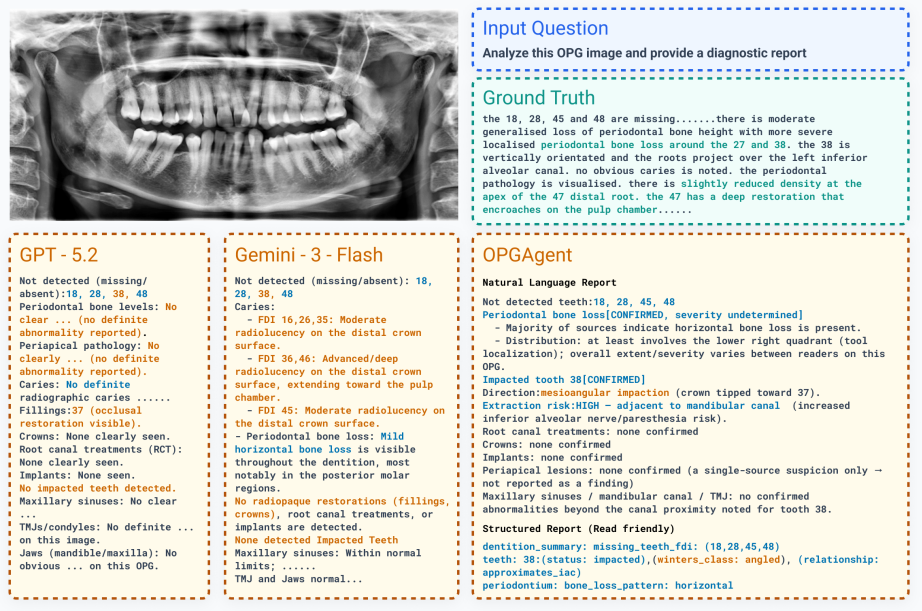

Fig. 2. Vision comparison. OPGAgent correctly identifies missing teeth, bone loss, and impacted tooth 38 via consensus, while GPT-5.2 and Gemini-3-Flash each miss or hallucinate findings.Blue for correct; Orange for wrong; Green for all undetected.

4 Conclusion

We present OPGAgent, a multi-tool agentic framework for auditable OPG interpretation that combines Hierarchical Evidence Gathering, a Specialized Toolbox, and a Consensus Subagent. Through phased analysis, multi-source voting, anatomical conflict resolution, and structured triple-based evaluation, OPGAgent produces more accurate and auditable dental reports than existing VLMs and agent frameworks. Ablation studies confirm that each module contributes to the overall performance.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bayraktar, Y., et al.: A novel deep learning-based perspective for tooth numbering and caries detection. Clinical Oral Investigations **28**, 178 (2024). doi:10.1007/s00784-024-05566-w
2. Cai, Z., et al.: DentalGPT: Incentivizing Multimodal Complex Reasoning in Dentistry. arXiv preprint arXiv:2512.11558 (2025)
3. Chen, J., et al.: Detecting and Evaluating Medical Hallucinations in Large Vision Language Models. In: ICLR (2025). arXiv:2406.10185
4. Fraser, J.: Artificial intelligence and dental panoramic radiographs: where are we now? British Dental Journal (2024). doi:10.1038/s41432-024-00978-9
5. Ghimire, A., et al.: When CNNs Outperform Transformers and Mambas: Revisiting Deep Architectures for Dental Caries Segmentation. arXiv preprint arXiv:2511.14860 (2025)
6. Golden, D., Pilgrim, R., et al.: MedGemma: Open models for health AI development. Google Research (2025). `https://developers.google.com/health-ai-developer-foundations/medgemma`
7. Grünhagen, T., et al.: Automating Dental Condition Detection on Panoramic Radiographs: Challenges, Pitfalls, and Opportunities. Diagnostics **14**(20), 2336 (2024). doi:10.3390/diagnostics14202336
8. Hamamci, I.E., et al.: DENTEX: Dental Enumeration and Tooth Pathosis Detection Benchmark for Panoramic X-rays. arXiv preprint arXiv:2305.19112 (2023)
9. Hao, J., et al.: Towards better dental AI: A multimodal benchmark and instruction dataset for panoramic X-ray analysis. arXiv preprint arXiv:2509.09254 (2025)
10. Hao, J., et al.: OralGPT-Omni: A Versatile Dental Multimodal Large Language Model. arXiv preprint arXiv:2511.22055 (2025)
11. Ismail, A.I., et al.: The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. Community Dentistry and Oral Epidemiology **35**(3), 170–178 (2007). doi:10.1111/j.1600-0528.2007.00347.x
12. ISO 3950:2016. Dentistry—Designation system for teeth and areas of the oral cavity. International Organization for Standardization (2016)
13. Kim, Y., et al.: MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. arXiv preprint arXiv:2404.15155 (2024)
14. Kurt-Bayrakdar, S., et al.: Detection of periodontal bone loss patterns and furcation defects from panoramic radiographs using deep learning algorithm. BMC Oral Health **24**, 155 (2024). doi:10.1186/s12903-024-03896-5

15. Li, C., et al.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In: NeurIPS 2023 Datasets and Benchmarks Track (2023). arXiv:2306.00890
16. Li, F., et al.: Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation. In: CVPR (2023). arXiv:2206.02777
17. Ma, J., et al.: Segment Anything in Medical Images. Nature Communications **15**, 654 (2024). doi:10.1038/s41467-024-44824-z
18. Ørstavik, D., Kerekes, K., Eriksen, H.M.: The periapical index: A scoring system for radiographic assessment of apical periodontitis. Dental Traumatology **2**(1), 20–34 (1986). doi:10.1111/j.1600-9657.1986.tb00119.x
19. Sohrabniya, F., et al.: Exploring a decade of deep learning in dentistry: A comprehensive mapping review. Clinical Oral Investigations **29**, 143 (2025). doi:10.1007/s00784-025-06216-5
20. Tang, X., et al.: MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In: Findings of ACL, pp. 599–621 (2024). arXiv:2311.10537
21. Tonetti, M.S., Greenwell, H., Kornman, K.S.: Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. Journal of Periodontology **89**(S1) (2018). doi:10.1002/JPER.18-0006
22. van Nistelrooij, N., et al.: Combining public datasets for automated tooth assessment in panoramic radiographs. BMC Oral Health **24**, 387 (2024). doi:10.1186/s12903-024-04129-5
23. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint arXiv:2402.13616 (2024)
24. Wang, Y., et al.: MICCAI STS 2024 Challenge: Semi-Supervised Instance-Level Tooth Segmentation in Panoramic X-ray and CBCT Images. arXiv preprint arXiv:2511.22911 (2025)
25. Wang, Z., et al.: MedAgent-Pro: Towards Evidence-based Multi-modal Medical Diagnosis via Reasoning Agentic Workflow. arXiv preprint arXiv:2503.18968 (2025)
26. Yao, S., et al.: ReAct: Synergizing Reasoning and Acting in Language Models. In: ICLR (2023). arXiv:2210.03629
27. Zhang, H., et al.: DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In: ICLR (2023). arXiv:2203.03605
28. Zhao, L., et al.: MARINE: Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance. In: ICML (2025)