# Clinically Guided Rendering for Orthodontic Report Generation

Jiashuo Guo[1], Jin Hao[2], Kuo Feng Hung[2], Feng Wang[1] ✉
1. College Division of Oral 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
✉ Email: diawang@hku.hk

## I. Introduction


Fig.1 Frontal, lateral, and occlusal views provide complementary appearance information.

- Orthodontic assessment requires joint interpretation of standardized intraoral photographs and three-dimensional intraoral scan (IOS) meshes.
- Existing Photographs provide soft-tissue and surface appearance, whereas IOS captures arch morphology and occlusal relationships.
- However, conventional multimodal large language models are designed for two-dimensional visual inputs and cannot directly process mesh geometry.
- We convert registered maxillary and mandibular IOS meshes into clinically meaningful 2D renderings and fuse them with photographs for automated structured orthodontic report generation.

## II. Methodology

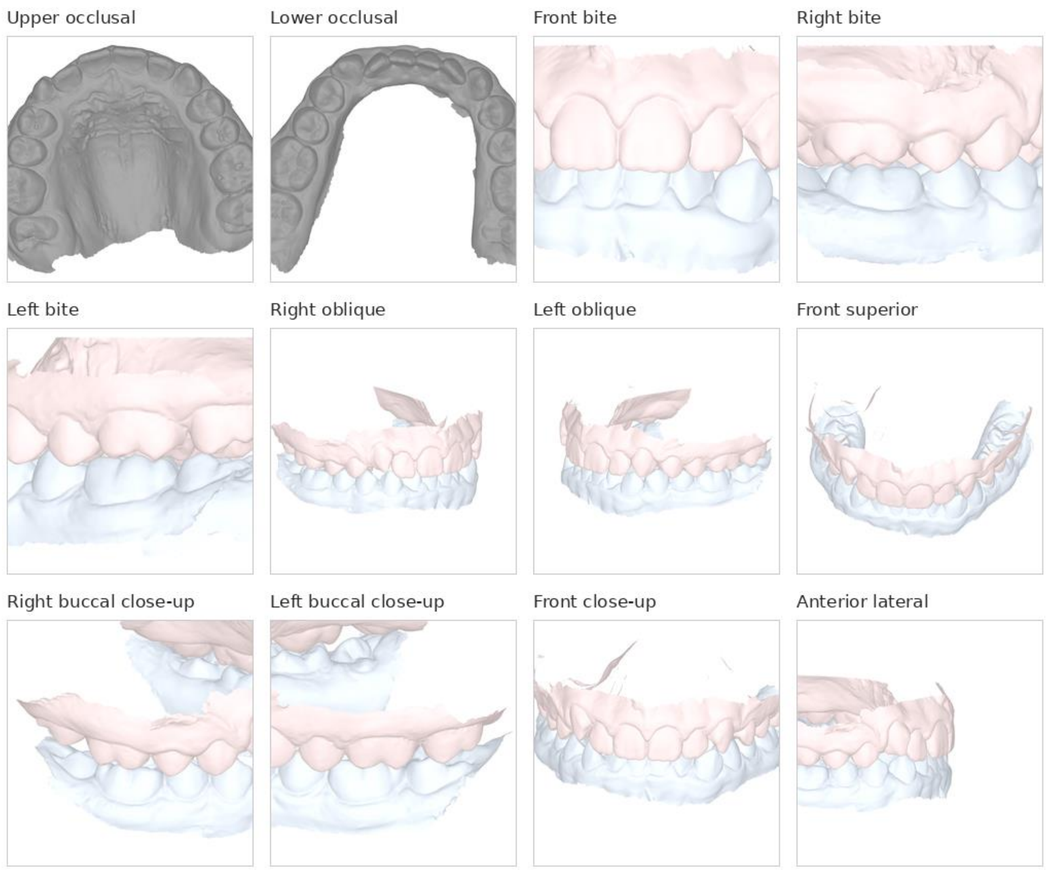


Fig.2 Twelve Clinically Targeted IOS Views.

- **Anatomical Normalization:** Registered maxillary and mandibular meshes are aligned to a canonical occlusal coordinate system for consistent rendering across cases.

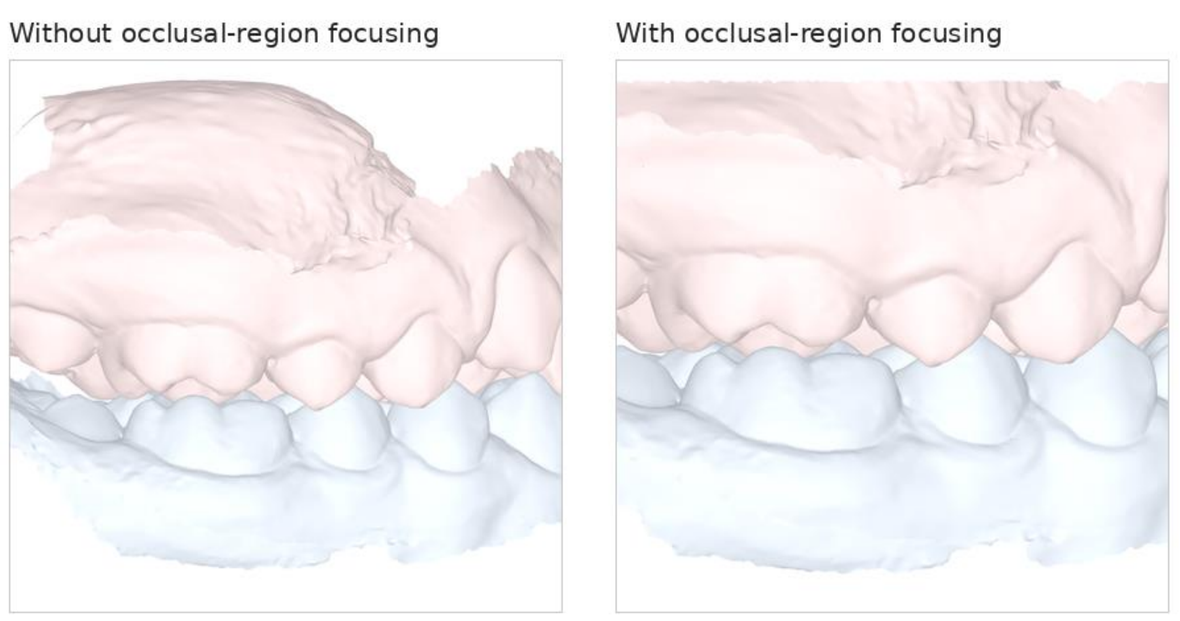


Fig.3 Occlusal-Region Focusing.

- **Clinically Targeted Rendering:** Twelve orthographic views cover occlusal morphology, standard bite relationships, oblique assessment, and regional close-ups.
- **Multimodal Adaptation:** Up to five intraoral photographs and the 12 renderings are arranged in a fixed order. The combined visual sequence is used to fine-tune OralGPT-Omni with LoRA.

## III. Experimental Results

- Our system achieves the highest Clinical F1 (0.3691) and final score (0.3811), exceeding the strongest zero-shot baseline by 0.1524. GPT-4o, GPT-5.5, and Qwen3-VL-8B were evaluated zero-shot using the same multimodal inputs and task instruction.

| 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** |

Table1. Comparison with general purpose MLLMs.

| 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** |

Table2. Ablation results for rendering and training configurations.

- Twelve targeted views provide a modest improvement over eight views, increasing the final score from 0.2976 to 0.3004.

- Occlusal-region focusing produces the largest visual-input gain, raising Clinical F1 from 0.2869 to 0.3272 and the final score to 0.3363. The full system reaches a final score of 0.3811.