

Segmentation-Guided 3D Dental ROI Transformer for Structured Clinical Report Generation from CBCT

Changkai Ji¹[0009-0007-7090-7360], Yusheng Liu¹[0009-0004-2624-9223], Yuxian Jiang¹[0009-0002-7689-5333], and Lisheng Wang¹[0000-0003-3234-7511]

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University,
Shanghai 200240, People’s Republic of China
{changkaiji, lswang}@sjtu.edu.cn

Abstract. The MICCAI STSR 2026 Challenge Task 3 aims to derive clinically relevant information and structured dental reports from three-dimensional cone-beam computed tomography (CBCT). To address the limited labeled data and the complex spatial relationships among multiple teeth, we propose a segmentation-guided and FDI-aligned three-dimensional dental ROI Transformer. An nnU-Net-based tooth segmentation model is first employed to localize individual teeth. Instead of applying the segmentation as a hard mask, the proposed method crops the complete dentition together with surrounding periapical tissues, periodontal structures, neighboring teeth, and alveolar bone. A 3D residual convolutional network extracts a global dentition representation and intermediate tooth-level ROI features. Each tooth token integrates its local visual feature, FDI positional embedding, tooth-presence information, and segmentation-derived reliability descriptors. A Tooth Transformer and attention-pooling module are then used to model inter-tooth relationships and obtain a dentition-level representation. The global and tooth-level features are fused for multi-task prediction of tooth notations, diagnosis categories, treatment actions, medications, and auxiliary clinical information. The predicted entities are converted into the seven required report fields using a deterministic template-based module. On the validation set, our method achieved a total score of 10.08 and ranked second in Task 3, demonstrating the effectiveness of segmentation-guided volumetric and tooth-level representation learning for structured dental CBCT reporting. Our code are available at <https://github.com/duola-wa/MICCAI-2026-STSR-Task-3>.

Keywords: Structured clinical prediction · Tooth segmentation · Medical report generation

1 Introduction

The MICCAI STSR 2026 Challenge Task 3 focuses on multimodal dental analysis using the MMDental dataset, which pairs three-dimensional cone-beam computed tomography (CBCT) images with detailed expert medical records. In contrast to conventional dental image analysis tasks that primarily concentrate on

anatomical segmentation, this task extends CBCT analysis toward higher-level clinical understanding, including diagnosis prediction, treatment reasoning, and structured report generation.

CBCT has become an important imaging modality in dental diagnosis and treatment planning because it provides detailed three-dimensional representations of teeth, tooth roots, alveolar bone, and surrounding maxillofacial structures [19,9,6]. In clinical practice, dentists must examine a large number of CBCT slices, identify abnormalities involving individual teeth, and integrate these findings with the patient’s chief complaint, oral examination, medical history, and follow-up documentation. This process requires substantial professional experience and careful cross-referencing between imaging findings and clinical records. Moreover, the interpretation and documentation of complex dental cases may be time-consuming and subject to inter-observer variation, particularly when multiple teeth, diagnoses, and treatment procedures are involved [12,7].

Deep learning methods have demonstrated considerable potential in dental image analysis, including tooth segmentation, root canal segmentation, and lesion detection [4,1,8,5]. Convolutional neural networks and Transformer-based models can automatically learn discriminative representations from medical images and have increasingly been applied to clinical classification and report generation [14]. However, most existing dental artificial intelligence studies rely on a single data modality. The limited availability of three-dimensional paired CBCT images and comprehensive clinical records has restricted the development of models capable of connecting volumetric dental evidence with patient-level diagnostic and treatment information. The MMDental dataset provides an opportunity to investigate this clinically relevant image-to-record prediction problem.

Despite this opportunity, automated clinical prediction from dental CBCT remains challenging. First, a complete CBCT examination may contain hundreds of high-resolution slices, resulting in substantial memory and computational requirements for true three-dimensional modeling. Sampling a limited number of two-dimensional slices reduces computational cost but may discard the continuous spatial relationships among the tooth crown, root, root apex, periodontal tissues, adjacent teeth, and alveolar bone. Second, many dental findings are associated with specific tooth positions. A patient-level global representation alone may identify an overall abnormality but fail to determine which tooth is clinically involved. Third, tooth localization obtained from an automatic segmentation model may contain missing regions, fragmented components, or incorrect labels. Directly masking the CBCT using imperfect segmentation results may consequently remove diagnostically important tissues outside the tooth boundary. Finally, the limited number of labeled cases and the highly imbalanced distributions of diagnoses, treatments, and medications create a substantial risk of overfitting.

To address these challenges, we propose a segmentation-guided three-dimensional dental ROI Transformer for structured clinical prediction from CBCT images. Our method first employs an nnU-Net-based tooth segmentation model to segment individual teeth and establish a standardized FDI anatomical coordinate

system [3,13]. Instead of using the segmentation result as a hard image mask, we crop a three-dimensional dentition region containing the complete dental arch and its surrounding tissues, thereby retaining potentially relevant information from the tooth roots, periapical regions, periodontal structures, neighboring teeth, and alveolar bone. A three-dimensional residual convolutional network then extracts multi-scale volumetric features from the original CBCT and the tooth segmentation prompt [2]. The main contributions of this work are summarized as follows:

- We propose a segmentation-guided true three-dimensional framework that preserves the complete dental arch and surrounding anatomical context, avoiding the loss of volumetric continuity associated with sparse 2.5-dimensional slice sampling.
- We introduce an anatomically aligned tooth-slot representation that combines tooth-level 3D ROI features with global dental-arch context. FDI positional embeddings, segmentation-quality gating, and segmentation dropout are incorporated to improve robustness to imperfect tooth segmentation.
- We formulate dental report generation as a multi-task structured prediction problem. The model jointly predicts tooth notations, diagnosis categories, tooth–diagnosis associations, treatment actions, and medications, which are subsequently converted into complete clinical reports through a constrained report-rendering module.
- The proposed method achieved second place in Task 3 of the MICCAI STSR 2026 Challenge, demonstrating its effectiveness on the official MMDental multimodal dental analysis benchmark.

2 Proposed Method

2.1 Framework Overview

As shown in Fig. 1, we propose a segmentation-guided and FDI-aligned three-dimensional dental ROI Transformer for structured clinical prediction from CBCT images. The framework consists of segmentation-guided preprocessing, global and tooth-level feature extraction, tooth relationship modeling, and multi-task structured prediction.

The framework follows a hierarchical representation strategy that is consistent with the clinical interpretation process of dental CBCT. Global volumetric features provide information regarding the overall dentition, maxillofacial structures, and the spatial distribution of abnormalities, whereas tooth-level features preserve localized evidence associated with individual anatomical positions. These two types of information are complementary. Relying only on global representations may obscure small tooth-specific abnormalities, while using isolated local regions may ignore contextual relationships among adjacent teeth and surrounding tissues. Therefore, the proposed framework jointly models patient-level and tooth-level information within a unified three-dimensional architecture.

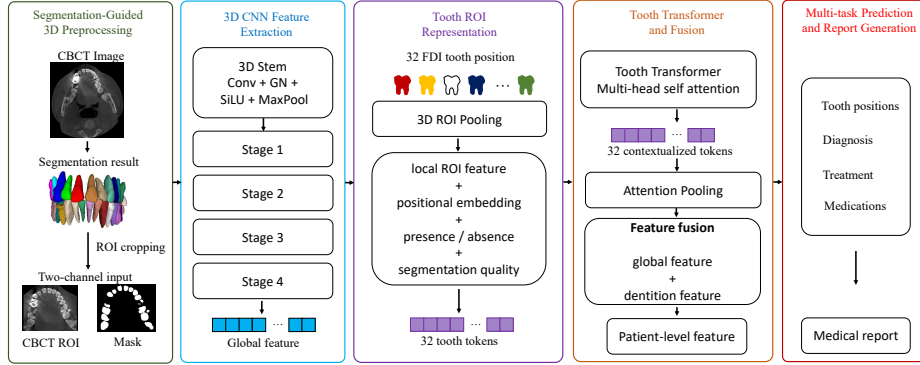


Fig. 1. Overview of the proposed framework. Global CBCT features and FDI-aligned tooth-level features are jointly modeled for structured clinical prediction and report generation.

2.2 Segmentation-Guided 3D Preprocessing

An nnU-Net-based model is first used to segment individual teeth and map them to 32 permanent-tooth FDI positions. Connected-component filtering is applied to remove segmentation noise, after which the union of the detected tooth regions is used to crop the complete dentition and surrounding tissues. Unlike hard masking, this strategy preserves diagnostically relevant structures outside the tooth boundaries, including periapical tissues, periodontal structures, neighboring teeth, and alveolar bone.

The segmentation results are used primarily as anatomical localization prompts rather than as definitive diagnostic regions. This distinction is important because clinically relevant findings may extend beyond the segmented tooth surface. For example, periapical lesions, periodontal changes, alveolar bone defects, and treatment-related structures may be located outside the predicted tooth mask. By using the union of tooth regions only to determine the spatial extent of the dentition ROI, the proposed preprocessing strategy reduces irrelevant background information while preserving surrounding anatomical context. The expanded ROI also provides additional tolerance to moderate localization errors produced by the automatic segmentation model.

Each tooth token combines its local three-dimensional ROI feature, FDI positional embedding, tooth-presence information, and segmentation-quality descriptor. This descriptor encodes tooth presence, relative component volume, bounding-box occupancy, connected-component purity, normalized spatial extent, and centroid location. The 32 tooth tokens are processed by a Transformer encoder to model contextual relationships among adjacent teeth, contralateral teeth, different dental quadrants, and the upper and lower dental arches. Segmentation-quality gating adaptively reduces the contribution of unreliable local features, while segmentation dropout prevents the network from depending excessively on potentially imperfect tooth localization.

2.3 Multi-Task Structured Prediction

The global CBCT representation and the Transformer-derived dentition representation are concatenated and fused to form a unified patient-level feature. This design allows the model to jointly capture global maxillofacial conditions and localized abnormalities associated with specific tooth positions. Multiple prediction heads are then used to predict clinically relevant tooth positions, diagnosis categories, tooth–diagnosis associations, treatment actions, medications, and auxiliary patient information.

The prediction tasks share the same fused representation but use separate task-specific output heads. This design enables the model to learn common visual patterns while maintaining the flexibility required by different clinical targets. Tooth-position prediction emphasizes spatial localization, diagnosis prediction focuses on pathological appearance, and treatment and medication prediction require higher-level associations between imaging findings and clinical decisions. Joint optimization allows these related tasks to provide complementary supervision. In particular, tooth–diagnosis association prediction encourages the network to connect a diagnostic category with its corresponding anatomical position rather than predicting only patient-level labels.

Class-balanced binary cross-entropy is adopted for the multi-label prediction tasks to alleviate the influence of label imbalance. The overall training objective is defined as a weighted combination of the task-specific losses, enabling the shared visual representation to be optimized using complementary clinical supervision. Because the frequency of clinical labels varies substantially, a single unweighted objective may be dominated by common diagnoses or treatment categories. The class-balanced formulation increases the contribution of relatively infrequent positive labels and reduces the tendency of the network to predict only frequently observed categories. The shared backbone and tooth-level modules are therefore trained under multiple forms of clinical supervision, while each output head remains specialized for its corresponding prediction target.

GPT is used as an offline annotation assistant to convert free-text clinical records into structured training labels. The extracted labels are stored before model training and used as supervision for the corresponding prediction tasks. GPT is not involved in neural network optimization, validation, report generation, or deployment inference.

2.4 Structured Report Generation

The predicted tooth notations, diagnosis categories, treatment actions, and medications are converted into the seven report fields required by the challenge through a deterministic template-based report-rendering module. Each predicted clinical entity is mapped to a predefined clinical phrase and inserted into the corresponding report field.

The constrained reporting process ensures a consistent output format and limits the generated content to information supported by the structured predictions.

3 Experiments and Results

3.1 Dataset and Assessment Metrics

The MICCAI STSR 2026 Challenge Task 3 is based on the MMDental dataset, which pairs three-dimensional CBCT images with expert medical records. The training set contains 50 labeled CBCT volumes with corresponding clinical records and 200 additional unlabeled CBCT volumes. The validation set contains 50 labeled cases. In the present study, the supervised three-dimensional model was trained using the 50 labeled training cases [18,15,17].

The free-text medical records contain information regarding tooth notations, diagnoses, treatment procedures, medications, and follow-up recommendations.

3.2 Implementation details

The computational environment and software dependencies used in our experiments are summarized in Table 1. The implementation was developed using Python and PyTorch.

Table 1. System Configuration

Ubuntu version	Ubuntu 24.04 LTS
CPU	Intel(R) Xeon(R) Platinum 8352S CPU @ 2.20GHz
RAM	503 GB
GPU	1 NVIDIA GeForce RTX 4090 (24G)
CUDA version	12.4
Programming language	Python 3.9.19
Deep learning framework	PyTorch (torch 1.12.1, torchvision 0.19.1)
Codes available at	https://github.com/duola-wa/MICCAI-2026-STs-Task-3

Preprocessing Settings. The individual-tooth segmentation results generated by nnU-Net were first mapped to 32 permanent-tooth FDI slots. The bounding box of each tooth was expanded by 10 mm to construct its local context region. The union of all reliable tooth regions was expanded by 15 mm to obtain the complete dentition ROI.

Network Configuration. The three-dimensional feature extractor contained a convolutional stem followed by four compact three-dimensional residual stages. The numbers of output channels were 24, 48, 96, and 192, respectively. The second-stage feature map was used for tooth-level ROI feature extraction, whereas the deepest feature map was processed by global average pooling and linear projection to obtain a 256-dimensional global representation.

For each of the 32 permanent-tooth FDI slots, three-dimensional ROI pooling was performed on the second-stage feature map. The resulting local representation was combined with an FDI positional embedding, a tooth-presence embedding, and segmentation-derived geometric information. The tooth tokens

were processed by a one-layer Transformer encoder with eight attention heads and a token dimension of 256. Attention pooling was then used to aggregate the contextualized tooth tokens into a 256-dimensional dentition representation.

Training Settings. We adopted five-fold patient-level cross-validation. In each fold, 40 cases were used for training and 10 cases were used for validation.

The network was optimized using AdamW with a weight decay of 1×10^{-4} . The learning rates for the three-dimensional backbone and global prediction modules were set to 1×10^{-4} , while the learning rate for the tooth-level ROI and Transformer modules was set to 3×10^{-4} .

Each fold was trained for a maximum of 300 epochs. Early stopping was applied with a patience of 15 epochs, and the checkpoint with the best validation performance was retained. Automatic mixed-precision training was employed to reduce GPU memory consumption. Data augmentation included random intensity scaling, intensity shifting, Gaussian noise, and segmentation dropout. Segmentation dropout was applied with a probability of 0.35, during which the tooth-mask channel, tooth bounding boxes, and segmentation-derived descriptors were removed while the original CBCT intensity channel was retained.

Inference Settings. During inference, the logits predicted by the five cross-validation models were averaged before applying the sigmoid function. Task-specific thresholds estimated from the out-of-fold predictions were then used to decode tooth notations, diagnoses, treatment actions, and medications. The decoded clinical entities were converted into the seven official report fields using a deterministic template-based report-rendering module.

3.3 Results and Analysis

Quantitative Performance. The official validation performance of the proposed method is summarized in Table 2. The proposed framework obtained an overall score of 10.08 and ranked second in Task 3 of the MICCAI STSR 2026 Challenge.

Table 2. Validation performance on MICCAI STSR 2026 Challenge Task 3.

Diagnosis Accuracy	Treatment Plan Record	Completeness	Total Score
5.29	3.12	1.68	10.08

The quality-aware representation and segmentation dropout were introduced to reduce the influence of imperfect tooth localization. When the segmentation result contains fragmented components or inaccurate tooth regions, the segmentation-derived descriptors reduce the contribution of unreliable local visual features. Meanwhile, segmentation dropout encourages the network to retain prediction ability through the global CBCT branch.

Qualitative Segmentation Results. Although tooth segmentation is not the final prediction target of Task 3, it provides the anatomical prior required for

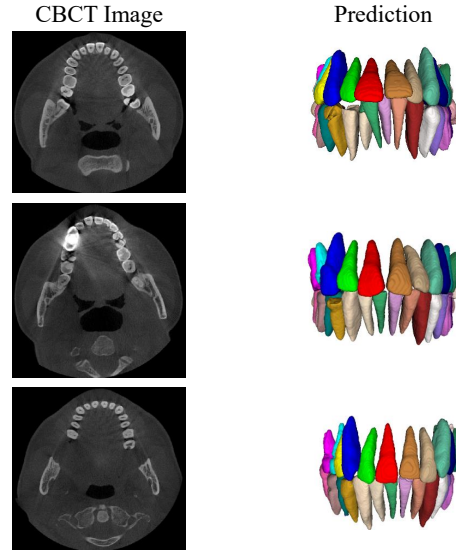


Fig. 2. Qualitative visualization of the nnU-Net tooth segmentation results. For each representative case, the left column shows the original CBCT image and the right column shows the corresponding individual-tooth segmentation result. Different colors indicate different tooth instances and FDI positions.

dentition localization and tooth-level ROI construction. Representative segmentation results are shown in Fig. 2. For each case, the left column presents the original CBCT image, while the right column shows the corresponding individual-tooth segmentation result generated by nnU-Net.

4 Conclusion

In this paper, we presented a segmentation-guided framework for structured clinical prediction and report generation from dental CBCT images. The proposed method uses nnU-Net tooth segmentation to localize the complete dentition while preserving surrounding anatomical structures, and combines a three-dimensional convolutional encoder with tooth-level ROI representations and a Tooth Transformer to jointly model global dentition context and localized abnormalities associated with specific FDI tooth positions. Multiple prediction heads are employed to estimate tooth notations, diagnosis categories, treatment actions, medications, and auxiliary clinical information, which are subsequently converted into the seven required report fields through a deterministic template-based rendering module. On the official validation set of MICCAI STSR 2026 Challenge Task 3, our method achieved a total score of 10.08 and ranked second. These results demonstrate the potential of segmentation-guided volumetric and tooth-level representation learning for dental CBCT-based structured reporting.

Acknowledgements

This work was developed and evaluated within the ODIN 2026 challenge cluster, including the STS, ToothFairy, and Bite2Text challenges. We thank the challenge organizers for providing the datasets, benchmark, and evaluation framework [17,10,11,16].

Disclosure of Interests

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

References

1. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., Van Nistelrooij, N., Van Lierop, P., Xi, T., Liu, Y., et al.: Segmenting the inferior alveolar canal in cbct volumes: the toothfairy challenge. *IEEE Transactions on Medical Imaging* (2024)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
3. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486* (2018)
4. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: Two-stage semi-supervised nnu-net framework for tooth segmentation in cbct images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 100–109. Springer (2024)
5. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: A two-stage semi-supervised nnu-net model for automated tooth segmentation in panoramic x-ray images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 91–99. Springer (2024)
6. Ji, C., Liu, Y., Jiang, Y., Wang, L.: Efficient nnu-net for tooth and root canal segmentation in cbct. In: *International Workshop on Oral and Dental Image Analysis*. pp. 171–180. Springer (2025)
7. Jiang, Y., Wang, X., Yang, T., Ji, C., He, L., Liu, Y., Wang, W., Liu, M., Shi, J., Guo, H., et al.: Morphology prior enhanced teeth segmentation for high-resolution oral scans. *IEEE Journal of Biomedical and Health Informatics* (2025)
8. Lin, Z., Liu, Y., Wu, J., Wang, D.H., Zhang, X.Y., Zhu, S.: Multi-modal pre-post treatment consistency learning for automatic segmentation and evaluation of the circle of willis. *Computerized Medical Imaging and Graphics* **122**, 102521 (2025)
9. Liu, Y., Xin, R., Yang, T., Wang, L.: Inferior alveolar nerve segmentation in cbct images using connectivity-based selective re-training. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 3–12. Springer (2024)

10. Lumetti, L., Di Bartolomeo, M., Pellacani, A., Anesi, A., Grana, C., Bolelli, F., et al.: Ontology-grounded structured prediction for dental cbct reporting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2026 (2026)
11. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F., et al.: Do multimodal llms understand intraoral dental data? dataset, platform, and baselines. In: Computer Vision–ECCV 2026 (2026)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
13. Van Nistelrooij, N., Krämer, L., Kempers, S., Beyer, M., Bolelli, F., Xi, T., Bergé, S., Heiland, M., Maier-Hein, K.H., Vinayahalingam, S., et al.: Toothseg: Robust tooth instance segmentation and numbering in cbct using deep learning and self-correction. *IEEE Journal of Biomedical and Health Informatics* (2026)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
15. Wang, Y., Chen, X., Qian, D., Ye, F., Wang, S., Zhang, H.: Semi-supervised Tooth Segmentation: First MICCAI Challenge, SemiToothSeg 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, vol. 14623. Springer Nature (2024)
16. 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)
17. 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)
18. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Zhang, Q., Wu, C., et al.: Sts miccai 2023 challenge: grand challenge on 2d and 3d semi-supervised tooth segmentation. *arXiv preprint arXiv:2407.13246* (2024)
19. Zhang, Y., Ye, F., Chen, L., Xu, F., Chen, X., Wu, H., Cao, M., Li, Y., Wang, Y., Huang, X.: Children’s dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Scientific Data* **10**(1), 380 (2023)