Clinically Guided Rendering for Orthodontic Report Generation

Jiashuo Guo[1], Jing Hao[2], Kuo Feng Hung[2], and Feng Wang[1*]

[1] Division of Oral and Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China
[2] Division of Applied Oral Sciences & Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China
diawang@hku.hk

Abstract. Automatic orthodontic report generation requires the joint interpretation of standardized intraoral photographs and three-dimensional intraoral scans (IOS). However, most multimodal large language models (MLLMs) are designed for two-dimensional visual inputs, making it difficult to incorporate IOS geometry without a dedicated 3D encoder. We present an anatomy-aware rendering and visual fusion framework that converts maxillary and mandibular IOS meshes into clinically meaningful 2D representations. The framework establishes a standardized occlusal coordinate system, renders twelve views targeting whole-arch morphology and local occlusal relationships, and focuses the combined bite views on the dentition before integrating the renderings with intraoral photographs. This design enables an existing dental-specialized MLLM to utilize complementary 3D geometric and 2D appearance information. Under the official Bite2Text evaluation against photograph-variant references, occlusal-region focusing raises the final score of the controlled twelve-view system from 0.3004 to 0.3363. Continuing fine-tuning from this focused model with photograph-report targets and applying topic-gated recall completion further raises the final score to 0.3811, outperforming the zero-shot GPT-4o, GPT-5.5, and Qwen3-VL-8B baselines. These results demonstrate that clinically guided 3D-to-2D rendering and reference-target alignment provide an effective and practical interface between IOS data and existing dental MLLMs.

Keywords: Orthodontic reporting · Intraoral scans · Multi-view rendering · Multimodal LLMs

1 Introduction

Multimodal large language models (MLLMs) are widely used for medical image understanding and report generation [13, 5]. Dental AI has similarly progressed from multimodal benchmarks and instruction data [8] to specialized MLLMs such as OralGPT-Omni [9], visual-tool-using models [6], and interactive agents [7]. However, open oral-maxillofacial imaging resources remain heterogeneous in modality, licensing, and ethical constraints [10]. Orthodontic reporting depends not only on appearance cues in photographs, but also on 3D arch

morphology and occlusal relationships from intraoral scans (IOS). The MICCAI 2026 ODIN Bite2Text challenge formalizes this setting as a multimodal report generation task: a model must generate a structured English orthodontic report from maxillary and mandibular IOS meshes and standardized intraoral photographs [15].

Existing MLLM visual encoders mainly take 2D images and cannot consume IOS meshes directly. Training a dedicated 3D encoder makes it hard to reuse dental knowledge already learned by 2D MLLMs. Rendering IOS into multi-view images is more practical, but scanner orientation varies across centers, generic views may miss local occlusal findings, and low-relevance structures such as the palatal vault waste visual tokens. How to preserve clinically discriminative 3D information under a limited visual context is therefore a key step in connecting IOS data to 2D MLLMs.

We cast IOS as anatomy-aware 2D views for an existing dental MLLM. We estimate a canonical occlusal frame from the registered arches, render twelve clinically targeted views, focus the bite views on the occlusal band, and combine the renderings with intraoral photographs to fine-tune OralGPT-Omni with LoRA [11]. Figs. 1 and 2 illustrate the rendered views and the focusing effect for one case. On the Bite2Text validation set, our best system reaches a final score of 0.3811, above zero-shot GPT-5.5 (0.2287), GPT-4o (0.2211), and Qwen3-VL-8B (0.2007) under the same official evaluation protocol [16, 12, 20].

Our main contributions are as follows:

- We present a 3D–2D multimodal framework for orthodontic report generation that connects IOS geometry to an existing dental MLLM without introducing a task-specific 3D encoder.
- We design twelve-view rendering and occlusal-region focusing to emphasize clinically relevant arch morphology and occlusion under a limited token budget, and combine them with intraoral photographs as a unified visual input.
- We validate the framework on Bite2Text against proprietary and open-source zero-shot MLLMs and controlled rendering ablations.

2 Method

2.1 Framework Overview

Given maxillary and mandibular IOS meshes $\mathcal{M}_u = (\mathcal{V}_u, \mathcal{F}_u)$, $\mathcal{M}_l = (\mathcal{V}_l, \mathcal{F}_l)$ and intraoral photographs $\mathcal{P} = \{P_i\}_{i=1}^{N_p}$ ($N_p \leq 5$), we generate a structured English orthodontic report $\mathbf{y} = (y_1, \ldots, y_T)$. The pipeline has three stages: anatomical coordinate normalization from the occluded arches; twelve fixed 2D views with occlusal-region focusing on combined bite views; and fixed-order assembly of photographs and renderings $\mathcal{I} = [\mathcal{P}; \mathcal{R}]$ ($\mathcal{R} = \{R_j\}_{j=1}^{12}$) for fine-tuning of OralGPT-Omni. The same preprocessing is used at training, evaluation, and inference.

2.2 Anatomical Coordinate Normalization

IOS data from different centers do not follow a consistent export orientation. In some cases, the upper and lower arches are separated along the Z-axis, whereas in others they are separated along the Y-axis. Thus, identical camera directions may yield different anatomical views. Since Bite2Text arches are already aligned in occlusion, we estimate the anatomical frame from their relative positions without additional ICP registration.

We define the origin as the mean of all vertices:

$$\mathbf{c} = \frac{\sum_{\mathbf{v}\in\mathcal{V}_u} \mathbf{v} + \sum_{\mathbf{v}\in\mathcal{V}_l} \mathbf{v}}{|\mathcal{V}_u| + |\mathcal{V}_l|}. \tag{1}$$

Let $\mathbf{d} = \boldsymbol{\mu}_u - \boldsymbol{\mu}_l$ denote the difference between the maxillary and mandibular centroids. The coordinate axis corresponding to the largest absolute component of $\mathbf{d}$ is selected as the occlusal axis. Its sign is set to point from the mandible toward the maxilla, producing $\mathbf{e}_z$. Among the two remaining axes, the axis with the larger extent in the joint bounding box is assigned as the left–right axis $\mathbf{e}_x$, while the other becomes the anteroposterior axis $\mathbf{e}_y$.

The sign of the anteroposterior axis is determined from the arch shape. The anterior arch is closed near the dental midline, whereas the posterior segments separate toward the two sides. We therefore set the end containing a higher proportion of near-midline vertices as the anterior direction. The left–right sign is adjusted to ensure that the rotation matrix $\mathbf{R} = [\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z]$ satisfies $\det(\mathbf{R}) > 0$, and each vertex is transformed as $\mathbf{v}' = (\mathbf{v} - \mathbf{c})\mathbf{R}$. This procedure provides consistent superior–inferior, anteroposterior, and left–right axes for fixed-camera rendering. Since absolute left–right handedness cannot be uniquely recovered from mirror-symmetric geometry alone, we retain the scanner's native handedness and use the same convention throughout training and inference.

2.3 Clinically Targeted Multi-view Rendering

We construct twelve fixed IOS views: two per-arch occlusal views for arch form, crowding, and spacing; three standard bite views for midlines, overbite, and bilateral occlusion; three oblique and anterosuperior views for posterior exposure and transverse or vertical relationships; and four regional close-ups for molar or canine Angle class, overbite, and overjet.

All views are rendered using an orthographic camera to avoid perspective-related changes in relative dental geometry. The camera frustum is adapted to the bounding box of the target mesh or local region, with a margin factor of 1.08. Each view is rendered at 768×768 pixels on a white background. In the combined bite views, the maxillary mesh is shown in a warm color and the mandibular mesh in a cool color to make the occlusal interface easier to distinguish. For the close-up views, the camera center and scale are determined from the local bounding box of either the anterior region or one of the posterior buccal segments rather than from the complete arches. Fig. 1 shows the twelve views for one case after occlusal-region focusing.

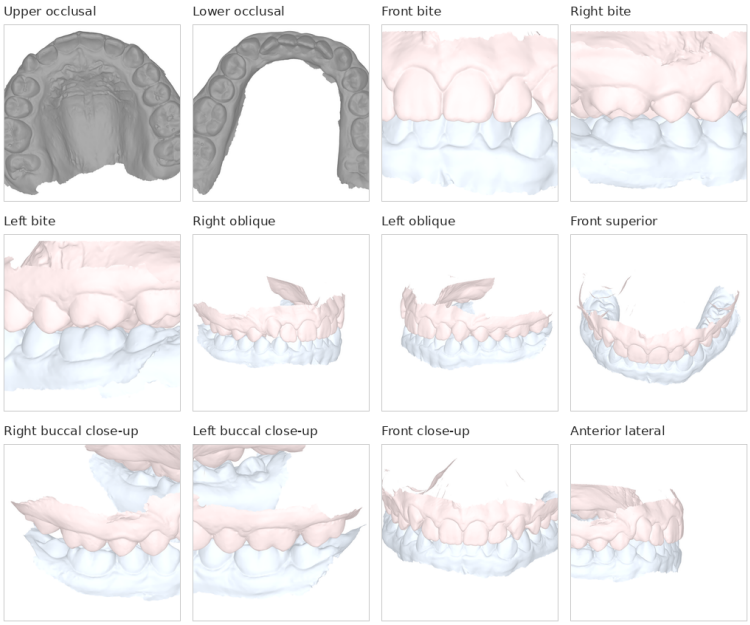

Fig. 1. Twelve clinically targeted IOS views after occlusal-region focusing.

2.4 Occlusal-region Focusing

In unprocessed frontal and lateral bite views, the maxillary palatal vault and mandibular base may occupy a large part of the image, leaving fewer pixels for the crowns and occlusal interface after image resizing (Fig. 2, left). To reduce this use of the visual budget, we retain only a mesh band around the occlusal plane for views that contain both arches. After coordinate normalization, the maxillary occlusal surface lies near the minimum Z value of the upper mesh, while the mandibular occlusal surface lies near the maximum Z value of the lower mesh. The retained vertices are therefore defined as

$$\begin{aligned} \widetilde{\mathcal{V}}_u &= \{\mathbf{v} \in \mathcal{V}_u \mid z(\mathbf{v}) \leq z^u_{\min} + b\}\,, \\ \widetilde{\mathcal{V}}_l &= \{\mathbf{v} \in \mathcal{V}_l \mid z(\mathbf{v}) \geq z^l_{\max} - b\}\,, \end{aligned} \tag{2}$$

where $b = 16\,\mathrm{mm}$. A triangular face is retained only if all three of its vertices satisfy the corresponding condition. As a safeguard against incomplete meshes or unexpected scan units, the original mesh is used when fewer than 15% of its vertices would be retained. The clipping operation is not applied to the separate maxillary and mandibular occlusal views, which preserve the complete arch geometry. Fig. 2 compares a right lateral bite view before and after focusing.

The rendered images are stored at 768×768 pixels and are limited by the visual processor to at most 327,680 pixels per image before being passed to the

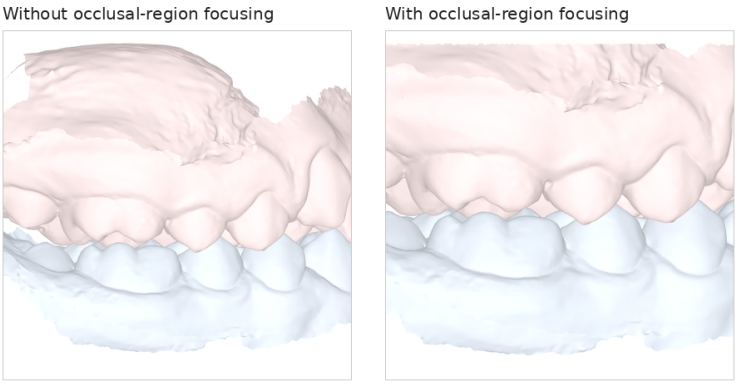

Fig. 2. Right lateral bite view before and after occlusal-region focusing.

model. Occlusal-region focusing therefore does not increase the number of input images or the visual-token limit. Instead, it allocates a larger proportion of the fixed pixel budget to the crowns and occlusal interface.

2.5 Multimodal Fusion and Reference-target Alignment

Appearance and geometric cues are fused within the 2D MLLM by representing each case as an ordered visual sequence $\mathcal{I} = [\mathcal{P}; \mathcal{R}]$. Up to five intraoral photographs occupy the first slots, followed by the twelve IOS renderings in a fixed anatomical order. This consistent ordering supplies implicit view identity and allows the model to associate photographic surface appearance with the corresponding 3D arch morphology and occlusal relationships.

Supervision is aligned with the photograph-report variant used by the official evaluation. Starting from the twelve-view checkpoint with occlusal-region focusing, we continue fine-tuning on photograph-variant reports while leaving the multimodal input unchanged. At inference, a topic-gated completion step appends up to four corpus-derived sentences covering crowding, transverse relationships, gingival status, and plaque or calculus only when the generated report does not already mention the corresponding topic.

3 Experiments

3.1 Dataset and Evaluation Metrics

We conduct experiments on the MICCAI 2026 ODIN Bite2Text dataset [15]. The released data contain approximately 1,000 orthodontic cases. Each case provides registered maxillary and mandibular IOS meshes in occlusion, standardized intraoral photographs, and clinician-authored English reports. After completeness

checks, 991 patient-level cases are split into 961 for training and 30 for validation. Multiple observer reports are treated as separate targets, yielding 1,455 IOS-report training samples. The official photograph-reference evaluation is performed on the 29 validation cases with an available photograph-variant report. Samples sharing the same visual input are grouped at the case level, and the model generates one report per case.

We adopt the official Bite2Text scoring functions. Report similarity is measured using BLEU-4 [17] and METEOR [1]; their arithmetic mean defines the captioning score S_{cap}. Clinical correctness is evaluated using RadFact logical F1 [2]. RadFact decomposes the generated and reference reports into clinical statements and evaluates bidirectional entailment. Denoting the harmonic mean of logical precision and recall by S_{clin}, the official final score is

$$S_{\text{final}} = 0.8S_{\text{clin}} + 0.2S_{\text{cap}}. \tag{3}$$

Evaluation protocol. The release contains two report streams: an IOS variant (`reports_ios_en`, approximately 82 words and focused on occlusal findings) and, when available, a photograph variant (`reports_intraoral-photo_en`, approximately 125 words and additionally covering soft tissue and oral hygiene). All results in this paper are recomputed against the photograph variant used by the challenge evaluation. BLEU-4 and METEOR are calculated with the official NLTK-based implementation, while clinical correctness is computed with the official `radfact_lite` pipeline using GPT-4o-mini. Thus, every reported row uses the same 29 cases, reference variant, and metric implementation.

3.2 Implementation Details

Our main model is OralGPT-Omni, which is based on Qwen2.5-VL-7B [9]. The vision tower and vision–language projector are frozen, and LoRA is applied to the language model. We use a LoRA rank of 16, a scaling factor of 32, and a dropout rate of 0.05. The per-device batch size is 1 with eight gradient-accumulation steps, giving an effective batch size of 8. The initial learning rate is 1×10^{-4}, with cosine scheduling and a 3% warm-up ratio. Training uses bfloat16 precision and gradient checkpointing. Rendering ablations are trained for up to three epochs. For reference-target alignment, we start from the focused twelve-view model and continue training on 843 photograph-report samples for up to two epochs with a reduced learning rate of 5×10^{-5}. Training is performed on one NVIDIA A100 80GB GPU.

GPT-4o and GPT-5.5 are evaluated through zero-shot API inference [12, 16], while Qwen3-VL-8B is evaluated locally without fine-tuning [20]. All three baselines receive the same multimodal input and report instruction, without example reports or additional clinical information.

3.3 Comparison with General Purpose MLLMs

Table 1 compares our method with proprietary and open-source general purpose MLLMs. GPT-4o, GPT-5.5, and Qwen3-VL-8B directly generate reports

Table 1. Comparison with general purpose MLLMs.

Model	Adaptation	Clinical F1	BLEU-4	METEOR	Final
GPT-4o	Zero-shot	0.2341	0.0607	0.2775	0.2211
GPT-5.5	Zero-shot	0.2457	0.0383	0.2837	0.2287
Qwen3-VL-8B	Zero-shot	0.2109	0.0548	0.2650	0.2007
OralGPT-Omni (ours)	**LoRA SFT**	**0.3691**	**0.3389**	**0.5193**	**0.3811**

Table 2. Representative report comparison. Blue marks findings supported by the reference; red marks errors.

OralGPT-Omni (ours)	**GPT-5.5 (zero-shot)**
The patient presents a **transverse constriction of the maxilla in the absence of crossbite**. From a vertical standpoint, there is a **deep bite**. Sagittally, there is a **full bilateral Class II** molar and canine relationship with **increased overjet**. The **dental midlines are deviated** relative to each other. The **Curve of Spee** and the **Curve of Wilson are increased**. There is **mild crowding in the upper arch** and **moderate crowding in the lower arch**. The gingivae are inflamed. Plaque/calculus is present in the posterior sectors.	**Correct transverse relationships**; **no posterior crossbite is present**. **Overbite is within normal limits**. Right side: Class I molar relationship; Class I canine relationship. Left side: **Class I molar relationship**; **Class I canine relationship**. **Overjet is within normal limits**. Upper and lower dental midlines are **approximately coincident**. The **Curve of Spee is within normal limits**; the **Curve of Wilson is within normal limits**. The maxillary arch has **moderate anterior crowding**; the mandibular arch has **mild anterior crowding**. Other visible findings include fissure staining/sealants on posterior teeth, mild plaque/staining, and no clear cavitated carious lesions.

in a zero-shot setting, while our final system uses LoRA SFT, reference-target alignment, and topic-gated completion.

Our method achieves a final score of 0.3811, exceeding GPT-5.5, GPT-4o, and Qwen3-VL-8B by 0.1524, 0.1600, and 0.1804, respectively. The zero-shot models generate semantically reasonable orthodontic descriptions, but their wording and organization differ substantially from the reference reports, leading to BLEU-4 scores at or below 0.061. Their clinical F1 scores are also lower than that of our adapted dental model. These results show that zero-shot general-purpose MLLMs do not fully meet the requirements for both clinical correctness and structured orthodontic reporting.

Table 2 uses a case tied for our highest official per-case clinical F1. Our report captures the reference-supported deep bite, increased overjet, midline deviation, increased curves, and upper/lower crowding distribution, while missing the localized crossbite and overstating the Class II relationship as bilateral. GPT-5.5 instead normalizes most findings and reverses the crowding severity between arches.

3.4 Ablation Study

Table 3 studies the visual input and training designs under the same data split, image-pixel limit, and LoRA configuration.

In the clean single-variable comparison, increasing from eight to twelve views raises the final score from 0.2976 to 0.3004. The gain is driven by captioning

Table 3. Ablation results for rendering and training configurations.

Configuration	Clinical F1	BLEU-4	METEOR	Caption	Final
8 rendered views	0.2896	0.2402	0.4193	0.3297	0.2976
12 rendered views (controlled)	0.2869	0.2674	0.4420	0.3547	0.3004
12 views with per-image text legends	0.3133	0.2210	0.4040	0.3125	0.3131
12 views with occlusal-region focusing	0.3272	0.2844	0.4610	0.3727	0.3363
Focused Reference SFT with Topic Gating	**0.3691**	**0.3389**	**0.5193**	**0.4291**	**0.3811**

(0.3297 to 0.3547), while clinical F1 remains similar and decreases slightly from 0.2896 to 0.2869. Thus, additional targeted views provide a modest benefit.

Occlusal-region focusing produces the largest visual-input improvement, increasing clinical F1 from 0.2869 to 0.3272 and the final score from 0.3004 to 0.3363. Per-image legends produce a mixed effect: they improve clinical F1 and final score relative to the unlabeled twelve-view setting but reduce captioning from 0.3547 to 0.3125 and remain below occlusal-region focusing. These findings support retaining fixed view order while allocating more of the pixel budget to the crowns and occlusal interface.

3.5 Reference-target Alignment

The controlled rendering experiments above are trained with IOS-variant reports so that only the visual input changes. However, the official evaluation uses the longer photograph-report variant. To reduce this training–evaluation mismatch without discarding the occlusal knowledge learned from all IOS reports, we initialize from checkpoint-400 of the focused twelve-view model and continue LoRA fine-tuning on the 843 training samples with photograph-variant targets. This preserves the focused visual input while aligning report style and clinical coverage with the evaluation target.

As shown in Table 3, reference-aligned SFT followed by topic-gated completion raises clinical F1 from 0.3272 to 0.3691 and the final score from 0.3363 to 0.3811 over its focused IOS-supervised source model. The sequential improvement shows that concentrating the visual budget on the occlusal region and aligning language supervision with the photograph-report target provide compatible gains.

3.6 Limitations and scope.

Our contribution is primarily a clinically guided rendering and adaptation interface rather than a new multimodal architecture. Validation on 29 cases may limit generalizability across sites and patient populations. GPT-4o and GPT-5.5 are proprietary API-based systems whose parameters and training procedures cannot be matched to our fine-tuned model; thus, all baselines were evaluated zero-shot with the same multimodal inputs and task instruction. Larger multi-site evaluations and more controlled baselines would strengthen the evidence within the applied challenge-track scope.

4 Conclusion

We presented an anatomy-aware multi-view rendering and visual fusion framework for orthodontic report generation. Without introducing a dedicated 3D encoder, the framework standardizes IOS meshes in an occlusal coordinate system, converts the maxillary and mandibular arches into twelve clinically meaningful 2D views, and combines occlusal-region focusing with standardized intraoral photographs. This design enables a dental-specialized MLLM to jointly utilize 3D arch morphology, occlusal relationships, and 2D surface appearance. Under the unified official evaluation, occlusal-region focusing produces the largest visual-input gain. Continuing reference-aligned SFT from this focused model and applying topic-gated completion further raises the final score to 0.3811. The resulting system outperforms the zero-shot GPT-4o, GPT-5.5, and Qwen3-VL-8B baselines. These results establish the effectiveness of reusing existing dental MLLMs through clinically driven 3D-to-2D visual representations and task-aligned language supervision for automated orthodontic report generation.

Acknowledgments. This work was developed and evaluated within the ODIN 2026 challenge cluster, including the STS, ToothFairy, and Bite2Text challenges. We acknowledge the associated challenge and dataset work [19, 14, 15, 18, 4, 3].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. 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 (2024), https://arxiv.org/abs/2406.04449
3. 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., Kr"amer, 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., M"uller, H., Mai, H., Dang, X., Bhandary, S., Grosu, R., Berg"e, S., Anesi, A., Grana, C.: Multi-structure segmentation in cbct volumes: The toothfairy2 challenge. Medical Image Analysis p. 104095 (2026). https://doi.org/10.1016/j.media.2026.104095
4. 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 Workshop - MICCAI 2025 (Sep 2025)

5. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: HuatuoGPT-Vision: Towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
6. Fan, Y., Hao, J., Chen, H., Bao, J., Shao, Y., Liang, Y., Hung, K.F., Tang, H.: Oralgpt-plus: Learning to use visual tools via reinforcement learning for panoramic x-ray analysis. CVPR 2026 (2026)
7. Hao, J., Dai, S., Zhang, Y., Liang, Y., Wu, J., Bao, J., Fan, Y., Ye, Z., Sun, Y., Zhang, X., et al.: Oralagent: Integrating reasoning, tools, and knowledge for interactive dental image analysis. arXiv preprint arXiv:2605.27378 (2026)
8. Hao, J., Fan, Y., Sun, Y., Guo, K., Lin, L., Yang, J., Ai, Q.Y.H., Wong, L.M., Tang, H., Hung, K.F.: Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. NeurIPS 2025 (2025)
9. Hao, J., Liang, Y., Lin, L., Fan, Y., Zhou, W., Guo, K., Ye, Z., Sun, Y., Zhang, X., Yang, Y., et al.: Oralgpt-omni: A versatile dental multimodal large language model. CVPR 2026 (2025)
10. Hao, J., Nalley, A., Yeung, A.W.K., Tanaka, R., Ai, Q.Y.H., Lam, W.Y.H., Shan, Z., Leung, Y.Y., AlHadidi, A., Bornstein, M.M., et al.: Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: a systematic review. npj Digital Medicine **8**(1), 412 (2025)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. 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. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28541–28564 (2023)
14. 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)
15. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F.: Do multimodal llms understand intraoral dental data? dataset, platform, and baselines. In: European Conference on Computer Vision (ECCV) (2026)
16. OpenAI: GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/ (2026), accessed July 2026
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
18. Wang, Y., Li, Z., Wu, C., Liu, J., Zhang, Y., Ni, J., Luo, Q., Chen, J., Zhang, H., Liu, J., et al.: MICCAI STS 2024 challenge: Semi-supervised instance-level tooth segmentation in panoramic x-ray and cbct images. Medical Image Analysis p. 103986 (2026)
19. 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**, 112049 (2026)
20. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)