Do Multimodal LLMs Understand Intraoral Dental Data? Dataset, Platform, and Baselines

Luca Lumetti[1], Federico Rizzo[2], Francesca Cremonini[2], Ettore Candeloro[1], Luca Lombardo[2], Costantino Grana[1], and Federico Bolelli[1,✉]

[1] University of Modena and Reggio Emilia, Italy
[2] University of Ferrara, Italy

Abstract. Progress in dental computer vision is limited by the absence of large-scale multimodal datasets that jointly capture 3D intraoral geometry and 2D appearance across diverse clinical settings. Existing resources are typically unimodal, which hinders robust cross-modal learning and generalization. We assemble and release a multi-center dataset of 1,000 patients comprising 2,000 registered upper/lower intraoral scans, 5,000 paired intraoral photographs, and 2,403 clinician-authored reports. This combination links detailed 3D dental geometry with complementary 2D evidence, supporting occlusal and orthodontic analysis. Moreover, to enable scalable and privacy-preserving acquisition and annotation across distributed centers, we introduce an open platform that supports multimodal ingestion and structured labeling. Experiments indicate that state-of-the-art multimodal models fail to generate clinically faithful reports, motivating geometry-aware adaptation. We therefore propose IOS-Qwen, which fuses a PointTransformer 3D encoder with Qwen3-VL to generate structured, point-cloud-conditioned reports. Together, the dataset, the platform, and the baselines establish a foundation for multimodal dental AI research. Code is publicly released.[3]

Keywords: Multimodal Dental Dataset · Intraoral Scans · Multimodal LLMs

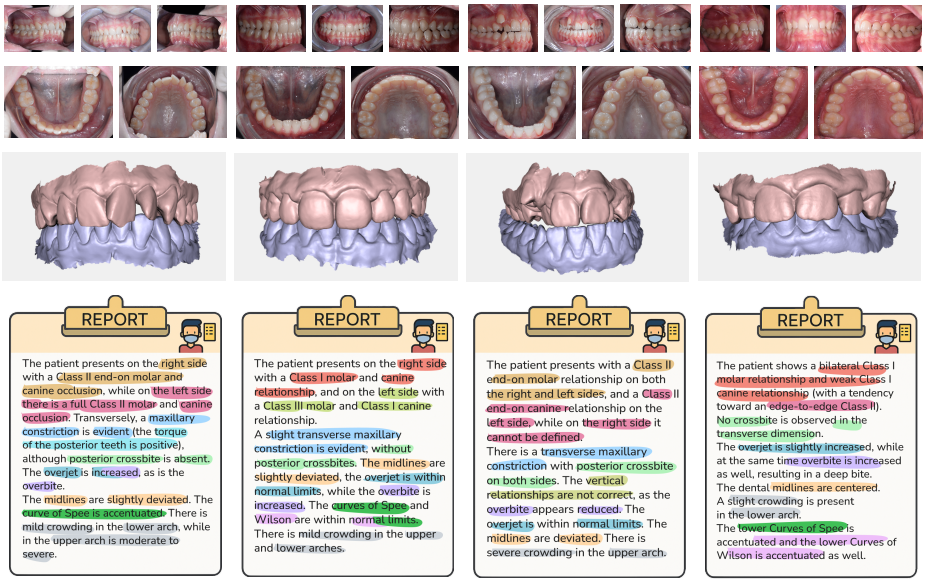
1 Introduction

Dentistry-focused computer vision is expanding rapidly, driven by the increasing availability of public datasets, benchmarks, and challenges [8–11, 18, 19, 21, 77, 80]. Nevertheless, it lags behind more established medical-imaging domains, such as chest X-ray [14, 37] and brain MRI [3, 4, 57, 62], where large-scale resources have enabled years of progress—including a growing body of work on report generation [54,61,70,79].

Orthodontic assessments rely on two complementary modalities: intraoral photographs (IOP) capture appearance (e.g., color and soft tissue), while intraoral scans (IOS) capture high-fidelity 3D geometry and occlusion. Yet, progress toward multimodal orthodontic report generation is hindered by the lack of data: large-scale datasets rarely provide patient-level pairs of IOP with occlusion-preserving, registered upper/lower IOS and clinician-authored reports. As a result, current models either learn from unimodal signals or from loosely aligned modalities, limiting cross-modal reasoning and generalization.

[3] https://github.com/AImageLab-zip/IOS-Report

✉ Corresponding author: federico.bolelli@unimore.it



(a) Patient 1979 **(b)** Patient 1982 **(c)** Patient 1983 **(d)** Patient 1992

Fig. 1: Multimodal intraoral samples from the proposed dataset. For each patient, from top to bottom, are displayed: *(i)* five RGB photos, representing right buccal, frontal, left buccal, lower occlusal, and upper occlusal intraoral photos, *(ii)* a screenshot from the 3D IOS with registered upper (pink) and lower (violet) scans, and *(iii)* the clinician-generated free-text report.

On the geometric side, intraoral 3D datasets such as Teeth3DS [7] and Poseidon3D [41] have recently enabled benchmarks for teeth/arch segmentation, mesh labeling, and morphology analysis. These collections typically provide surface meshes or point clouds acquired by professional scanners and are well-suited for geometry-based learning. However, they often have limited sample sizes, homogeneous clinical conditions, and sparse coverage of scanner vendors or acquisition settings; critically, they do not provide paired photometric information that reflects the real intraoral appearance.

Conversely, intraoral photographic datasets—e.g., AlphaDent [71], FDTooth [48], the Annotated Caries Dataset [26], and the Gingivitis Image Collection [25]—offer complementary RGB observations annotated for tooth boundaries or disease presence. Yet, by construction, they lack the geometric precision and spatial correspondence needed to reason about 3D anatomy; models trained solely on such 2D data struggle to generalize to tasks like occlusal analysis and cross-modal synthesis.

Only a few efforts begin to bridge modalities, e.g., [48] pairs photos with CBCT, but they do not include registered surface-level intraoral scans. To the best of our knowledge, there is no dataset that simultaneously provides registered upper and lower IOS together with corresponding intraoral photos and paired clinical reports, collected under multi-center, multi-vendor, and multi-protocol conditions.

We *assemble and release the first-of-its-kind multi-center dataset of 1,000 patients* that pairs occlusion-preserving, mutually registered upper/lower intraoral scans with a

standardized set of five intraoral photographs per patient, along with clinician-authored reports for a total of 2,000 IOS meshes, 5,000 intraoral photos, and 2,403 clinician-authored reports (Fig. 1). Unlike prior dental resources that are modality-specific or lack surface-level registration, our dataset provides a shared geometric frame for arch-level and occlusal reasoning while grounding photometric cues (color/soft tissue) in the corresponding 3D anatomy, enabling cross-modal learning and report generation from realistic orthodontic inputs. The dataset spans multiple centers and scanner vendors, and includes report text aligned to a clinically validated template, making it a reproducible benchmark for multimodal dental AI.

High-quality, multi-center datasets require more than ad-hoc aggregation: they demand a backbone for secure data upload, curation, and annotation. Automated pipelines streamline ingestion, de-identification, and pre-annotation, allowing clinicians to focus on expert labeling and meaningful reporting. Embedding these processes from the outset ensures both scale and consistency across sites [28, 76]. To address this gap and enable scalable, consistent multi-center acquisition and annotation of such multimodal orthodontic data, we developed YGGDRASIL,[4] a *web-based privacy-preserving infrastructure that ingests multimodal orthodontic data*, such as Intraoral Scans (IOS), Intraoral Photographs (IOP), Cone-Beam Computed Tomography (CBCT), Orthopantomography images (OPG), Periapical X-ray (PA), and Cephalometric radiographs (CEPH), and runs automatic modality-aware inference out-of-the-box: IOS multi-class segmentation and landmark detection [8], IOS/IOP occlusal classification [12], and CBCT multi-class segmentation pipelines [51]. Automatic predictions pre-populate findings to accelerate report drafting, and the interface supports speech-to-text (ASR) dictation for efficient authoring. The application is open-source,[5] supports collaborative labeling across globally distributed centers, and enforces privacy by design, i.e., de-identification at ingestion. The implementation is modality-agnostic and easy to extend to other domains (e.g., brain tumor, laparoscopy, or whole-slide histopathology).

Additionally, we *perform an extensive evaluation* showing that state-of-the-art general-purpose multimodal models struggle to produce clinically faithful reports on these data, underscoring the need for geometry-aware adaptation. We therefore present a simple modeling recipe, named IOS-Qwen, which augments the Qwen3-VL [1] family with a 3D point encoder: a PointTransformer module [81, 86] processes registered IOS point clouds into geometry tokens, which are fused with the vision-language backbone. It is trained to output template-conformant drafts with clinically meaningful slots (tooth- and arch-level fields, plus standardized narrative segments).

Paper Contributions. In summary, this paper contributes to the literature with *(i)* a multi-center dataset comprising 1,000 patients and pairing intraoral photographs, registered upper/lower IOS, and clinical reports, enabling multimodal supervision (Fig. 1); *(ii)* an open-source platform for multimodal data ingestion, integrated with deep learning pipelines, 2D and 3D viewers with labeling tools, and speech-to-text reporting, used to build our dataset; *(iii)* a systematic evaluation showing that state-of-the-art general-purpose multimodal models fail to generate clinically aligned reports on this

[4] Taking its name from the cosmic tree of Norse mythology, YGGDRASIL represents a unified platform that connects diverse research fields, datasets, users, and workflows.

[5] `https://github.com/AImageLab-zip/Yggdrasil`

Table 1: Summary of multimodal dental and oral imaging datasets combining multiple image modalities and/or audio/text. The table lists their key properties, including imaging modalities, public availability, license, and access type. We identify with the symbol "∼" datasets for which access is indirect and requires author validation/acceptance.

Dataset Name	Year	Country	Patients	Modalities	Public	License	Access Type
Tufts Dental Database [60]	2021	USA	Unknown	OPG (1,000) Eye-tracking (1,000) Think-aloud (1,000)	✓	Unknown Non-commercial	Direct Download
Multimodal Dental Dataset [33]	2024	China	169	CBCT (329) OPG (8) PA (16,203)	∼	PhysioNet License 1.5.0	Requires Review
3D Multimodal Dental Dataset [44]	2024	China	289	CBCT (289) IOS (289)	✓	CC BY 4.0	Direct Download
MMDental [77]	2025	China	660	CBCT (660) Reports (660)	✓	CC BY 4.0	Direct Download
FDTooth [48]	2025	China	241	CBCT (241) IOP (241)	✓	PhysioNet License 1.5.0	Requires Credentials
Ours	2026	Europe	1,000	IOS (2,000) IOP (5,000) Reports (2,403)	✓	CC BY-SA 4.0	Requires Credentials

distribution, quantifying the current literature gap and motivating geometry-aware methods; and *(iv)* a lightweight fusion of a PointTransformer with Qwen3-VL, yielding point cloud-conditioned, template-guided draft report generation for orthodontics.

2 Related Work

In this section, four strands of prior work are surveyed: *(i)* general vision-language foundations, *(ii)* medical adaptations, *(iii)* collection and annotation platforms, and *(iv)* datasets for intraoral imaging. Existing representative methods and evaluation protocols are outlined, and the proposed dataset and platform are positioned in relation to these.

Vision-Language Models. Self- and semi-supervised learning on web-scale image-text corpora has enabled multimodal foundation models that couple vision with language for cross-modal reasoning. Among existing models, CLIP aligns images and text via contrastive learning for broad zero-shot transfer [64], while vision-only DINO/DINOv2 [15, 58] learn strong representations. SAM [39] provides promptable zero-shot segmentation on ViT backbones [24], with SAM2 [66] adding video memory. In parallel, scalable LLMs—GPT-3 [13], the recent LLaMA family [30, 72, 73], and PaLM [17]—offer strong in-context and instruction-following capabilities.

Coupling vision encoders with LLMs yields vision-language models (VLMs) for captioning, visual question answering (VQA), retrieval, and visual reasoning. LLaVA [47] and LLaVA-1.5 [46] connect CLIP to Vicuna [16] via an MLP projector with visual-instruction tuning; LLaVA-NeXT [43] further incorporates dynamic high-resolution inputs and richer document/chart understanding using Mistral-7B [35].

Recent open VLMs include Qwen2.5-VL [2], which integrates a redesigned ViT with Qwen2.5 [84], and Qwen3-VL [1], which improves temporal-spatial alignment for long-video understanding. At a larger scale, GPT-4o [34] unifies text, audio, image, and video, and GPT-5 [69] introduces dynamic routing between fast general-purpose

and deeper reasoning models, while OneLLM [31] extends this trend with a unified framework aligning diverse modalities, including 3D point clouds, to language.

Multimodal AI in Medical and Oral Imaging. Medical AI has progressed from CNNs to multimodal foundation models that couple images and text, enabling broad transfer across detection, segmentation, retrieval, VQA, captioning, and report generation [45,49]. Representative medical systems, like MedGemma [68], LLaVA-Med [42], and Med-PaLM M [74], adapt LLM-ViT stacks to clinical data, while MedSAM [53] and domain encoders like HAI-DEF [38] provide promptable segmentation and reusable features for downstream tasks. Dentistry is beginning to follow suit: DentVLM [56] and Oral-GPT [32] specialize VLMs for 2D oral imaging and dialogue/reporting; geometry-aware and segmentation-focused efforts such as ChatIOS [82] and Tooth-ASAM [78] extend to 3D IOS, CBCT, and panoramic radiographs. Yet explicit dental report generation remains nascent: recent datasets/benchmarks fine-tune Qwen-VL to produce full diagnostic reports [32, 52], and ontology-grounded CBCT reporting has recently been explored [50], but most work targets 2D radiographs rather than tightly registered multimodal inputs with 3D CBCT and/or IOS data, highlighting the need for geometry-aware, template-guided VLMs for oral imaging and reporting [75].

Dataset Collection and Annotation Frameworks. In radiology and digital pathology, multi-center datasets are built from composable stacks rather than single platforms: XNAT [55] provides a scalable, role-aware DICOM backbone; OHIF [23] and MONAI Label [22] add viewer-integrated, clinician-in-the-loop annotation with AI pre-labels. Recent literature distills the key building blocks for such infrastructures: interoperable archives and viewers, de-identification, traceable curation, labeling workflows, and catalogs of open-source tools (e.g., 3D Slicer, QuPath, Orthanc) [5, 27, 28, 36, 76] alongside large-scale infrastructures for AI in medical imaging [40].

Dentistry is falling behind in this trajectory. Most studies rely on pre-curated data with little account of multi-center governance or end-to-end annotation stacks; TSegLab, for example, is task-specific on IOS [67]. Public datasets are fragmented and modality-specific; to our knowledge, no vendor-agnostic, open-source platform supports collaborative multi-center, multimodal aggregation with integrated AI-assisted annotation and report drafting—existing tools cover only subsets and require substantial glue code [28].

Our work addresses this gap with a privacy-first, containerized web platform—role-based governance, de-identification at ingestion, and 2D/3D viewers—used to assemble our proposed dataset.

Multimodal Datasets in Oral Imaging. In the dental/oral imaging domain, the availability of publicly accessible multimodal datasets remains limited, but recent contributions are emerging (Tab. 1). Tufts Dental Database [60] is one of the first attempts in the field; it pairs 1,000 panoramic radiographs with radiologist eye-tracking and think-aloud audio—useful for modeling diagnostic reasoning itself.

Since 2024, releases have accelerated and diversified. Huang *et al*. [33] contribute a clinically matched triad—CBCT, panoramic OPG, and CBCT-derived periapicals—across 169 patients, addressing earlier gaps in multimodal alignment and pathology breadth. Li *et al*. [44] then link 3D CBCT with intraoral scans, enabling cross-modal registration and anatomy modeling. In 2025, MMDental goes beyond images to pair 660 CBCT studies of the teeth with expert medical records (initial and follow-up notes), enabling

image-to-report learning at the patient level. Parallel task-focused sets emerge too: FD-Tooth couples intraoral photographs with corresponding CBCT for 241 patients to study fenestration/dehiscence, and MMOral curates 20,563 panoramic X-rays with 1.3M instruction pairs plus a 100-image benchmark to test VLMs. In this latter case, images are collected from other datasets and not publicly released.

Together, recent corpora trace a clear shift from single-view radiography to richer combinations of 3D-2D imagery and image-text. However, they rarely model the practical reporting workflow and, to our knowledge, none simultaneously provide registered upper/lower 3D intraoral scans, paired intraoral photographs, and clinician-authored reports collected across multiple centers, vendors, and protocols. In this paper, we address this gap by introducing a dataset of 1,000 patients that unifies these modalities and establishes a realistic benchmark for multimodal dental AI.

3 The Proposed Dataset

One of the main contributions of this paper is the introduction of a multimodal maxillofacial dataset comprising *1,000 patients*. For every patient, we provide a pair of mutually registered intraoral scans—high-resolution meshes for the upper and lower arches—aligned according to the patient's occlusion and stored in STL format (*2,000 scans* in total), five intraoral RGB photographs (*5,000 images overall*), and a clinician-authored report (*2,403 reports*), written as free text in the clinician's native language, along with its English translation and a template-standardized version (Sec. 3.3).

3.1 Inclusion Criteria

Patients were randomly included without filtering to reflect real clinical distributions. Specifically, patients were randomly sampled from routine orthodontic assessments across participating centers, without stratification by malocclusion type, dentition status, treatment stage, or scanner vendor. The resulting cohort spans a broad age range (16-99 years) and exhibits a sex distribution that is approximately representative of the contributing clinics ($\sim$58% female, $\sim$42% male). No post-hoc rebalancing was performed. Minimal eligibility constraints were applied to ensure data usability: *(i)* age $\geq$ 16 at acquisition, *(ii)* successful ingestion and de-identification, and *(iii)* availability of the core study package—mutually registered upper/lower IOS and an associated clinician-authored report. Intraoral photographs were always available.

Cases were excluded only when essential data were missing or unusable, e.g., failed occlusion-preserving registration between arches, severe acquisition artifacts preventing reliable visualization/annotation, or incomplete uploads that could not be curated into a consistent patient record. This design allows the capture of the natural heterogeneity of real-world orthodontic practice, including multi-center and multi-vendor variability.

3.2 Intraoral Acquisitions: 3D Scans and 2D Photos

Intraoral Scans. The dataset comprises *1,000 pairs* of intraoral scans in STL format, with registered meshes for the maxillary and mandibular arches (pink and violet elements shown in Fig. 1, respectively). Scans are *spatially aligned* to preserve the true

Table 2: Summary of imaging modalities, their average resolutions or 3D properties, and the number of samples included in the initial release of the dataset.[6]

Modality	Type	Average Size	Statistics (avg.)	#Samples
IOS	3D Mesh	$65 \times 40 \times 35$ mm	114,096.37 vertices	2,000
IOP	2D Image	$3{,}926 \times 2{,}545$ px	$\sim$10MP res.	5,000
Reports	Text	–	–	2,403
Patients	–	–	–	1,000

occlusal relationship of the patient and are transformed into a common RAS (Right-Anterior-Superior) reference frame, where axes point toward the patient's right, anterior, and superior directions. Bases, when present, were automatically removed to retain only the gingival and dental structures. Scans were acquired using different scanner families, including Carestream and 3Shape TRIOS, to capture variability in acquisition technologies. Statistics are reported in Tab. 2.

Intraoral Photos. For each patient, *five* standardized *intraoral RGB photographs* are also available (*5,000 images* in total), categorized by view: *frontal*, *left buccal*, *right buccal*, *upper occlusal*, and *lower occlusal* (Fig. 1). Images were captured with a variety of high-resolution RGB cameras (e.g., smartphone cameras, Canon EOS 90D, Sony α7 III) with a resolution of $\geq$ 24MP (6,000$\times$4,000 pixels), then cropped and rotated to isolate and align the anatomical portion of interest. These photographs complement the IOS geometry by contributing photometric appearance: they enable clinicians not only to assess tooth alignment and occlusion, but also to check soft-tissue status (e.g., gingival and mucosa health), identify early pathological changes (such as oral tumors or lesions), and enable 2D-3D correspondence learning and cross-modal supervision.

3.3 Clinical Reports and Normalization

Our reports are authored by clinicians either as typed free-text (Fig. 1) or dictated speech. For dictated reports, we use OpenAI's Whisper model [65] for speech-to-text conversion. The transcription is generated in real time, and the annotators are tasked with verifying its correctness and correcting any errors, typos, or mistranslations. The free-text reports, although not templated, follow the standardized internal reporting protocol routinely used by the involved centers in orthodontic practice and are released with the dataset to ensure transparency and reproducibility. Reports are released in both the clinician's original language and in English, along with a *structured template* version automatically generated from the clinician-authored free-text.

Template Taxonomy. As mentioned, to standardize reporting and evaluation, all the experiments described in this paper adopt a structured template—covering tooth- and arch-level fields, plus short, standardized narrative segments—so that model outputs populate clinically meaningful slots, thereby reducing hallucination surface area, yielding consistent gains over strong baselines, and enabling fast, reliable human oversight. The

[6] The platform introduced in Sec. 4 and used to collect and label the dataset is designed for continuous expansion to include more patients, acquisition machines, and centers over time.

Template

Overbite:
Normal | Increased | Decreased | Open bite

Crowding:
Severity: Absent | Mild | Moderate | Severe
Location: Upper arch | Lower arch | Both arches
Modifiers: With spacing | Diastemas | Aligned arches

Molar/Canine Occlusion (L/R):
Class I | Class II head-to-head | Class II full | Class III | Not assessable due to missing tooth

Curve of Spee:
Normal | Increased | Decreased | Reversed

Curve of Wilson:
Degree: Normal | Increased | Decreased | Reversed
Location: Lower arch | Upper arch | Right side | Left side

Midlines:
Centered | Deviated

Transverse Relationship:
Severity: Normal | Mild | Moderate | Severe
Type: Contraction | Constriction of maxilla
Modifier: With | Without
Crossbite: Crossbite | Posterior crossbite
Location: Bilateral | Right side | Left side

Fig. 2: Structured template.

System Prompt

You are a professional orthodontist specialized in writing structured reports. Based on the free-text report provided by the user, author a structured report following the provided template: `[TEMPLATE]`

Follow these rules:

- Consider only the input free-text report, not any prior reports or external information.
- If the information cannot be determined from the free-text, answer "unknown".
- Output only the value, without repeating the field name or adding any explanations.
- Replace [numbers] with the actual tooth numbers, using FDI notation.
- Do not put any symbol or brackets in the report.

User Prompt

On the right, the patient shows a Class II molar and canine relationship, while the left side is Class I. The arches do not relate normally transversely, with a fissured left premolar. Overjet is increased at the lateral incisors, the dental midlines are slightly off, and the lower curve of Spee is accentuated.

Fig. 3: Structured template generation prompts.

proposed structure focuses on clinically salient fields identified by expert orthodontists. The list of fields and allowed values in the template is shown in Fig. 2.

Both the original clinician-authored free-text reports (in the original language and an English translation) and the templated, structured version derived from the free-text are released for standardized benchmarking. The structured version is generated by *gpt-5-latest*, conditioned on the template (Fig. 2) and the prompting rules (Fig. 3). To avoid treating model outputs as ground truth annotations, we explicitly treat the templated report as a *derived annotation layer*. Each automatically templated report is reviewed by the clinician who authored the original free-text report, and mapping-related discrepancies are corrected during this verification step.

Annotator Expertise. Four orthodontic specialists (>5 years of experience) curated the labels following a two-step validation protocol: in the first stage, each report is drafted, automatically converted to the structured template, and verified by one specialist. During the second stage, a separate specialist validates the annotations, and any discrepancies are resolved through consensus to ensure labeling consistency and accuracy.

Inter- and Intra-Rater Clinical Agreement. Since at least two independent reports per patient are available from different annotators, and in some cases two reports per patient were drafted by the same clinician, we are able to compute and report inter- and intra-rater clinical agreement. We refer the reader to Sec. 5, where the evaluation metrics and the corresponding agreement scores are reported.

3.4 Ethics and Availability

Ethical Approval. Ethical approval for all procedures, protocols, and data release was granted by the Comitato Etico di Ateneo di Ferrara under Approval 262/2025/Oss/UniFe.

Availability. The dataset is available under CC BY-SA 4.0 license through a dedicated download platform at `https://ditto.ing.unimore.it/bite2text`.

4 Multimodal Data Management

The second contribution of the paper is the introduction of a containerized, privacy-preserving web platform that manages the full lifecycle of maxillofacial data across multiple clinical centers. Heterogeneous sources—IOP, registered IOS of both arches, CBCT, OPG, and CEPH—are consolidated in a single patient-centric repository with tools for visualization, labeling, modality-aware inference, and template-guided report drafting. The stack is Django-based and deployed as cooperating Docker services, which simplifies replication across sites with differing IT constraints while preserving a uniform audit and provenance model. Benefits w.r.t. the current literature approaches [28, 75, 76] include *(i)* consistent de-identification and governance at ingestion, *(ii)* a unified labeling/indexing that aligns 3D geometry with radiographic and textual findings, *(iii)* clinician-in-the-loop review in the browser, and *(iv)* scalable, modular inference decoupled from the web application, enabling annotations and model outputs to be stored as versioned, immutable, provenance-linked records. Although designed for maxillofacial data, its modular design makes it extensible to additional modalities and use cases. Moreover, although the proposed dataset focuses on IOS/IOP and reports, the application already supports all orthodontic/maxillofacial imaging modalities.

The application code is released as open-source with this paper, and a live version is available at `https://yggdrasil.unimore.it`. Application screenshots and implementation details are provided in the *Supplementary Material*.

4.1 Privacy-First Ingestion and Governance

Clinician browsers connect to a Django backend that exposes the application with user and role-based access control. Metadata, indexing, and audit logs live in a relational database. Ingestion emphasizes de-identification from the outset: configurable DICOM scrubbing removes sensitive tags; photographs and meshes are stripped of metadata and normalized; pseudonyms are generated, and linkage to hospital identifiers is kept outside the research-export boundary. The system performs conformance/quality checks—e.g., voxel spacing and slice ordering for CBCT; manifoldness and vertex-count bounds for IOS; expected views for photographs—and produces thumbnails for quick browsing. All imports, views, edits, approvals, and exports are recorded in an append-only audit table. Exports include a manifest of data versions and model hashes to ensure reproducibility.

4.2 Data Model and Interactive Tooling

Each pseudonymized patient includes one or more modalities. IOS are stored as meshes or point clouds with explicit registration metadata for upper and lower arches, enabling direct reasoning about occlusion. All assets share a common index based on FDI tooth numbers and arch labels, which simplifies alignment of geometric outputs (e.g., per-tooth segmentation) with textual slots in the report. Annotations—like segmentation masks, landmarks, per-tooth tags—and model outputs—with confidence scores and versioned weights—are immutable and linked to inputs for precise reproducibility. Visualization is browser-based: the 2D viewer supports synchronized multiplanar navigation, window/level, measurements, overlays, and cephalometric landmarking; the 3D viewer

renders IOS with surface painting and per-tooth overlays, tracks edit versions, and computes inter-rater agreement. Structured report information can thus be cross-validated with model outputs, e.g., a suspected missing tooth in the draft report can be checked against both the 3D IOS and the corresponding radiographs.

4.3 Modular Inference and Scalable Deployment

Pipeline execution is modality-aware and asynchronous. When a study satisfies a model's input contract (e.g., both arches registered for occlusion classification), the backend enqueues a job to a task-specific runner. Runners are GPU-enabled containers that declare typed inputs/outputs and resource needs, allowing the scheduler to batch compatible jobs and utilize accelerators efficiently. Outputs are returned as overlays, tooth-level labels, or structured fields that populate a template-guided reporting interface. Rather than free text, clinicians receive a draft with predefined slots (tooth presence/absence, restorations, occlusion class, arch-level findings, global comments); provenance badges indicate whether each value originated from a model or a human edit. The same interface supports voice dictation via ASR, with parsed text highlighted and mapped into slots for clinical confirmation. Operationally, the platform runs on a single machine via Docker Compose or on Kubernetes for larger installations. Runners remain self-contained, so updating or adding a model does not require changes to the web app or database schema. New modalities can be plugged in by simply providing a validator and a viewer adapter.

4.4 Clinical Workflow Efficacy

We assess workflow-level efficacy via *clinical report generation time* recorded during routine-style annotation. Orthodontic specialists repeatedly reported on a randomly sampled subset of 100 patients under two conditions: a baseline workflow using the standard visualizer and our platform with its integrated automatic pipelines (i.e., segmentation, occlusal classification, and speech-to-text tool) enabled. In this pilot evaluation, the mean reporting time decreased from 300 ± 120 s to 60 ± 30 s per patient, i.e., $5\times$ speed-up.

5 Experiments

As a final contribution, we present an empirical evaluation of the performance of existing multimodal large language models (MLLMs) in generating orthodontic *occlusal reports* from the modalities provided in our dataset. Our goal is *not* to propose a new state-of-the-art model, but rather to quantify the out-of-the-box performance of widely used MLLMs on a clinically structured report-generation task, and establish a trainable, modality-specific reference baseline for intraoral scans that future work can build upon.

Two Evaluation Tracks. We separate experiments into two tracks based on modality support. *(A) IOP-only off-the-shelf evaluation:* we benchmark existing general and medical MLLMs using intraoral photographs only, *without any retraining*, using a shared prompt and a fixed `<field:value>` output template for automatic scoring (Fig. 2); *(B) IOS-only reference baseline:* because most existing MLLMs do not accept 3D mesh inputs and 3D-capable MLLMs are typically designed for sparse object-level

point clouds (e.g., a few thousand points) rather than high-resolution full-arch intraoral surfaces [63,83], they cannot be directly applied to IOS without aggressive downsampling that would remove fine-grained occlusal geometry. Therefore, we evaluate IOS-based report generation separately and introduce a trainable IOS-specific baseline (IOS-Qwen) to serve as a reference point for future extensions on this dataset.

Inputs and Dataset. For the IOP-only track, each case includes five standardized intraoral photographs (frontal, right buccal, left buccal, maxillary occlusal, mandibular occlusal). We use a fixed patient-level random train/test split of 700/300. The *off-the-shelf MLLMs are evaluated only on the 300-patient test set*, and are *never trained/fine-tuned* on our data. The 700-patient training split is used *only* to train the IOS-specific baseline (IOS-Qwen) in the IOS-only track.

5.1 Models

Off-the-Shelf MLLMs (IOP-Only). We evaluate general and medical MLLMs in a strictly *zero-shot* setting on IOP: Qwen3-VL-2B/4B Instruct [1], Qwen2.5-VL-7B Instruct [2], LLaVA-1.6-Mistral-7B [43], MedGemma-4B/27B-IT [68], GPT-5.2 [69], and Gemini 3 Flash [29]. All models use the *same* prompt, template definition, and decoding settings; no fine-tuning or adaptation is performed. Exact prompts used are available in the source code.

IOS-Qwen: An IOS-Specific Trainable Reference Baseline (IOS-Only). Because off-the-shelf MLLMs are not directly applicable to high-resolution IOS geometry, we introduce IOS-Qwen as a *trainable reference baseline* for IOS-based report generation (Fig. 4). IOS-Qwen integrates IOS point clouds through a PointTransformer encoder and a lightweight projector into Qwen3-VL-4B, enabling future work to compare against a reproducible geometry-aware pipeline on this dataset. More specifically, we explore three different training strategies: (♣) *frozen LLM, with a joint training of the encoder and projector*. In this case, PointTransformer and the projector are initialized from scratch and trained to generate the ground-truth report, with Qwen frozen. In the second configuration, (♦) we *pre-train the PointTransformer with per-field classification heads*: only fields with discrete values are retained and encoded as multi-class labels. Token features are attention-pooled into a single embedding, which is fed to the MLP heads. We optimize cross-entropy loss until convergence. Heads are then replaced with the projector that maps PointTransformer to Qwen dimensionality, and tokens are fed into the language model. The point encoder is trained jointly with the projector using the same approach as above. Finally, (♠) the third strategy involves *fine-tuning Qwen using LoRA* while also training the projector; the base model and the point encoder are frozen.

5.2 Prompting and Decoding

In our IOS-only experiments, we consider two prompting regimes, i.e., *one-shot* or single pass (†), and *field-by-field* or multi-turn (‡). In the former case, the model is tasked to output all fields, one per line, in one response, strictly following the `<field: value>` template and by listing all the required fields in the prompt. In the latter, instead, the user prompt specifies a single field for which a value is required, and the model outputs only

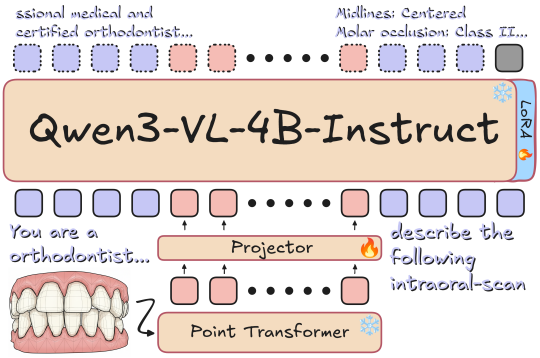


Fig. 4: An illustration of our IOS-Qwen. The intraoral scan is encoded by the PointTransformer, and the resulting geometry tokens are concatenated with the prompt. Token-by-token generation is performed to generate the full report.

the value for the requested field. One inference per field is required. This substantially improves *coverage* (Sec. 5.3) by reducing formatting and omission errors, but increases computational requirements by a factor of $\sim$9. Exact prompts, decoding hyperparameters (temperature, top-p, max tokens), and the per-field decoding script are available in the source-code repository.

5.3 Metrics

We employ both standard caption metrics—BLEU-1, ROUGE-L (P/R/F1), METEOR, and Sentence-BERT cosine similarity (SBERT-sim)—as well as a clinically specific metric, RadFact [6]. We also propose two ad hoc task-specific scores aligned with the employed template. Indeed, as noted in prior work, traditional metrics often fail to capture the semantic similarity and diversity of clinical text [6, 83].

Standard Metrics. Among the selected metrics, BLEU-1 measures unigram overlap, emphasizing lexical precision but ignoring order and synonyms. ROUGE-L scores the longest common subsequence (often reported as Precision (P), Recall (R), and F1 score), capturing coverage and relevance. METEOR adds stemming, synonym matching, and word-order penalties, improving sensitivity to meaning. SBERT-sim computes cosine similarity between Sentence-BERT embeddings, directly assessing semantic similarity beyond surface overlap. However, all these metrics were originally developed for general-purpose text generation tasks such as machine translation or summarization. They are not specifically tailored to the medical or dental domain, where factual accuracy, clinical correctness, and diagnostic completeness are crucial. As a result, they may not fully reflect the clinical validity of the generated reports [20, 59, 85].

Clinical Metrics. RadFact [6] is an LLM-based evaluation framework for comparing a generated report to a ground-truth report at the sentence level. It treats sentences from one report as hypotheses and sentences from the other as premises, and uses an LLM[7] to judge

[7] We employed GPT-5.2 as in the original formulation.

whether each hypothesis is logically entailed by the premises. From these entailment decisions, RadFact reports a suite of precision/recall metrics: logical precision, i.e., the fraction of generated sentences entailed by the ground-truth (it penalizes hallucinations); and logical recall, i.e., the fraction of ground-truth sentences entailed by the generated report (it penalizes omissions). The harmonic mean of these two values is reported as the *RadFact-F1* score. Although originally proposed for chest X-rays, RadFact is not limited to that modality since it provides a general methodology for comparing lists of findings.

Custom Task-Specific Scores. Let K be the set of ground-truth fields for a case and $\hat{K}$ the set of fields predicted by a model. We define *Coverage* as the fraction of required keys found in the prediction, i.e., $|K \cap \hat{K}|/|K|$. *Accuracy*, instead, averages per-field correctness where a missing key is treated as a score of 0; for constrained slots (e.g., molar/canine occlusion) we match against canonical label sets with light normalization (e.g., capitalization, *"Class 2"*$\rightarrow$*"Class II"*); for compositional slots (e.g., crowding and transverse) we compute the number of matching words between the predicted and ground-truth value strings. Fields explicitly marked as *unknown* in the GT are ignored in both coverage and accuracy. This is similar in spirit to unigram-overlap measures such as the BLEU-1 metric but covers task-specific edge cases.

Double-Blind Clinical Preference Testing. Clinicians are shown two generated reports with the corresponding visual data and asked to choose the better one. Following LMArena-style ranking,[8] scores are computed with a Bradley-Terry model over 1,000 battles balanced across models.

5.4 Results

Tab. 3 summarizes performance on the held-out test set of 300 patients. We stress that IOP-only and IOS-only results are reported as *separate tracks*; since inputs differ, they are not intended as a single unified state-of-the-art ranking.

IOP-Only (Off-the-Shelf). Among zero-shot MLLMs evaluated on intraoral photographs, Qwen3-VL-4B-Instruct and MedGemma variants achieve the strongest overall scores, although slot omissions and formatting errors remain common. Despite fluent text (high SBERT-sim), slot-level fidelity varies across models, indicating that generic semantic similarity does not guarantee clinically correct field values.

IOS-Only (Reference Baseline). Separately, IOS-Qwen provides a trainable, geometry-conditioned reference baseline for intraoral scan inputs. Across the evaluated IOS-Qwen training strategies, we observe consistent but modest gains in template fidelity, suggesting that learning to condition on IOS geometry is feasible and providing a reproducible starting point for future work on this dataset.

Clinical Agreement and Double-Blind Clinical Preference. As a reference, the last two rows of Tab. 3 show the inter- and intra-rater clinical agreement computed on templated reports. Disagreement concerns borderline categories (e.g., "Class I" *vs.* "Mild Class II" cases), reflecting clinical subjectivity. Moreover, the last two columns of the same table report the Elo-style scores and the win rates from the double-blind clinical preference study described in Sec. 5.3.

[8] `https://arena.ai/`

Table 3: Results on occlusal report generation with 1,000 bootstrap samples, including double-blind clinical preference as both Elo-style scores and win rates, and inter- and intra-rater clinical agreement. *Top blocks:* IOP-only evaluation of off-the-shelf MLLMs in a zero-shot setting (no retraining). *Bottom blocks:* IOS-only trainable reference baselines using intraoral scans. Due to different input modalities, the two blocks should be interpreted as separate tracks rather than a direct model ranking. ♣, ♦, and ♠ identify different training strategies while † and ‡ identify different generation approaches described in Sec. 5.1 and Sec. 5.2, respectively.

	Model	BLEU-1↑	ROUGE-L (P)↑	ROUGE-L (R)↑	ROUGE-L (F1)↑	METEOR↑
IOP-Only	Qwen3-VL-4B-Instruct [1]	0.757±0.087	0.774±0.080	0.726±0.100	0.746±0.083	0.777±0.088
	LLaVA-hf_LLaVA-v1.6-Mistral-7B-hf [43]	0.678±0.058	0.705±0.061	0.732±0.086	0.715±0.061	0.710±0.073
	Google_MedGemma-4B-IT [68]	0.654±0.101	0.815±0.075	0.623±0.101	0.702±0.080	0.682±0.092
	Google_MedGemma-27B-IT [68]	0.696±0.110	**0.824±0.067**	0.676±0.105	0.722±0.080	0.662±0.098
	Qwen3-VL-2B-Instruct [1]	0.609±0.190	0.564±0.180	0.742±0.075	0.619±0.147	0.725±0.083
	Qwen2.5-VL-7B-Instruct [2]	0.527±0.096	0.703±0.066	0.542±0.106	0.597±0.081	0.561±0.099
	GPT-5.2 [69]	0.657±0.170	0.740±0.161	0.619±0.167	0.670±0.160	0.665±0.174
	Gemini 3 Flash [29]	0.549±0.244	0.784±0.120	0.551±0.190	0.632±0.162	0.593±0.207
IOS-Only	IOS-Qwen ♠ ‡	**0.781±0.082**	0.772±0.074	**0.754±0.083**	**0.763±0.056**	**0.790±0.079**
	IOS-Qwen ♠ †	0.719±0.096	0.703±0.087	0.672±0.096	0.687±0.065	0.714±0.090
	IOS-Qwen ♦ †	0.455±0.124	0.485±0.102	0.405±0.104	0.441±0.075	0.541±0.111
	IOS-Qwen ♣ †	0.448±0.132	0.645±0.114	0.409±0.108	0.501±0.088	0.527±0.116
	Inter-rater agreement	0.800±0.086	0.820±0.095	0.820±0.095	0.820±0.077	0.845±0.079
	Intra-rater agreement	0.880±0.088	0.900±0.090	0.900±0.090	0.900±0.078	0.915±0.075

	Model	Accuracy↑	Coverage↑	SBERT-sim↑	RadFact F1↑	Elo↑	Win Rate↑
IOP-Only	Qwen3-VL-4B-Instruct [1]	0.580±0.156	0.996±0.026	0.928±0.027	0.358±0.210	934	0.419
	LLaVA-hf_LLaVA-v1.6-Mistral-7B-hf [43]	0.541±0.107	**1.000±0.000**	0.942±0.020	0.316±0.156	870	0.324
	Google_MedGemma-4B-IT [68]	0.514±0.160	**1.000±0.000**	0.887±0.044	0.218±0.257	1016	0.554
	Google_MedGemma-27B-IT [68]	0.544±0.160	**1.000±0.000**	0.907±0.046	0.339±0.209	894	0.375
	Qwen3-VL-2B-Instruct [1]	0.439±0.116	**1.000±0.000**	0.926±0.031	0.238±0.143	862	0.324
	Qwen2.5-VL-7B-Instruct [2]	0.399±0.140	0.638±0.049	0.785±0.029	0.368±0.196	943	0.432
	GPT-5.2 [69]	0.447±0.186	0.749±0.229	0.884±0.152	0.318±0.192	989	0.514
	Gemini 3 Flash [29]	0.411±0.202	0.741±0.239	0.863±0.081	0.339±0.212	998	0.514
IOS-Only	IOS-Qwen ♠ ‡	**0.632±0.110**	**1.000±0.000**	**0.959±0.020**	**0.372±0.182**	**1048**	**0.583**
	IOS-Qwen ♠ †	0.512±0.170	**1.000±0.000**	0.881±0.041	0.233±0.154	1030	0.541
	IOS-Qwen ♦ †	0.411±0.160	0.765±0.190	0.823±0.060	0.308±0.186	988	0.501
	IOS-Qwen ♣ †	0.372±0.160	0.743±0.220	0.788±0.066	0.282±0.191	973	0.472
	Inter-rater agreement	0.740±0.140	0.920±0.108	0.950±0.034	0.650±0.150	—	—
	Intra-rater agreement	0.860±0.127	0.950±0.091	0.970±0.029	0.810±0.161	—	—

5.5 Qualitative Evaluation

Tab. 4 presents a qualitative comparison of the best-performing architectures. Here, we provide insights into how different models handle the structured report generation task, highlighting the specific impact of geometric awareness. For all models, the input prompt remains consistent with the template definitions. Again, all off-the-shelf models use intraoral photographs as input in a zero-shot setting, while IOS-Qwen is the only model that leverages a modality-specific encoder trained via a LoRA adapter as previously described. While general-purpose MLLMs (e.g., MedGemma and Qwen3-VL) often default to generic descriptions (“Normal,” “Class I”) or hallucinate missing teeth regardless of the input images, IOS-Qwen improves the identification of specific orthodontic conditions and aligns closely with the ground truth. It is worth noting that MedGemma-27B consistently predicted “Class I” for the Occlusion category for all patients. For the “Missing Teeth” category, all models, except LLaVA-1.6-7B, which occasionally mispredicts missing teeth, returned “None” for all patients. By leveraging

Table 4: Qualitative comparison of report generation. We compare the Ground Truth (GT) against representative baselines and our proposed IOS-Qwen. Green text indicates clinically accurate findings that match the GT; orange text indicates an incomplete but correct clinical finding; red text indicates significant hallucinations or diagnostic errors. Note that IOS-Qwen is the only model to correctly identify the Class II relationship and the increased overbite.

Feature	Ground Truth	MedGemma-27B	LLaVA-1.6-7B	Qwen3-VL-4B	IOS-Qwen ♠ ‡
Overbite	Increased	Normal	Decreased	Normal	Increased
Crowding	Moderate in the lower arch and severe in the upper arch	Mod-to-severe in both arches	Mild-to-moderate (upper)	Absent in the upper arch with aligned arches	Moderate in the lower arch
Occlusion (Molar)	R: Class II H-to-H L: Class II full	R: Class I L: Class I	R: Class II full L: Class II full	R: Class I L: Class I	R: Class II H-to-H L: Class II full
Occlusion (Canine)	R: Class II H-to-H L: Class II full	R: Class I L: Class I	R: Class II full L: Class II full	R: Class I L: Class I	R: Class II H-to-H L: Class II full
Curve of Spee	Increased	Normal	Normal	Normal	Increased
Curve of Wilson	Normal in the upper arch	Normal	Normal in the upper arch	Normal in the lower arch	Normal
Midlines	Slightly deviated	Centered	Centered	Centered	Deviated
Transverse	Moderate constriction of the maxilla without posterior crossbite	Moderate constriction of the maxilla without crossbite	Moderate [contraction of the maxilla without crossbite or posterior crossbite]	Moderate constriction of the maxilla without crossbite	Moderate constriction of the maxilla without crossbite
Missing Teeth	None	None	Teeth [12] are absent	None	None

3D information, our Qwen3-VL-4B adaptation, IOS-Qwen, improves both occlusion and missing-tooth predictions.

6 Conclusion

We presented a 1,000-patient multimodal dataset for dental report generation—spanning paired IOS and five IOPs per patient—and showed that current MLLMs struggle to produce clinically faithful drafts under this distribution. To seed progress, we introduced IOS-Qwen, which fuses 3D point-cloud features with Qwen3-VL and establishes a simple, reproducible baseline. The dataset was created with an open-source web application that supports multi-center ingestion, de-identification, AI-assisted labeling, and template-guided reporting, enabling scalable contributions and rigorous evaluation.

This work lays the foundation for a long-term global effort. The open infrastructure enables privacy-preserving, multi-center contributions; the dataset defines a concrete benchmark; and the geometry-aware baseline provides a starting point for community progress in multimodal dental AI.

Acknowledgements

This project has received funding from Fondazione di Modena, through the FAR 2024 (E93C24002080007), and from MUR, under the NRRP "Fit4MedRob-Fit for Medical Robotics" (PNC0000007).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report. arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL Technical Report. arXiv:2502.13923 (2025)
3. Baid, U., Dorent, R., Malec, S., Pytlarz, M., Su, R., Wijethilake, N., Bakas, S., Crimi, A. (eds.): Brain Tumor Segmentation, and Cross-Modality Domain Adaptation for Medical Image Segmentation, Lecture Notes in Computer Science, vol. 14669 (2024)
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(1) (2017)
5. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W.: QuPath: Open source software for digital pathology image analysis. Scientific Reports **7**(1) (2017)
6. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M., Srivastav, S., Gong, J., Codella, N.C.F., Falck, F., Oktay, O., Lungren, M.P., Wetscherek, M.T., Alvarez-Valle, J., Hyland, S.L.: MAIRA-2: Grounded Radiology Report Generation. arXiv:2406.04449 (2024)
7. Ben-Hamadou, A., Neifar, N., Rekik, A., Smaoui, O., Bouzguenda, F., Pujades, S., Boyer, E., Ladroit, E.: Teeth3DS+: An Extended Benchmark for Intraoral 3D Scans Analysis. arXiv:2210.06094 (2022)
8. Ben-Hamadou, A., Smaoui, O., Rekik, A., Pujades, S., Boyer, E., Lim, H., Kim, M., Lee, M., Chung, M., Shin, Y.G., Leclercq, M., Cevidanes, L., Prieto, J.C., Zhuang, S., Wei, G., Cui, Z., Zhou, Y., Dascalu, T., Ibragimov, B., Yong, T.H., Ahn, H.G., Kim, W., Han, J.H., Choi, B., van Nistelrooij, N., Kempers, S., Vinayahalingam, S., Strippoli, J., Thollot, A., Setbon, H., Trosset, C., Ladroit, E.: 3DTeethSeg'22: 3D Teeth Scan Segmentation and Labeling Challenge. arXiv:2305.18277 (2023)
9. Bolelli, F., Lumetti, L., van Nistelrooij, N., Vinayahalingam, S., Di Bartolomeo, M., Marchesini, K., Pellacani, A., Candeloro, E., Rosati, G., Xi, T., Isensee, F., Kirchhoff, Y., Krämer, L., Rokuss, M., Ulrich, C., Maier-Hein, K., Jiang, Y., Liu, Y., Wang, L., Wang, H., Chen, S., Cui, Z., Shi, P., Pan, Z., Liang, X., Ma, Q., Konukoglu, E., Wodzinski, M., Müller, H., Mai, H., Dang, X., Bhandary, S., Grosu, R., Bergé, S., Anesi, A., Grana, C.: Multi-Structure Segmentation in CBCT Volumes: the ToothFairy2 Challenge. Medical Image Analysis (2026)
10. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., R. Rokuss, M., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles Ballester, V., Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the Inferior Alveolar Canal in CBCT Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging (2024)

11. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volumes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
12. Borghi, L., Lumetti, L., Cremonini, F., Rizzo, F., Grana, C., Lombardo, L., Bolelli, F.: Bits2Bites: Intra-oral Scans Occlusal Classification. In: Oral and Dental Image Analysis - ODIN, MICCAI Workshop (2025)
13. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems **33** (2020)
14. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66** (2020)
15. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: IEEE/CVF International Conference on Computer Vision (2021)
16. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality (2023), https://lmsys.org/blog/2023-03-30-vicuna/
17. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: PaLM: Scaling Language Modeling with Pathways. Journal of Machine Learning Research **24**(240) (2023)
18. Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., Minafra, P., Anesi, A., Grana, C.: Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. IEEE Access (2022)
19. Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., Grana, C.: Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
20. Delbrouck, J.B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C.: Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In: Findings of the Association for Computational Linguistics: EMNLP 2022 (2022)
21. Di Bartolomeo, M., Pellacani, A., Bolelli, F., Cipriano, M., Lumetti, L., Negrello, S., Allegretti, S., Minafra, P., Pollastri, F., Nocini, R., Colletti, G., Chiarini, L., Grana, C., Anesi, A.: Inferior Alveolar Canal Automatic Detection with Deep Learning CNNs on CBCTs: Development of a Novel Model and Release of Open-Source Dataset and Algorithm. Applied Sciences (2023)
22. Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., Roth, H.R., Vercauteren, T., Xu, D., Dogra, P., Ourselin, S., Feng, A., Cardoso, M.J.: MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images. Medical Image Analysis **95** (2024)
23. Doran, S.J., Al Sa'd, M., Petts, J.A., Darcy, J., Alpert, K., Cho, W., Escudero Sanchez, L., Alle, S., El Harouni, A., Genereaux, B., Ziegler, E., Harris, G.J., Aboagye, E.O., Sala, E.,

Koh, D.M., Marcus, D.: Integrating the OHIF Viewer into XNAT: Achievements, Challenges and Prospects for Quantitative Imaging Studies. Tomography **8**(1) (2022)
24. Dosovitskiy, A.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 (2020)
25. Duy, H.B., Hue, T.T., Son, T.M., Nghia, L.L., Lan, L.T.H., Duc, N.M., Son, L.H.: A dental intraoral image dataset of gingivitis for image captioning. Data in Brief **57** (2024)
26. Faizan Ahmed, S.M., Ghori, M.H., Khalid, A., Nooruddin, A., Adnan, N., Lal, A., Umer, F.: Annotated intraoral image dataset for dental caries detection. Scientific Data **12**(1) (2025)
27. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R.: 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magnetic Resonance Imaging **30**(9) (2012)
28. Galbusera, F., Cina, A.: Image annotation and curation in radiology: an overview for machine learning practitioners. European Radiology Experimental **8**(1) (2024)
29. Google DeepMind: Gemini 3 Flash - Model Card. Tech. rep., Google DeepMind (2025)
30. Grattafiori, A., et al.: The Llama 3 Herd of Models. arXiv:2407.21783 (2024)
31. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: OneLLM: One Framework to Align All Modalities with Language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
32. Hao, J., Fan, Y., Sun, Y., Guo, K., Lizhuo, L., Yang, J., Ai, Q., Wong, L., Tang, H., Hung, K.: Towards Better Dental AI: A Multimodal Benchmark and Instruction Dataset for Panoramic X-ray Analysis. In: Advances in Neural Information Processing Systems (2025)
33. Huang, Y., Liu, W., Yao, C., Miao, X., Guan, X., Lu, X., Liang, X., Ma, L., Tang, S., Zhang, Z.: A multimodal dental dataset facilitating machine learning research and clinic services. Scientific Data **11**(1) (2024)
34. Hurst, A., et al.: GPT-4o System Card. arXiv:2410.21276 (2024)
35. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B. arXiv:2310.06825 (2023)
36. Jodogne, S., Bernard, C., Devillers, M., Lenaerts, E., Coucke, P.: Orthanc - A lightweight, restful DICOM server for healthcare and medical research. In: IEEE 10th International Symposium on Biomedical Imaging (2013)
37. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1) (2019)
38. Kiraly, A.P., Baur, S., Philbrick, K., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., George, N., Jamil, F., Tang, J., Bailey, K., Ahmed, F., Goel, A., Ward, A., Yang, L., Sellergren, A., Matias, Y., Hassidim, A., Shetty, S., Golden, D., Azizi, S., Steiner, D.F., Liu, Y., Thelin, T., Pilgrim, R., Kirmizibayrak, C.: Health AI Developer Foundations. arXiv:2411.15128 (2024)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: IEEE/CVF international conference on computer vision (2023)
40. Kondylakis, H., Kalokyri, V., Sfakianakis, S., Marias, K., Tsiknakis, M., Jimenez-Pastor, A., Camacho-Ramos, E., Blanquer, I., Segrelles, J.D., López-Huguet, S., Barelle, C., Kogut-Czarkowska, M., Tsakou, G., Siopis, N., Sakellariou, Z., Bizopoulos, P., Drossou, V., Lalas, A., Votis, K., Mallol, P., Marti-Bonmati, L., Alberich, L.C., Seymour, K., Boucher, S., Ciarrocchi, E., Fromont, L., Rambla, J., Harms, A., Gutierrez, A., Starmans, M.P.A., Prior, F., Gelpi, J.L., Lekadir, K.: Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. European Radiology Experimental **7**(1) (2023)

41. Kubík, T., Španěl, M.: LMVSegRNN and Poseidon3D: Addressing Challenging Teeth Segmentation Cases in 3D Dental Surface Orthodontic Scans. Bioengineering **11**(10) (2024)
42. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Advances in Neural Information Processing Systems **36** (2023)
43. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895 (2024)
44. Li, X.: 3D multimodal dental dataset based on CBCT and oral scan. In: Figshare (2024)
45. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42** (2017)
46. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
47. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems (2023)
48. Liu, K., Elbatel, M., Chu, G., Shan, Z., Sum, F.H.K.M.H., Hung, K.F., Zhang, C., Li, X., Yang, Y.: FDTooth: Intraoral Photographs and CBCT Images for Fenestration and Dehiscence Detection. Scientific Data **12**(1) (2025)
49. Lu, Y., Wang, A.: Integrating language into medical visual recognition and reasoning: A survey. Medical Image Analysis **102** (2025)
50. Lumetti, L., Di Bartolomeo, M., Pellacani, A., Anesi, A., Grana, C., Bolelli, F.: Ontology-Grounded Structured Prediction for Dental CBCT Reporting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026)
51. Lumetti, L., Tan, Z.Q., Borghi, L., van Nistelrooij, N., Rosati, G., Addison, O., Li, Y., Vinayahalingam, S., Grana, C., Bolelli, F.: ToothFairy3: Scaling CBCT Maxillofacial Segmentation to 77 Classes with U-Mamba2. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026)
52. Lv, H., Haq, I., Du, J., Ma, J., Zhu, B., Dang, X., Liang, C., Du, R., Zhang, Y., Saqib, M.: A benchmark multimodal oro-dental dataset for large vision-language models. arXiv:2511.04948 (2025)
53. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images. Nature Communications **15**(1) (2024)
54. Marchesini, K., Carpentiero, O., Del Gaudio, L., Farioli, F., Cucchiara, R., Grana, C., Cuculo, V., Bolelli, F.: ReportX: The BraTS Clinical Report Dataset. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026)
55. Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L.: The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics **5**(1) (2007)
56. Meng, Z., Hao, J., Dai, X., Feng, Y., Liu, J., Feng, B., Wu, H., Gai, X., Zhu, H., Hu, T., Wu, Y., Xu, H., Li, J., Xiao, J., Liu, X., Zhou, J.T., Zhu, F., Zhao, Z., Xia, L., Fang, B., Sun, J., Wu, J., Liu, Z.: DentVLM: A Multimodal Vision-Language Model for Comprehensive Dental Diagnosis and Enhanced Clinical Practice. arXiv:2509.23344 (2025)
57. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G.,

Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10) (2014)

58. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 (2023)
59. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: GREEN: Generative Radiology Report Evaluation and Error Notation. In: Findings of the association for computational linguistics: EMNLP 2024 (2024)
60. Panetta, K., Rajendran, R., Ramesh, A., Rao, S.P., Agaian, S.: Tufts Dental Database: A Multimodal Panoramic X-Ray Dataset for Benchmarking Diagnostic Systems. IEEE Journal of Biomedical and Health Informatics **26**(4) (2021)
61. Pang, T., Li, P., Zhao, L.: A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine **22**(1) (2023)
62. Pipoli, V., Saporita, A., Marchesini, K., Grana, C., Ficarra, E., Bolelli, F.: IM-Fuse: A Mamba-based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025 (2025)
63. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: ShapeLLM: Universal 3D Object Understanding for Embodied Interaction. In: European Conference on Computer Vision (2024)
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (2021)
65. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. In: International Conference on Machine Learning (2023)
66. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714 (2024)
67. Rekik, A., Ben-Hamadou, A., Smaoui, O., Bouzguenda, F., Pujades, S., Boyer, E.: TSegLab: Multi-stage 3D dental scan segmentation and labeling. Computers in Biology and Medicine **185** (2025)
68. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S.A., Baby, S.M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D.F., Kirmizibayrak, C., Pilgrim, R., Golden, D., Yang, L.: MedGemma Technical Report. arXiv:2507.05201 (2025)
69. Singh, A., et al.: OpenAI GPT-5 System Card. arXiv:2601.03267 (2026)

70. Song, X., Zhang, X., Ji, J., Liu, Y., Wei, P.: Cross-modal Contrastive Attention Model for Medical Report Generation. In: Proceedings of the 29th International Conference on Computational Linguistics (2022)
71. Sosnin, E.I., Vasilev, Y.L., Solovyev, R.A., Stempkovskiy, A.L., Telpukhov, D.V., Vasilev, A.A., Amerikanov, A.A., Romanov, A.Y.: AlphaDent: A dataset for automated tooth pathology detection. arXiv:2507.22512 (2025)
72. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 (2023)
73. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 (2023)
74. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., Virmani, S., Semturs, C., Mahdavi, S.S., Green, B., Dominowska, E., y Arcas, B.A., Barral, J., Webster, D., Corrado, G.S., Matias, Y., Singhal, K., Florence, P., Karthikesalingam, A., Natarajan, V.: Towards Generalist Biomedical AI. NEJM AI **1**(3) (2024)
75. Uribe, S.E., Issa, J., Sohrabniya, F., Denny, A., Kim, N., Dayo, A., Chaurasia, A., Sofi-Mahmudi, A., Büttner, M., Schwendicke, F.: Publicly Available Dental Image Datasets for Artificial Intelligence. Journal of Dental Research **103**(13) (2024)
76. Vahdati, S., Khosravi, B., Mahmoudi, E., Zhang, K., Rouzrokh, P., Faghani, S., Moassefi, M., Tahmasebi, A., Andriole, K.P., Chang, P., Farahani, K., Flores, M.G., Folio, L., Houshmand, S., Giger, M.L., Gichoya, J.W., Erickson, B.J.: A Guideline for Open-Source Tools to Make Medical Imaging Data Ready for Artificial Intelligence Applications: A Society of Imaging Informatics in Medicine (SIIM) Survey. Journal of Imaging Informatics in Medicine **37**(5) (2024)
77. Wang, C., Zhang, Y., Wu, C., Liu, J., Huang, X., Wu, L., Wang, Y., Feng, X., Lu, Y., Wang, Y.: MMDental-A multimodal dataset of tooth CBCT images with expert medical records. Scientific Data **12**(1) (2025)
78. Wang, P., Gu, H., Sun, Y.: Tooth segmentation on multimodal images using adapted segment anything model. Scientific Reports **15**(1) (2025)
79. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
80. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Dan, R., Zhang, Q., et al.: MICCAI 2023 STS Challenge: A retrospective study of semi-supervised approaches for teeth segmentation. Pattern Recognition **170** (2026)
81. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. Advances in Neural Information Processing Systems **35** (2022)
82. Wu, Y., Zhang, Y., Wu, Y., Zheng, Q., Li, X., Chen, X.: ChatIOS: Improving automatic 3-dimensional tooth segmentation via GPT-4V and multimodal pre-training. Journal of Dentistry **157** (2025)

83. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: PointLLM: Empowering Large Language Models to Understand Point Clouds. In: European Conference on Computer Vision (2024)
84. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 Technical Report. arXiv:2412.15115 (2025)
85. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y.: Evaluating progress in automatic chest X-ray radiology report generation. Patterns **4**(9) (2023)
86. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

Do Multimodal LLMs Understand Intraoral Dental Data? Dataset, Platform, and Baselines

Supplementary Material

7 Reproducibility and Training Details

We provide the specific implementation details for the IOS-Qwen architecture, data preprocessing, and the three-stage training pipeline used to integrate 3D intraoral scan (IOS) geometry with the Qwen3-VL backbone.

7.1 Data Preprocessing and Augmentation

Point Cloud Generation. Raw intraoral scan (IOS) meshes (STL) of the upper and lower arches are first oriented to a common reference system. The two meshes are then combined into a single point cloud, which is centered at the origin and normalized to lie within the unit sphere. At model input time, we apply Farthest Point Sampling (FPS) with a random starting point to downsample the point cloud to $N = 32{,}768$ points.

Augmentation. During training, we apply geometric augmentations to the point clouds:

- FPS of $N = 32{,}768$ with a random starting point;
- random rotations ($\pm 30^\circ$ around X/Y/Z axes);
- random jitter (Gaussian noise $\sigma = 0.01$);
- random translation (± 0.1);
- scaling (random from $0.95\times$ to $1.05\times$).

7.2 Model Architecture

3D Encoder. We utilize a PointTransformer-based encoder with an embedding dimension of 384, 12 transformer layers, and 6 attention heads per layer. The encoder utilizes point grouping (size 32) to generate 512 semantic geometry tokens from the input point cloud.

Multimodal Projection. A trainable Multi-Layer Perceptron (MLP) projector maps the 384-dimensional output of the PointTransformer into the embedding space of the LLM.

LLM Backbone. We utilize Qwen3-VL-4B-Instruct. To adapt the model to the orthodontic domain, we employ Low-Rank Adaptation (LoRA) on the language model's linear layers (attention Q, K, V, O, and MLP projections).

7.3 Training Pipeline

Our training strategy consists of three distinct stages to ensure robust geometric understanding before semantic alignment. These stages are described below.

Stage 1: Encoder Pre-training (Classification). To ground the 3D encoder in dental features, we first pre-train the PointTransformer (without the LLM) on a multi-task

Table 5: Hyperparameters for IOS-Qwen Training.

Parameter	Stage 1 (Pre-train)	Stage 2 (Align)	Stage 3 (LoRA)
Learnable Components	Encoder	Projector	Projector, LoRA
Optimizer	AdamW	AdamW	AdamW-8bit
Batch Size (Effective)	32	64	16
Learning Rate	$1e{-4}$	$3e{-4}$	$1e{-5}$ (LoRA)
Weight Decay	$1e{-3}$	$1e{-3}$	$1e{-2}$ (Proj)
Epochs	200	50	15
Warmup	25 epochs	200 steps	10 steps
Max Tokens	–	1024	512

classification objective. The encoder is trained to predict different clinical attributes (using those that could be easily standardized into a few classes and dropping the others, which were more verbose) using a Focal Loss ($\gamma = 3.0$). This stage runs for 200 epochs with a learning rate of 1×10^{-4}.

Stage 2: Modality Alignment. We freeze the pre-trained PointTransformer and the Qwen3 backbone, training only the projector. This aligns the geometry tokens with the LLM's text embedding space. The model is trained on the report generation task using a next-token prediction NLL loss.

Stage 3: Instruction Tuning (LoRA). In the final stage, we fine-tune the projector and the Qwen3-VL LoRA adapters ($r = 8, \alpha = 8$). The PointTransformer remains frozen to preserve the geometric features learned during Stage 1. This stage optimizes for the generation of structured clinical findings following the template described in the paper.

7.4 Hyperparameters and Hardware

All models were implemented in PyTorch using the HuggingFace `transformers` and `unsloth` libraries. Training was conducted on NVIDIA L40S GPUs. We used BFloat16 precision and gradient checkpointing to optimize memory usage. Detailed hyperparameters are listed in Table 5.

Inference. For generation, we use beam search (`num_beams=4`) with a temperature of 0.1 and a repetition penalty of 1.1 to ensure clinical consistency.

8 Web Application

8.1 System Architecture and Implementation

The system is implemented as a containerized web application composed of multiple cooperating services with health checks and persistent volumes:

- a Django web service providing authenticated views, a graphical user interface for visualizing different data modalities, and REST API endpoints;
- asynchronous worker services that handle modality-specific pipelines, including preprocessing, inference, and postprocessing tasks. Each worker runs in a GPU-enabled container configured with its supported data types and the required resources;
- a relational database used for metadata storage, indexing, and maintaining append-only audit logs.

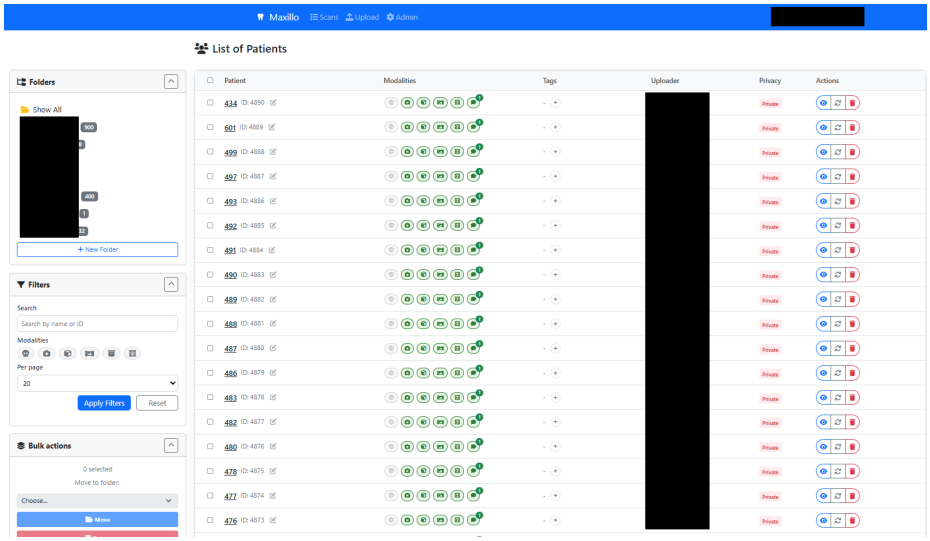


Fig. 5: Current project interface. The sidebar provides folder management and filters; the table shows anonymized subjects with modality availability, tags, uploader, privacy status, and role-aware actions. Identifiers are anonymized, and any headers are cropped for double-blind review.

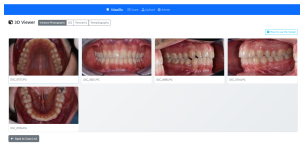
Worker containers operate independently and are automatically discoverable. A centralized queue manages job dispatching and load balancing across workers. Creating a new Docker container for a pipeline is simplified by using a predefined template, requiring only the implementation of the main logic.

8.2 Interactive Visualization

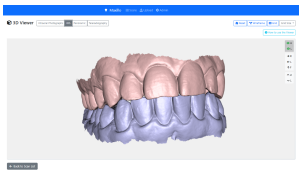
A 2D viewer provides synchronized multiplanar navigation for 3D CBCT volumes, with settings for window/level, measurements, and overlay channels to show all the available class labels. A 3D viewer uses Three.js to render IOS meshes with per-tooth overlays (Fig. 6). The other modalities are displayed as simple RGB images.

8.3 Screenshots

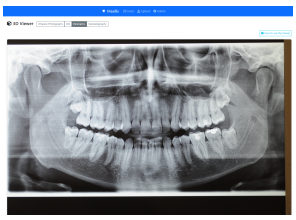
List of Patients Interface (Fig. 5). The listing view aggregates anonymized subjects in a tabular layout with role-aware actions. A collapsible sidebar on the left exposes *(i)* a hierarchical folder tree for organizing cases/datasets and *(ii)* filter controls, including free-text search, per-modality toggles, and pagination size. The central table displays, per row: a checkbox for bulk operations; an anonymous patient identifier; a compact modality strip where icons indicate the presence of CBCT, intraoral scans (upper/lower), panoramic radiograph, cephalometric radiograph, and photographs (green = correctly processed, yellow = processing, red = error during processing, grey = missing); optional tags; the uploader (anonymized in the figure); a privacy flag (default *Private*); and action buttons for opening the study, re-running the pipelines, or deleting a record (the latter



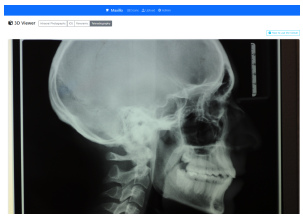
(a) Intraoral photographs (IOP).



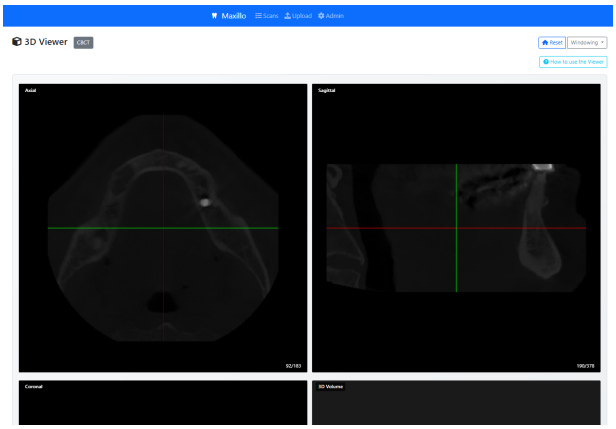
(b) Intraoral scan (IOS) meshes.



(c) Panoramic radiograph (OPG).



(d) Lateral teleradiography (CEPH).



(e) CBCT viewer with synchronized MPR panes.

Fig. 6: Example visual interfaces across modalities. (a) gallery of intraoral photographs with per-image thumbnails; (b) interactive 3D intraoral scan (IOS) viewer with separate arches; (c) panoramic radiograph (OPG); (d) lateral cephalometric radiograph; (e) cone-beam CT (CBCT) viewer showing synchronized axial/sagittal/coronal multiplanar reconstruction panes.

gated by permissions and audit-logged). Bulk actions at the bottom of the sidebar enable moving selected patients across folders while preserving provenance.

Example Visualizations. Fig. 6 illustrates the modality-specific viewers used in the platform. The interface supports (a) grid browsing of intraoral photographs; (b) 3D inspection of intraoral scans with separate upper/lower arches and view controls; (c–d) 2D radiographic review for panoramic and lateral cephalometric images; and (e) volumetric CBCT exploration with synchronized multiplanar panes.

Single-Study Details. Fig. 7 consolidates curation, metadata, and export controls for one study. The *Bite Classification* card supports AI suggestions and manual override across sagittal, vertical, and transverse axes with explicit midline entries. The *Voice Captions* area enables recording or editing of structured notes; saved entries appear in a time-ordered feed with modality badges (e.g., CBCT) and timestamps. The *Scan Settings* section surfaces pseudonymous subject/study identifiers, acquisition date, uploader (censored), visibility, dataset linkage, folder assignment, and tag editing; changes are permission-gated and audit-logged. The *NIfTI Metadata* panel provides orientation (e.g., RAS), data type, volume dimensions, voxel spacing, and an editable affine matrix, ensuring downstream pipelines receive spatially consistent inputs. Finally, *File Management* organizes downloadable raw and processed files for exports.

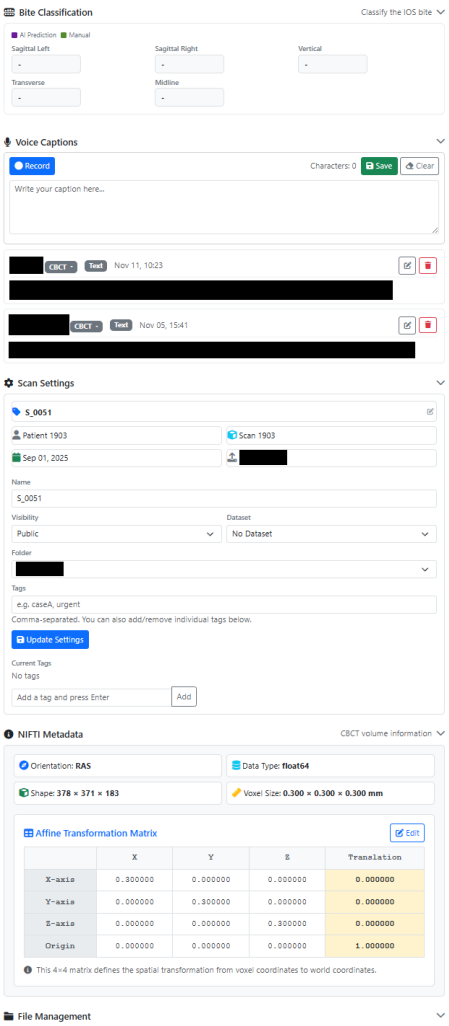


Fig. 7: Details view for a single study. From top to bottom: *(i) Bite Classification* with AI/Manual modes and fields for sagittal (left/right), vertical, transverse, and midline axes; *(ii) Voice Captions* panel with record/edit controls and a chronological feed of prior notes, each tagged with the originating modality and timestamp; *(iii) Scan Settings* summarizing pseudonymous identifiers (subject and study), acquisition date, uploader (censored), visibility, dataset linkage, folder assignment, and tag management; *(iv) NIfTI Metadata*; and *(v) File Management*.