POCKET-DENTIST: BENCHMARKING COMPACT VISION-LANGUAGE MODELS FOR DENTAL IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Oral diseases affect approximately 3.7 billion people, yet access to specialist screening remains limited for many people. As smartphones become more widely available, compact vision–language models (VLMs) could help extend dental image assessment to settings with limited access to specialists. However, it remains unclear how well VLMs perform across dental imaging modalities and clinical tasks, or whether capable models can run within mobile hardware constraints. We introduce Pocket-Dentist, a large-scale curated dental imaging dataset, together with a comprehensive, deployment-aware benchmark for evaluating vision–language models across dental imaging modalities and clinical tasks. It includes over 50,800 images from more than 5,900 patients, spanning panoramic, periapical, and cephalometric radiographs and intraoral photographs, and covers four task types and 11 sub-tasks. We study dental-task performance across 13 representative VLMs under zero-shot and few-shot prompting, and examine the effects of fine-tuning the open-weight models. Our results show that, for zero-shot and few-shot settings, no VLM performs significantly better than others across all tasks. However, after fine-tuning, compact 4B VLMs such as Qwen3.5-4B become competitive with larger VLMs. We also observe that, even trained on large scale medical data, state-of-the-art medical VLMs provides no consistent advantage on dental understanding domain. To support practical deployment, we further proposed an on-device framework and deployed best performing VLMs. Our benchmark suggests that compact VLMs is practical for on-device use in dental image understanding. Our project page is available at https://anonymous.4open.science/r/Pocket-Dentist-Benchmark-Anonymous-62B3.

1 INTRODUCTION

Oral diseases are among the most prevalent chronic health conditions worldwide, affecting approximately 3.7 billion people (Bernabe et al., 2025). In remote and underserved communities, patients often lack access to basic oral health screening and specialist diagnosis. Meanwhile, smartphones are becoming more common in regions where specialist dental services remain scarce. This raises a natural question. *Can compact vision-language models being developed to assess oral health, explain their findings, and run on devices people use every day?*

Dental care relies on heterogeneous visual evidence, including panoramic radiographs, intraoral photographs, periapical radiographs, and cephalometric images, with outputs spanning multiple formats. Yet current dental AI evaluation largely studies these pieces in isolation: datasets are usually modality-specific, task definitions are not standardized across resources, and performance may not transfer consistently across datasets or imaging conditions (Uribe et al., 2024; Li et al., 2026). Evaluating models for use on everyday devices also requires understanding fine-tuning needs, structured-output reliability, memory requirements, and inference latency alongside task accuracy.

To assess compact VLMs on dental tasks and their feasibility for deployment, we introduce Pocket-Dentist, a multimodal benchmark and deployment-aware evaluation framework. We curated large cohort dental datasets comprising 50,800 images from 5,900 patients and converted their annotations into image-grounded questions with defined answer formats and task-specific scoring rules,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

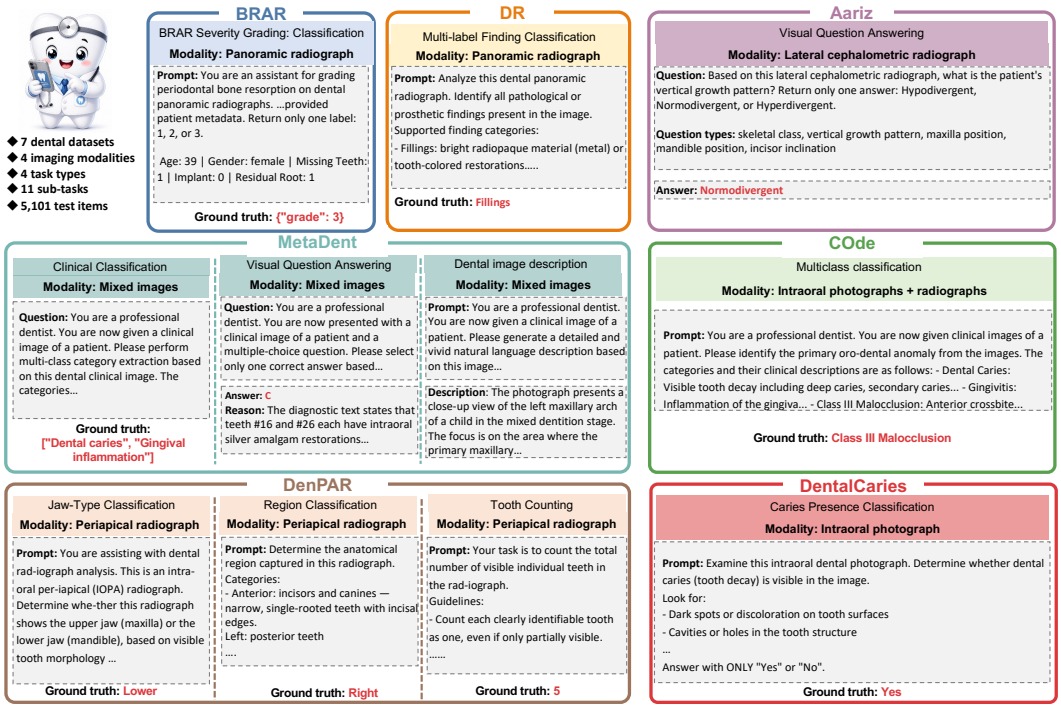


Figure 1: Overview of the curated Pocket-Dentist benchmark datasets and task formulations. We curated a large cohort consists of seven dental datasets, including BRAR, DR, Aariz, MetaDent, COde, DenPAR, and DentalCaries, across four imaging modalities, four task types, 11 sub-tasks, and 5,101 test items. Each panel shows a representative prompt, image modality, and ground-truth target, illustrating how heterogeneous dental resources are converted into standardized vision–language tasks including VQA, classification, captioning, and tooth counting.

as summarized in Figure 1. We evaluate 13 VLMs in zero-shot and few-shot settings, fine-tune open-weight models under a uniform LoRA-SFT training budget, and deployed three best performing fine-tuned compact VLMs on a consumer smartphone to measure latency and memory. We also invited dentists to review the model's predictions and reasoning and assess how closely they aligned with their clinical interpretations. Figure 2 shows the pipeline from data preparation and task design to fine-tuning and mobile deployment.

Our main contributions are:

- We introduce Pocket-Dentist, a multimodal dental vision–language benchmark that unifies seven previously isolated datasets, covering more than 5,900 patients, 50,800 images, four imaging modalities, and four task types. Under a shared interface, Pocket-Dentist enables comparison across heterogeneous dental modalities and clinical output formats.

- We propose a deployment-aware evaluation framework that jointly assesses dental VLMs under zero-shot, few-shot, and fine-tuning settings, together with mobile inference efficiency. We further deploy Pocket-Dentist-4B (LoRA-fine-tuned Qwen3.5-4B, the best fine-tuned model in our benchmark) on a consumer smartphone for fully local inference. We also invited dentists to review the model's predictions and reasoning and assess how closely they aligned with their clinical interpretations.

- Using this framework, we identify two empirical findings that are difficult to uncover through single-dataset evaluations. First, under the same dental LoRA-SFT budget, prior medical-domain specialization does not provide a consistent advantage. Second, larger parameter scale does not necessarily translate into better performance after dental-domain fine-tuning.

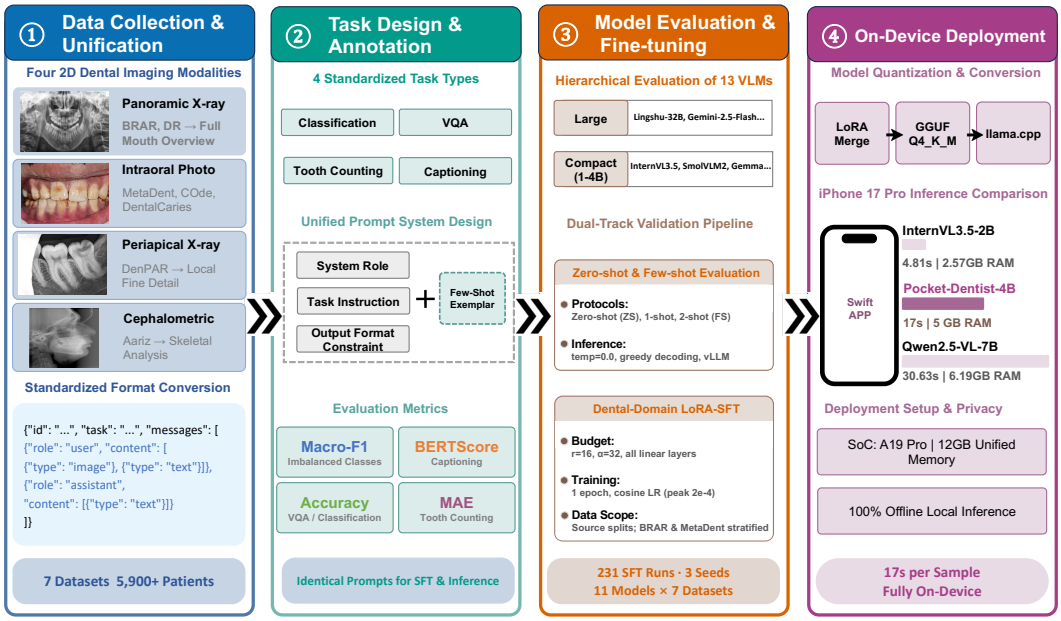


Figure 2: Deployment-aware evaluation pipeline of Pocket-Dentist. The pipeline consists of four stages: (1) curating heterogeneous dental datasets into a unified multimodal benchmark; (2) converting source annotations into shared prompt–response task formats; (3) evaluating VLMs under zero-shot, few-shot, and LoRA-SFT settings; and (4) measuring on-device latency and memory of three LoRA-fine-tuned, quantized models on a consumer smartphone.

2 Related Work

Medical VLMs. Medical vision–language models have progressed from task-specific medical VQA datasets and benchmarks for medical visual question answering (He et al., 2020; Zhang et al., 2023) to more general-purpose medical vision–language models such as LLaVA-Med (Li et al., 2023) and Med-Flamingo (Moor et al., 2023) that leverage large-scale medical image–text pretraining. Recent work in medical AI has also emphasized clinical alignment, safety, and human evaluation as prerequisites for reliable deployment (Singhal et al., 2023). However, medical VLM evaluation remains focused on radiology and pathology, with limited coverage of dental imaging.

Dental AI benchmarks. Recent dental AI research on caries has mainly focused on task-specific image-based models (Noor Uddin et al., 2025), while recent clinical evaluation has noted the potential of multimodal integration to support more comprehensive diagnosis (Zhang et al., 2024). MetaDent (Li et al., 2026) introduced a dental imaging benchmark covering VQA, captioning, and classification from expert-annotated intraoral photographs, while BRAR (Xia et al., 2025) contributed panoramic radiograph annotations for periodontal bone loss grading. However, prior work has not unified heterogeneous dental datasets into a deploy-aware VLM benchmark spanning multiple modalities, task types, prompting and fine-tuning settings, and deployment efficiency.

On-device deployment. Recent work has increasingly studied language models under mobile and edge constraints, including benchmark-level analyses of model capability, latency, and memory (Lu et al., 2025; Wang et al., 2025b). In healthcare, HealthSLM-Bench (Wang et al., 2025b) further evaluated small language models on wearable health tasks together with smartphone deployment efficiency. Extending this direction to multimodal models, LQA (Wang et al., 2026) proposed a lightweight quantized-adaptive framework for deploying VLMs on edge devices. However, existing work largely focuses on text-based health applications or general-domain VLM deployment, leaving the deployment behavior of dental VLMs underexplored. Pocket-Dentist brings deploy-aware evaluation to dental VLMs by jointly assessing dental task performance, fine-tuning behavior and smartphone inference within a unified benchmark.

3 POCKET-DENTIST

3.1 DATA CURATION

Pocket-Dentist brings together seven public available dental health datasets, summarized in Figure 1. The datasets cover common 2D dental imaging modalities, including panoramic, periapical, and cephalometric radiographs, as well as intraoral photographs. The annotations of these datasets provide references that can be posed as prompts and verified automatically. Together, the datasets cover four imaging modalities and four task types. Eleven of the 12 evaluations use reference labels (i.e., refers to model predictions) released with or verified by the source datasets. We also invited dentists to review the model's predictions and reasoning and assess how closely they aligned with their clinical interpretations (see §5.5). Detailed datasets information, license, tasks description and data split are discussed below:

- **COde** (Lv et al., 2026) (CC BY 4.0) is the largest dataset in the benchmark by patient scale, comprising data from 4,800 patients aged 10–90 years with 8,775 dental checkups collected over eight years of longitudinal follow-up, totaling approximately 50,000 intraoral photographs and 8,056 radiographs (panoramic, periapical, and CBCT-derived 2D slices). We use 6,349 / 703 / 600 train / val / test check-ups for pathology classification (Macro F1).
- **MetaDent** (Li et al., 2026) (CC BY-NC 4.0) is a large-scale dental image resource with 60,669 images from clinical, public, and web sources. We use its expert-curated subset of 2,588 dental images, each annotated with a semi-structured meta-label describing imaging perspective, anatomical focus, and observed abnormalities. The three sub-tasks are VQA, multi-label classification (18 clinical finding categories), and clinical captioning, yielding 7,238 / 906 / 2,301 train / val / test items. MetaDent is the only dataset in the benchmark that simultaneously covers VQA, generative captioning, and classification.
- **BRAR** (Xia et al., 2025) (CC BY 4.0) comprises 1,104 panoramic radiographs from 1,104 patients collected at Shanghai Stomatological Hospital (adults aged 18–75). Bone resorption severity is graded into three levels (Grade 1/2/3) using the age-normalized BRAR index introduced with that release, with annotations verified by experienced periodontists. The public release contains 988 of these radiographs, all of which we use (691 / 148 / 149 train / val / test, stratified by grade).
- **Aariz** (Khalid et al., 2025) (CC BY 4.0) provides 1,000 lateral cephalometric radiographs, each annotated by clinical experts with 29 cephalometric landmarks. We use the official 700/150/150 split of all 1,000 radiographs and derive cephalometric VQA (750 test items: five question types on 150 radiographs).
- **DenPAR** (Rasnayaka et al., 2025) (CC BY 4.0) contains 1,000 periapical radiographs, each accompanied by metadata describing arch type, anatomical orientation, and per-tooth FDI notation. We derive three sub-tasks: arch classification (upper/lower), site classification (anterior/left/right), and tooth counting (200 test images $\times$ 3 sub-tasks).
- **DentalCaries** (Faizan Ahmed et al., 2025) (CC BY 4.0) reports 6,313 intraoral photographs; the release we obtained holds 6,266 (5,011 / 627 / 628 train / val / test, official split), used for binary caries presence classification (628 test).
- **DR** (IMT Kaggle Team, 2023) (CC BY-SA 4.0) is a community-contributed collection of 1,269 panoramic X-rays for multi-label finding classification across four categories (1,075/121/73 train/val/test).

3.2 DATA PROCESSING AND TASK FORMULATION

We convert all seven datasets into task-specific image–prompt pairs with text responses. This common format supports four task types (VQA, classification, captioning, and tooth counting) and allows us to fine-tune the models using the same LoRA-SFT protocol. Details of task processing and formulation are provided in Appendix A.2, and evaluation metrics are defined in Appendix A.3.

3.3 Training and Decoding Protocol

Let each benchmark task be represented as a set of image–prompt–target triples $\mathcal{D}_t = \{(x_i, p_i, y_i)\}_{i=1}^{n_t}$, where x_i is a dental image, p_i is the structured task prompt, and y_i is the reference output. For a VLM with parameters θ, zero-shot decoding predicts

$$\hat{y}_i^{\mathrm{ZS}} = \arg\max_y p_\theta(y \mid x_i, p_i). \tag{1}$$

Few-shot evaluation augments the query with a fixed exemplar set $\mathcal{E}_k = \{(x_j, p_j, y_j)\}_{j=1}^{k}$, shared across models for each task:

$$\hat{y}_i^{\mathrm{FS}} = \arg\max_y p_\theta(y \mid \mathcal{E}_k, x_i, p_i), \qquad k \in \{1, 2\}. \tag{2}$$

For LoRA instruction tuning, only the low-rank adapter parameters ϕ are optimized while the base model parameters θ remain frozen:

$$\phi^* = \arg\min_\phi \sum_t \sum_{(x_i, p_i, y_i) \in \mathcal{D}_t} -\log p_{\theta+\Delta_\phi}(y_i \mid x_i, p_i). \tag{3}$$

The fine-tuned model then predicts $\hat{y}_i^{\mathrm{SFT}} = \arg\max_y p_{\theta+\Delta_{\phi^*}}(y \mid x_i, p_i)$.

Models are evaluated under three task-adaptation settings. Zero-shot and few-shot evaluations cover all 13 models, while LoRA instruction tuning covers the 11 open-weight models. In the zero-shot (ZS) setting, models receive structured prompts without task-specific training and use greedy decoding for deterministic predictions. In the few-shot (FS) setting, following Brown et al. (2020), we evaluate 1-shot and 2-shot in-context learning using fixed exemplars shared across models. We report 1-shot and 2-shot results in Table 2 and Table 5 (Appendix A.1), respectively. We limit evaluation to two shots because the context window must accommodate both exemplar images and the query image. Further, HealthSLM-Bench (Wang et al., 2025b) found that additional examples do not consistently improve performance across health-domain tasks, a pattern we also observe in Section 5.1.

4 Experimental Setup

4.1 Models

We benchmark 13 vision–language models in two tiers: (i) large-scale VLMs ($\geq$7B), including three open-weight models, two of which are medical-domain-specialized models (Lingshu-32B and MedMO-8B-Next) and one general-purpose (Qwen2.5-VL-7B (Bai et al., 2025)), and two closed-source APIs (Gemini-2.5-Flash, Gemini-3.1-Flash-Lite); and (ii) compact VLMs ($\leq$4B), including seven general-purpose models (Qwen3.5-4B, Qwen3-VL-4B, gemma-4-E4B-it (Google DeepMind, 2026; Gemma Team et al., 2024), SmolVLM2-2.2B (Marafioti et al., 2025), InternVL3.5-2B (Wang et al., 2025a), gemma-4-E2B-it, InternVL3.5-1B) and one medical-domain-specialized model (medgemma-4b-it). This model pool is chosen to support deployment-aware comparison across parameter scale, access type, and prior domain specialization. Training details and hyperparameters are listed in Appendix A.4, and the training loss of all 11 fine-tuned models are shown in Figure 4.

4.2 On-Device Deployment Setup

To assess systems-level feasibility for on-device deployment, we deploy three of the evaluated models on an iPhone 17 Pro (A19 Pro SoC, 12 GB Unified Memory) to quantify the efficiency–capability frontier: Pocket-Dentist-4B (Qwen3.5-4B + LoRA, the best fine-tuned model of Table 3), InternVL3.5-2B + LoRA (the most compact deployable model), and Qwen2.5-VL-7B + LoRA (7B baseline). All three are the LoRA-merged checkpoints of Section 3.3 from the same LoRA-SFT protocol, so the three rows of Table 4 differ only in the base model. The iPhone 17 Pro was selected because it is the lowest-cost Apple device whose 12 GB Unified Memory can accommodate the 7B baseline (6.19 GB).

All three models are stored in GPT-Generated Unified Format (GGUF) and use Q4_K_M post-training quantization (approximately 4-bit). Inference is performed with llama.cpp (Gerganov,

Table 1: Performance of large-scale and compact VLMs under zero-shot (ZS) setting across all seven datasets. **GRA / GRF**: bone resorption grading (accuracy / macro-F1); **FND**: finding classification; **VQA**: visual question answering; **CAP**: captioning; **PTH**: pathology classification; **JAW**: upper vs. lower jaw; **LOC**: imaged region; **CNT**: tooth counting (mean absolute error, in teeth); **CAR**: caries presence. Best result is **bold**, second-best is underlined; Ranks use three-decimal precision; equal displayed values may receive different marks.

	Model	BRAR		DR	MetaDent			Aariz	COde	DenPAR			DentalCaries
		GRA (↑)	GRF (↑)	FND (↑)	VQA (↑)	CAP (↑)	FND (↑)	VQA (↑)	PTH (↑)	JAW (↑)	LOC (↑)	CNT (↓)	CAR (↑)
Large VLMs (ZS)	Lingshu-32B	0.49	0.39	0.50	**0.69**	0.18	0.33	0.32	0.48	0.59	**0.55**	0.88	0.56
	MedMO-8B-Next	0.23	0.18	0.44	0.19	0.06	0.06	0.30	0.27	0.61	0.26	2.90	0.59
	Qwen2.5-VL-7B	0.27	0.17	0.39	0.63	0.15	0.18	0.28	0.50	0.40	0.35	1.01	**0.63**
	gemini-2.5-flash	0.21	0.23	0.55	0.65	0.14	0.26	0.29	**0.58**	**0.99**	0.53	0.47	0.54
	gemini-3.1-flash-lite	0.43	**0.41**	**0.67**	0.67	**0.20**	**0.45**	0.30	0.51	**0.99**	0.54	0.50	0.47
Compact VLMs (ZS)	Qwen3.5-4B	0.26	0.25	0.45	0.67	0.19	0.33	0.28	0.51	0.49	0.50	0.43	0.40
	Qwen3-VL-4B	0.44	0.37	0.32	0.61	**0.20**	0.29	0.30	0.54	0.44	0.19	**0.42**	0.63
	gemma-4-E4B-it	0.56	0.24	0.39	0.60	0.18	0.25	0.33	0.51	0.40	0.52	0.52	0.43
	medgemma-4b-it	0.44	0.33	0.49	0.59	0.14	0.13	**0.44**	0.27	0.40	0.20	0.89	0.52
	SmolVLM2-2.2B	**0.56**	0.35	0.53	0.49	0.10	0.16	0.30	0.09	0.60	0.13	0.89	0.44
	InternVL3.5-2B	0.53	0.24	0.02	0.50	0.15	0.19	0.38	0.35	0.40	0.19	4.36	0.38
	gemma-4-E2B-it	**0.56**	0.24	0.25	0.58	0.15	0.19	0.37	0.50	0.27	0.19	0.73	0.61
	InternVL3.5-1B	0.16	0.19	0.55	0.54	0.12	0.15	0.39	0.26	0.60	0.30	1.38	0.61

Table 2: Performance under 1-shot (FS-1) setting across all seven datasets. Abbreviations are as in Table 1. Best result is **bold**, second-best is underlined; Ranks use three-decimal precision; equal displayed values may receive different marks.

	Model	BRAR		DR	MetaDent			Aariz	COde	DenPAR			DentalCaries
		GRA (↑)	GRF (↑)	FND (↑)	VQA (↑)	CAP (↑)	FND (↑)	VQA (↑)	PTH (↑)	JAW (↑)	LOC (↑)	CNT (↓)	CAR (↑)
Large VLMs (FS-1)	Lingshu-32B	0.20	0.13	0.54	0.62	0.18	0.09	0.40	0.52	0.66	0.18	0.89	0.39
	MedMO-8B-Next	0.17	0.10	0.53	0.55	0.13	0.03	0.27	0.27	0.60	0.17	1.27	0.58
	Qwen2.5-VL-7B	0.30	0.23	0.26	0.56	0.17	0.18	0.33	0.49	0.40	0.29	1.18	0.58
	gemini-2.5-flash	0.29	0.28	0.56	0.67	0.18	0.33	0.31	**0.54**	0.96	**0.60**	0.42	0.55
	gemini-3.1-flash-lite	0.40	**0.37**	**0.59**	**0.67**	0.21	**0.46**	0.35	0.47	**1.00**	0.59	0.46	**0.62**
Compact VLMs (FS-1)	Qwen3.5-4B	0.34	0.28	0.33	0.64	0.19	0.15	0.37	0.47	0.60	0.29	**0.38**	0.36
	Qwen3-VL-4B	0.20	0.13	0.46	0.61	**0.22**	0.28	0.40	0.51	0.60	0.33	0.64	0.51
	gemma-4-E4B-it	0.51	0.26	0.46	0.58	0.19	0.24	**0.41**	0.51	0.36	0.53	0.48	0.53
	medgemma-4b-it	0.17	0.10	0.50	0.57	0.17	0.14	**0.41**	0.28	0.60	0.24	0.90	0.47
	SmolVLM2-2.2B	0.25	0.18	0.53	0.51	0.13	0.14	0.30	0.06	0.60	0.15	0.89	0.36
	InternVL3.5-2B	0.55	0.24	0.40	0.57	0.19	0.17	0.37	0.25	0.39	0.17	1.76	0.51
	gemma-4-E2B-it	**0.57**	0.27	0.26	0.56	0.18	0.20	0.39	0.43	0.37	0.29	0.57	0.60
	InternVL3.5-1B	0.04	0.09	0.53	0.56	0.17	0.03	0.33	0.12	0.60	0.17	0.91	0.36

2023) through a native Swift package using the Metal GPU API. Models are sideloaded into the app's Documents directory and loaded into Unified Memory for hardware-accelerated inference. More deployment details can be founded at Appendix A.5.

5 RESULTS

5.1 ZERO-SHOT AND FEW-SHOT PERFORMANCE

Tables 1–3 present the full results across all 13 models, seven datasets, and three evaluation settings.

Zero-shot performance. We observe that no model is consistently best across the 12 evaluations (Table 1). The Gemini APIs lead in several areas, while both large and compact open-weight models achieve the highest scores elsewhere. Performance varies substantially by evaluation, with no consistent overall leader.

Few-shot performance. Moving from 1-shot (Table 2) to 2-shot (*cf.* Appendix A.1 Table 5), most models show marginal or mixed improvement. On several datasets, 2-shot performance is comparable to or lower than 1-shot performance. For example, gemma-4-E2B-it drops from 0.20 (FS-1) to 0.16 (FS-2) on MetaDent FND, while Qwen3-VL-4B drops from 0.46 to 0.39 on DR FND. This observation is consistent with HealthSLM-Bench (Wang et al., 2025b), which found that the benefits of additional in-context examples may vary by task.

Table 3: Performance under LoRA instruction tuning (SFT) across all seven datasets. Every cell is the seed-42 run; the three-seed mean and spread of the same cells are in Table 6. Abbreviations are as in Table 1. Best result is **bold**, second-best is underlined; Ranks use three-decimal precision; equal displayed values may receive different marks.

	Model	BRAR		DR	MetaDent			Aariz	COde	DenPAR			DentalCaries
		GRA (↑)	GRF (↑)	FND (↑)	VQA (↑)	CAP (↑)	FND (↑)	VQA (↑)	PTH (↑)	JAW (↑)	LOC (↑)	CNT (↓)	CAR (↑)
Large VLMs (LoRA)	Lingshu-32B	0.56	**0.44**	0.66	0.80	0.26	0.11	0.39	**0.81**	0.81	0.47	0.84	0.92
	MedMO-8B-Next	0.32	0.29	0.37	0.76	0.24	0	0.38	0.81	0.42	0.13	1.18	0.64
	Qwen2.5-VL-7B	0.42	0.38	0.57	0.78	0.26	0.10	0.38	0.80	0.99	0.71	0.30	0.91
Compact VLMs (LoRA)	Qwen3.5-4B	**0.59**	0.42	**0.67**	**0.80**	**0.27**	**0.37**	**0.45**	0.81	**1.00**	0.71	**0.21**	**0.94**
	Qwen3-VL-4B	0.56	0.43	0.64	0.75	0.27	0.34	0.38	0.79	**1.00**	**0.88**	0.25	0.93
	gemma-4-E4B-it	0.48	0.37	0.56	0.75	0.27	0.23	0.38	0.81	0.99	0.83	0.23	0.91
	medgemma-4b-it	0.48	0.39	0.62	0.74	0.26	0.27	0.38	0.80	0.98	0.57	0.44	0.91
	SmolVLM2-2.2B	0.37	0.32	0.49	0.65	0.26	0.13	0.38	0.62	**1.00**	0.64	0.33	0.92
	InternVL3.5-2B	0.51	0.42	0.60	0.66	0.26	0.26	0.38	0.80	**1.00**	0.74	0.27	0.93
	gemma-4-E2B-it	0.44	0.38	0.57	0.73	0.25	0.16	0.38	0.79	0.98	0.74	0.47	0.91
	InternVL3.5-1B	0.24	0.18	0.63	0.63	0.26	0.13	0.38	0.79	0.99	0.69	0.30	0.91

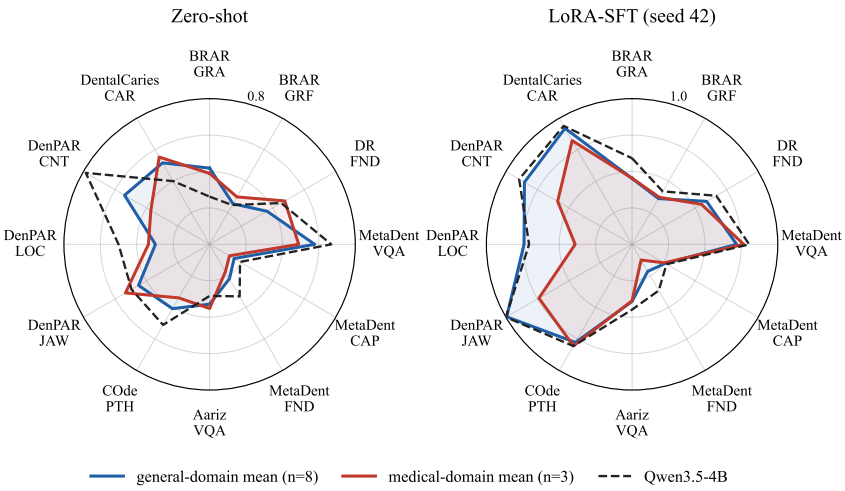


Figure 3: Per-column mean of the three medical-domain models (red) and the eight general-domain open-weight models (blue), with Qwen3.5-4B (dashed), zero-shot (left) and after LoRA instruction tuning (right, seed 42). DenPAR CNT is visualized as $1 - \mathrm{MAE}/2$.

5.2 LoRA Instruction Tuning

Effect of dental-domain fine-tuning. Under LoRA instruction tuning (see Table 3), compact VLMs improve substantially across the benchmark. Qwen3.5-4B achieves the strongest seed-42 results on several columns, including DR FND (0.67), MetaDent FND (0.37), DenPAR CNT (0.21), and DentalCaries CAR (0.94). Across the three training seeds, the main gains on the headline metrics remain consistent. A more detailed robustness analysis is provided in Appendix A.6.

Comparison with larger open-weight models. The comparison between compact and large models is stable across different training seeds. Specifically, Qwen3.5-4B varies by no more than 0.045 on any of the four headline metrics (DR FND, MetaDent CAP, MetaDent FND, and DentalCaries CAR), and by no more than 0.025 on three of them (Appendix A.6). Within each seed, none of the 12 paired bootstrap comparisons shows a significant advantage for a larger model after Holm correction. As shown in Table 3, Qwen3.5-4B scores higher than the strongest of the three larger open-weight models on 10 of the 12 sub-tasks. Lingshu-32B leads on the remaining two, BRAR GRF and COde PTH. Under the same low-cost LoRA budget, larger parameter scale offers no consistent advantage on the dental tasks evaluated here.

Role of prior medical-domain specialization. The three medical-domain models, medgemma-4b-it, Lingshu-32B, and MedMO-8B-Next, all perform worse than the two best fine-tuned models, Qwen3.5-4B and Qwen3-VL-4B. Lingshu-32B ranks first among open-weight models zero-shot, but that lead does not persist after uniform dental-domain fine-tuning. The mean performance of

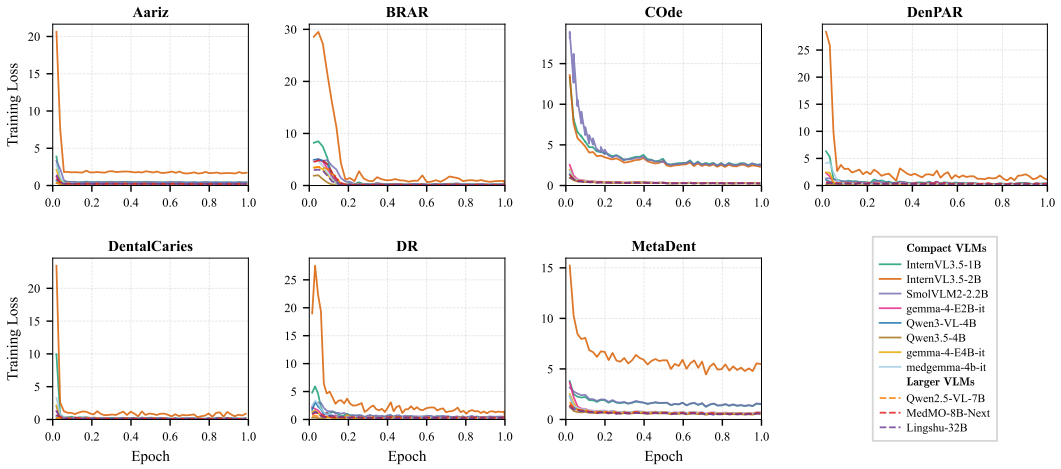


Figure 4: Training loss curves for LoRA instruction tuning on seven dental datasets for the 11 open-weight VLMs in Table 3. Solid and dashed lines denote compact ($\leq$4B) and larger ($\geq$7B) VLMs, respectively.

Table 4: On-device deployment efficiency of VLMs on iPhone 17 Pro (MetaDent dataset, N=10 per task $\times$ 3 tasks).

Dataset	Model	TTFT (s) ↓	ITPS (t/s) ↑	OET (s) ↓	OTPS (t/s) ↑	Total (s) ↓	CPU (%) ↓	RAM (GB) ↓
MetaDent	Pocket-Dentist-4B	1.80	209.02	15.20	9.97	17.00	47.92	5.09
	InternVL3.5-2B	**0.52**	**734.07**	**4.29**	**27.95**	**4.81**	62.25	**2.57**
	Qwen2.5-VL-7B	3.31	106.17	27.31	7.70	30.63	**44.83**	6.19

the medical-domain models exceeds that of the general-domain open-weight models on six of 12 sub-tasks zero-shot and four of 12 after fine-tuning (Figure 3).

5.3 On-Device Deployment

Table 4 reports the inference latency and memory usage of three models deployed on an iPhone 17 Pro (12 GB unified memory) using Metal-accelerated on-device inference (Appendix A.5). The goal is to assess whether benchmark-competitive compact VLMs can satisfy the latency and memory constraints of mobile inference. Compared with the fine-tuned 7B baseline, Pocket-Dentist-4B reduces per-sample latency from 30.63 s to 17.00 s and memory usage from 6.19 GB to 5.09 GB. InternVL3.5-2B is faster and more memory-efficient, requiring only 4.81 s and 2.57 GB, but achieves lower benchmark performance under the same LoRA-SFT setting. These results highlight a clear efficiency–capability trade-off, with Pocket-Dentist-4B providing a practical balance between dental-domain performance and mobile resource use.

As inference is performed entirely on device, patient images do not need to be uploaded to the cloud, and the system can operate without network connectivity. This reduces potential data exposure and helps preserve patient privacy. The on-device evaluation interface is shown in Figure 6 of Appendix A.5.

5.4 Illustrative Cases

Figure 5 shows three test-set cases in which the LoRA-fine-tuned Qwen3.5-4B is correct and the substantially larger Lingshu-32B is not. Asked whether prominent black triangles are visible between the teeth (MetaDent VQA; ground truth *true*), Lingshu-32B misses the finding. Asked what best describes the alignment of the right maxillary anterior teeth (MetaDent VQA; ground truth *misaligned*), it answers *well-aligned*, the opposite of the visible finding. Counting the visible teeth in a DenPAR periapical radiograph (ground truth 4), it answers 6 (5 and 5 under the two other seeds). The cases illustrate three residual error types after LoRA-SFT: false negatives for present findings, categorical answers that contradict the visible finding, and over-counting in tooth enumeration. Appendix A.7 shows one further case for each task type.

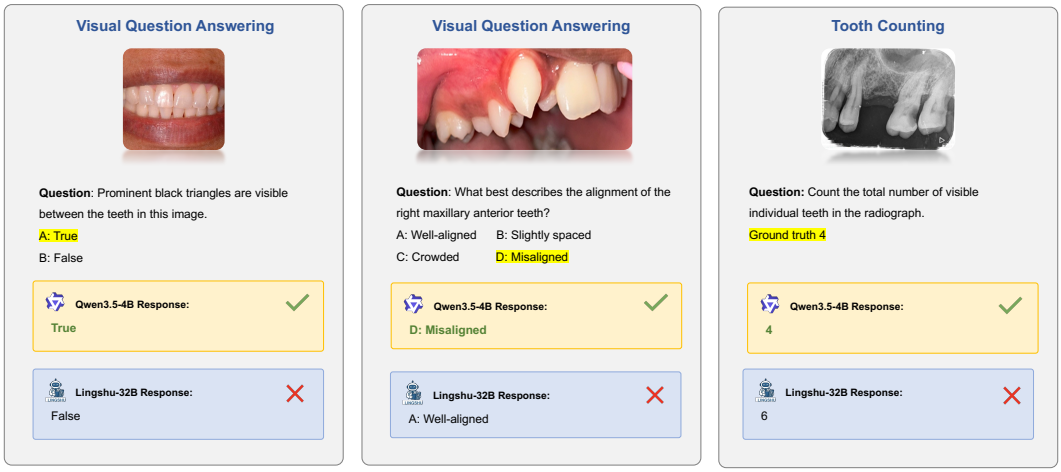


Figure 5: Illustrative cases, selected *post hoc* on outcome: each is a test-set case in which the LoRA-fine-tuned Qwen3.5-4B is correct and Lingshu-32B is incorrect under the formal (seed-42) run, and both outcomes are unchanged under the two additional training seeds (Appendix A.6). Because the selection rule conditions on the conclusion, these cases are not evidence of relative accuracy; the quantitative comparison is Table 3.

5.5 Dental Expert Evaluation

Since all curated datasets include reference labels, we can measure how closely generated text matches them. However, this does not establish whether a model interpreted the input image correctly or whether its answer is clinically appropriate. We therefore invited four dentists to review the images alongside the model outputs and assess whether the answers were supported by the visible findings. Further details of this assessment are provided in Appendix 5.5.

Results show that 11 of the 12 evaluations use labels released with or verified by their source datasets. Only MetaDent CAP lacked expert review, so we conducted a blinded evaluation of one fine-tuned model's outputs on this task. Four practising dentists assessed a stratified sample of 82 test outputs against the corresponding images. Among the 51 outputs in the highest BERTScore tertile, 76% (39/51) were judged correct or partially correct, including 27% (14/51, 95% CI [0.17, 0.41]) that were fully correct and usable in a clinical record without amendment. Across the full test set, the stratified estimate for outputs judged correct or usable after minor amendment was 58% (95% CI [0.46, 0.70]). These findings provide clinician-based evidence about the quality of the generated descriptions.

6 Limitations

Each model is fine-tuned and evaluated within a single dataset's splits; therefore, transfer across institutions, imaging devices, and datasets is not assessed. The benchmark is further limited to single-turn tasks on 2D images. Finally, on-device measurements assess efficiency only and are limited to a single phone and 30 MetaDent samples.

7 Conclusion

In conclusion, we introduced Pocket-Dentist, a deployment-aware benchmark that brings diverse dental imaging tasks into a common evaluation setting. It enables systematic comparison of model scale, prior medical specialization, and dental-domain fine-tuning, while connecting model capability to mobile deployment constraints. Our results show that low-cost dental-domain fine-tuning can make compact 4B VLMs competitive with substantially larger open-weight models, while prior medical-domain specialization provides no consistent advantage under the same LoRA-SFT budget. Together with our on-device prototype, these findings highlight the potential of compact, domain-fine-tuned VLMs for specialized applications under practical deployment constraints. Future work should extend this evaluation to broader hardware platforms and expert assessment, and to clinical validation.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript and to assist with code review. All AI-assisted text and code-review suggestions were checked and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work introduces no new patient data. All seven datasets are public releases used under their Creative Commons licences (stated with each dataset in Section 3.1). Reference labels are those released with each source dataset; task targets are derived from them by deterministic conversion rules, and we produced no clinical annotation and reviewed none. The benchmark is not a clinical instrument: every score in this paper measures agreement with a dataset's reference labels, and none establishes efficacy, safety, or fitness for diagnosis.

REPRODUCIBILITY STATEMENT

We plan to release the code used in this work upon acceptance, including scripts for converting source annotations into benchmark prompt–response pairs and data splits, prompts and few-shot exemplars, evaluation and scoring code, and LoRA training configurations. The seven source datasets will remain distributed through their original repositories under the licences described in Section 3.1. We hope that these resources will enable the community to verify our results and extend the benchmark to new models, tasks, and devices.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. arXiv preprint, 2025. https://doi.org/10.48550/arXiv.2502.13923.

Eduardo Bernabe, Wagner Marcenes, Rizwan Suliankatchi Abdulkader, Lucas Guimarães Abreu, Saira Afzal, Fadwa Naji Alhalaiqa, et al. Trends in the global, regional, and national burden of oral conditions from 1990 to 2021: A systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, 405(10482):897–910, March 2025. ISSN 01406736. doi: 10.1016/S0140-6736(24)02811-3.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Nips '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-7138-2954-6.

Syed Muhammad Faizan Ahmed, Muhammad Huzaifa Ghori, Aamna Khalid, Ayesha Nooruddin, Niha Adnan, Abhishek Lal, et al. Annotated intraoral image dataset for dental caries detection. *Scientific Data*, 12(1):1297, July 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05647-9.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, et al. Gemma: Open Models Based on Gemini Research and Technology. arXiv preprint, 2024. https://doi.org/10.48550/arXiv.2403.08295.

Georgi Gerganov. llama.cpp: Llm inference in C/C++. https://github.com/ggerganov/llama.cpp, 2023. GitHub repository.

Google DeepMind. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Last updated: 2026-04-17.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ Questions for Medical Visual Question Answering. arXiv preprint, 2020. https://doi.org/10.48550/arXiv.2003.10286.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

IMT Kaggle Team. Dental radiography dataset. Kaggle, 2023. URL https://www.kaggle.com/datasets/imtkaggleteam/dental-radiography.

Muhammad Anwaar Khalid, Kanwal Zulfiqar, Ulfat Bashir, Areeba Shaheen, Rida Iqbal, Zarnab Rizwan, et al. A Benchmark Dataset for Automatic Cephalometric Landmark Detection and CVM Stage Classification. *Scientific Data*, 12(1):1336, July 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05542-3.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. arXiv preprint, 2023. https://doi.org/10.48550/arXiv.2306.00890.

M.-X. Li, W.-H. Deng, Z.-X. Wu, C.-X. Jin, J.-M. Wu, Y. Han, et al. MetaDent: Labeling Clinical Images for Vision-Language Models in Dentistry. *Journal of Dental Research*, pp. 00220345261424242, March 2026. ISSN 0022-0345, 1544-0591. doi: 10.1177/00220345261424242.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Wei Liu, et al. Demystifying Small Language Models for Edge Deployment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14747–14764, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.718.

Haoxin Lv, Ijazul Haq, Jin Du, Jiaxin Ma, Binnian Zhu, Xiaobing Dang, Chaoan Liang, Ruxu Du, Yingjie Zhang, and Muhammad Saqib. A benchmark multimodal oro-dental dataset for large vision-language models. *Scientific Data*, 13(1):990, April 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07342-9.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, et al. SmolVLM: Redefining small and efficient multimodal models. arXiv preprint, 2025. https://doi.org/10.48550/arXiv.2504.05299.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, et al. Med-flamingo: A multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, December 2023.

Rithesh Murthy, Liangwei Yang, Juntao Tan, Tulika Manoj Awalgaonkar, Yilun Zhou, Shelby Heinecke, et al. MobileAIBench: Benchmarking LLMs and LMMs for On-Device Use Cases. arXiv preprint, 2024. https://doi.org/10.48550/arXiv.2406.10290.

Ayesha Noor Uddin, Syed Ahmed Ali, Abhishek Lal, Niha Adnan, Syed Muhammad Faizan Ahmed, and Fahad Umer. Applications of AI-based deep learning models for detecting dental caries on intraoral images – a systematic review. *Evidence-Based Dentistry*, 26(1):71–72, March 2025. ISSN 1462-0049, 1476-5446. doi: 10.1038/s41432-024-01089-1.

Sumudu Rasnayaka, Dhanushka Leuke Bandara, Amali Jayasundara, Ruwan Jayasinghe, Chathura Wimalasiri, Piumal Rathnayake, et al. DenPAR: Annotated Intra-Oral Periapical Radiographs Dataset for Machine Learning. *Scientific Data*, 12(1):1615, October 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05906-9.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06291-2.

S.E. Uribe, J. Issa, F. Sohrabniya, A. Denny, N.N. Kim, A.F. Dayo, et al. Publicly Available Dental Image Datasets for Artificial Intelligence. *Journal of Dental Research*, 103(13):1365–1374, December 2024. ISSN 0022-0345, 1544-0591. doi: 10.1177/00220345241272052.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint, 2025a. https://doi.org/10.48550/arXiv.2508.18265.

Xin Wang, Ting Dang, Xinyu Zhang, Vassilis Kostakos, Michael J. Witbrock, and Hong Jia. HealthSLM-Bench: Benchmarking Small Language Models for Mobile and Wearable Healthcare Monitoring. arXiv preprint, 2025b. https://doi.org/10.48550/arXiv.2509.07260.

Xin Wang, Hong Jia, Hualin Zhou, Sheng Guang Wang, Yu Zhang, Ting Dang, and Tao Gu. LQA: A Lightweight Quantized-Adaptive Framework for Vision-Language Models on the Edge. arXiv preprint, 2026. https://doi.org/10.48550/arXiv.2602.07849.

Yiru Xia, Zhi Li, Zhikai Lin, Shuai Wang, Yaqi Wang, and Yufeng Xie. BRAR-anchored multimodal dataset of panoramic radiographs for periodontal bone resorption grading. *Scientific Data*, 13(1): 89, December 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-06400-y.

Jing-Wen Zhang, Jie Fan, Fang-Bing Zhao, Bing'er Ma, Xiao-Qing Shen, and Yuan-Ming Geng. Diagnostic accuracy of artificial intelligence-assisted caries detection: A clinical evaluation. *BMC Oral Health*, 24(1):1095, September 2024. ISSN 1472-6831. doi: 10.1186/s12903-024-04847-w.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, et al. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv preprint, 2023. https://doi.org/10.48550/arXiv.2305.10415.

A Appendix

A.1 Performance under 2-shot (FS-2) setting

Table 5: Performance under 2-shot (FS-2) setting across all seven datasets. Abbreviations are as in Table 1. '–' denotes no valid prediction. Best result is **bold**, second-best is underlined; Ranks use three-decimal precision; equal displayed values may receive different marks.

	Model	BRAR		DR	MetaDent			Aariz	COde	DenPAR			DentalCaries
		GRA (↑)	GRF (↑)	FND (↑)	VQA (↑)	CAP (↑)	FND (↑)	VQA (↑)	PTH (↑)	JAW (↑)	LOC (↑)	CNT (↓)	CAR (↑)
Large VLMs (FS-2)	Lingshu-32B	0.38	0.32	0.53	0.64	0.20	0.10	0.35	0.53	0.84	0.54	0.72	0.43
	MedMO-8B-Next	**0.56**	0.24	0.53	0.56	0.08	0.03	0.21	0.37	0.60	0.21	1.25	**0.65**
	Qwen2.5-VL-7B	0.34	0.28	0.29	0.56	0.15	0.13	0.31	0.47	0.40	0.27	1.33	0.63
	gemini-2.5-flash	0.28	0.27	**0.63**	**0.67**	0.18	0.31	0.33	0.53	0.97	0.57	0.43	0.53
	gemini-3.1-flash-lite	0.35	0.35	0.59	**0.67**	**0.21**	**0.45**	0.27	0.49	**0.99**	**0.67**	0.43	0.60
Compact VLMs (FS-2)	Qwen3.5-4B	0.45	**0.40**	0.54	0.65	0.19	0.29	0.39	0.48	0.53	0.53	**0.34**	0.40
	Qwen3-VL-4B	0.31	0.30	0.39	0.59	0.21	0.29	0.30	**0.59**	0.61	0.43	0.57	0.59
	gemma-4-E4B-it	0.40	0.29	0.39	0.60	0.18	0.24	**0.40**	0.53	0.41	0.59	0.47	0.50
	medgemma-4b-it	**0.56**	0.36	0.53	0.55	0.17	0.10	0.28	0.38	0.39	0.19	1.62	0.50
	SmolVLM2-2.2B	**0.56**	0.24	0.53	0.51	0.11	0.16	0.30	–	0.60	0.24	0.89	0.56
	InternVL3.5-2B	0.53	0.29	0.54	0.59	0.19	0.20	0.29	0.16	0.40	0.19	1.80	0.49
	gemma-4-E2B-it	**0.56**	0.24	0.29	0.56	0.17	0.16	0.29	0.50	0.33	0.21	0.90	0.59
	InternVL3.5-1B	0.55	0.24	0.53	0.54	0.15	0.03	0.29	0.11	0.60	0.17	2.00	0.63

A.2 Data processing and task formulation

COde. Each sample presents multiple clinical images (intraoral photographs and radiographs) alongside patient information (chief complaint, present illness history). The task is pathology classification: the model identifies the primary oro-dental anomaly from six categories (Dental Caries, Gingivitis, Pulpitis, Class III Malocclusion, Tooth Loss, Tooth Structure Loss), evaluated via Macro F1.

MetaDent. Three task-specific prompt–response formats are instantiated from the expert-authored meta-labels: VQA (one image–question pair per sample), multi-label classification (18 clinical finding categories), and clinical captioning (free-text description) (Li et al., 2026).

BRAR. We formulate a multimodal radiograph-plus-metadata classification task: each prompt presents the panoramic image together with patient-level metadata (age, gender, missing teeth, implants, residual roots, functional tooth count) and requests a single severity grade as a JSON object. To ensure reliable evaluation, we normalize predictions using a deterministic parser.

For the remaining four datasets, we apply analogous prompt standardisation.

Aariz pairs each lateral cephalometric radiograph with a clinical VQA prompt, evaluated by exact-match Accuracy. The five VQA labels (skeletal class, growth pattern, maxilla and mandible position, incisor inclination) are derived deterministically from the senior orthodontists' landmark set of the release.

DR aggregates per-image CSV annotations—discarding the original bounding-box coordinates—to produce image-level multi-label classification targets across four finding categories (Cavity, Fillings, Impacted Tooth, Implant), evaluated via Macro F1. The prompt asks only for the names of the categories present, and evaluation compares the predicted label set with the reference set; no spatial coordinates are requested or scored.

DenPAR formulates three sub-tasks from periapical radiographs: jaw identification (maxillary vs. mandibular, Accuracy), anatomical site classification (Anterior / Left / Right, Macro F1), and tooth count estimation (MAE).

DentalCaries formulates caries presence classification from intraoral photographs (binary Yes/No, Accuracy).

A.3 Evaluation Metrics

This section provides formal definitions and clinical rationale for the task metrics in Tables 1–3 and Table 5; the deployment metrics are defined in Appendix A.5. Arrows in column headers indicate optimization direction: (↑) higher is better; (↓) lower is better.

BRAR Accuracy / F1 (GRA / GRF, ↑) evaluate three-class classification performance (Grade 1/2/3) for periodontal bone resorption grading in panoramic radiographs. Accuracy measures the proportion of correctly classified samples, while Macro F1 gives equal weight to each class and is therefore more informative under class imbalance.

DR Macro F1 (FND, ↑) evaluates image-level multi-label classification across four finding categories (Cavity, Fillings, Impacted Tooth, Implant) in panoramic radiographs. Macro F1 is the unweighted mean of the per-class F1 scores (label-wise, scikit-learn; a class absent from both prediction and reference contributes 0), so every finding category weighs the same regardless of its prevalence. In clinical practice, correctly identifying the types of findings present in a radiograph is a prerequisite for appropriate treatment planning.

VQA Accuracy (↑) denotes the proportion of correctly answered visual questions, computed as exact-match accuracy against the reference answers. In dental diagnostics, a higher VQA accuracy directly translates to more reliable triage recommendations—an incorrect answer (e.g., misidentifying a lesion as "normal") could delay critical treatment.

Captioning BERTScore F1 (CAP, ↑) calculates a semantic similarity score between generated captions and reference descriptions, computed via contextual embeddings. Unlike surface-level n-gram overlap (BLEU/ROUGE), BERTScore captures paraphrase equivalence, which is essential in clinical reporting where lexical variation is high but semantic fidelity is paramount. Higher scores indicate that the model's narrative more faithfully conveys the clinical findings present in the image.

Classification Macro F1 (FND, ↑) is the unweighted mean over the 18 clinical finding categories of the per-class F1. Macro F1 is preferred over raw accuracy, and over support-weighted F1, because the MetaDent label distribution is heavily long-tailed: a model that predicts only the majority class achieves high accuracy and a high support-weighted F1 but clinically dangerous low recall on rare pathologies (e.g., periapical abscess); giving every class the same weight makes that recall visible.

Aariz VQA Accuracy (↑) evaluates the Aariz cephalometric dataset and measures exact-match correctness on clinical questions about skeletal class, growth pattern, maxilla / mandible position, and incisor inclination from lateral cephalometric radiographs. The table value is computed over all 750 items (five question types on 150 radiographs); the per-type accuracies are in the released metric files. The reference labels are derived from the source landmark annotations.

COde Classification Macro F1 (PTH, ↑) evaluates pathology classification on the COde dataset, where each item is one check-up with its images, as the unweighted mean of the per-class F1 over the six pathology categories.

DenPAR Arch Accuracy (JAW, ↑) measures exact-match classification of the imaged jaw (upper vs. lower; two classes). The column is saturated: always answering *Lower* scores 0.605 and nine of the eleven fine-tuned models reach 0.96–1.00, so it is excluded from the headline comparisons. DenPAR Site Macro F1 (LOC, ↑) evaluates anatomical site identification as the unweighted mean of the per-class F1 over the three sites. DenPAR Tooth Count MAE (CNT, ↓) measures the mean absolute error between the predicted and ground-truth tooth count per image; lower values indicate more accurate counting.

DentalCaries Presence Accuracy (CAR, ↑) evaluates binary caries presence classification (present vs. absent) on intraoral photographs, computed as exact-match accuracy against expert annotations. The source release carries bounding boxes, but the benchmark neither prompts for nor scores them, so the column measures finding presence, not detection.

What each column can and cannot separate. Two columns need a reading note. DenPAR JAW is a two-class task on which 7 of the 11 fine-tuned models reach at least 0.99; it is saturated and takes no part in any capability claim. MetaDent CAP is compressed: after LoRA-SFT all 11 cells lie between 0.24 and 0.27, so the column separates caption style from image content only weakly.

A.4 HYPERPARAMETERS

For LoRA instruction tuning (SFT), we apply LoRA (Hu et al., 2022) with rank r=16 and α=32, targeting all linear layers. Training runs for one epoch with cosine learning rate scheduling (peak lr=2×10^{-4}, warmup ratio 0.1), effective batch size 16, and gradient checkpointing. All task-specific LoRA configurations share identical hyperparameters across all seven datasets, so differences across models reflect fine-tuning behavior under a controlled low-cost budget rather than task-specific tuning. Every reported SFT run is the checkpoint at the last step of that single epoch; no checkpoint is selected on validation loss. All tasks use the standard next-token prediction loss; the only class weighting in the training code is an inverse-frequency weight over the three BRAR grades, which is a no-op for every other dataset. Inference uses greedy decoding for every open-weight model; for the two Gemini APIs greedy decoding cannot be enforced; both are queried at temperature 0, Gemini-2.5-Flash with the provider's default thinking budget and Gemini-3.1-Flash-Lite at the lowest thinking level (`thinking_level=minimal`), a setting the provider does not guarantee to suppress reasoning tokens entirely, although its token accounting reported none in our runs. All experiments are conducted on NVIDIA H100 96 GB GPUs using PyTorch 2.x and the Hugging Face `transformers` library.

A.5 DEPLOYMENT SETUP AND METRICS

A.5.1 DEPLOYMENT FRAMEWORK AND PIPELINE

The on-device prototype is a native iOS application (Swift/SwiftUI) running on an iPhone 17 Pro (Apple A19 Pro SoC, 12 GB unified memory, iOS 27.0). All inference runs locally without network access. The pipeline consists of four stages.

Model preparation. Each LoRA-merged checkpoint is converted with the `llama.cpp` conversion tools into two GGUF files: the language model quantized to Q4_K_M (4-bit) and the vision encoder with its projector (`mmproj`) kept in FP16. The pair is copied to the app's Documents directory over USB.

Inference engine. The app integrates `llama.cpp` (build b8851) through the `LocalLLMClient` Swift package. The runtime memory-maps the GGUF files, offloads all transformer layers and the vision encoder to the GPU via Metal, and applies the chat template stored in the GGUF. We use a 4,096-token context, sampling with temperature 0.1, and at most 512 generated tokens.

Image input and prompting. In evaluation mode, the app processes the 30 bundled MetaDent test images individually using the same task prompts as in the server-side evaluation. In interactive mode the user selects an image from the photo library, the image is downscaled to at most 1,024 px on its longer side, and a free-form question is entered.

Output and logging. Generated tokens are streamed to the interface. For every sample the app records time-to-first-token, generation time, output throughput, CPU load, and peak resident memory, and exports them as JSON; Table 4 is computed from these files. The model is reloaded before each sample so that every measurement starts from an empty KV cache. The application source code is available at https://anonymous.4open.science/r/Pocket-Dentist-Benchmark-Deployment-Anonymous-7555.

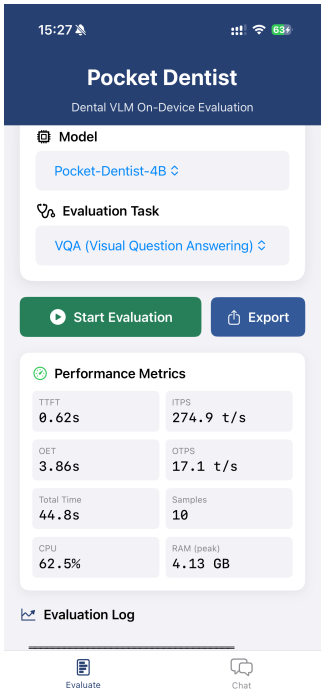


Figure 6: Pocket-Dentist iOS app running Pocket-Dentist-4B on an iPhone 17 Pro.

A.5.2 EFFICIENCY METRICS

Following MobileAIBench (Murthy et al., 2024), we adopt:

- TTFT (s): Time-to-first-token—latency from prompt submission to first output token.

- ITPS (t/s): Estimated input tokens per second. The app does not expose the prompt token count, so ITPS is approximated as the number of words in the text prompt $\times$ 1.3 divided by TTFT; image tokens are not included.
- OTPS (t/s): Output tokens per second—generation throughput.
- OET (s): Output evaluation time—wall-clock time for complete response generation.
- Total Time (s): End-to-end per-sample latency (TTFT + OET).
- CPU (%): Average CPU utilization during inference.
- RAM (GB): Peak memory allocation during model inference.

All values in Table 4 are means over the 30 MetaDent samples, pooled across the three tasks rather than reported per task (Wang et al., 2025b).

A.6 Seed Variance and Paired Tests

Training-Seed Robustness. Every SFT configuration was trained with three LoRA seeds (42 / 56 / 3407), and Table 6 reports the mean and range across the three runs. Across the headline metrics, the largest range across all models is 0.150, while Qwen3.5-4B varies by at most 0.045. The main tables report the original seed-42 runs; seeds 56 and 3407 are used only as robustness checks.

Paired-Comparison Robustness. Across all three training seeds, the compact Qwen3.5-4B is not significantly worse than the larger models. Table 7 compares Qwen3.5-4B with three large open-weight models using item-level paired bootstrap tests (B=2,000), with Holm correction over 12 comparisons per seed. The tests are performed separately for seeds 42, 56, and 3407. After correction, 8, 9, and 11 of the 12 comparisons significantly favour Qwen3.5-4B, respectively. These results show no consistent advantage from larger parameter scale under the same LoRA-SFT budget.

Table 6: LoRA SFT results over three seeds (42/56/3407). Each cell reports the three-seed mean, with the range across seeds (max − min) in parentheses. The four headline metrics are DR FND, MetaDent CAP, MetaDent FND and DentalCaries CAR. CNT is lower-is-better; abbreviations as in Table 1.

	BRAR		DR	MetaDent			Aariz	COde	DenPAR			DentalCaries
Model	GRA	GRF	FND	VQA	CAP	FND	VQA	PTH	JAW	LOC	CNT	CAR
Lingshu-32B	0.54 (0.05)	0.49 (0.08)	0.61 (0.08)	0.80 (0.02)	0.26 (0)	0.15 (0.07)	0.39 (0.01)	0.82 (0.01)	0.93 (0.19)	0.73 (0.41)	0.51 (0.51)	0.91 (0.02)
MedMO-8B-Next	0.30 (0.22)	0.24 (0.24)	0.32 (0.15)	0.76 (0.02)	0.24 (0.01)	0 (0)	0.39 (0)	0.81 (0.01)	0.47 (0.20)	0.17 (0.07)	1.14 (0.16)	0.64 (0)
Qwen2.5-VL-7B	0.35 (0.13)	0.34 (0.10)	0.54 (0.11)	0.78 (0)	0.26 (0)	0.12 (0.07)	0.39 (0.02)	0.80 (0)	0.99 (0.01)	0.74 (0.30)	0.33 (0.06)	0.91 (0)
Qwen3.5-4B	0.36 (0.42)	0.28 (0.32)	0.66 (0.03)	0.81 (0.01)	0.28 (0.01)	0.34 (0.04)	0.43 (0.04)	0.80 (0.02)	0.99 (0.02)	0.69 (0.31)	0.21 (0.01)	0.94 (0)
Qwen3-VL-4B	0.49 (0.21)	0.40 (0.08)	0.64 (0.03)	0.76 (0.02)	0.27 (0)	0.31 (0.07)	0.39 (0.01)	0.80 (0.02)	0.99 (0.03)	0.56 (0.63)	0.26 (0.02)	0.93 (0.01)
gemma-4-E4B-it	0.47 (0.02)	0.42 (0.09)	0.61 (0.08)	0.76 (0.02)	0.26 (0.01)	0.24 (0.01)	0.38 (0)	0.80 (0.04)	0.99 (0.02)	0.82 (0.01)	0.26 (0.05)	0.91 (0.01)
medgemma-4b-it	0.40 (0.38)	0.33 (0.39)	0.60 (0.05)	0.74 (0)	0.26 (0.01)	0.25 (0.03)	0.39 (0.03)	0.80 (0.01)	0.96 (0.07)	0.54 (0.05)	0.42 (0.08)	0.92 (0.01)
SmolVLM2-2.2B	0.38 (0.05)	0.34 (0.05)	0.46 (0.06)	0.68 (0.05)	0.26 (0)	0.14 (0.04)	0.39 (0)	0.63 (0.01)	1.00 (0.01)	0.71 (0.20)	0.31 (0.05)	0.92 (0)
InternVL3.5-2B	0.43 (0.19)	0.39 (0.18)	0.59 (0.04)	0.71 (0.09)	0.26 (0.01)	0.24 (0.03)	0.40 (0.06)	0.80 (0.02)	0.99 (0.03)	0.78 (0.07)	0.26 (0.03)	0.92 (0.02)
gemma-4-E2B-it	0.46 (0.06)	0.33 (0.08)	0.52 (0.07)	0.72 (0.01)	0.26 (0.01)	0.16 (0)	0.38 (0)	0.78 (0.03)	0.99 (0.01)	0.68 (0.14)	0.37 (0.18)	0.91 (0.01)
InternVL3.5-1B	0.24 (0.14)	0.17 (0.13)	0.60 (0.06)	0.67 (0.07)	0.25 (0.01)	0.16 (0.08)	0.40 (0.04)	0.79 (0.02)	0.99 (0)	0.74 (0.10)	0.33 (0.07)	0.91 (0.01)

Table 7: Paired bootstrap comparison of Qwen3.5-4B with the three large open-weight models on the four headline metrics, evaluated separately for each SFT seed (B=2,000). Δ is Qwen3.5-4B minus the comparison model; positive values favour Qwen3.5-4B. p_{Holm} is the Holm-adjusted two-sided bootstrap p-value over the 12 comparisons within each seed.

Metric	vs.	seed	Δ	95% CI	p_{Holm}
DR FND	Lingshu-32B	42	+0.011	[-0.096, +0.103]	0.787
DR FND	Lingshu-32B	56	+0.089	[+0.015, +0.157]	0.078
DR FND	Lingshu-32B	3407	+0.050	[-0.018, +0.116]	0.157
DR FND	MedMO-8B-Next	42	+0.298	[+0.212, +0.379]	0.012
DR FND	MedMO-8B-Next	56	+0.306	[+0.220, +0.382]	0.012
DR FND	MedMO-8B-Next	3407	+0.422	[+0.359, +0.471]	0.012
DR FND	Qwen2.5-VL-7B	42	+0.096	[+0.031, +0.162]	0.018
DR FND	Qwen2.5-VL-7B	56	+0.087	[+0.010, +0.162]	0.078
DR FND	Qwen2.5-VL-7B	3407	+0.173	[+0.098, +0.237]	0.012
MetaDent FND	Lingshu-32B	42	+0.259	[+0.201, +0.307]	0.012
MetaDent FND	Lingshu-32B	56	+0.150	[+0.108, +0.194]	0.012
MetaDent FND	Lingshu-32B	3407	+0.161	[+0.116, +0.214]	0.012
MetaDent FND	MedMO-8B-Next	42	+0.366	[+0.311, +0.402]	0.012
MetaDent FND	MedMO-8B-Next	56	+0.321	[+0.266, +0.361]	0.012
MetaDent FND	MedMO-8B-Next	3407	+0.337	[+0.283, +0.375]	0.012
MetaDent FND	Qwen2.5-VL-7B	42	+0.271	[+0.220, +0.321]	0.012
MetaDent FND	Qwen2.5-VL-7B	56	+0.160	[+0.110, +0.218]	0.012
MetaDent FND	Qwen2.5-VL-7B	3407	+0.228	[+0.175, +0.273]	0.012
MetaDent CAP	Lingshu-32B	42	+0.010	[-0.002, +0.024]	0.244
MetaDent CAP	Lingshu-32B	56	+0.010	[-0.002, +0.021]	0.092
MetaDent CAP	Lingshu-32B	3407	+0.018	[+0.010, +0.026]	0.012
MetaDent CAP	MedMO-8B-Next	42	+0.030	[+0.018, +0.040]	0.012
MetaDent CAP	MedMO-8B-Next	56	+0.036	[+0.024, +0.047]	0.012
MetaDent CAP	MedMO-8B-Next	3407	+0.050	[+0.041, +0.060]	0.012
MetaDent CAP	Qwen2.5-VL-7B	42	+0.014	[+0.002, +0.024]	0.084
MetaDent CAP	Qwen2.5-VL-7B	56	+0.020	[+0.008, +0.031]	0.012
MetaDent CAP	Qwen2.5-VL-7B	3407	+0.020	[+0.013, +0.028]	0.012
DentalCaries CAR	Lingshu-32B	42	+0.021	[+0.003, +0.040]	0.087
DentalCaries CAR	Lingshu-32B	56	+0.037	[+0.014, +0.057]	0.012
DentalCaries CAR	Lingshu-32B	3407	+0.038	[+0.019, +0.059]	0.012
DentalCaries CAR	MedMO-8B-Next	42	+0.299	[+0.258, +0.338]	0.012
DentalCaries CAR	MedMO-8B-Next	56	+0.299	[+0.256, +0.341]	0.012
DentalCaries CAR	MedMO-8B-Next	3407	+0.298	[+0.258, +0.339]	0.012
DentalCaries CAR	Qwen2.5-VL-7B	42	+0.032	[+0.014, +0.049]	0.018
DentalCaries CAR	Qwen2.5-VL-7B	56	+0.032	[+0.013, +0.049]	0.012
DentalCaries CAR	Qwen2.5-VL-7B	3407	+0.033	[+0.014, +0.054]	0.012

A.7 Extended Case Studies

This section shows further test-set cases, one per task type, comparing the LoRA-fine-tuned Qwen3.5-4B (Pocket-Dentist-4B) with Lingshu-32B fine-tuned under the same LoRA budget. As for Figure 5, the cases are selected post hoc on outcome from the test items on which Qwen3.5-4B gives the reference answer and Lingshu-32B does not under all three training seeds (42, 56, 3407); they illustrate failure modes and support no claim about relative accuracy. No dentist assessed these outputs.

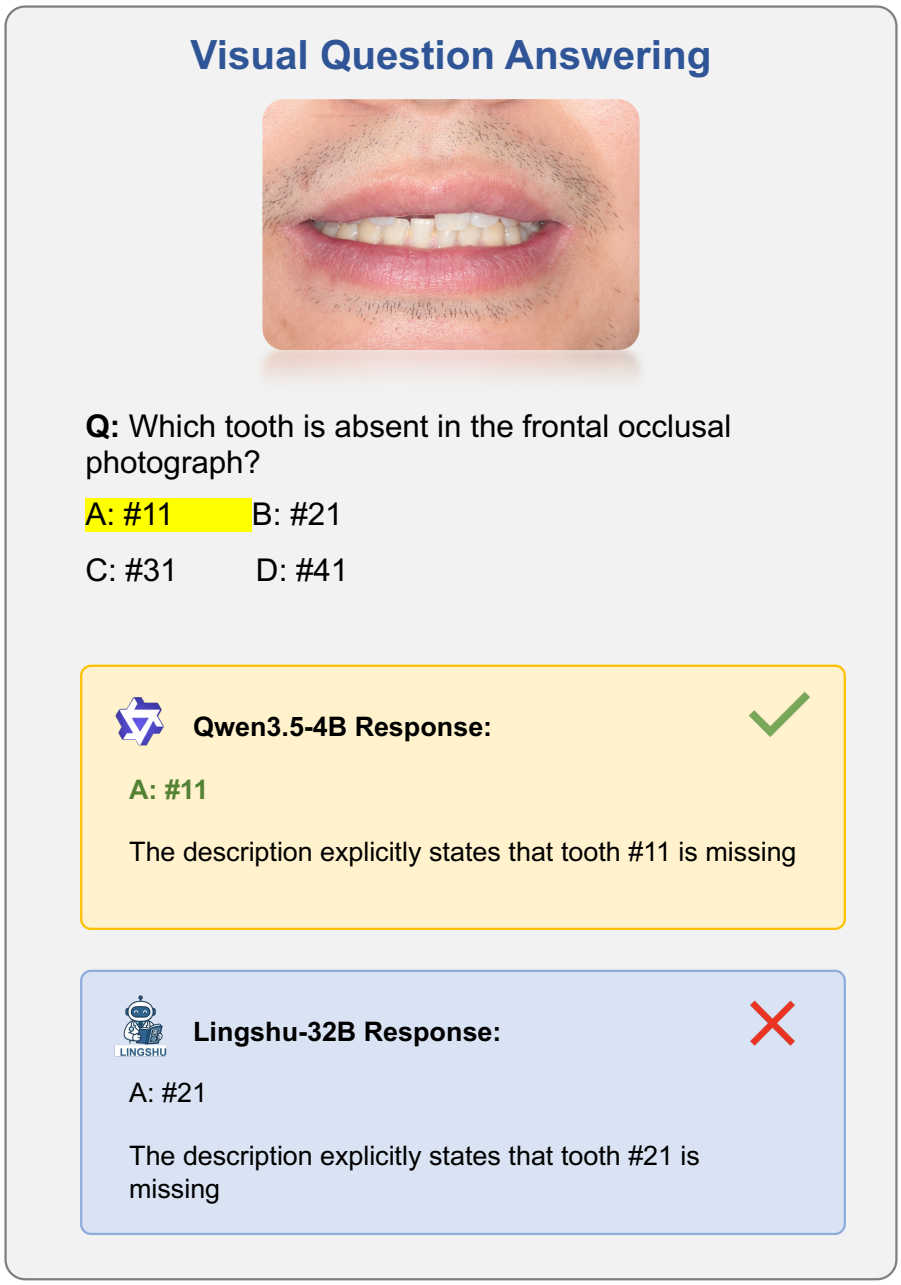


Figure 7: **Visual question answering (MetaDent).** Asked which tooth is absent in a frontal photograph, Lingshu-32B answers #21 and Qwen3.5-4B #11, the reference, under all three seeds.

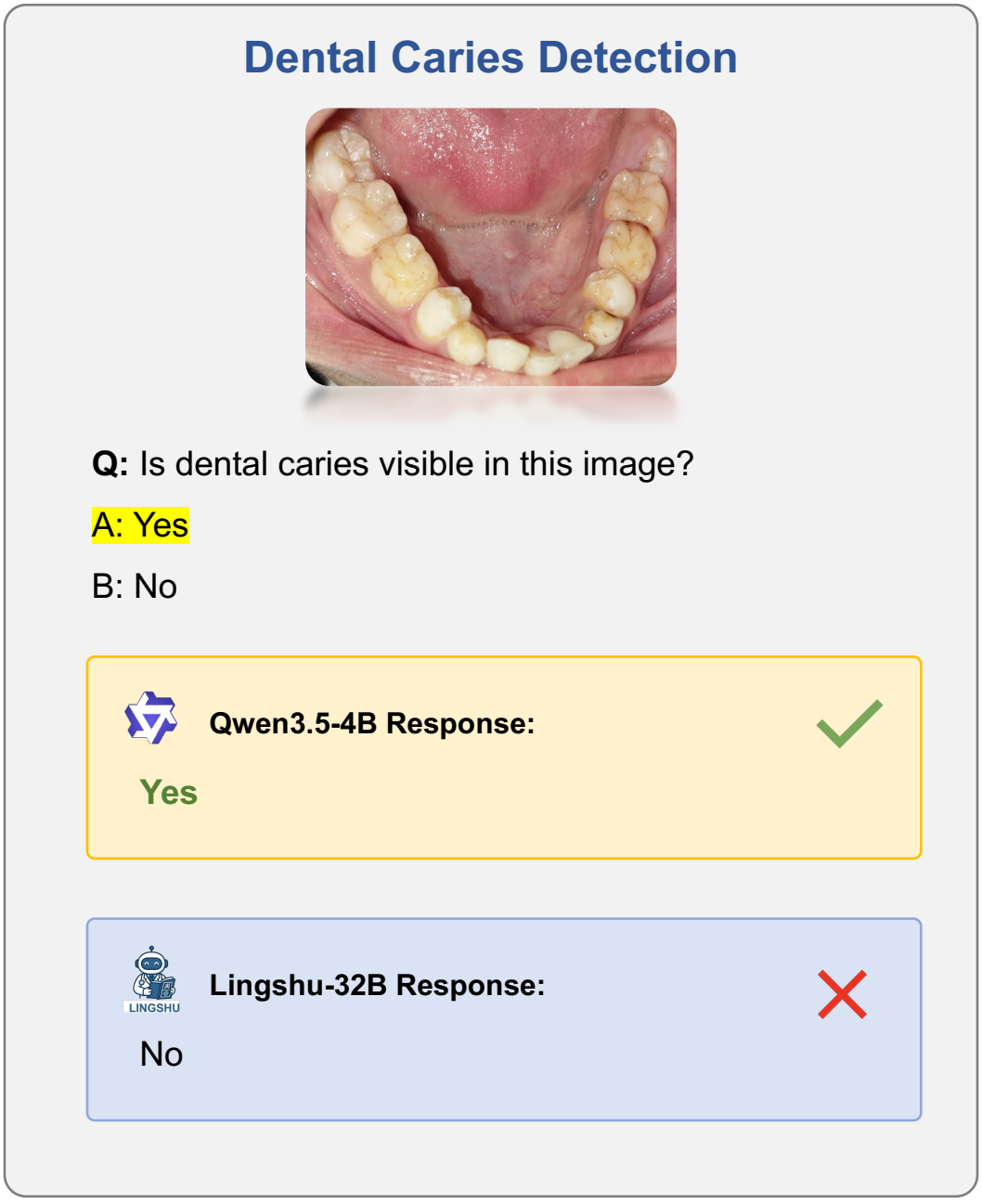


Figure 8: **Caries presence (DentalCaries).** A mandibular occlusal photograph with cavitated occlusal lesions on the molars on both sides; Lingshu-32B answers "No" and Qwen3.5-4B "Yes" under all three seeds.

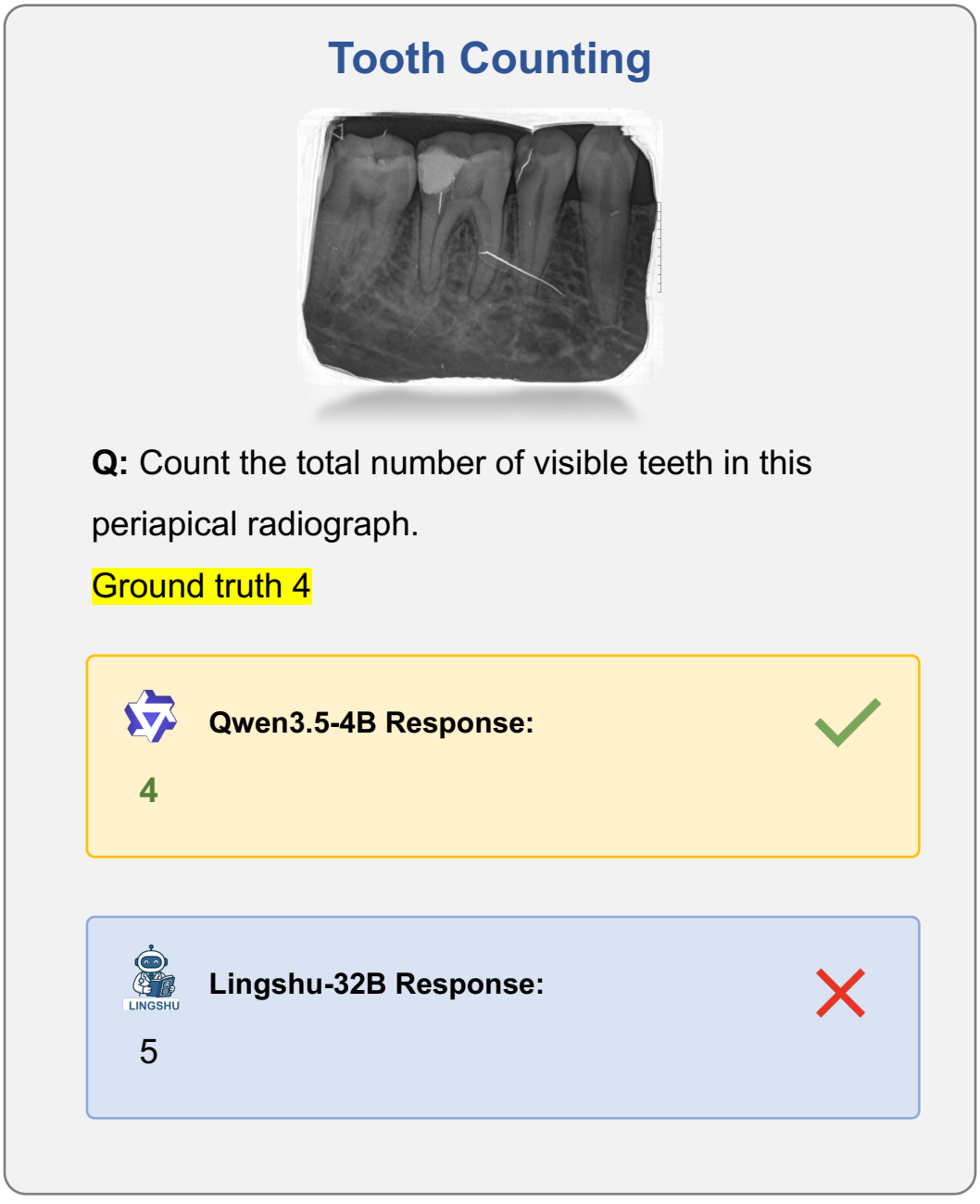


Figure 9: **Tooth counting (DenPAR).** Four teeth are visible in the periapical radiograph; Lingshu-32B counts five under all three seeds, Qwen3.5-4B four.

Clinical Image Captioning



Prompt:

Generate a detailed and vivid natural language description based on this clinical image. The description should clearly describe the imaging direction, the main subject of the image, and all observed abnormalities.

Qwen3.5-4B Response: ✓

The photograph presents a frontal intraoral view of the lower dental arch, captured with a mouth opener ... The lower incisors and canines are prominently displayed, with the central incisors (#31 and #41) and the lateral incisors (#32 and #42) showing a noticeable degree of crowding ... there is a distinct black discoloration of the gingiva in the region of the lower anterior segment ...

Lingshu-32B Response: ✗

The photograph presents a frontal view of the maxillary dentition, captured with the patient's mouth opened wide enough to expose the full complement of upper teeth ... the central incisors (#11 and #21) positioned symmetrically in the midline and the lateral incisors (#12 and #22) flanking them ... a noticeable area of black pigmentation on the labial gingiva adjacent to the right canine (#13) ...

Figure 10: **Clinical captioning (MetaDent).** The mandibular anterior teeth behind a mouth opener; Lingshu-32B describes the maxillary dentition and places the gingival pigmentation next to tooth #13, Qwen3.5-4B describes the lower arch. Excerpts; BERTScore F1 against the reference is higher for Qwen3.5-4B under all three seeds.