RESEARCH ARTICLE

Multimodal oral image fusion for intelligent early detection of dental caries

Lefei Li[1], Guangyao Hu[2, *]

[1]School of Stomatology, Beihua University, Jilin City, Jilin, China. [2]Department of Stomatology, Affiliated Hospital of Beihua University, Jilin City, Jilin, China.

Received: June 5, 2026; **accepted:** October 5, 2026.

Early caries detection is an important task in computer-aided dental diagnosis, aiming to identify subtle lesions before irreversible progression. However, existing methods often struggle with low lesion contrast, small lesion size, weak correspondence between X-ray and intraoral red–green–blue (RGB) images, and false positives caused by lesion-like artifacts. This study proposed the Adaptive Gated Cross-Modal Former (AGCM-Former), a tooth-structure-guided multimodal framework, for lesion-level early caries detection. The framework encoded tooth position, tooth surface, and adjacency relationships into structured queries and aggregated local evidence from available modalities around tooth-position–tooth-surface candidate regions, reducing dependence on strict pixel-level registration. Artifact-aware supervision, representation correlation constraints, and dynamic curriculum learning were further introduced to improve D1/D2 lesion detection and suppress hard-negative false positives. Experiments on the self-constructed multimodal dataset C showed that AGCM-Former achieved 72.4% mAP@0.5:0.95, 81.7% D1/D2 Recall, and an 8.9% artifact false-positive (FP) rate, outperforming the Caries Detection Transformer (Caries-DETR)-style baseline by 3.8% and 5.5% in mAP and recall, respectively. These results demonstrated that AGCM-Former provided a practical and robust solution for multimodal early caries detection.

Keywords: early caries detection; multimodal oral imaging; lesion-level detection; tooth-surface-aware query; hard-negative artifact; deformable attention.

[*]**Corresponding author:** Guangyao Hu, Department of Stomatology, Affiliated Hospital of Beihua University, Jilin City, Jilin 132011, China. Phone: +86 19989690195. Email: guangyaohuedu@phc.ac.cn.

Introduction

Dental caries is one of the most prevalent chronic oral diseases and remains a major challenge in global oral health management. Early detection is particularly important because initial lesions may be controlled or arrested through preventive intervention before irreversible structural damage occurs [1, 2]. In computer-aided dental diagnosis, the objective is therefore not only to determine whether caries is present but also to accurately localize subtle lesions and characterize their anatomical extent. Early caries may appear as slight enamel demineralization, weak radiolucency, white-spot changes, or subtle proximal abnormalities. These manifestations are typically small, low-contrast, and poorly demarcated, making reliable recognition highly dependent on clinical expertise and imaging quality [3]. In the lesion-depth classification, D1 and D2 represent lesions limited to the outer and inner enamel, respectively, while D3 and D4

indicate lesions extending into the outer and inner dentin, respectively. Accordingly, D1/D2 lesions are considered the primary target for early caries detection, while D3/D4 lesions represent more advanced disease stages. Conventional caries assessment relies mainly on visual examination, radiographic interpretation, and complementary optical imaging [4]. Radiographs and intraoral red–green–blue (RGB) images provide complementary diagnostic information with X-ray images revealing internal demineralization and proximal radiolucency and RGB images capturing visible surface characteristics including discoloration, white-spot lesions, fissure morphology, and restoration-related changes [5, 6]. However, neither modality alone can fully characterize early caries because radiographs may overlook surface-level morphological changes, while RGB images may fail to detect subsurface or proximal lesions [7]. Moreover, imaging artifacts such as tooth overlap, saliva, reflection, exposure variation, fissure shadows, cervical radiolucency, and restoration margins may obscure lesions or generate lesion-like patterns. Recent clinical investigations have therefore continued to report difficulties in identifying early and approximal caries, particularly when lesion boundaries are ambiguous [8].

Deep learning has substantially advanced computer-aided caries assessment. Convolutional neural network (CNN)-based methods have been applied to image-level, tooth-level, and candidate-level caries classification in radiographic and intraoral images [9]. Although useful for screening, classification models provide limited lesion localization and anatomical information. Object detection approaches including YOLO-family detectors, Faster R-CNN, RetinaNet, and EfficientDet have further enabled lesion-level localization for interproximal, occlusal, secondary, and severity-aware caries detection [10, 11]. More recently, Detection Transformer (DETR) and DETR with Improved deNoising anchOr boxes (DINO) based detectors have demonstrated the potential of transformer-based architectures for fine-grained lesion localization [12–16]. These developments indicate that query-based detection frameworks provide a promising foundation for precise dental lesion localization. Segmentation and multi-task learning approaches provide complementary regional information by estimating lesion or tooth structures at the pixel level. U-Net, Mask R-CNN, and related segmentation frameworks have been explored for caries and other dental abnormalities [17]. However, early caries often exhibits weak contrast and uncertain boundaries, making dense annotation difficult and susceptible to inter-observer variability. Small differences in lesion contours may also substantially influence Dice similarity coefficient (Dice) and intersection over union (IoU). Therefore, although segmentation is valuable for detailed analysis, lesion-level detection remains more practical for weakly expressed early caries with ambiguous boundaries. Multimodal learning provides a potential solution because radiographic and intraoral RGB images capture complementary pathological characteristics. Existing multimodal medical-image fusion strategies have explored image-level, feature-level, decision-level, multibranch, attention-based, and transformer-based integration [18–20]. However, many of these approaches rely on strong spatial correspondence between modalities. In routine dental practice, X-ray and RGB images are frequently acquired using different devices, viewpoints, scales, and acquisition times. Cross-modal correspondence is primarily established through patient identity, tooth position, and tooth-surface information rather than strict pixel-level registration. Therefore, conventional fusion strategies may combine anatomically unrelated regions and weaken representations of small early lesions. Recent weakly aligned multimodal detection studies have demonstrated that deformable attention can establish adaptive cross-modal correspondence without explicit image registration. CF-Deformable Detection Transformer (CF-Deformable DETR) introduces alignment-flexible feature aggregation for weakly aligned visible-infrared detection [21]. However, these approaches are mainly

developed for generic object detection and do not explicitly model dental anatomical relationships such as tooth position, tooth surface, and adjacent-tooth contact. Dental anatomical structure provides important prior knowledge for caries detection. Proximal lesions are closely associated with mesial and distal contact regions, occlusal lesions frequently occur around pits and fissures, and cervical regions contain both true lesions and lesion-like radiolucencies. Previous studies on tooth segmentation, numbering, and panoramic-radiograph interpretation have demonstrated the value of structural information for dental image analysis [22–24]. Nevertheless, existing lesion detectors generally treat caries as ordinary visual objects and lack explicit anatomical constraints, which may reduce sensitivity to subtle but clinically plausible lesions. False-positive suppression remains another major challenge. Early caries may share visual characteristics with cervical radiolucency, fissure shadows, reflections, restoration margins, prosthetic structures, and normal anatomical overlap. Although focal loss, hard-example mining, uncertainty modeling, and confidence calibration strategies can improve difficult-sample learning, they do not explicitly distinguish early lesions from recurrent artifact patterns. Consequently, increasing sensitivity may unintentionally introduce additional false-positive detections.

Taken together, four major challenges remain unresolved in lesion-level early caries detection. First, most existing approaches rely on single-modal images and cannot fully exploit complementary radiographic and surface information. Second, current multimodal fusion methods commonly assume stronger spatial correspondence than is available in clinical dental imaging. Third, general detection architectures lack explicit modeling of tooth position, tooth surface, and adjacent-tooth relationships. Fourth, effective discrimination between subtle D1/D2 lesions and lesion-like artifacts remains insufficiently addressed. These limitations are particularly important because clinically useful early caries detection requires both high sensitivity and acceptable false-positive control.

To address these challenges, this study proposed the Adaptive Gated Cross-Modal Former (AGCM-Former), a tooth-structure-guided multimodal framework, for lesion-level early caries detection under tooth-level weak correspondence. AGCM-Former aimed to integrate complementary X-ray and intraoral RGB evidence without requiring strict pixel-level registration while incorporating anatomical priors and artifact information into the detection process. The framework introduced tooth-position–tooth-surface-aware queries to guide candidate generation, candidate-region-centered cross-modal deformable attention to aggregate local multimodal evidence, and artifact-aware learning with representation constraints and dynamic curriculum optimization to improve D1/D2 lesion discrimination. By combining weakly corresponding multimodal evidence with dental anatomical priors, AGCM-Former extended early caries detection beyond conventional single-modal and registration-dependent approaches. The proposed framework provided a potential computational solution for robust lesion localization under realistic dental imaging conditions, where modality heterogeneity, anatomical variability, and lesion-like artifacts remained unavoidable.

Materials and methods

Dataset, image annotation, and preprocessing

Dataset C, a self-constructed weakly paired multimodal dental imaging dataset consisting of bitewing radiographs and intraoral RGB images, was employed for this study. The data were retrospectively collected from clinical dental imaging records at the Department of Stomatology, Affiliated Hospital of Beihua University, Jilin City, Jilin, China and included 1,286 patients, 8,742 images, and 4,368 annotated carious lesions. The multimodal images were associated at the patient, tooth-position, and tooth-surface levels and were randomly divided into training, validation, and

test sets at the patient level with a ratio of 70%:10%:20%. An independent, non-overlapping validation cohort (C-ext) was constructed from a separate retrospective patient cohort at the same institution. C-ext was acquired using different imaging devices and acquisition conditions from Dataset C and contained no overlapping patients or images. Importantly, C-ext was used only for independent evaluation and was not involved in training, hyperparameter optimization, model selection, or threshold adjustment. Lesion annotations were independently performed by three dentists and reviewed by a senior dental specialist. The inter-observer agreement was $\kappa = 0.86$ for lesion presence, $\kappa = 0.79$ for lesion severity classification, and $\kappa = 0.82$ for artifact categorization. Therefore, D1/D2 labels represented expert-consensus clinical annotations rather than histopathological confirmation. Hard-negative regions were explicitly annotated and included cervical radiolucency, restoration margins, fissure shadows, specular reflections, and other anatomical structures that could visually resemble carious lesions. All images were resized to 1,024 × 1,024 pixels before model input. RGB images were normalized using ImageNet statistics. Radiographic images were normalized to [0, 1] and enhanced using contrast-limited adaptive histogram equalization (CLAHE) implemented with the OpenCV library (https://opencv.org/). Standard geometric and photometric augmentations were applied only to the training set. During multimodal training, modality dropout was introduced while ensuring that at least one modality remained available for each sample. No augmentation, optimization, or threshold adjustment was performed using the test or C-ext datasets. The study procedures were approved by the Ethics Committee of the Affiliated Hospital of Beihua University (Jilin City, Jilin, China).

Multimodal early-caries detection formulation

Early caries detection was formulated as a lesion-level multimodal object detection problem under tooth-level weak correspondence. Given complementary dental radiographic and intraoral RGB images from the same patient or anatomically associated tooth region, the model predicted lesion localization, caries severity category, and corresponding anatomical attributes including tooth position and tooth surface. Unlike conventional multimodal detection tasks that assumed pixel-level registration, the images in this study were associated through patient-level, tooth-position-level, and tooth-surface-level correspondence. The model was designed to establish adaptive cross-modal interactions without requiring spatially aligned image pairs. For each input sample, the available modalities and their corresponding structural information were provided to guide multimodal feature aggregation as below.

$$Y_i^{det}=\{(b_{i,k},c_{i,k},t_{i,k},u_{i,k})\}_{k=1}^{K_i} \quad (1)$$

where each instance contained lesion location, caries category, and anatomical attributes. Caries severity labels followed the clinical classification adopted in this study, where D1/D2 represented early enamel lesions and D3/D4 represented dentinal involvement. To incorporate dental structural priors and artifact-related information, the learning objective jointly optimized lesion detection and auxiliary structural prediction as follows.

$$f_\vartheta(X_i,A_i,\pi_i,G_i)\rightarrow\hat{Y}_i \quad (2)$$

where the objective combined the primary lesion detection loss with structural attribute supervision and artifact-aware constraints when available, which enabled the model to exploit complementary multimodal evidence while reducing responses to anatomically implausible lesion-like regions.

AGCM-Former architecture

The AGCM-Former framework was designed to address two major challenges in early caries detection including heterogeneous evidence from weakly corresponding imaging modalities

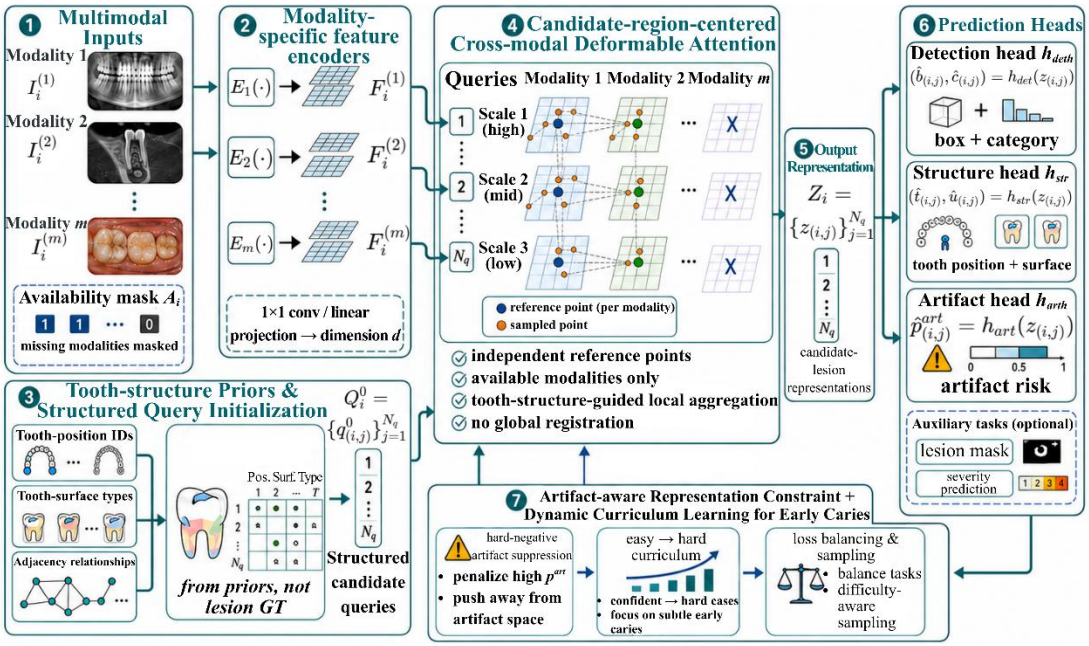


Figure 1. Overall framework of AGCM-Former.

and the difficulty of distinguishing subtle lesions from anatomically similar artifacts. Instead of relying on direct image registration or pixel-level fusion, AGCM-Former performed structure-guided multimodal feature interaction based on dental anatomical relationships (Figure 1). For each available modality, a modality-specific backbone extracted multiscale visual representations. The extracted features from different modalities were projected into a shared embedding space to enable subsequent cross-modal interaction. Features from unavailable modalities were masked during fusion to maintain robustness under incomplete multimodal input. The multi-scale feature representation was formulated as below.

$$F_i^{(m)}=\{F_{i,l}^{(m)}\}_{l=1}^{L}, F_{i,l}^{(m)}\in \mathbb{R}^{d\times H_{m,l}\times W_{m,l}} \quad (3)$$

where the multilevel features preserved complementary information from radiographic and intraoral RGB images including internal demineralization evidence and surface-level visual characteristics. To incorporate clinically meaningful anatomical information, AGCM-Former introduced structure-guided detection queries, which were initialized from tooth position, tooth surface, and adjacent-tooth relationships rather than lesion annotations or manually defined lesion boxes. The query generation process provided anatomical guidance while avoiding dependence on ground-truth lesion locations during inference as follows.

$$Q_i^0=\{q_{i,j}^0\}_{j=1}^{N_q} \quad (4)$$

Based on these structure-guided queries, candidate-region-centered cross-modal deformable attention was applied to aggregate informative evidence from available modalities. Instead of enforcing pixel-wise correspondence, the attention mechanism adaptively sampled local regions associated with potential lesion locations, which enabled the model to integrate complementary X-ray and RGB information around clinically relevant regions such as proximal contact areas, occlusal fissures, and cervical surfaces. The fused query representations were subsequently processed by task-specific prediction heads. The primary detection head predicted lesion localization and caries category, while auxiliary heads estimated

tooth position, tooth surface attributes, and artifact risk. These structural and artifact-related predictions provided additional supervision for separating early caries lesions from visually similar but clinically irrelevant regions.

Tooth-position–tooth-surface-aware query initialization

The anatomy-aware query initialization module introduced dental structural information into lesion detection through soft anatomical priors. Early caries lesions were not randomly distributed within dental images but frequently occurred in clinically meaningful regions including proximal contact surfaces, occlusal pits and fissures, and cervical areas. Incorporating tooth-level anatomical information provided the detector with additional contextual guidance for identifying plausible lesion locations, particularly for subtle early lesions with weak visual contrast. The Fédération Dentaire Internationale (FDI) two-digit permanent tooth numbering system and predefined tooth-surface categories were encoded as anatomical descriptors to provide tooth-level contextual information for query initialization as follows.

$$T = \{11,\ldots,17,21,\ldots,27,31,\ldots,37,41,\ldots,47\}$$
$$U = \{\text{occlusal, mesial, distal, buccal, lingual, cervical}\} \quad (5)$$

Third molars were excluded because of their inconsistent clinical visibility and frequent absence in routine dental imaging. A visible-tooth indicator was introduced to account for missing teeth, incomplete fields of view, and occlusal conditions. Queries associated with unavailable anatomical regions were masked during inference. For each anatomical candidate, a query representation was generated by combining tooth-position, tooth-surface, and anatomical-relation embeddings with a learnable query component, which allowed the detector to associate visual evidence with clinically meaningful dental structures rather than relying solely on appearance-based localization as below.

$$q_{i,j}^{surf} = W_q[E_t(\tau_j); E_u(\upsilon_j); E_r(\rho_j); e_j] \quad (6)$$

The query representation combined learnable embeddings for tooth position, tooth surface, and relation type with a learnable query vector. To capture proximal caries patterns, additional pairwise queries were constructed for neighboring teeth. These queries represented adjacent-tooth relationships and contact-side surfaces, enabling the model to consider inter-tooth anatomical context during multimodal feature aggregation as follows.

$$Q_i^0 = \{q_{i,j}^{surf}\} \cup \{q_{i,ab}^{pair} | (t_a, t_b) \in E_i\} \quad (7)$$

The paired queries encoded the identifiers of adjacent teeth, their corresponding contact-side surfaces, and their anatomical relationship. Importantly, query initialization relied only on automatic tooth localization, tooth numbering, or tooth-level weak metadata. Manual lesion bounding boxes, lesion masks, lesion categories, and annotations from the test set were not involved in query construction, ensuring that anatomical guidance did not introduce lesion-level information leakage.

Candidate-region-centered stochastic cross-modal feature aggregation

Candidate-region-centered cross-modal feature aggregation integrated complementary local evidence from radiographic and RGB modalities around anatomically defined candidate regions. Since the two modalities might exhibit different acquisition geometries, imaging conditions, and spatial representations, pixel-level correspondence was not assumed. Instead, modality-specific reference regions were independently estimated according to the associated tooth position, tooth surface, and anatomical relationship. For each anatomical query, a modality-specific candidate reference center was estimated to represent the expected lesion-related region associated with the corresponding tooth and surface rather than a manually annotated lesion location as below.

$$c_{i,j}^{(m)}=\begin{cases}\Psi(\hat{B}_{i,t}^{(m)},u_j) & \text{surface query}\\ \Psi_{pair}(\hat{B}_{i,t_a}^{(m)},\hat{B}_{i,t_b}^{(m)}) & \text{tooth-pair query}\\ \sigma(W_c^{(m)}q_{i,j}^0) & \text{unreliable tooth box}\end{cases} \quad (8)$$

The candidate center was derived from automatically predicted tooth regions, predefined surface representations, and adjacent-tooth contact relationships when applicable. The modality-specific reference location at each decoder layer was subsequently updated as follows.

$$p_{i,j}^{(m,r)}=\sigma\left(W_p^{(m)}z_{i,j}^{(r-1)}+\phi(c_{i,j}^{(m)})\right) \quad (9)$$

Because reference locations were generated independently for each modality, the proposed aggregation did not require rigid registration or shared pixel coordinates between radiographic and RGB images. The correspondence was established through anatomical semantics rather than direct spatial alignment. Around each modality-specific reference region, multiscale local features were sampled using deformable offsets, which allowed the model to capture lesion-related appearance variations while maintaining consistency with the anatomical candidate structure. The modality weights served only as internal feature fusion coefficients and should not be interpreted as estimates of clinical reliability or diagnostic confidence. To improve robustness against local appearance variation and modality-specific noise, AGCM-Former further introduced a Bayesian-inspired Bootstrap Transformer (BBT). During training, bootstrap-based stochastic perturbations were applied to the contribution of sampled local features as below.

$$\tilde{\alpha}_{i,j,n}=\frac{\lambda_{i,j,n}\alpha_{i,j,n}}{\sum_{n'=1}^{N}\lambda_{i,j,n'}\alpha_{i,j,n'}+\epsilon},\lambda_{i,j}\sim\text{Dirichlet}(\alpha) \quad (10)$$

The BBT output was obtained by weighted aggregation of the sampled local features. The stochastic perturbation exposed the model to diverse local evidence configurations during training and encouraged more stable multimodal feature aggregation. Consequently, the model was less dependent on a single modality-specific feature response when handling ambiguous early lesions.

Artifact-aware uncertainty-guided learning

Artifact-aware representation regularization and uncertainty-guided curriculum learning were introduced to improve discrimination between subtle early lesions and lesion-like hard-negative regions. Early carious lesions and imaging artifacts might exhibit similar low-level intensity, texture, or radiolucent patterns, particularly around cervical regions, fissures, restoration margins, and areas affected by illumination or exposure variation. The objective was therefore not to enforce complete semantic separation, but to encourage lesion-related and artifact-related predictions to rely on more discriminative task-specific representations. In addition, D1/D2 lesions were typically smaller and less contrasted than advanced lesions, motivating controlled emphasis on difficult positive early-lesion samples during training. The fused query representation was first mapped into caries-oriented and artifact-oriented task-specific projections as follows.

$$z_{i,j}^{car}=W_{car}z_{i,j},z_{i,j}^{art}=W_{art}z_{i,j} \quad (11)$$

Rather than assuming that these projections were fully disentangled, AGCM-Former applied a representation-correlation regularization term that penalized excessive linear dependence between the caries-oriented and artifact-oriented representations as below.

$$L_{rep}=\frac{1}{|\Omega_{sup}|}\sum_{(i,j)\in\Omega_{sup}}\left(\frac{(z_{i,j}^{car})^{\top}z_{i,j}^{art}}{\|z_{i,j}^{car}\|_2\|z_{i,j}^{art}\|_2+\epsilon}\right)^2 \quad (12)$$

This regularization encouraged the two prediction branches to retain complementary task information while reducing feature overlaps associated with recurrent artifact patterns. It did not impose statistical independence or guarantee complete semantic disentanglement between caries and artifact representations. For

candidate regions with explicit artifact annotations, an auxiliary artifact prediction head estimated the probability of an artifact-like response, and the corresponding supervision term was calculated as follows.

$$\hat{p}_{i,j}^{art}=\sigma(w_{art}^{\top}z_{i,j}^{art}+b_{art}),\mathrm{L}_{art}=\mathrm{BCE}(\hat{p}_{i,j}^{art},y_{i,j}^{art}) \quad (13)$$

The artifact branch was trained only from the annotated artifact categories and was used as an auxiliary learning signal rather than as an independent clinical diagnosis to provide explicit supervision for recurrent lesion-like structures that might otherwise produce false-positive detections. To characterize prediction variability during training, a stochastic latent-feature approximation was further applied to the fused query representation. The latent query feature was sampled as below.

$$q_{\phi}(r_{i,j}|z_{i,j})=N(\mu_{i,j},\mathrm{diag}(\sigma_{i,j}^{2})),r_{i,j}^{(s)}=\mu_{i,j}+\sigma_{i,j}\odot\epsilon^{(s)} \quad (14)$$

Repeated stochastic samples were propagated through the prediction heads to estimate variability in lesion classification and localization. The resulting uncertainty measure was used internally for sample weighting and curriculum construction and should not be interpreted as a calibrated probability of diagnostic correctness or as a direct estimate of clinical uncertainty. Lesion size, local image contrast, and the estimated predictive variability were subsequently combined to construct the early-caries dynamic curriculum (EDC) weight defined below.

$$\omega_{i,j}=1+\rho\cdot\mathrm{Norm}(d_{i,j}^{size}+d_{i,j}^{con}+d_{i,j}^{unc}) \quad (15)$$

The EDC weight was applied only to positive queries matched with reference D1/D2 lesions. Smaller, lower-contrast, or more uncertain early-lesion positives could therefore receive progressively greater training emphasis, whereas D3/D4 lesions and other positive categories retained their standard loss weights. Importantly, hard-negative artifact samples were not automatically upweighted because of high uncertainty or classification difficulty. They remained supervised through the artifact-aware branch and the standard negative-sample learning objective. This separation prevented difficult artifact regions from being treated as equivalent to difficult early-caries positives during curriculum learning.

Training objective and implementation

During training, Hungarian matching was used to establish one-to-one correspondence between predicted queries and reference lesion annotations. The matching cost combined lesion classification, L1 bounding-box regression, and generalized intersection-over-union (GIoU) terms as follows.

$$C_{match}=\lambda_{cls}C_{cls}+\lambda_{box}C_{box}+\lambda_{giou}C_{giou} \quad (16)$$

After matching, the detection objective was calculated over the assigned positive and negative queries. For positive queries corresponding to D1/D2 lesions, the early-caries dynamic curriculum (EDC) weight was incorporated into the detection loss as follows.

$$\mathrm{L}_{det}^{edc}=\frac{1}{|\Omega_{car}^{+}|}\sum_{(i,j)\in\Omega_{car}^{+}}\omega_{i,j}\left(\mathrm{L}_{cls}^{(ij)}+\lambda_{box}\mathrm{L}_{box}^{(ij)}+\lambda_{giou}\mathrm{L}_{giou}^{(ij)}\right)+\lambda_{bg}\mathrm{L}_{bg} \quad (17)$$

The EDC weighting was restricted to matched D1/D2-positive queries and did not alter the weighting of hard-negative artifact regions, which maintained the separation between early-lesion emphasis and artifact-aware supervision. The stochastic latent representation was regularized toward a Gaussian prior using the Kullback–Leibler (KL) divergence as follows.

$$\mathrm{L}_{kl}=\frac{1}{N_q}\sum_{j=1}^{N_q}D_{KL}\left[q_{\phi}(r_{i,j}|z_{i,j})\|N(0,I)\right] \quad (18)$$

The complete training objective combined the primary detection loss with available auxiliary supervision, artifact-related learning, representation-correlation regularization, and variational regularization as follows.

$$\mathrm{L}_{total}=\mathrm{L}_{det}^{edc}+\gamma_{i}^{mask}\lambda_{mask}\mathrm{L}_{mask}+\gamma_{i}^{sev}\lambda_{sev}\mathrm{L}_{sev}+\gamma_{i}^{str}\lambda_{str}\mathrm{L}_{str}+\gamma_{i}^{art}\lambda_{art}\mathrm{L}_{art}+\lambda_{rep}\mathrm{L}_{rep}+\beta\mathrm{L}_{kl} \quad (19)$$

The corresponding loss coefficients were set to $\lambda_{\text{mask}} = 1.0$, $\lambda_{\text{sev}} = 0.5$, $\lambda_{\text{str}} = 0.5$, $\lambda_{\text{art}} = 0.5$, $\lambda_{\text{rep}} = 0.1$ and $\beta = 0.01$. Auxiliary loss terms were included only when the corresponding annotations were available. AGCM-Former used a six-layer encoder and a six-layer decoder. Features from four spatial scales were projected into a shared 256-dimensional representation space with eight attention heads used in the transformer layers. The structured query set contained 194 queries including 168 tooth-surface queries and 26 adjacent-tooth pair queries. Four sampling points $(K = 4)$ were used for candidate-region-centered deformable feature aggregation. All input images were resized to $1{,}024 \times 1{,}024$ pixels. Model parameters were optimized using AdamW for 100 epochs with a batch size of 4. The initial learning rate was 1×10^{-4}, and the backbone learning rate was 1×10^{-5}. Weight decay was set to 1×10^{-4}, and gradient clipping was applied with a maximum norm of 0.1. Linear learning-rate warm-up was used for the first five epochs followed by cosine annealing. Modality dropout was applied only during training, while ensuring that at least one imaging modality remained available for every sample.

Comparison with representative conventional detectors

All experiments were performed on an Ubuntu 22.04 workstation equipped with an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. The implementation was developed using Python 3.10 (https://www.python.org/) and PyTorch (https://pytorch.org/) with CUDA and cuDNN acceleration. Five independent random seeds were used to characterize training variability. During inference, the batch size was set to 1. Structured queries were generated using automatic tooth localization and numbering or available tooth-level weak metadata. Manual lesion bounding boxes, lesion masks, lesion categories, and test-set lesion annotations were not used for query initialization. When uncertainty estimation was required, predictions from repeated stochastic forward passes were averaged to obtain the final output, while the resulting predictive variability was retained only as an internal model-reliability indicator and was not interpreted as clinical diagnostic confidence. AGCM-Former was compared with representative conventional detectors, a detection–segmentation model, Transformer-based detectors, and multimodal fusion baselines. To ensure comparability, all models used the same patient-level training, validation, and test partitions, input resolution, validation-based decision-threshold selection, and evaluation protocol. The same training schedule was also used whenever compatible with the original architecture. The conventional detection baselines included Faster R-CNN and YOLO11. Faster R-CNN was implemented following the original region-proposal-network formulation described by Ren *et al*. [25], while YOLO11 and YOLO11-seg were implemented using the official Ultralytics software release (https://docs.ultralytics.com/models/yolo11/). YOLO11 was used as the conventional one-stage detector, and YOLO11-seg was included as the detection–segmentation comparator. For Transformer-based detection, Deformable Detection Transformer (Deformable DETR) was implemented [26]. A Caries-DETR-style model was additionally implemented as a dental structure-aware comparison. This baseline retained a DETR-based detection framework while incorporating tooth-related structural information but did not include the complete candidate-region-centered stochastic cross-modal aggregation, artifact-aware representation regularization, or uncertainty-guided curriculum learning used in AGCM-Former. The multimodal comparisons included Early Fusion, Late Fusion, and Vanilla Cross-Attention. Early Fusion combined modality information before high-level prediction, whereas Late Fusion combined independently generated modality-specific predictions. Vanilla Cross-Attention provided an attention-based multimodal comparator without the proposed candidate-region-centered sampling and anatomy-guided fusion strategy. These multimodal baselines were implemented within the same PyTorch experimental pipeline to

maintain consistent data processing, optimization, and evaluation procedures. Single-modality controls were also constructed using X-ray-only and RGB-only inputs to determine the contribution of complementary multimodal information. Faster R-CNN and Deformable DETR were reimplemented in PyTorch according to their original published architectures [25, 26]. Caries-DETR-style, Early Fusion, Late Fusion, Vanilla Cross-Attention, the X-ray-only and RGB-only controls, and the parameter-matched, query-matched, and auxiliary-head controlled variants were also implemented within the same PyTorch experimental pipeline.

Evaluation metrics

Detection performance was evaluated using mean average precision over intersection-over-union (IoU) thresholds from 0.50 to 0.95 (mAP@0.5:0.95), D1/D2 recall, and artifact false-positive (FP) rate. mAP@0.5:0.95 was used to assess overall lesion localization and classification performance across multiple IoU thresholds. D1/D2 recall quantified the proportion of reference D1/D2 lesions correctly detected and was used as the primary measure of sensitivity to early caries. For recall-based analyses, a lesion was considered correctly detected when the predicted lesion category matched the reference annotation, and the bounding-box IoU was at least 0.50. Artifact FP rate was used to evaluate false-positive responses in predefined hard-negative regions including cervical radiolucency, restoration margins, fissure shadows, specular reflections, and anatomically similar non-carious structures. Each annotated hard-negative region was treated as one evaluation unit, and multiple detections associated with the same region were counted once. A lower artifact FP rate therefore indicated better discrimination between true carious lesions and lesion-like structures. Robustness under missing modalities, image-quality degradation, and cross-modal spatial perturbation was quantified by the reduction in mAP relative to the corresponding unperturbed condition, denoted as $\Delta\mathrm{mAP}$. A smaller $\Delta\mathrm{mAP}$ indicated greater robustness to the evaluated perturbation. For natural occurring cross-modal correspondence differences, absolute mAP, D1/D2 recall, and artifact FP rate were additionally reported to characterize performance across different levels of multimodal inconsistency. Computational efficiency was assessed using the number of model parameters, floating-point operations (FLOPs), end-to-end inference time, GPU memory consumption, and frames per second (FPS). All efficiency measurements were obtained using the same hardware configuration, input resolution, inference batch size, and measurement protocol to ensure comparability among models.

Statistical analysis

Statistical analyses were performed using Python 3.10 and SciPy (https://scipy.org/). Training variability was characterized across five independent random seed runs and was reported as mean ± standard deviation (SD). These random-seed repetitions were used only to quantify optimization variability and were not treated as independent clinical observations for inferential statistical testing. The patient was used as the statistical unit for between-model comparisons. Prespecified pairwise comparisons were conducted between AGCM-Former and the principal comparator models including Deformable DETR, Vanilla Cross-Attention, and Caries-DETR-style. Patient-level paired bootstrap resampling with 10,000 repetitions was used for each comparison. In each bootstrap replicate, patients were sampled with replacement, and the between-model difference in the corresponding evaluation metric was recalculated while preserving the paired predictions for each sampled patient. Two-sided *P* values were obtained from the paired bootstrap distributions under a null between-model difference of zero. To account for multiple prespecified comparisons within the same performance metric, *P* values were adjusted using the Holm procedure. Statistical inference was therefore based on patient-level paired comparisons rather than differences among random-seed averages. Inter-observer

agreement for lesion presence, lesion severity classification, and artifact categorization was assessed using Fleiss' κ coefficient. All statistical tests were two-sided, and a Holm-adjusted *P* value less than 0.05 was considered statistically significant. Effect estimates were reported for the prespecified principal comparisons, and a Holm-adjusted *P* value less than 0.05 was considered statistically significant.

Results

Overall detection performance on Dataset C and C-ext

On Dataset C, AGCM-Former achieved the highest overall detection performance with an mAP@0.5:0.95 of 72.4 ± 0.8% and a D1/D2 recall of 81.7 ± 1.2%, while maintaining the lowest artifact FP rate of 8.9 ± 0.6%. Its recall was 74.9 ± 0.9% for D1 lesions and 85.3 ± 0.6% for D2 lesions, indicating that the improvement in early-lesion detection was observed for both enamel-depth categories. Caries-DETR-style was the strongest comparator on Dataset C, achieving a mAP of 68.6 ± 1.1%, a D1/D2 recall of 76.2 ± 1.3%, and an artifact FP rate of 12.4 ± 1.0%. Relative to this baseline, AGCM-Former showed absolute improvements of 3.8% in mAP and 5.5% in D1/D2 recall, together with a 3.5% reduction in artifact FP rate. Compared with Deformable DETR, the corresponding differences were 7.4% in mAP, 9.6% in D1/D2 recall, and 5.7% in artifact FP rate. Performance decreased across all methods on C-ext compared with Dataset C. Nevertheless, AGCM-Former remained the best-performing model, achieving a mAP of 65.8 ± 1.6%, a D1/D2 recall of 74.2 ± 2.0%, and an artifact FP rate of 12.1 ± 1.4%. Caries-DETR-style achieved 61.5 ± 1.8%, 69.4 ± 2.2%, and 15.6 ± 1.6%, respectively. AGCM-Former therefore retained absolute differences of 4.3% in mAP, 4.8% in D1/D2 recall, and 3.5% in artifact FP rate relative to Caries-DETR-style (Table 1). These descriptive results showed that the ranking advantage of AGCM-Former was preserved in the independent same-institution cohort acquired under different imaging and device conditions.

Stratified performance across anatomical regions and artifact categories

Descriptive stratified analysis showed that, for AGCM-Former, recall was 81.2 ± 1.2% for proximal lesions, 81.5 ± 1.3% for fissure lesions, and 78.1 ± 1.4% for cervical lesions. The corresponding values for Caries-DETR-style were 75.1 ± 1.6%, 76.4 ± 1.5%, and 72.9 ± 1.7%, respectively. Thus, the absolute recall differences in favor of AGCM-Former were 6.1% for proximal lesions, 5.1% for fissure lesions, and 5.2% for cervical lesions. Artifact-related false-positive responses were also lower for AGCM-Former across all predefined hard-negative categories. The artifact FP rates were 9.8 ± 0.9% for cervical radiolucency, 9.5 ± 0.8% for restoration margins, 8.7 ± 0.9% for fissure shadows, 9.2 ± 1.0% for intraoral reflections, and 7.8 ± 0.7% for anatomically similar normal regions. For Caries-DETR-style, the corresponding rates were 14.5 ± 1.3%, 13.2 ± 1.2%, 12.8 ± 1.1%, 12.1 ± 1.3%, and 9.6 ± 0.9%, respectively. The largest absolute reduction in artifact FP rate was observed for cervical radiolucency as 4.7% followed by fissure shadows' 4.1%, restoration margins' 3.7%, intraoral reflections' 2.9%, and anatomically similar normal regions' 1.8%. These descriptive subgroup results indicated that the performance differences were observed across multiple anatomical regions and hard-negative categories rather than being concentrated in a single subgroup.

Statistical comparison with prespecified principal baselines

On Dataset C, AGCM-Former showed absolute differences of +3.8% in mAP, +5.5% in D1/D2 recall, and −3.5% in artifact FP rate relative to Caries-DETR-style. The corresponding differences relative to Vanilla Cross-Attention were +5.2%, +6.7%, and −3.9%, whereas those relative to Deformable DETR were +7.4%, +9.6%, and −5.7%, respectively. The prespecified patient-level comparisons that reached statistical significance remained significant after Holm correction (adjusted $P < 0.05$). On C-ext, AGCM-Former showed the same direction of performance differences with higher mAP and

Table 1. Main performance comparison on Dataset C and C-ext.

Method	Type	Dataset C mAP	D1 Recall	D2 Recall	D1/D2 Recall	Artifact FP Rate	C-ext mAP	C-ext D1/D2 Recall	C-ext Artifact FP Rate
Faster R-CNN	Two-stage detector	56.7 ± 1.8	59.8 ± 2.7	70.6 ± 2.1	65.2 ± 2.5	18.9 ± 1.9	49.8 ± 2.4	58.7 ± 3.1	22.4 ± 2.5
YOLOv11	One-stage detector	60.4 ± 1.5	63.2 ± 2.2	72.0 ± 1.9	67.6 ± 2.1	16.7 ± 1.6	53.0 ± 2.2	61.1 ± 2.9	20.5 ± 2.2
YOLOv11-seg	Detection + segmentation	62.1 ± 1.3	64.7 ± 2.0	74.3 ± 1.7	69.5 ± 1.9	15.8 ± 1.5	55.2 ± 2.1	62.9 ± 2.8	19.6 ± 2.1
Deformable DETR	Transformer detector	65.0 ± 1.2	63.9 ± 2.2	78.4 ± 1.5	72.1 ± 1.6	14.6 ± 1.3	58.1 ± 1.9	65.8 ± 2.4	17.9 ± 1.8
Early Fusion	Input / feature fusion	64.6 ± 1.4	67.2 ± 2.0	78.6 ± 1.7	72.3 ± 1.8	13.9 ± 1.4	56.5 ± 2.0	64.1 ± 2.6	18.6 ± 1.9
Late Fusion	Decision fusion	64.5 ± 1.3	68.0 ± 1.9	78.4 ± 1.6	73.2 ± 1.7	13.5 ± 1.1	57.2 ± 2.0	64.8 ± 2.5	17.5 ± 1.7
Vanilla Cross-Attention	Multimodal attention	67.2 ± 1.3	69.4 ± 1.8	80.9 ± 1.4	75.0 ± 1.5	12.8 ± 1.2	60.4 ± 1.9	68.2 ± 2.3	16.4 ± 1.6
Caries-DETR-style	Structure-aware detector	68.6 ± 1.1	68.4 ± 1.8	82.0 ± 1.4	76.2 ± 1.3	12.4 ± 1.0	61.5 ± 1.8	69.4 ± 2.2	15.6 ± 1.6
AGCM-Former	Proposed framework	72.4 ± 0.8	74.9 ± 0.9	85.3 ± 0.6	81.7 ± 1.2	8.9 ± 0.6	65.8 ± 1.6	74.2 ± 2.0	12.1 ± 1.4

D1/D2 recall and a lower artifact FP rate than the principal comparators.

Computational efficiency

The computational efficiency of the evaluated detection models showed that AGCM-Former achieved the highest detection accuracy and D1/D2 recall among the compared methods but required greater computational resources and exhibited lower inference throughput than the lighter conventional detectors. The model was therefore positioned in the higher-accuracy and higher-computational-cost region of the accuracy–efficiency comparison. The efficiency difference was more evident when compared with lightweight one-stage detection models, whereas the computational gap was smaller relative to the Transformer-based and multimodal detection baselines, which indicated that the improved early-caries detection performance of AGCM-Former was accompanied by additional computational overhead rather than being obtained without an increase in model complexity (Figure 2a). The end-to-end inference-time decomposition further showed that the additional latency of AGCM-Former was primarily associated with model forward propagation and structured-query-related processing. Other components contributed comparatively less to the total inference time. Thus, although AGCM-Former required greater computational resources than the lighter baselines, the additional inference cost was concentrated mainly in the model components responsible for structured multimodal feature processing (Figure 2b).

Robustness to missing modalities

multimodalcreased when one imaging modality was unavailable at inference, although the magnitude of degradation differed among the multimodal methods. For AGCM-Former, mAP decreased by 3.5 ± 0.6% when RGB images were unavailable and by 4.2 ± 0.7% when radiographic images were unavailable. Under random modality missingness, the mAP reduction was 3.7 ± 0.6% with a D1/D2 recall of 76.9 ± 1.3%. Under the same random modality-missing condition, mAP decreased by 11.8 ± 1.3% for Early Fusion, 7.4 ± 0.9% for Late Fusion, 8.3 ± 1.0% for Vanilla

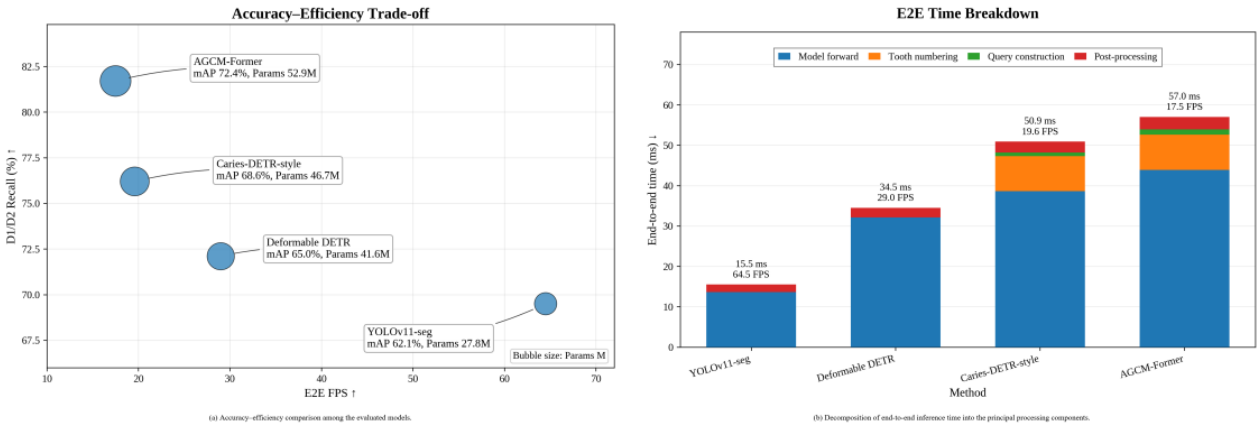


Figure 2. Computational efficiency and inference-time decomposition of the evaluated detection models. **(a)** Accuracy–efficiency comparison among the evaluated models. **(b)** Decomposition of end-to-end inference time into the principal processing components.

Cross-Attention, and 7.1 ± 0.8% for Caries-DETR-style. The corresponding D1/D2 recalls were 63.7 ± 2.1%, 67.1 ± 1.8%, 68.6 ± 1.7%, and 70.4 ± 1.6%, respectively. AGCM-Former therefore showed the smallest mAP reduction and the highest D1/D2 recall among the evaluated multimodal methods under random modality missingness.

Performance on low-quality images

Detection performance decreased for all evaluated models on the predefined low-quality subset. For AGCM-Former, mAP decreased from 72.9 ± 1.0% on the normal-quality subset to 62.6 ± 2.3% on the low-quality subset, corresponding to a reduction of 10.3%. The corresponding reductions were 13.4% for YOLO11-seg, 12.6% for Deformable DETR, and 11.3% for Caries-DETR-style. On the low-quality subset, AGCM-Former achieved a D1/D2 recall of 71.2 ± 2.4%, compared with 66.8 ± 2.5% for Caries-DETR-style, 61.9 ± 2.8% for Deformable DETR, and 58.4 ± 3.1% for YOLO11-seg. Its artifact FP rate was 12.8 ± 1.8%, compared with 16.5 ± 2.0%, 19.4 ± 2.2%, and 21.7 ± 2.5% for the three comparator models, respectively. Among the evaluated methods, AGCM-Former therefore retained the highest mAP and D1/D2 recall and the lowest artifact FP rate on the low-quality subset.

Robustness to cross-modal spatial misalignment

The results of controlled spatial-perturbation experiments demonstrated that translation, rotation, and scaling were independently applied to one imaging modality to increase cross-modal spatial inconsistency. Performance decreased as perturbation magnitude increased for all evaluated fusion methods. Across the tested perturbation levels, AGCM-Former showed a smaller reduction in mAP than Early Fusion and Vanilla Cross-Attention (Figure 3a). The performance after stratification according to naturally occurring cross-modal correspondence showed that, in cases classified as having weak correspondence inconsistency, AGCM-Former achieved a mAP of 74.1 ± 0.9%, a D1/D2 recall of 83.0 ± 1.1%, and an artifact FP rate of 8.2 ± 0.6%. Under moderate inconsistency, the corresponding values were 71.8 ± 1.1%, 80.9 ± 1.4%, and 9.1 ± 0.8%, respectively. For cases with obvious correspondence inconsistency, mAP decreased to 67.3 ± 1.6% and D1/D2 recall to 75.6 ± 2.0%, while the artifact FP rate increased to 11.8 ± 1.2% (Figure 3b). The descriptive stratified results therefore showed progressively lower lesion-detection performance and higher artifact-related false-positive rates as cross-modal correspondence became less consistent.

Feature-space analysis

Feature-space analysis was performed to examine the distribution of the final decoder query representations learned by Deformable DETR, Caries-DETR-style, and AGCM-Former. For qualitative visualization, the decoder query

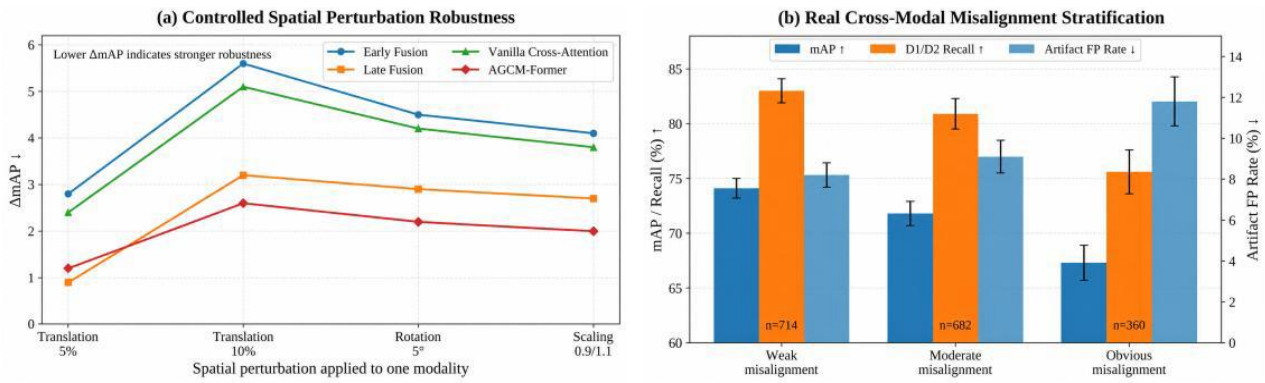


Figure 3. Robustness to cross-modal spatial inconsistency. (**a**) Changes in detection performance under controlled translation, rotation, and scaling of one imaging modality. (**b**) Detection performance stratified by naturally occurring levels of cross-modal correspondence inconsistency.

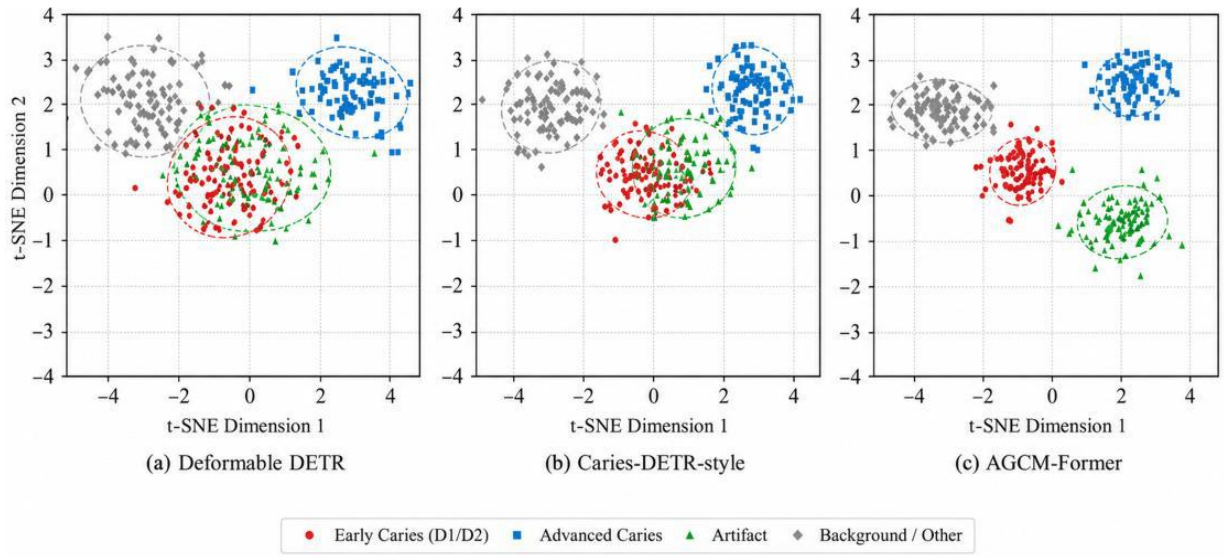


Figure 4. Feature-space visualization of decoder query embeddings. t-SNE projections of final decoder query representations from Deformable DETR (**a**), Caries-DETR-style (**b**), and AGCM-Former (**c**), showing the relative distributions of caries related and hard-negative artifact samples.

embeddings extracted from the test samples were projected into two dimensions using t-distributed stochastic neighbor embedding (t-SNE). The projected representations from Deformable DETR showed substantial overlap between early-caries samples and hard-negative artifact regions. Caries-DETR-style showed greater separation between the two groups, although partial overlap remained. In comparison, the AGCM-Former projection contained more compact lesion-related clusters and visually clearer separation between caries-related and artifact-related representations (Figure 4). Quantitative feature-space measurements showed a consistent descriptive pattern. The Silhouette Score increased from 0.214 ± 0.018 for Deformable DETR and 0.287 ± 0.014 for Caries-DETR-style to 0.346 ± 0.012 for AGCM-Former. The corresponding Davies–Bouldin Index decreased from 2.31 ± 0.12 and 1.91 ± 0.09 to 1.54 ± 0.07, respectively. Caries–artifact cosine similarity was also lower for AGCM-Former (0.25 ± 0.02) than for Caries-DETR-style (0.37 ± 0.03) and Deformable DETR (0.46 ± 0.04). These descriptive measurements indicated greater compactness and lower representational

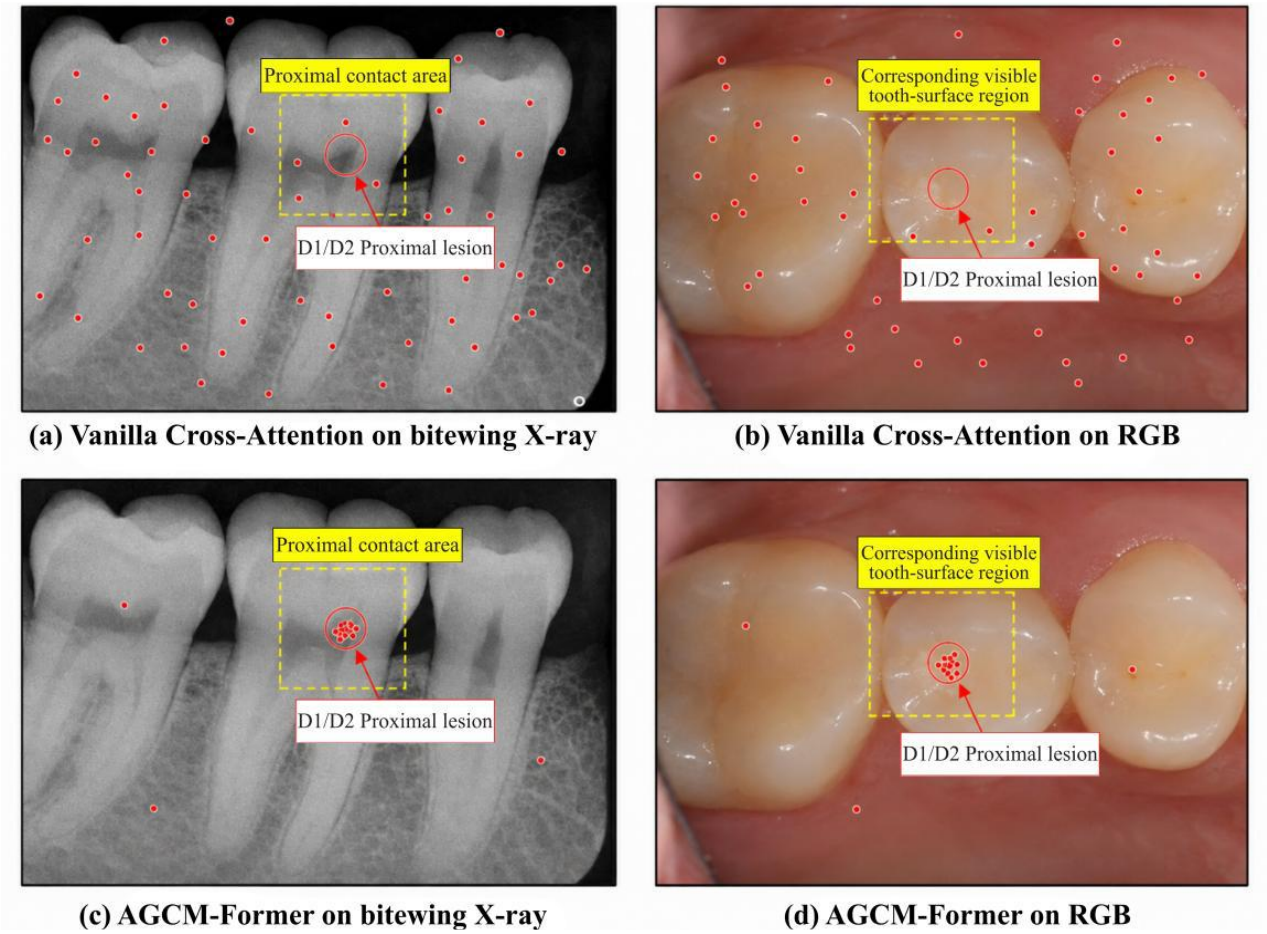


Figure 5. Attention visualization of a representative D1/D2 proximal caries case. Comparison of Vanilla Cross-Attention (**a and b**) and AGCM-Former (**c and d**) on the corresponding bitewing radiograph and intraoral RGB image.

overlap between lesion-related and artifact-related samples in the learned feature space. Because t-SNE is a nonlinear dimensionality-reduction method, the two-dimensional projections were used only for visualization and were not interpreted as independent evidence of class separability. The quantitative feature-space measurements provided complementary characterization of the learned representations rather than proof of complete semantic disentanglement.

Attention and qualitative case analysis

Candidate-region-centered sampling points were projected back onto the corresponding multimodal images to examine the spatial distribution of the learned attention. A representative D1/D2 proximal caries case from Dataset C was applied to compare Vanilla Cross-Attention and AGCM-Former on the corresponding bitewing radiograph and intraoral RGB image, where Vanilla Cross-Attention exhibited a relatively diffuse response around the proximal contact region, whereas AGCM-Former showed more concentrated sampling around the anatomically relevant proximal area in both modalities (Figure 5). The resulting AGCM-Former response was more closely aligned with the reference proximal lesion region, illustrating how anatomy-guided candidate-centered aggregation could focus multimodal evidence around a clinically plausible lesion location. This representative example was consistent with the higher D1/D2 recall reported for AGCM-Former and proximal-lesion results. However, the qualitative example itself did not provide

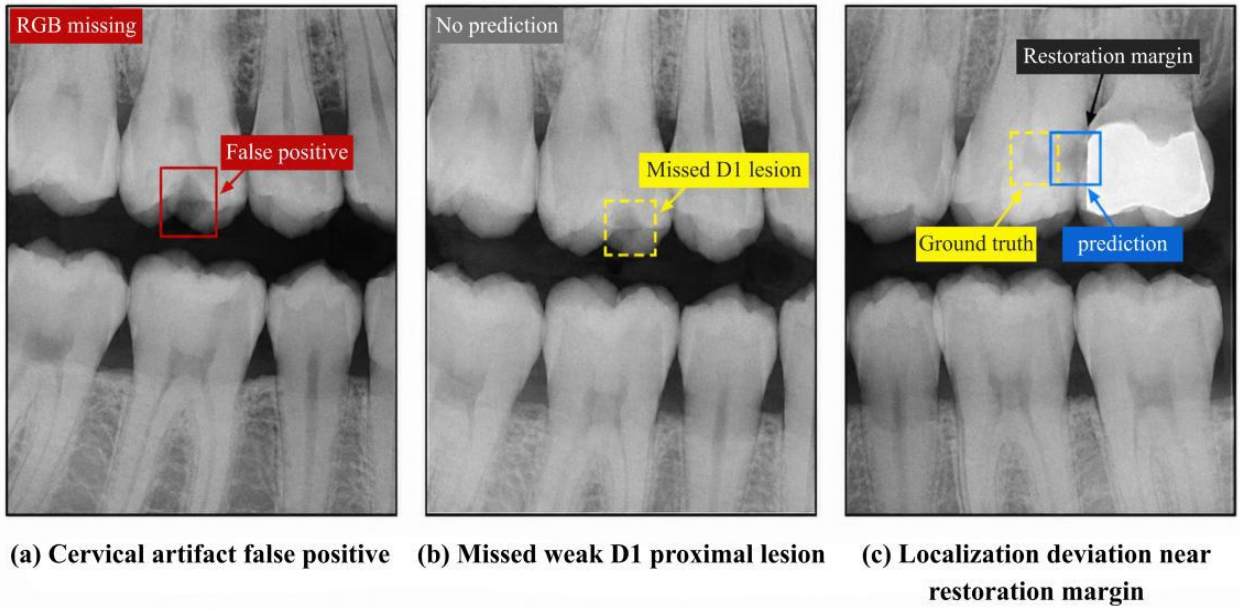


Figure 6. Representative failure cases of AGCM-Former. Examples of a cervical artifact false positive (**a**), a missed subtle D1 proximal lesion (**b**), and localization deviation near a restoration margin (**c**).

statistical evidence of model superiority. Attention visualization reflected the spatial allocation of learned model features rather than a direct clinical explanation of individual predictions. The visualization should therefore be interpreted as qualitative evidence of model behavior and not as evidence of causal or clinically validated interpretability.

Attention stability analysis

Attention stability was further evaluated under repeated stochastic forward passes using attention entropy, sampling-point displacement, and the coefficient of variation of modality weights. Lower values indicated more concentrated or less variable attention behavior across repeated stochastic evaluations. AGCM-Former showed an attention entropy of 0.53 ± 0.03 compared with 0.71 ± 0.04 for Vanilla Cross-Attention. Mean sampling-point displacement was 4.1 ± 0.5 for AGCM-Former and 6.8 ± 0.7 for Vanilla Cross-Attention, while the modality-weight coefficient of variation decreased from 0.36 ± 0.04 to 0.24 ± 0.03. Thus, AGCM-Former showed lower variability in both spatial sampling and modality weighting under the evaluated stochastic perturbations. These results provided complementary evidence that the proposed attention mechanism produced more consistent internal feature-allocation patterns than Vanilla Cross-Attention under the evaluated conditions. They should, however, be interpreted as measurements of model-internal stability rather than as direct evidence of clinical interpretability.

Error and failure case analysis

Residual failure patterns were manually reviewed to identify the main conditions associated with incorrect predictions. The largest failure categories were cervical artifact-related false positives (27.6%) and missed subtle D1 lesions (24.1%) followed by missing-modality cases (18.7%), low-quality images (15.3%), and ambiguous annotation boundaries (14.3%). Representative examples showed that cervical artifacts occasionally produced lesion-like responses, whereas subtle D1 lesions were mainly missed because of weak contrast and small lesion extent (Figure 6). Incomplete multimodal information, poor image quality, and uncertain lesion boundaries also contributed to localization or classification errors.

Ablation and sensitivity analysis

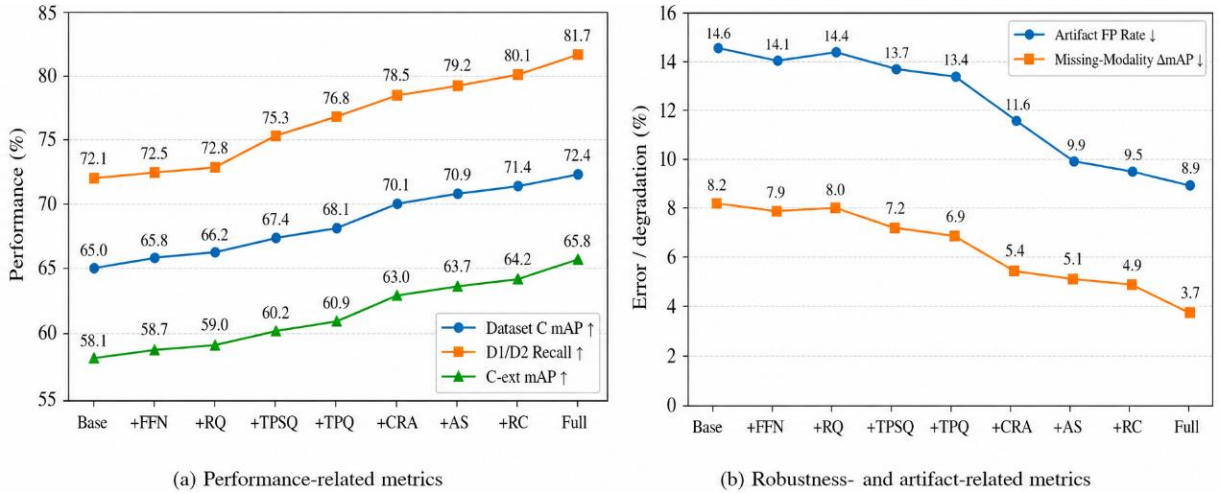


Figure 7. Overall module ablation of AGCM-Former. (**a**) detection performance. (**b**) artifact control and missing-modality robustness.

Ablation analyses were performed primarily on Dataset C to examine the contribution of the principal components of AGCM-Former. C-ext mAP was additionally evaluated in the overall module analysis. Unless otherwise specified, all variants used the same data partitions, input resolution, optimization protocol, validation-based threshold selection, and evaluation procedure.

(1) Overall module contribution and controlled variants

The overall ablation results showed that removing the anatomy-aware structured queries reduced early-caries recall, whereas replacing candidate-region-centered cross-modal attention with simpler fusion strategies resulted in larger performance degradation under cross-modal perturbation and missing-modality conditions. Removing artifact-aware learning increased artifact-related false-positive responses, while removing early-caries dynamic curriculum (EDC) weighting reduced early-lesion sensitivity. Parameter-matched, query-matched, and auxiliary-head controls produced only limited changes compared with the corresponding baseline variants. In contrast, the complete AGCM-Former achieved the most favorable overall combination of Dataset C mAP, D1/D2 recall, C-ext mAP, artifact FP rate, and missing-modality robustness (Figure 7). These controlled comparisons indicated that the observed performance differences could not be explained by model size, query number, or auxiliary supervision alone.

(2) Structured-query and cross-modal fusion ablations

The structured-query analysis showed progressive improvement as anatomical information was incorporated into query initialization. Random queries achieved a mAP of 66.2 ± 1.2%, whereas tooth-position-only and tooth-position-plus-surface queries achieved 67.4 ± 1.1% and 68.1 ± 1.2%, respectively. The full structured-query design increased mAP to 72.4 ± 0.8%. Proximal recall increased from 70.2 ± 1.8% with random queries to 81.2 ± 1.2% with the complete design. Fissure and cervical recalls increased from 74.6 ± 1.6% and 69.5 ± 1.9% to 81.5 ± 1.3% and 78.1 ± 1.4%, respectively. The largest gain in proximal detection was consistent with the inclusion of adjacency-aware tooth-pair information. The cross-modal fusion results showed that Early and Late Fusion demonstrated smaller improvements under weak cross-modal correspondence, while Vanilla Cross-Attention improved detection performance but remained more sensitive to spatial perturbation and artifact-related interference. Candidate-region-centered cross-modal attention provided a more favorable combination of mAP, D1/D2 recall,

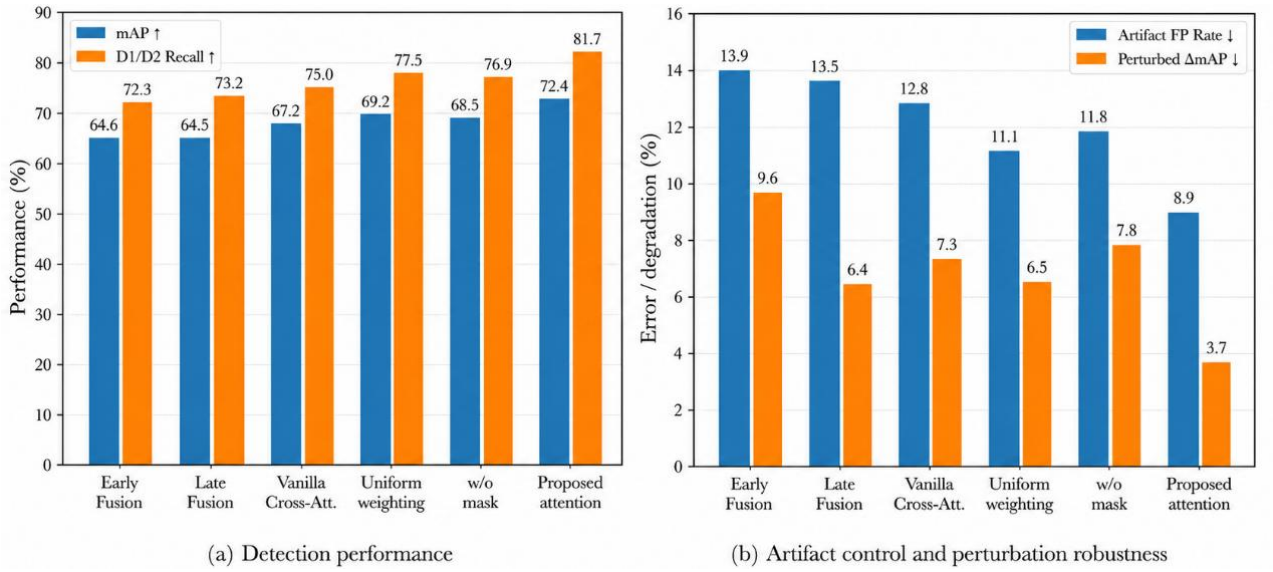


Figure 8. Ablation comparison of multimodal fusion strategies. (**a**) mAP and D1/D2 recall. (**b**) artifact FP rate and perturbation robustness.

artifact FP rate, and perturbation robustness than the alternative fusion strategies (Figure 8).

(3) EDC and sampling-point sensitivity

Removing EDC reduced mAP from 72.4 ± 0.8% to 70.8 ± 1.1% and D1 recall from 74.9 ± 0.9% to 70.2 ± 1.6%. Focal Loss and online hard example mining (OHEM) increased D1 recall to 72.4 ± 1.5% and 75.1 ± 1.5%, respectively, but their artifact FP rates increased to 10.4 ± 0.9% and 11.2 ± 1.0% compared with 8.9 ± 0.6% for the selected EDC setting. The selected EDC configuration also achieved the highest D2 recall of 85.3 ± 0.6%, providing the most favorable overall balance between early-lesion sensitivity and artifact control among the evaluated hard-sample learning strategies. Sampling-point sensitivity showed a similar accuracy as efficiency trade-off. With K = 2, mAP and D1/D2 recall were 70.6 ± 1.1% and 78.9 ± 1.4%, respectively, with an artifact FP rate of 10.4 ± 0.9%. Increasing K to 4 improved these values to 72.4 ± 0.8%, 81.7 ± 1.2%, and 8.9 ± 0.6%, respectively. Further increasing K to 6 or 8 produced only marginal changes in detection performance while reducing forward throughput from 22.8 ± 0.8 FPS at K = 4 to 20.3 ± 0.7 and 18.1 ± 0.6 FPS, respectively. Therefore, K = 4 was retained as the default setting because it provided the most favorable balance between detection performance, artifact control, and computational efficiency. Overall, the ablation analyses supported complementary contributions from anatomy-aware query initialization, candidate-region-centered cross-modal attention, artifact-aware learning, and EDC optimization. The sensitivity experiments further showed that the selected query and sampling configurations provided a practical balance between accuracy, robustness, and computational cost.

Discussion

This study proposed AGCM-Former for lesion-level early caries detection in tooth-level weakly corresponding multimodal oral images. Unlike single-modal detectors or direct multimodal fusion approaches, AGCM-Former did not require strict pixel-level registration between radiographic and intraoral RGB images. Instead, AGCM-Former incorporated anatomical structural information to guide candidate-region-centered multimodal evidence aggregation under weak cross-modal correspondence. Recent studies have similarly emphasized the

importance of evaluating artificial intelligence-based caries detection under independent patient cohorts and realistic acquisition conditions rather than relying exclusively on internal test settings [27, 28]. The results supported the importance of incorporating anatomical context into early-caries detection. Early lesions are often small, low contrast, and strongly dependent on their anatomical location, while proximal, occlusal/fissure, and cervical regions are associated with different lesion appearances and hard-negative patterns. The structured tooth-position–tooth-surface queries restricted candidate generation to anatomically plausible regions, and the ablation results showed that progressively incorporating tooth position, surface information, and adjacent-tooth relationships was associated with improved region-specific detection performance. These observations were consistent with previous studies highlighting the value of anatomical and structural information in dental image analysis [22–24]. Candidate-region-centered cross-modal attention addressed another difficulty of multimodal oral imaging, the absence of reliable pixel-level correspondence. Radiographic and RGB images may differ in viewpoint, acquisition geometry, field of view, and imaging conditions, making direct concatenation or globally aligned fusion susceptible to irrelevant or mismatched information. AGCM-Former instead estimated modality-specific reference locations and aggregated local evidence around anatomically defined candidate regions. The robustness and fusion-ablation experiments showed that this design was less sensitive to spatial perturbation and missing-modality conditions than the evaluated conventional fusion strategies. This finding was broadly consistent with recent multimodal studies using cross-guided or attention-based mechanisms to integrate heterogeneous information [29–31]. Artifact-aware representation learning and early-caries dynamic curriculum (EDC) weighting were associated with complementary improvements in early-lesion sensitivity and false-positive control. Early lesions can share visual characteristics with cervical radiolucency, fissure shadows, restoration margins, and other hard-negative structures. The ablation analysis showed that removing EDC reduced early-lesion sensitivity, whereas alternative hard-sample strategies produced less favorable artifact false-positive rates. Feature-space analysis also showed lower caries–artifact representation similarity for AGCM-Former, while attention analysis indicated more spatially concentrated and stable feature allocation. Nevertheless, t-SNE projections and attention maps should be interpreted as analyses of model behavior rather than clinically validated explanations. Likewise, the uncertainty estimates generated by stochastic inference represented model-level predictive variability and should not be interpreted directly as clinical diagnostic confidence. The study had several limitations. It was retrospective and was primarily based on a self-constructed dataset from a single institution. Although C-ext provided an independent, non-overlapping validation cohort acquired at the same institution using different imaging devices and acquisition conditions, it did not constitute multicenter external validation. Further, D1/D2 lesion severity and artifact annotations were based on expert consensus rather than histopathological confirmation, which might introduce uncertainty in subtle early-lesion boundaries. Furthermore, AGCM-Former relied on automatic tooth localization/numbering or tooth-level weak structural information for query initialization. Errors in upstream tooth localization might propagate to lesion detection. The additional structured-query and multimodal processing introduced greater computational cost than lightweight detectors, and further optimization would be required for resource-constrained or real-time deployment [32, 33]. Future work should include prospective multicenter validation, dentist-involved reader studies, more standardized annotation protocols for subtle early lesions and hard-negative structures, stronger missing-modality and image-quality adaptation, and lightweight model development. Within the scope of the present retrospective experiments, AGCM-Former

provided a structured approach for detecting early caries from weakly corresponding multimodal oral images by combining anatomy-aware query initialization, local cross-modal evidence aggregation, artifact-aware learning, and EDC optimization. Further prospective and multicenter evaluation is required before its clinical utility and generalizability can be established.

Acknowledgements

This study was supported by 2018 and 2025 Beihua University University-level Youth Education and Teaching Reform Research Project (Grant No. XJQN2018131 and XJQN20250008), 2023 and 2024 Scientific Research Project of the Jilin Provincial Department of Education (Grant No. JJKH20230084KJ and JJKH20240098KJ).

References

1. Molyneux LE, Banerjee A. 2024. Minimum intervention oral care: Staging and grading dental carious lesions in clinical practice. Br Dent J. 237(6):457-463.
2. Bayati M, Alizadeh Savareh B, Ahmadinejad H, Mosavat F. 2025. Advanced AI-driven detection of interproximal caries in bitewing radiographs using YOLOv8. Sci Rep. 15(1):4641.
3. Albano D, Galiano V, Basile M, Di Luca F, Gitto S, Messina C, *et al*. 2024. Artificial intelligence for radiographic imaging detection of caries lesions: A systematic review. BMC Oral Health. 24(1):274.
4. Tassery H, Miletic I, Turkun LS, Sauro S, Gurgan S, Banerjee A, *et al*. 2024. Preventive management of carious lesions: From non-invasive to micro-invasive operative interventions. Br Dent J. 236(8):603-610.
5. Pérez de Frutos J, Holden Helland R, Desai S, Nymoen LC, Langø T, Remman T, *et al*. 2024. AI-Dentify: Deep learning for proximal caries detection on bitewing x-ray-HUNT4 oral health study. BMC Oral Health. 24(1):344.
6. Yoon K, Jeong HM, Kim JW, Park JH, Choi J. 2024. AI-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation. J Dent. 141:104821.
7. Karakuş R, Öziç MÜ, Tassoker M. 2024. AI-assisted detection of interproximal, occlusal, and secondary caries on bite-wing radiographs: A single-shot deep learning approach. J Imaging Inform Med. 37(6):3146-3159.
8. Tichý A, Kunt L, Nagyová V, Kybic J. 2024. Automatic caries detection in bitewing radiographs—Part II: Experimental comparison. Clin Oral Investig. 28(2):133.
9. Xiong Y, Zhang H, Zhou S, Lu M, Huang J, Huang Q, *et al*. 2024. Simultaneous detection of dental caries and fissure sealant in intraoral photos by deep learning: A pilot study. BMC Oral Health. 24(1):553.
10. Zhang JW, Fan J, Zhao FB, Ma BE, Shen XQ, Geng YM. 2024. Diagnostic accuracy of artificial intelligence-assisted caries detection: A clinical evaluation. BMC Oral Health. 24(1):1095.
11. Carvalho BKG, Nolden EL, Wenning AS, Kiss-Dala S, Agócs G, Róth I, *et al*. 2024. Diagnostic accuracy of artificial intelligence for approximal caries on bitewing radiographs: A systematic review and meta-analysis. J Dent. 151:105388.
12. Tan ZQ, Pinho IF, Banks R, Chen X, Pontes LRA, Braga MM, *et al*. 2025. Caries detection in primary molars with bitewing radiographs through deep learning based-object detectors. Caries Res. 18:1–9.
13. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, *et al*. 2024. DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Pp. 16965-16974.
14. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, *et al*. 2023. DINO: DETR with Improved DeNoising Anchor Boxes for end-to-end object detection. In: International Conference on Learning Representations (ICLR). Available from: OpenReview.
15. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, *et al*. 2024. YOLOv10: Real-time end-to-end object detection. Adv Neural Inf Process Syst. 37:107984-108011.
16. Wang CY, Yeh IH, Mark Liao HY. 2024. YOLOv9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland. Pp. 1-21.
17. Liu Y, Cheng Y, Song Y, Cai D, Zhang N. 2024. Oral screening of dental calculus, gingivitis and dental caries through segmentation on intraoral photographic images using deep learning. BMC Oral Health. 24(1):1287.
18. Safari M, Fatemi A, Archambault L. 2023. MedFusionGAN: Multimodal medical image fusion using an unsupervised deep generative adversarial network. BMC Med Imaging. 23(1):203.
19. Wang W, He J, Liu H, Yuan W. 2024. MDC-RHT: Multi-modal medical image fusion *via* multi-dimensional dynamic convolution and residual hybrid transformer. Sensors. 24(13):4056.
20. Lin C, Chen Y, Feng S, Huang M. 2024. A multibranch and multiscale neural network based on semantic perception for multimodal medical image fusion. Sci Rep. 14(1):17609.
21. Fu H, Yuan J, Zhong G, He X, Lin J, Li Z. 2024. CF-Deformable DETR: An end-to-end alignment-free model for weakly aligned visible-infrared object detection. In: IJCAI. Pp. 758-766.
22. Hao J, Wong LM, Shan Z, Ai QYH, Shi X, Tsoi JKH, *et al*. 2024. A semi-supervised transformer-based deep learning framework for automated tooth segmentation and identification on panoramic radiographs. Diagnostics. 14(17):1948.
23. Zhou W, Lu X, Zhao D, Jiang M, Fan L, Zhang W, *et al*. 2024. A dual-labeled dataset and fusion model for automatic teeth segmentation, numbering, and state assessment on panoramic radiographs. BMC Oral Health. 24(1):1201.

24. Xu M, Wu Y, Xu Z, Ding P, Bai H, Deng X. 2023. Robust automated teeth identification from dental radiographs using deep learning. J Dent. 136:104607.
25. Ren S, He K, Girshick R, Sun J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 39(6):1137-1149.
26. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations (ICLR).
27. Schwarzmaier J, Frenkel E, Neumayr J, Ammar N, Kessler A, Schwendicke F, *et al*. 2024. Validation of an artificial intelligence-based model for early childhood caries detection in dental photographs. J Clin Med. 13(17):5215.
28. Adnan N, Faizan Ahmed SM, Das JK, Aijaz S, Sukhia RH, Hoodbhoy Z, *et al*. 2024. Developing an AI-based application for caries index detection on intraoral photographs. Sci Rep. 14(1):26752.
29. Lee S, Park J, Park J. 2024. CrossFormer: Cross-guided attention for multi-modal object detection. Pattern Recognit Lett. 179:144-150.
30. Shen J, Chen Y, Liu Y, Zuo X, Fan H, Yang W. 2024. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognit. 145:109913.
31. Yuan M, Wei X. 2024. C²former: Calibrated and complementary transformer for RGB-infrared object detection. IEEE Trans Geosci Remote Sens. 62:1-12.
32. Zong Z, Song G, Liu Y. 2023. DETRs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Pp. 6748-6758.
33. Cheng T, Song L, Ge Y, Liu W, Wang X, Shan Y. 2024. YOLO-World: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Pp. 16901-16911.