



OrthoScribe: Training-Free Multimodal Orthodontic Report Generation via Photographic Retrieval & 3D Geometry Correction

Sanket Kachole^{1,2}, Spyridon Bakas^{1,2,3,4}

1. Division of Computational Pathology, Dept. of Pathology and Laboratory Medicine, Indiana University School of Medicine (IUSM), Indianapolis, USA · 2. IU Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, USA · 3. Biostatistics and Health Data Science; Radiology and Imaging Sciences; Neurological Surgery, IUSM · 4. Luddy School of Informatics, Computing, and Engineering, IU Indianapolis, USA

Background



~50%

of the global population has malocclusion; every case needs a structured diagnostic report



2 modalities

3D scans → occlusion (overjet, overbite)
Photographs → gingiva, restorations, teeth



Challenge goal

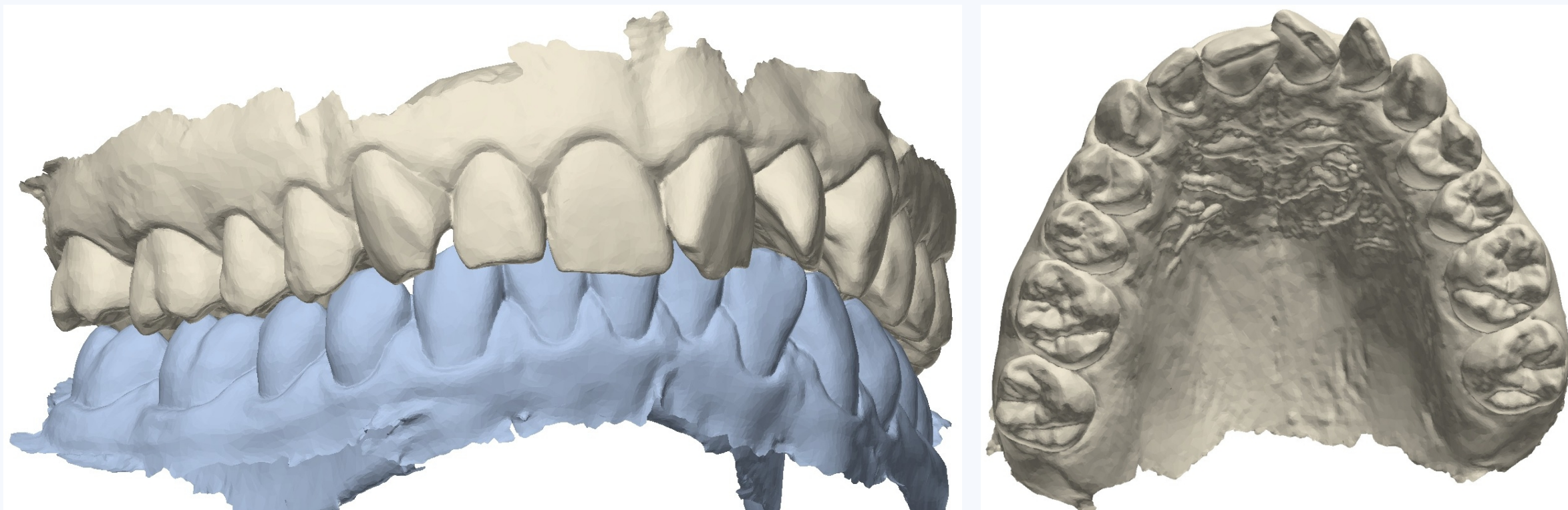
ODIN 2026 Bite2Text (Task 2): generate the orthodontic diagnostic report from a patient's 3D scans and intraoral photographs

Materials & Methods

Example case (F1979)



Intraoral photograph (1 of 5)



3D scans (upper + lower, in occlusion)

Expert orthodontic report

“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. ...”

Dataset: ODIN 2026 Bite2Text

997 cases released

872 with English photograph report

869 usable scan + photograph pairs

781 train / 88 validation

5,192

intraoral photographs (5 per case)

118 / 9

words / sentences per report

75%

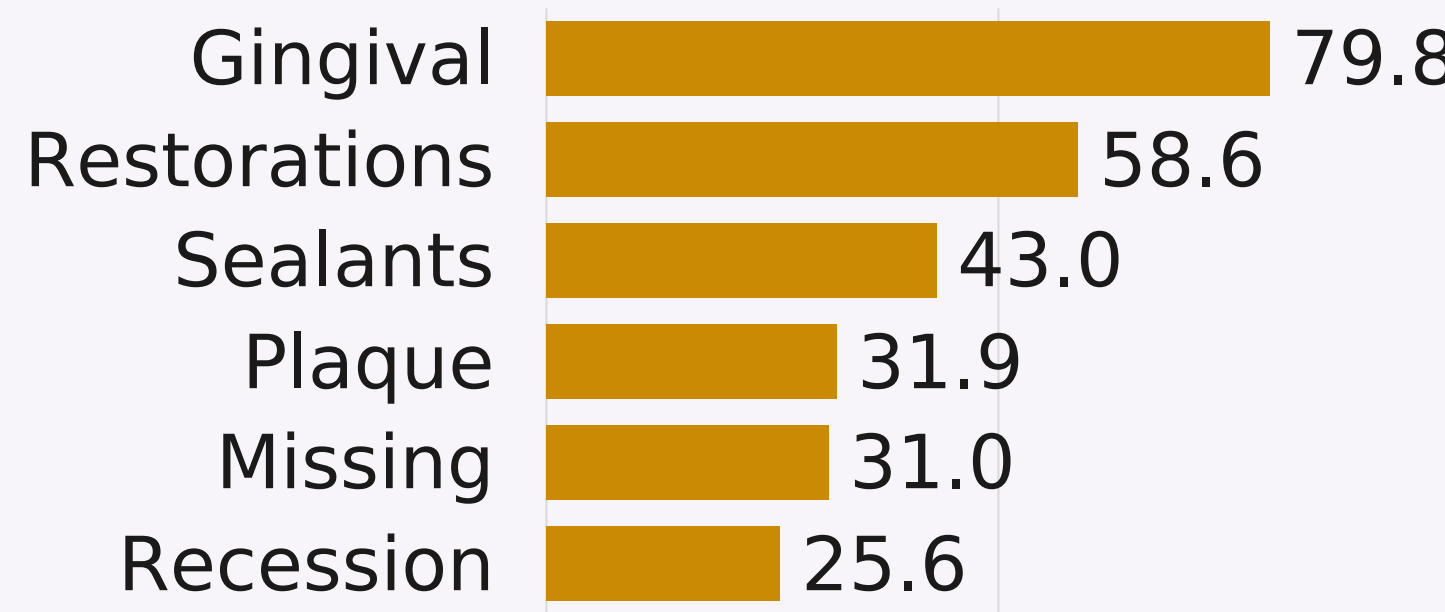
of reports name specific teeth

Descriptor prevalence in 872 photograph reports (%)

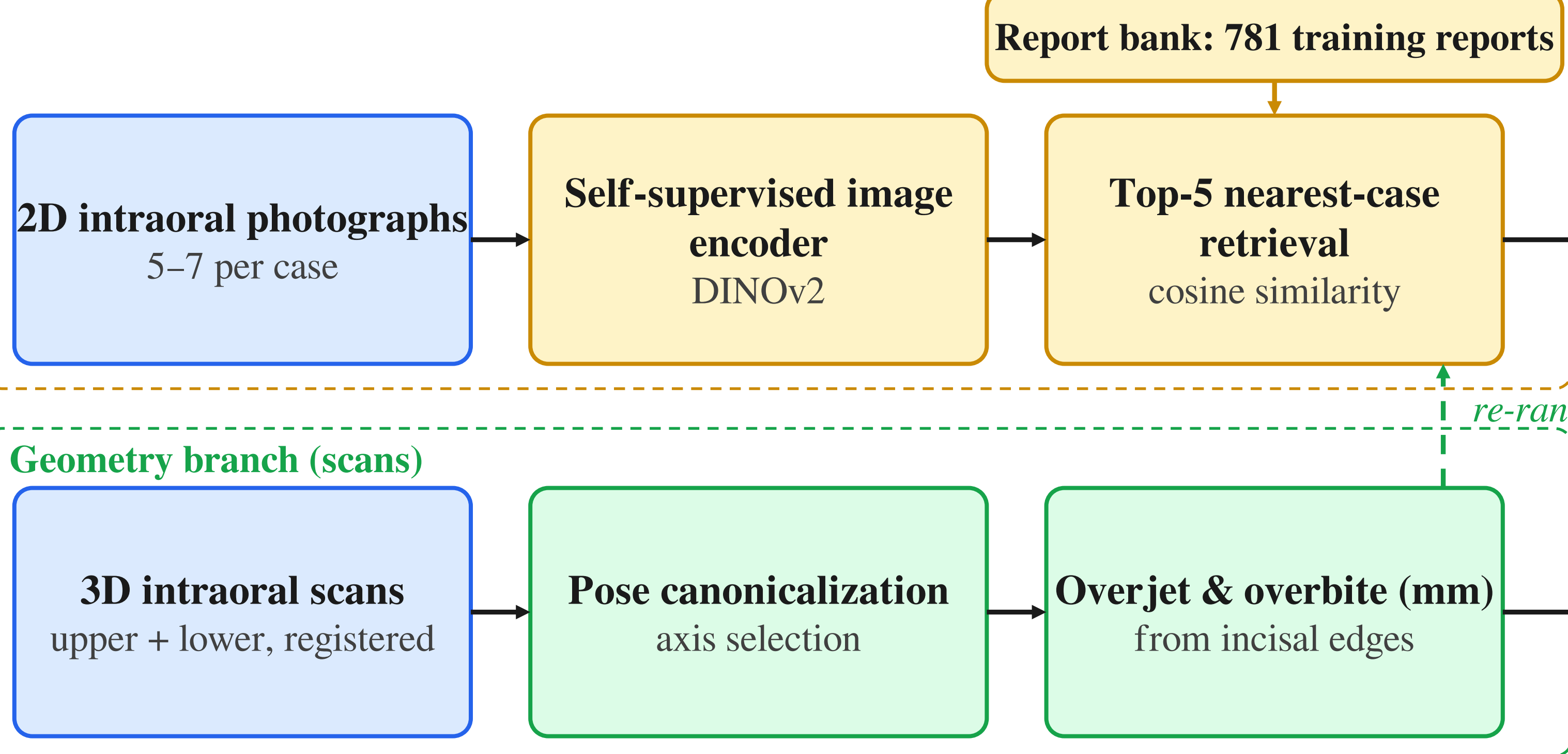
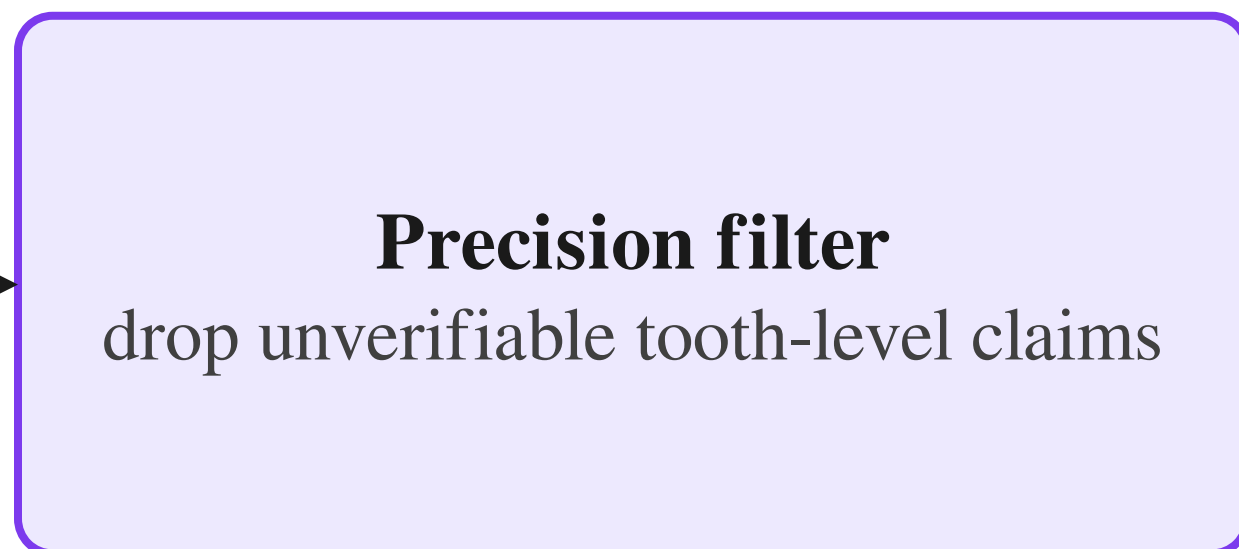
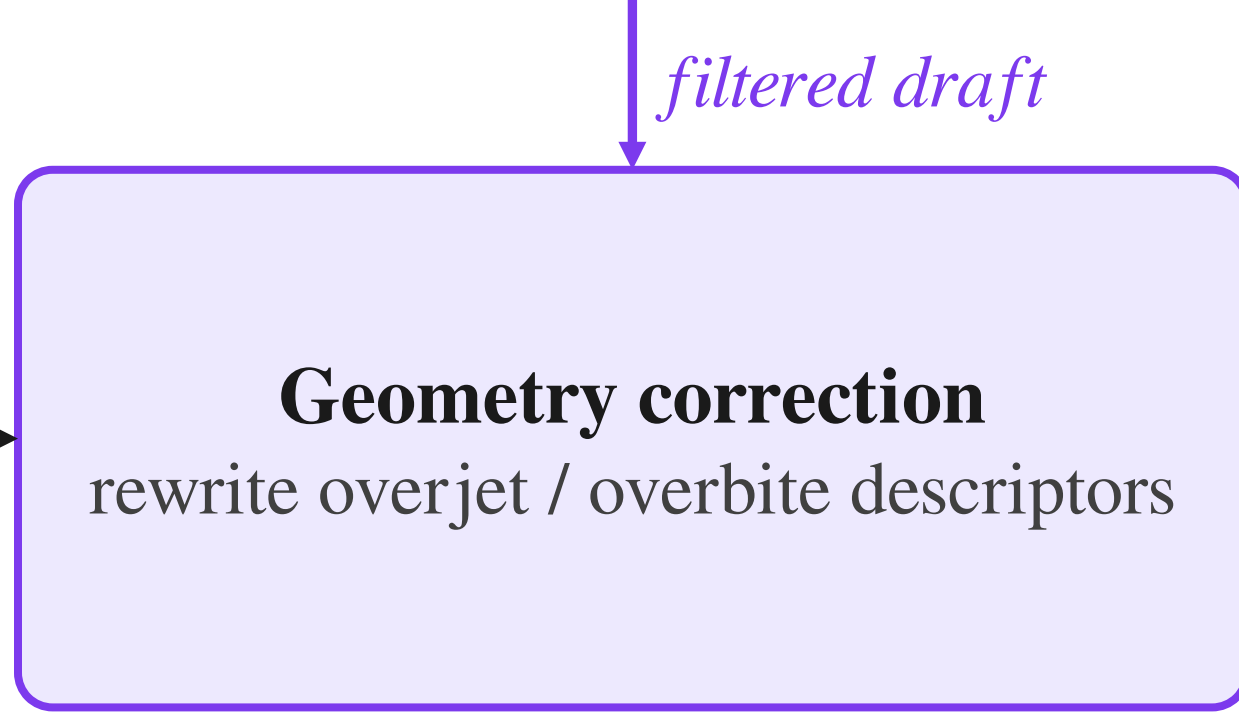
Occlusal (near-universal)



Photographic (variable)



Retrieval branch (photographs)



Generated report

Class I molar and canine relationship; overjet within normal limits; overbite reduced; midlines coincident; mild crowding in both arches; restorations present.

underlined = from 3D measurements

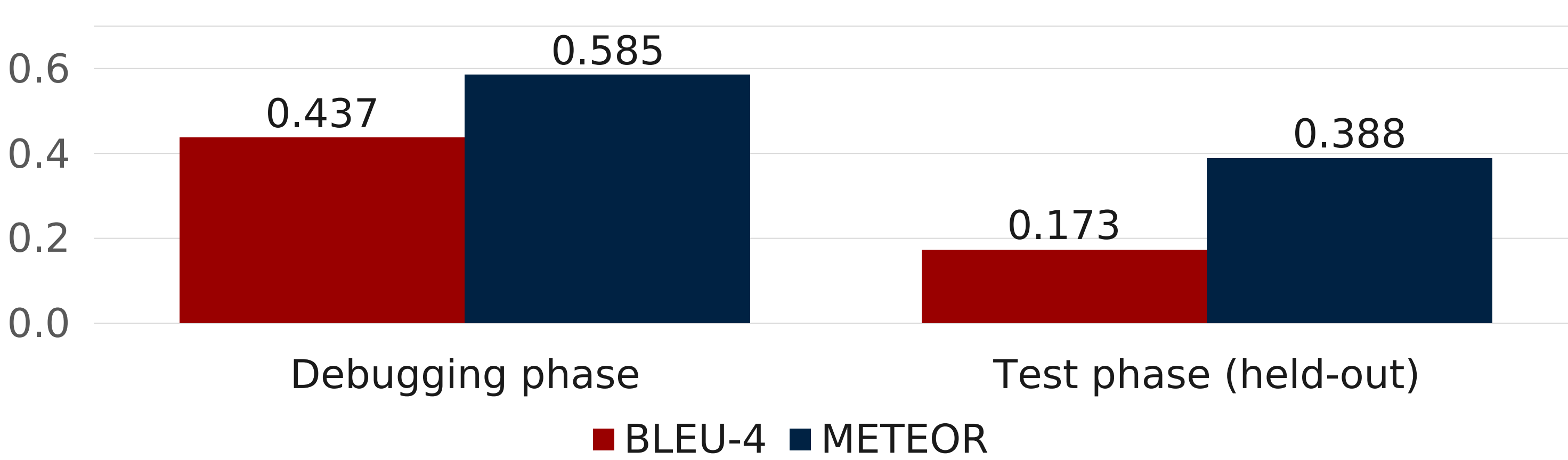
Results

Official Grand Challenge leaderboard (mean ± SD across cases)

Evaluation phase	BLEU-4	METEOR
Debugging phase	0.437 ± 0.402	0.585 ± 0.296
Test phase (held-out, multi-center)	0.173 ± 0.099	0.388 ± 0.109

Leaderboard shows lexical metrics only. Final ranking = $0.8 \times \text{RadFact F1} + 0.2 \times \text{mean}(\text{BLEU-4}, \text{METEOR})$; RadFact is computed offline by the organizers. Lower test-phase scores reflect generalization to unseen centers and acquisition conditions.

Leaderboard scores by phase



Conclusions

- Cross-modal division of labor: intraoral photographs retrieve a similar case's report as the draft, while the 3D scans measure overjet and overbite, re-rank the retrieval candidates, and overwrite the corresponding descriptors.
- Removing unverifiable tooth-level claims and correcting the measurable ones improves claim precision without losing recall, outperforming constant-report and fine-tuned vision-language baselines in validation factuality.
- Training-free and deployable within the challenge limits (Docker, no network, T4 GPU, under 5 min per case); lower lexical scores on the held-out multi-center test set show the cost of unseen centers and acquisition conditions.
- Limitations: dependence on the report bank, only two corrected descriptors, and no claim-level or expert evaluation on the test phase. Next: extend geometric correction to more descriptors and add orthodontist assessment.

References

- Lumetti, L., Rizzo, F., Cremonini, F., et al.: Do multimodal LLMs understand intraoral dental data? Dataset, platform, and baselines. In: ECCV (2026)
- Oquab, M., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
- Bannur, S., et al.: MAIRA-2: Grounded radiology report generation. arXiv:2406.04449 (2024)
- Bai, S., et al.: Qwen2.5-VL technical report. arXiv:2502.13923 (2025)
- Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. arXiv:2106.09685 (2021)
- Lombardo, L., et al.: Worldwide prevalence of malocclusion in the different stages of dentition: A systematic review and meta-analysis. Eur J Paediatr Dent 21(2), 115–122 (2020)
- Papineni, K., et al.: BLEU: A method for automatic evaluation of machine translation. In: ACL, pp. 311–318 (2002)
- Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop, pp. 65–72 (2005)

Contact:
skachole@iu.edu

