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], **Federico Bolelli**[1]

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

UNIMORE Università degli Studi di Modena e Reggio Emilia · Università degli Studi di Ferrara · AImage Lab University of Modena and Reggio Emilia · ECCV 2026 | MALMÖ | SEPT 8–13

Dataset · Code · Yggdrasil · Challenge

MOTIVATION

The problem. Existing dental datasets are unimodal or loosely paired—no large-scale resource links registered IOS with photos and reports in a multicenter setting.

The research question. *Can multimodal LLMs reliably understand intraoral data and generate clinically faithful orthodontic reports?*

Our approach. *(i)* a new 1,000-patient **multimodal** and **multicenter dataset**, *(ii)* Yggdrasil, an open acquisition platform for dental data, *(iii)* a benchmark of state-of-the-art MLLMs, *(iv)* IOS-Qwen—a geometry-aware baseline.

DATASET AT A GLANCE

- ✓ 1,000 orthodontic patients (age 16–99, ~58% F / 42% M)
- ✓ 2,000 registered upper/lower IOS meshes
- ✓ 5,000 paired intraoral photographs
- ✓ 2,403 clinician-authored reports

Acquisition diversity:

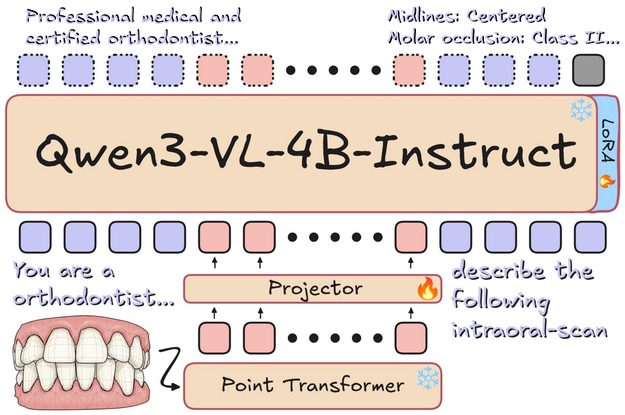
- ✓ Multicenter clinical acquisition
- ✓ Multiple scanner families (Carestream, 3Shape TRIOS)
- ✓ High-resolution RGB photography (≥24MP)
- ✓ Real-world orthodontic cases, no exclusion filtering

IOS-QWEN

A geometry-aware baseline for IOS report generation. A Point Transformer encoder turns the 3D scan into geometry tokens, fused into Qwen3-VL-4B to generate a template-conformant report.

Why? Off-the-shelf MLLMs cannot ingest full-arch, high-resolution IOS meshes.

Best recipe: LoRA fine-tuning of Qwen + projector (encoder frozen), field-by-field decoding.



STRUCTURED REPORT

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

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.

BEYOND TEXT SIMILARITY

STRUCTURED CORRECTNESS
ACCURACY
Are the predicted structured fields correct?

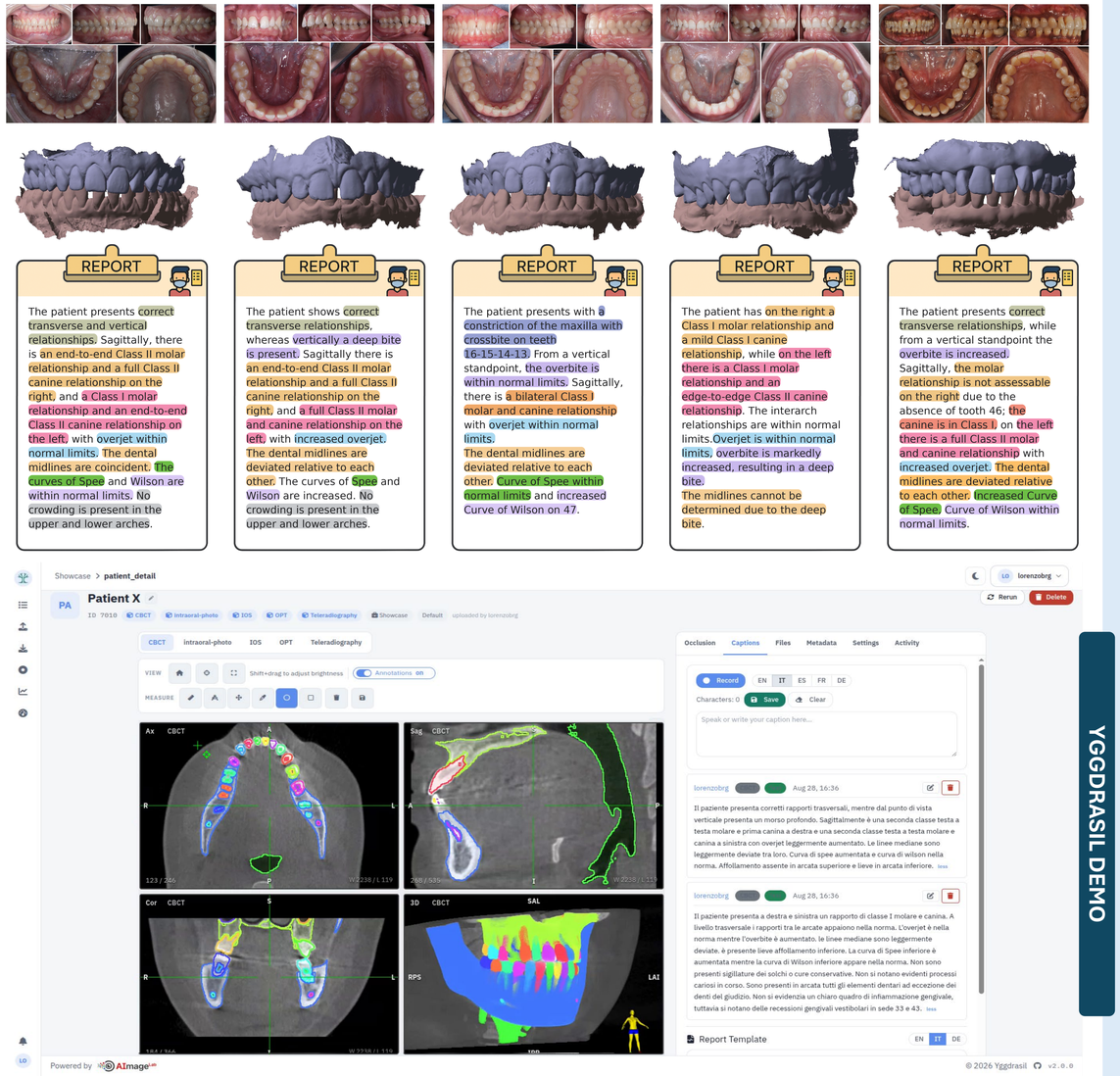
CLINICAL FACTUALITY
RADFACT F1 [1]
Are clinical statements supported?

SCHEMA COMPLETION
COVERAGE
How much of the required report schema is completed?

DOUBLE-BLIND A/B
CLINICIAN PREFERENCE
Which report do orthodontists prefer in a blinded pairwise comparison?

[1] Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., ... & Hyland, S. L. (2024). MAIRA-2: Grounded Radiology Report Generation. arXiv preprint arXiv:2406.04449.

DATASET DEMO



YGGDRASIL DEMO

BENCHMARK

Quantitative

	Model	BLEU-1↑	ROUGE-L (P)↑	ROUGE-L (R)↑	ROUGE-L (F1)↑	METEOR↑	Accuracy↑	Coverage↑	SBERT-sim↑	RadFact F1↑	Elo↑	W. Rate↑
IOP-Only	Qwen3-VL-4B-Instruct	0.757±0.087	0.774±0.080	0.726±0.100	0.746±0.083	0.777±0.088	0.580±0.156	0.996±0.026	0.928±0.027	0.358±0.210	934	0.419
IOP-Only	LLaVA-hf_LLaVA-v1.6-Mistral-7B-hf	0.678±0.058	0.705±0.061	0.732±0.086	0.715±0.061	0.710±0.073	0.541±0.107	**1.000±0.000**	0.942±0.020	0.316±0.156	870	0.324
IOP-Only	Google_MedGemma-4B-IT	0.654±0.101	0.815±0.075	0.623±0.101	0.702±0.080	0.682±0.092	0.514±0.160	**1.000±0.000**	0.887±0.044	0.218±0.257	1016	0.554
IOP-Only	Google_MedGemma-27B-IT	0.696±0.110	**0.824±0.067**	0.676±0.105	0.722±0.080	0.662±0.098	0.544±0.160	**1.000±0.000**	0.907±0.046	0.339±0.209	894	0.375
IOP-Only	Qwen3-VL-2B-Instruct	0.609±0.190	0.564±0.180	0.742±0.075	0.619±0.147	0.725±0.083	0.439±0.116	**1.000±0.000**	0.926±0.031	0.238±0.143	862	0.324
IOP-Only	Qwen2.5-VL-7B-Instruct	0.527±0.096	0.703±0.066	0.542±0.106	0.597±0.081	0.561±0.099	0.399±0.140	0.638±0.049	0.785±0.029	0.368±0.196	943	0.432
IOP-Only	GPT-5.2	0.657±0.170	0.740±0.161	0.619±0.167	0.670±0.160	0.665±0.174	0.447±0.186	0.749±0.229	0.884±0.152	0.318±0.192	989	0.514
IOP-Only	Gemini 3 Flash	0.549±0.244	0.784±0.120	0.551±0.190	0.632±0.162	0.593±0.207	0.411±0.202	0.741±0.239	0.863±0.081	0.339±0.212	998	0.514
IOS-Only	IOS-Qwen ♠ ‡	**0.781±0.082**	0.772±0.074	**0.754±0.083**	**0.763±0.056**	**0.790±0.079**	**0.632±0.110**	**1.000±0.000**	**0.959±0.020**	**0.372±0.182**	**1048**	**0.583**
IOS-Only	IOS-Qwen ♠ †	0.719±0.096	0.703±0.087	0.672±0.096	0.687±0.065	0.714±0.090	0.512±0.170	**1.000±0.000**	0.881±0.041	0.233±0.154	1030	0.541
IOS-Only	IOS-Qwen ♦ †	0.455±0.124	0.485±0.102	0.405±0.104	0.441±0.075	0.541±0.111	0.411±0.160	0.765±0.190	0.823±0.060	0.308±0.186	988	0.501
IOS-Only	IOS-Qwen ♣ †	0.448±0.132	0.645±0.114	0.409±0.108	0.501±0.088	0.527±0.116	0.372±0.160	0.743±0.220	0.788±0.066	0.282±0.191	973	0.472
	Inter-rater agreement	0.800±0.086	0.820±0.095	0.820±0.095	0.820±0.077	0.845±0.079	0.740±0.140	0.920±0.108	0.950±0.034	0.650±0.150	—	—
	Intra-rater agreement	0.880±0.088	0.900±0.090	0.900±0.090	0.900±0.078	0.915±0.075	0.860±0.127	0.950±0.091	0.970±0.029	0.810±0.161	—	—

Qualitative

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