

MMLVM

Geometry-Aware Multimodal Language Modeling for Orthodontic Report Generation

Juntao Han • Yunfei Zhou • Yufei Wang • Wenjun Tan

School of Computer Science and Engineering, Northeastern University, China
Key Laboratory of Medical Image Computing, Ministry of Education

Three specialists describe 11 findings. A fourth LoRA writes the report.
The report generator verifies preliminary findings against the original photographs and 3-D scans.

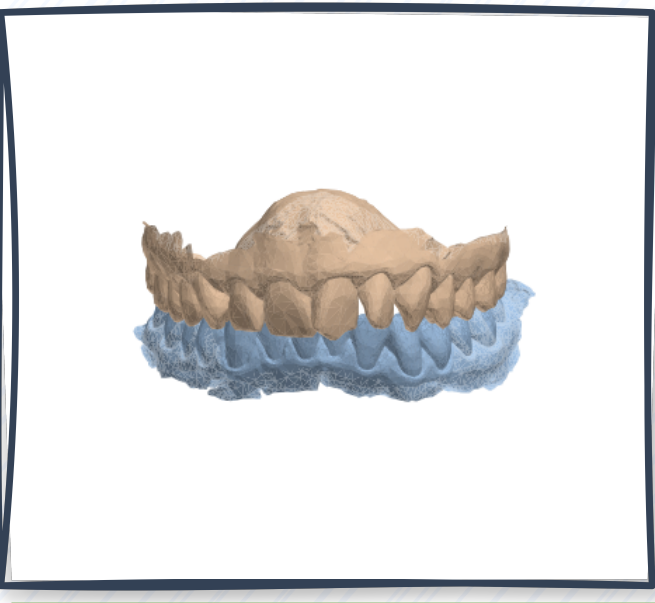
01 Official automatic challenge test results

RadFact precision	RadFact recall	RadFact F1	BLEU-4	METEOR	Composite score
0.4262	0.3960	0.3715	0.2902	0.4797	0.3742

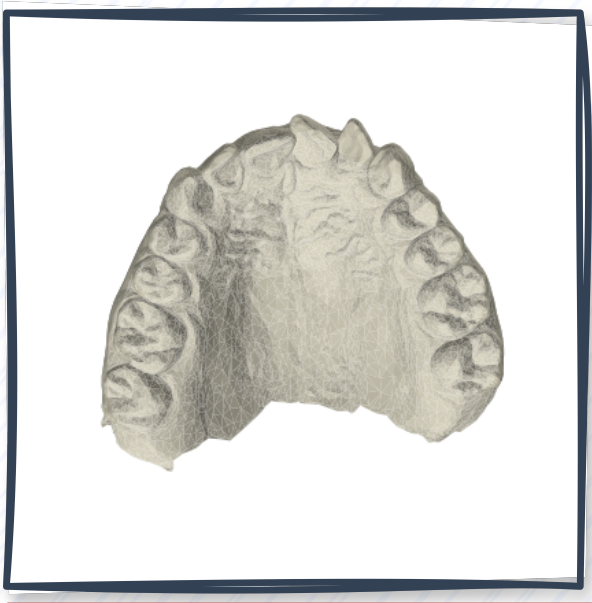
Final four-LoRA submission. RadFact Logical metrics; official scores supplied by the organizers.

02 Task & data: paired IOS geometry, photographs and expert reports

(a) 3-D IOS scans
Paired upper and lower surface meshes



Occluded arches



Upper arch



Lower arch

(b) Five complete intra-oral views
All TIFF frames are shown after proportional scaling; no image is cropped.



Frontal



Left buccal



Right buccal



Lower occlusal



Upper occlusal

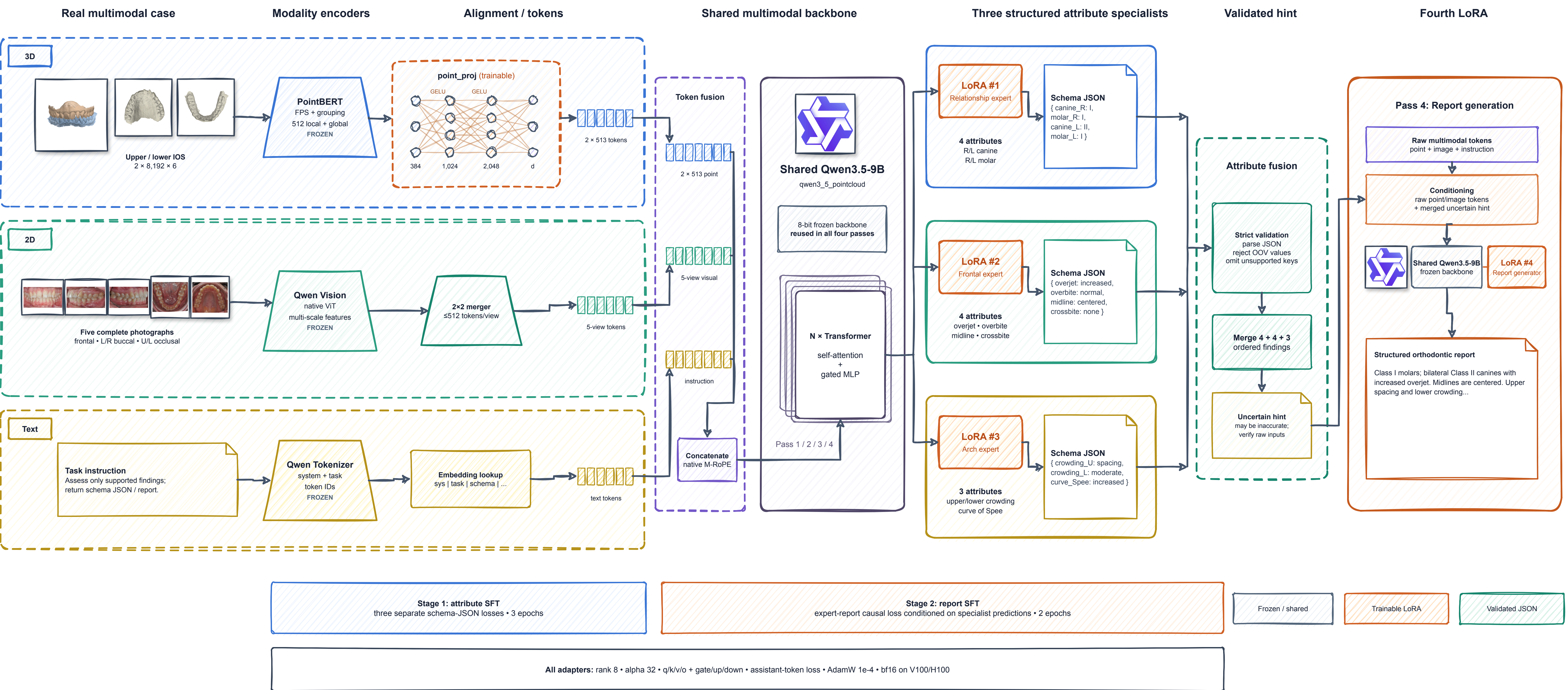
(c) Expert report
De-identified case-level target

The patient presents a slight constriction of the maxilla in the absence of crossbite. From a vertical standpoint, there is a deep bite. Sagittally, there is a weak Class I molar relationship and an edge-to-edge Class II canine relationship on the right, and a full Class II molar and canine relationship on the left with increased overjet. The midlines are centered with each other. The Curve of Spee and the Curve of Wilson are increased. There is moderate crowding in the upper arch and mild crowding in the lower arch.

Standard Sample Data

992 curated cases | 891 training / 98 validation / 3 qualitative cases
7,909 high-confidence attribute labels from photo-IOS report agreement (83.5% of 9,468 populated slots).

03 One shared backbone, four task-specific LoRA adapters



Real de-identified examples are shown; the three specialists and report generator reuse one frozen multimodal backbone.

Geometry + vision. Frozen PointBERT and the Qwen vision encoder; trainable geometry projector.
Inspect, then verify. Schema-validated specialist hints remain uncertain; the report adapter retains raw-modality access.
Deployment. Shared 8-bit backbone, approximately 15 GB peak memory; five-minute budget on a 24-GB A10G.
Limitations. Modest training cohort; missing or ambiguous views can cause correlated specialist errors.