

DMITIC: Dentition Defect Diagnosis via Multimodal Instruction-Tuning for CBCT Image Captioning

Anonymous ACL submission

Abstract

The rapid development of LLMs has brought powerful text generation capabilities, leading to significant improvements in image captioning tasks. Addressing the challenges in medical domains, such as limited data availability, complex recognition requirements, and difficult manual annotation, we innovatively extend image captioning to CBCT-based dentition defect diagnosis tasks. Unlike traditional approaches that use semantic segmentation or object detection methods to locate missing teeth, our method only requires standard CBCT images (both with or without missing teeth) as input. Through image-text combined instruction-tuning with our model that integrates CLIP and SAM into BLIP2, we can successfully extract missing tooth location information from CBCT images and provide assessments in textual form. This greatly enhances the ability to reveal clinical information and provides valuable diagnostic assistance to doctors. In terms of performance, our method outperforms both MSMedCap, which is specifically designed for medical imaging, and InstructBLIP, which is trained on general datasets. We have achieved state-of-the-art results in our pioneering approach of using image captioning for dentition defect diagnosis. The key raw data has been uploaded to Research Data Deposit (www.researchdata.org.cn), validating the authenticity of this paper with the RDD number: [REDACTED].

1 Introduction

In dental restoration therapies, Cone Beam Computed Tomography (CBCT) has become an indispensable imaging modality (Huang et al., 2022)(Wei et al., 2024). Its primary objective lies in precisely determining the three-dimensional anatomical locations of missing teeth. Although existing AI technologies achieve missing tooth area identification through semantic segmentation (Wei et al., 2024), traditional computer vision methods

can only obtain coordinate information and fail to interpret the anatomical correlations and pathological characteristics required for clinical reports. The integration of natural language processing (NLP) technology, through constructing mapping models between imaging features and clinical semantics, can surmount the limitations of isolated image analysis. This approach provides intelligent decision support for restorative treatments by incorporating spatial topology and biomechanical relationships.

In recent years, the rapid development of LLM has brought powerful text generation capabilities(Zhao et al., 2023). In this context, the development of image captioning technology has attracted extensive attention from the academic community. Image captioning provides a new technical path for intelligent interpretation of medical images by organically combining computer vision technology and natural language processing technology(Stefanini et al., 2022). Its core goal is to achieve an accurate description of images. In specific applications in the medical field, the input image is usually a radiological image of a patient with corresponding instructions, which can be in the form of a pre-defined set of finite categories or a dynamically generated sequence of words, thus generating a medical report for the clinical practice, revealing a wealth of clinical information, and providing valuable diagnostic assistance to doctors.

In clinical report drafting scenarios, radiologists typically synthesize examination requests from clinicians and patients' medical images to compose clinically valuable reports aligned with diagnostic objectives, thereby supporting diagnostic and therapeutic decision-making. Within this process, how to effectively utilize multimodal information (including textual descriptions and imaging features) to generate accurate and clinically meaningful diagnostic conclusions remains the central research topic in medical imaging artificial intelligence(Pesapane et al., 2023). Although recent ad-

vancements in language models have demonstrated remarkable potential in textual domains, critical challenges persist in specialized domains such as complex dental defect cases: On the one hand, it is difficult for models to identify and localize complex oral anatomical structures accurately (Sloan et al., 2024). It is difficult for purely text-supervised general-purpose models to efficiently capture the details present in an image due to the inherent ambiguity of the language and the varying granularity levels of textual descriptions. This limitation arises from the difficulty image encoders face in capturing subtle feature differences within localized regions of the image that exhibit fuzzy boundaries, noise, and poor contrast. The challenge of unstable generalization capabilities remains to improve model performance. On the other hand, In real clinical settings, patients often present with more complex oral conditions extending beyond simple single-tooth loss, such as multiple missing teeth combined with adjacent root pathologies, insufficient bone volume, or presence of pre-existing dental implants/prosthetic restorations. These complex factors manifest as structural overlaps and artifacts in CBCT images. When encountering challenging scenarios markedly differing from single-tooth loss contexts – particularly multiple missing teeth with intricate bone alterations – models may demonstrate recognition funnel effects. This phenomenon arises from inadequate differentiation and interpretation of superimposed anatomical features and their interactions, ultimately compromising accurate localization of missing tooth positions in diagnostic assessments.

During dataset establishment for tooth loss patterns, we optimized through three key aspects. First, we enhanced diversity by including complex cases with multiple missing teeth, residual roots, restorations, and implants. Second, we maintained balanced distribution between intact and missing teeth to reduce bias. Third, we implemented standardized annotation guidelines with multi-annotator cross-validation to ensure labeling quality.

This research proposes an interactive CBCT image interpretation method based on image captioning. DMITIC introduces a CBCT image without missing teeth as an example image, combined with text as instruction, enabling the model to correctly learn through comparison to extract detailed feature differences between CBCT images with missing and complete teeth to capture fine-grained infor-

mation. Specifically, our model contains a dual-encoder architecture: one ViT image encoder pre-trained using CLIP (Radford et al., 2021a) to extract overall information, and a segmentation model (SAM) (Kirillov et al., 2023a) guided encoder to capture fine-grained details. Both are instruction-tuned through fusion with example image instructions. By adopting unique pretraining strategies and hybrid semantic learning to simultaneously capture overall information and finer details in dental CBCT images. For the specific semantic segmentation detail features that CLIP lacks, SAM provides supplementary learning, and the functions of the two encoders are well-complemented and coordinated. We conducted experiments on various datasets (Lin et al., 2014) to evaluate our model, confirming the effectiveness of our proposed method.

To address the imaging description of dental defects in complex situations, this study proposes an interactive CBCT image interpretation method based on image captioning - DMITIC. The main advantages of DMITIC include:

- We designed an image-text combined instruction-tune mode that enhances feature learning through comparison in traditional image captioning tasks, transforming it into a VQA-like mode to compensate for the disadvantage of being unable to capture fine-grained features in image captioning tasks.
- We innovatively propose a novel task of using image captioning for dentition defect diagnosis in CBCT images, addressing the limited generalizability of traditional semantic segmentation approaches.
- We improved the dual-encoder architecture of MSMedCap and incorporated different forms of instruction at different stages, which can better retain the prompting and constraining effects of instruction, capturing position-granular information such as missing teeth in dental CBCT.
- Our proposed DMITIC demonstrates significantly improved performance in missing tooth position diagnosis on dental CBCT datasets compared to baseline models including BLIP2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and MSMedCap (Zhang et al., 2024).

187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236

2 Related Work

2.1 Current Status of Medical IMAGE CAPTION Research

In the field of medical image automatic analysis, diagnostic report generation faces unique challenges. Unlike general image description tasks, medical data acquisition is strictly regulated, and diagnostic reports typically consist of structurally complex, complete paragraphs. Visually prominent healthy organs in medical images may have low diagnostic relevance, while subtle pathological features require focused attention. Given the critical nature of medical diagnosis, report generation demands extremely high accuracy, as any omission of key information could lead to serious consequences. Nevertheless, the automatic generation of diagnostic reports still holds significant practical value, providing a preliminary reference for clinicians and effectively improving diagnostic efficiency. Current research primarily focuses on predicting simple pathological descriptions, with relatively less attention paid to modeling temporal change features and complex concepts. Existing medical image description methods predominantly employ contrastive learning for cross-modal pre-training. Although these methods demonstrate good performance in general image description tasks, their performance significantly deteriorates in medical scenarios. This performance gap mainly stems from the difficulty of general pre-trained models in effectively capturing fine-grained semantic information in medical images, which is often crucial for accurate diagnosis. Additionally, the inherent ambiguity and noise characteristics of medical images pose significant challenges for feature extraction. The recently proposed MSMedCap model adopts a dual-encoder architecture guided by SAM (Segment Anything Model), achieving simultaneous capture of both global features and local details in medical images through an innovative hybrid semantic learning strategy.

2.2 Related Work on Multimodal Learning

Multimodal learning (Ramachandram and Taylor, 2017) aims to exploit the complementary information between different modal data (e.g., image, text, audio, etc.) to enhance the model’s ability to understand and model multimodal tasks. BLIP (Li et al., 2022b) proposed a self-supervised visual-verbal contrast learning paradigm to achieve better graphic-text matching and cross-modal migra-

tion capabilities by minimizing the contrast loss between image and text and aligning cross-modal features in a shared semantic space. On this basis, BLIP-2 (Li et al., 2023) is an efficient and versatile visual-linguistic pre-training strategy that significantly reduces the computational cost by utilizing pre-trained frozen image encoders and Large Language Models (LLMs). BLIP-2 employs a lightweight Querying Transformer (Q-Former) to bridge the modal gap through a two-stage pre-training to bridge the modal gap. The first stage leads to visual-verbal representation learning from a frozen image encoder, and the second stage leads to visual-to-verbal generative learning from a frozen language model, thus achieving zero-sample image-to-text generative capability. Despite having far fewer parameters than existing methods, BLIP-2 achieves state-of-the-art performance on several visual-linguistic tasks, e.g., outperforming Flamingo80B by 8.7% on zero-sample VQAv2 while reducing the parameter count by a factor of 54.

2.3 Baseline Methods

Zhang et al. (Zhang et al., 2024) proposed a novel medical image caption generation model, MSMedCap, which adopts a dual-encoder architecture and a hybrid semantic learning strategy to capture the overall information and fine-grained details of medical images. MSMedCap consists of two image encoders: a ViT (Dosovitskiy, 2020) encoder based on CLIP pre-training for extracting the overall features and a SAM-based encoder for extracting the fine-grained features. MSMedCap contains two image encoders: a ViT encoder based on CLIP pre-training to extract overall features and a SAM-based encoder to extract fine-grained features. Given an input image, the two encoders encode it into different image embedding vectors. Next, a dual Query Transformer (Q-Former) is used to cross-attentionally align the output features of the two encoders to obtain the aligned features. Finally, the aligned features are spliced with textual cue embeddings and fed into a pre-trained language model to generate medical image descriptions.

3 Method

In this section, we first describe the model architecture, followed by the pre-training strategy and how to help the model better extract fine-grained features of CT image locations through image-text

237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285

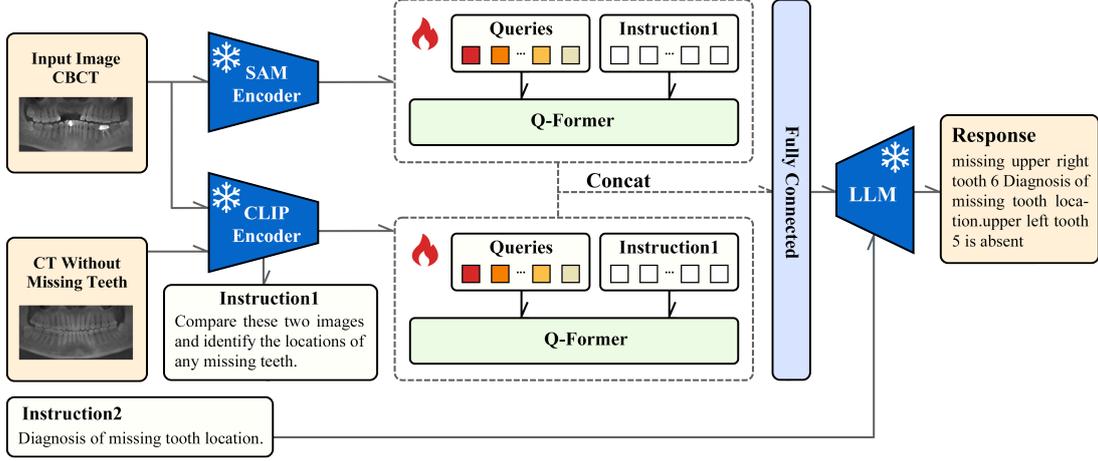


Figure 1: The DMITIC architecture is divided into two instructions: Instruction 1 is multimodal, where standard example images are processed through CLIP to extract image features, which are then transformed into text embeddings and combined with prompts. Instruction 2 is purely text-based.

combined instructions, thereby using Large Language Models (LLM) to generate medical image captions.

3.1 Model Architecture

As shown in Figure 1, similar to MSMedCap, we adopt dual encoders consisting of CLIP(Radford et al., 2021b) and SAM(Kirillov et al., 2023b) for feature extraction, where the example image only goes through one CLIP encoder for feature extraction. This is because the example image serves the same purpose as the SAM encoder - helping the model better extract detailed feature information when searching for missing tooth information, thus achieving effects that other pre-trained models cannot capture. Next, visual features are extracted from the frozen image encoders through dual Query Transformers (Q-Former)(Li et al., 2023).

The Q-Former output consists of K encoded visual vectors, one for each query embedding, which are then linearly projected and input to the frozen LLM. Like MSMedCap, before instruction tuning, Q-Former is pre-trained with image caption data in two phases. The first phase pre-trains Q-Former using frozen image encoders for visual-language representation learning. The second phase adapts Q-Former’s output as soft visual prompts for text generation using the frozen LLM. After pre-training, we fine-tune Q-Former through instruction tuning, where phase one uses instruction1, which combines non-missing tooth CBCT example images with prompts and instructions as input. The second phase uses text instructions alone, where LLM receives visual encoding from Q-Former and task

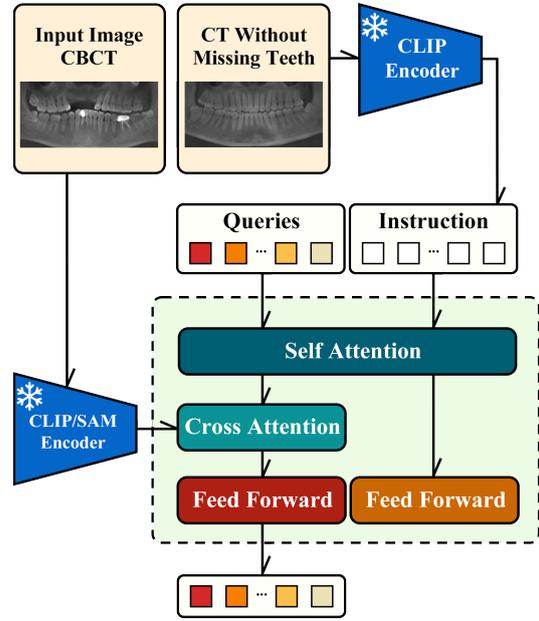


Figure 2: The Q-Former architecture, instruction 1 participates in the model’s stage 1 training phase, ensuring information extraction without affecting stage 2’s text generation.

instructions as input to generate missing tooth diagnoses.

3.2 Dual Image Encoder

We use two ViT encoders trained based on CLIP and SAM, namely f_{CLIP} and f_{SAM} , to encode image features. Image x is input to both encoders, producing two different sets of image embedding vectors:

$$v_{CLIP} \in R^{N \times C}, v_{SAM} \in R^{Q \times S} \quad (1)$$

$$v_{CLIP} = f_{CLIP}(x) \quad (2)$$

$$v_{SAM} = f_{SAM}(x) \quad (3)$$

where N and Q represent the number of feature vectors, C and S represent the dimension of each feature vector.

Dual Query Transformer (Q-Former): Features output from the dual encoders are processed through cross-attention through their respective Q-Formers (g_{CLIP} and g_{SAM}) to produce aligned features:

$$\tilde{v}_{CLIP} = g_{CLIP}(q_{CLIP}, v_{CLIP}) \quad (4)$$

$$\tilde{v}_{SAM} = g_{SAM}(q_{SAM}, v_{SAM}) \quad (5)$$

where $q_{CLIP}, q_{SAM} \in R^{M \times D}$ are two sets of learnable query vectors in Q-Former. Note that Q-Formers' output layers use linear projection layers.

3.3 Hybrid Semantic Pre-training

Models pre-trained using different methods produce different granularity and semantic information when extracting features from images. By leveraging the advantages of various pre-training methods, we adopted a training strategy that combines general image information with medical domain-specific image information, as shown in Figure 3.

In this phase, we trained Q-Former separately for CLIP and SAM. First, we froze the image encoders and input the diagnostic CBCT images into both CLIP and SAM encoders, while inputting the non-pathological CBCT examples only to the CLIP encoder for feature extraction. We then input both trainable Soft Queries and embedded image captions into Q-Former, connecting encoder-extracted image features to Q-Former through Cross Attention. Inspired by BLIP2, we optimized Q-Former for three objectives using corresponding masks in Self Attention to meet different requirements.

Our Q-Former optimization encompasses three key objectives. First, Image-Text Matching (ITM) focuses on classifying the relevance between image and text input pairs. Second, Image-based Text Generation (ITG) enables the generation of descriptive text based on image inputs. Finally, Image-Text Contrastive Learning (ITC) optimizes feature representation by minimizing distances between matching image-text pairs while maximizing distances for unrelated pairs.

Since extracting missing tooth location features requires combining general semantic information and fine-grained image details, our model training

process involves achieving hybrid semantic representation learning of CLIP and SAM. Given that CLIP excels at capturing more general semantic information, we aim to preserve this capability in our model. Therefore, we encode both input images through CLIP. In contrast, the SAM image encoder has been pre-trained on segmentation tasks, making it better at capturing fine-grained image details than CLIP. To capture medical image details like pixel-level semantics, we use a combination of general and medical datasets to train SAM's Q-Former. We demonstrate in subsequent experiments that this training strategy is more effective than other methods in maximizing feature diversity.

3.4 Caption Generation Using Frozen LLM

Vicuna-7B (Chiang et al., 2023) is used for generating medical captions. Vicuna is a decoder-only Transformer fine-tuned from LLaMA (Touvron et al., 2023). During visual-language instruction tuning, we initialize the model from the BLIP-2 checkpoint pre-trained using only COCO (Lin et al., 2014), and only fine-tune Q-Former's parameters while keeping image encoders and LLM frozen. Having completed hybrid semantic pre-training and through image-text instructions with example images, the model has already developed the ability to distinguish between the presence and absence of missing teeth, identify missing tooth locations, and effectively align with text. In this phase, we fine-tune the entire model on our collected DM-Tooth dataset using frozen LLM to generate medical image captions. We use Vicuna-7B as our LLM. Initially, we freeze all parameters of both image encoders and LLM, focusing only on training Q-Formers and linear projection layers. Through instructions, we help the model focus on features extracted by CLIP and SAM encoders. Finally, their respective Q-Formers and linear projection layers are connected and input to LLM. The model is trained using LLM loss.

4 Experimental Setup

4.1 Datasets

Our experiments utilize the publicly available COCO dataset (Lin et al., 2014) and a private dataset (Diagnosis of Missing Teeth dataset, DM-Tooth).

- **Public Dataset:** We use COCO for pre-training. While we employ the train2014 version, we do not use it for testing since it

428	contains no dental CBCT images and is not	477
429	aligned with our specific task.	478
430	• DMTooth Dataset: This research strictly ad-	479
431	heres to the Declaration of Helsinki and was	480
432	approved by the Ethics Committee. The	481
433	dataset comprises CBCT images and clinical	482
434	data from patients with dental defects between	483
435	July 2019 and October 2023. Inclusion criteria	484
436	were: (1) confirmed diagnosis of dental	485
437	defects; (2) age ≥ 18 years; (3) exclusion of	486
438	CBCT images with severe distortion due to	487
439	orthodontic treatment or metallic restorations.	488
440	Imaging data was acquired using two devices:	
441	NewTom (QR srl, Verona, Italy) and Care-	
442	stream Health CS 9300 (Carestream Health	
443	Inc, Rochester, NY, USA). Cone Beam Com-	
444	puted Tomography Reconstructed Panoramic	
445	(CRP) images were reconstructed using CS	
446	3D Image 3.4.3 software, with the mandibu-	
447	lar dental arch curve as the reference. Each	
448	CBCT image was accompanied by a standard	
449	radiological diagnosis report from oral radi-	
450	ologists, from which descriptions of missing	
451	tooth locations were extracted. The key raw	
452	data has been uploaded to Research Data De-	
453	posit (www.researchdata.org.cn), validating	
454	the authenticity of this paper with the RDD	
455	number: ██████████.	
456	4.2 Data Preprocessing and Dataset	
457	Construction	
458	DMTooth consists of 400 samples split in a 7:1:2	
459	ratio for training, validation, and testing. GPT4o	
460	was used to process standard radiological reports,	
461	transforming missing tooth location descriptions	
462	into different image-text pairs. Each image cor-	
463	responds to 1-2 texts with identical meaning but	
464	varied expressions to enhance training set diversity.	
465	The final training set contains 567 image-text pairs.	
466		
467	4.3 Evaluation Metrics	
468	We employ BLEU (Papineni et al., 2002), ME-	
469	TEOR (Banerjee and Lavie, 2005), ROUGE-	
470	L (Lin, 2004), CIDEr (Vedantam et al., 2015),	
471	BERTSCORE (Zhang et al., 2020) as evaluation	
472	metrics. In this task, we also need to evaluate the	
473	accuracy of missing tooth position detection. There-	
474	fore, under the specified answer format, we used	
475	accuracy and F1 scores to assess the precision of	
476	position detection. For ease of comparison, we	
	scaled the scores for each metric, as shown in Ta-	477
	ble 2.	478
	Higher metric scores indicate better quality of	479
	generated results. Our best model, termed DMITIC,	480
	is compared with state-of-the-art models MSMed-	481
	Cap, BLIP-2, and InstructBLIP, along with vari-	482
	ations using different instructions. We conduct 3	483
	training iterations on COCO dataset. For DMTooth,	484
	we perform 10 iterations for both hybrid semantic	485
	pre-training and captioning phases. Different ex-	486
	ample images were compared to select the most	487
	suitable one.	488
	4.4 Training and Hyperparameters	489
	Implementation, training, and evaluation were con-	490
	ducted using the LAVIS library (Li et al., 2022a).	491
	All models underwent instruction tuning for up	492
	to 60K steps with validation every 3K steps. For	493
	each model, the best checkpoint was selected for	494
	evaluation across all datasets. We use batch sizes	495
	of 128 for COCO pre-training and 8 for DM-	496
	Tooth, considering dataset sizes. Training em-	497
	ployes AdamW optimizer (Loshchilov, 2017) with	498
	$\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.05.	499
	Learning rate undergoes linear warmup from 10^{-8}	500
	to 10^{-5} in the first 5,000 steps, followed by cosine	501
	decay to 0. All models were trained on 4 Nvidia	502
	A40 GPUs, completing in 1.5 days.	503
	5 Results and Discussion	504
	5.1 Model evaluation result	505
	We evaluated the DMITIC model on the DMTooth	506
	dataset. We compared DMITIC with previous state-	507
	of-the-art models including MSMedCap, BLIP-2,	508
	and InstructBLIP. As shown in Table 1, to ensure	509
	fair evaluation, we also fine-tuned these three mod-	510
	els on DMTooth, and the results demonstrate that	511
	we achieved state-of-the-art performance on the	512
	DMTooth dataset. Considering the significantly	513
	low evaluation scores for other models, our pro-	514
	posed task introduces a novel challenge. Except for	515
	MSMedCap, which has been trained on medical-	516
	related datasets, the other models lack prior knowl-	517
	edge related to CBCT data and struggle to accu-	518
	rately identify missing tooth positions. This lack of	519
	domain-specific training makes it difficult for mod-	520
	els like BLIP-2 and InstructBLIP to generate the	521
	required responses, even after additional training.	522

Models	Bleu ($\times 10^3$)	METEOR ($\times 10^3$)	ROUGE L ($\times 10^2$)	CIDEr ($\times 10^3$)	BERT score ($\times 10^2$)
BLIP2	0.9	2.1	2.1	0.1	8.6
InstructBLIP	7.2	3.9	4.6	0.2	50.1
MSMedCap	3.4	10.3	11.0	0.7	65.8
DMITIC-OT	63.6	17.1	28.1	18.9	86.6
DMITIC-SA	38.7	13.9	20.6	15.6	85.9
DMITIC-NS	75.4	25.2	44.5	40.0	90.8
DMITIC-NC	5.1	3.4	4.7	0.2	55.6
DMITIC	80.2	35.3	52.3	39.8	93.2

Table 1: Comparison with Benchmarks Across Different Evaluation Metrics.

5.2 Ablation Study on Instruction Tuning

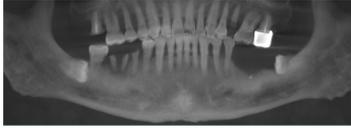
To validate the effectiveness of our image-text combined instruction approach and dual-encoder architecture, we conducted comparative experiments with several variants. The model naming conventions are shown in Table 1:

- DMITIC-OT: using text instruction only
- DMITIC-SA: using dual-encoder architecture only
- DMITIC-NS: using CLIP encoder only
- DMITIC-NC: using SAM encoder only

The final results reveal a significant performance drop when only SAM is used or when multimodal instructions are absent. SAM is fundamentally designed for segmentation tasks, and without the feature support from CLIP, it cannot perform effectively. This observation underscores our decision to use CLIP for feature extraction on the example images.

5.3 Qualitative Assessment

We qualitatively compared DMITIC with concurrent multimodal models (gpt-4o-2024-08-06(Brown, 2020), Llama-3.1-405B(Dubey et al., 2024), Gemini-1.5-Pro(Team et al., 2023)). For dental CBCT missing tooth diagnosis, these models cannot directly determine missing tooth locations from images alone. While they can generally provide location-based diagnoses following prompts, they often fail to identify correct positions. When provided with example images following our approach, GPT-4o could generate responses in the correct format, but only achieved partial accuracy in 1 out of 5 test cases.

[Input Image] 

[Input Instruction] Diagnosis of missing tooth location.

[Model Response]

DMITIC Lower left teeth 6 and 7 are absent.

MSMedCap the missing tooth is in the front of the person's mouth

GPT4o In the provided panoramic X-ray image, it appears that a tooth is missing on the lower left side of the jaw, likely in the molar region. Additionally, a dental implant is visible in this area, likely placed to restore the missing tooth.

Figure 3: Answer templates from different LLMs were examined. DMITIC was able to provide concise and accurate responses for missing tooth positions. Note: CBCT panoramic images are shown from the doctor’s perspective, but results are output from the patient’s perspective, so the bottom right of the image corresponds to the patient’s lower left.

Although all models can generate task-relevant responses, their text outputs tend to be overly complex with poor readability and cannot critically accurately identify missing tooth locations. Importantly, we argue that lengthy responses are not always desirable. Our DMITIC model typically provides more precise answers focusing on key location information, thanks to our image-text combined instruction tuning approach.

5.4 Accuracy Testing

For the task of missing tooth detection, accurate position determination is crucial. Simply generating image captions may lead to issues similar to concurrent multimodal models. For example, if a missing tooth is at position 6 but is identified as position 7, although the generated text can correctly identify the presence of a missing tooth and follows the positional judgment format, the position given is incorrect, making it an invalid generation. Therefore, we restricted the format of generated text through prompts and extracted only the missing tooth position information as labels to create new outputs for evaluation. We set "no missing tooth" as 0, and converted other positions to FDI tooth notation, then compared Accuracy and F1 scores. In our test set, there are 80 CBCT images, of which 53 contain missing teeth, totaling 95 missing teeth. The remaining 27 images without missing teeth are each counted as a single instance, as we aim to assess the model's detection capability. This brings the total count for evaluation to 122. The resulting accuracy and F1 scores are shown in Table 2. As observed, our accuracy surpasses that of other methods; however, the overall performance is still suboptimal, indicating room for further improvement.

Models	Accuracy	F1 scores
BLIP2	*	*
InstructBLIP	*	*
MSMedCap	*	*
DMITIC-OT	0.178	0.054
DMITIC-SA	0.111	0.022
DMITIC-NS	0.356	0.187
DMITIC-NC	*	*
DMITIC	0.667	0.635

Table 2: Accuracy and F1 scores. * indicates that the model does not provide any valid answers regarding accurate positions.

6 Conclusions

Building upon MSMedCap's demonstrated capability in capturing fine-grained features from medical datasets, we enhanced the SAM framework. We extended the image captioning task to missing tooth diagnosis, which requires specific location information and fine-grained feature detection. On our collected DMTooth dataset, we introduced a novel image-text combined instruction tuning approach that addresses previous models' limitations in han-

dling location information, enabling the model to generate location-specific text descriptions using detailed positional features. This method outperforms baseline models across all metrics, significantly improving output quality. Our successful integration addresses BLIP-2's limitations in providing professional and detailed medical diagnoses.

Currently, besides our proposed DMTooth dataset, other CBCT-related datasets focus on semantic segmentation or object detection of tooth positions. Leveraging such data for specific diagnostic text generation remains challenging. Additionally, medical dataset limitations prevent zero-shot learning capabilities. Future work will extend this approach to other dental conditions, such as caries and restorations.

Limitations

Although DMITIC demonstrates outstanding performance in describing complex dentition defects, the expansion of its capabilities to multi-description scenarios holds significant research value. Clinical cases often involve intricate combinations of multiple oral abnormalities (e.g., tooth loss combined with retained root fragments, dental implant artifacts, etc.), requiring the model to generate hierarchical descriptions integrating spatial topological relationships, pathological features, and biomechanical correlations. Future research could explore a dynamic multi-instruction tuning framework that enables adaptive prioritization of diagnostic subtasks based on CBCT feature saliency, potentially realized through an attention-based description path selection mechanism - allowing simultaneous disentanglement of overlapping anatomical features while maintaining contextual coherence. Additionally, developing temporal modeling capabilities for sequential CBCT scans would enhance clinical value in progressive defect tracking. To achieve these objectives, we plan to extend the DMTooth dataset with longitudinal multi-defect cases and develop description completeness evaluation metrics for composite pathologies. Such expansions will effectively bridge the technical gap between isolated defect descriptions and comprehensive diagnostic report generation, laying the foundation for constructing a multimodal diagnostic and therapeutic decision-making system in oral healthcare.

References

- 650 Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- 656 Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- 658 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- 666 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. **Instructblip: Towards general-purpose vision-language models with instruction tuning**.
- 670 Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- 673 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 678 Zelun Huang, Haoran Zheng, Junqiang Huang, Yang Yang, Yupeng Wu, Linhu Ge, and Liping Wang. 2022. **The construction and evaluation of a multi-task convolutional neural network for a cone-beam computed-tomography-based assessment of implant stability**. *Diagnostics*, 12(11):2673. Publisher: MDPI.
- 684 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, and Wan-Yen Lo. 2023a. **Segment anything**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.
- 690 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023b. **Segment anything**.
- 695 Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven Hoi. 2022a. Lavis: A library for language-vision intelligence. *arXiv preprint arXiv:2209.09019*.
- 699 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. **Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models**. In *International conference on machine learning*, pages 19730–19742. PMLR.
- 704 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. **Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation**. In *International conference on machine learning*, pages 12888–12900. PMLR.
- 709 Chin-Yew Lin. 2004. **Rouge: A package for automatic evaluation of summaries**. In *Text summarization branches out*, pages 74–81.
- 712 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. **Microsoft COCO: Common Objects in Context**. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.
- 720 I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- 722 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- 727 Filippo Pesapane, Priyan Tantrige, Paolo De Marco, Serena Carriero, Fabio Zugni, Luca Nicosia, Anna Carla Bozzini, Anna Rotili, Antuono Latronico, Francesca Abbate, and 1 others. 2023. Advancements in standardizing radiological reports: a comprehensive review. *Medicina*, 59(9):1679.
- 733 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021a. **Learning transferable visual models from natural language supervision**. In *International conference on machine learning*, pages 8748–8763. PMLR.
- 739 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- 746 Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108.
- 750 Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*.
- 754 Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

760 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-
761 Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan
762 Schalkwyk, Andrew M Dai, Anja Hauth, Katie Mil-
763 lican, and 1 others. 2023. Gemini: a family of
764 highly capable multimodal models. *arXiv preprint*
765 *arXiv:2312.11805*.

766 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier
767 Martinet, Marie-Anne Lachaux, Timothée Lacroix,
768 Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal
769 Azhar, and 1 others. 2023. Llama: Open and effi-
770 cient foundation language models. *arXiv preprint*
771 *arXiv:2302.13971*.

772 Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi
773 Parikh. 2015. [Cider: Consensus-based image de-
774 scription evaluation](#). In *Proceedings of the IEEE*
775 *conference on computer vision and pattern recogni-
776 tion*, pages 4566–4575.

777 Lifen Wei, Shuyang Wu, Zelun Huang, Yaxin Chen,
778 Haoran Zheng, and Liping Wang. 2024. [Autologous
779 Transplantation Tooth Guide Design Based on Deep
780 Learning](#). *Journal of Oral and Maxillofacial Surgery*,
781 82(3):314–324. Publisher: Elsevier.

782 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q.
783 Weinberger, and Yoav Artzi. 2020. [BERTScore:
784 Evaluating Text Generation with BERT](#). *arXiv*
785 *preprint*. ArXiv:1904.09675.

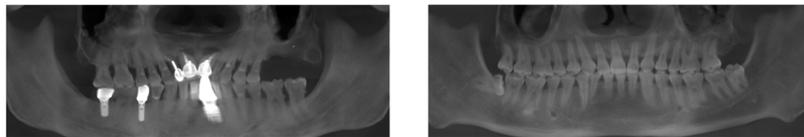
786 Zhenyu Zhang, Benlu Wang, Weijie Liang, Yizhi Li,
787 Xuechen Guo, Guanhong Wang, Shiyan Li, and
788 Gaoang Wang. 2024. [Sam-guided enhanced fine-
789 grained encoding with mixed semantic learning for
790 medical image captioning](#). In *ICASSP 2024-2024*
791 *IEEE International Conference on Acoustics, Speech*
792 *and Signal Processing (ICASSP)*, pages 1731–1735.
793 IEEE.

794 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang,
795 Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen
796 Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023.
797 A survey of large language models. *arXiv preprint*
798 *arXiv:2303.18223*.

799 A Appendix

Dental Diagnosis

[Input Image]



[Input Instruction] Diagnosis of missing tooth location.

[Model Response]

DMITIC
(Not using multimodal instruction)

upper right tooth 7 is absent

DMITIC
(Using multimodal instruction)

upper left teeth 6 and 7 are absent

Blip2

upper left corner of the mouth

Instructblip

the missing tooth is in the front of the person's mouth

MSMedCap

The image is a dental x-ray of a person's face, specifically focusing on the upper jaw and teeth. The x-ray reveals that there is a missing tooth in the upper jaw, which is indicated by the yellow arrow pointing to the location of the missing tooth.

Dental Diagnosis - No sample Image Introduced

[Input Image]



[Input Instruction] Diagnosis of missing tooth location.

[Model Response]

DMITIC
(Not using multimodal instruction)

upper right tooth 7 is absent

gpt-4o-2024-08-06 The panoramic X-ray image you provided shows several missing teeth, along with what appears to be dental implants. Here's a general assessment:

1. Missing Teeth:

- It looks like the posterior teeth in the lower jaw, particularly in the molar regions, may be missing and possibly replaced by implants.
- The upper jaw also shows some gaps that might indicate missing teeth.

2. Dental Implants:

- There are visible dental implants, likely replacing some of the missing teeth.
- The bright, cylindrical structures in the lower jaw are likely dental implants supporting crowns or bridges.

To provide a precise diagnosis, a dentist or radiologist would need to examine this X-ray carefully, along with a clinical examination and history, as specific identification of each missing tooth location requires detailed analysis.

Llama-3.1-405B

Based on the X-ray, it appears that the missing teeth are located in the upper front area of the mouth. The presence of two dental implants in this region suggests that these are the areas where the teeth are missing.

Gemini-1.5-Pro

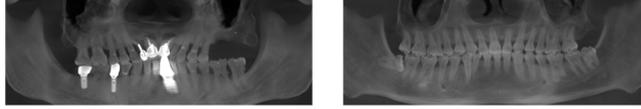
This panoramic dental X-ray shows several missing teeth. Specifically:

- Missing lower incisors: Two implants are clearly visible in the position of the lower central incisors. This indicates the patient's natural lower central incisors were missing and replaced with implants.
- Missing lower left lateral incisor: The space between the left central incisor implant and the left canine suggests a missing lower left lateral incisor. It's unclear from this image alone whether this tooth was extracted or never developed. There's no obvious implant or retained root present.
- Missing lower right second premolar: There is a distinct gap between the lower right first premolar and the lower right first molar. This indicates a missing lower right second premolar. Again, the reason for its absence (extraction, agenesis) isn't discernible from just this image.

It's important to note: A panoramic X-ray provides a good overview, but it's a 2D representation of a 3D structure. For a definitive diagnosis and treatment planning, other imaging techniques (like a CBCT scan) and a clinical examination are necessary. This interpretation is based solely on the provided image.

Dental Diagnosis - Sample Image Introduced

[Input Image]



[Input Instruction] Compared with those without missing teeth. Diagnosis of missing tooth location.

[Model Response]

DMITIC

(Using multimodal instruction)

upper left teeth 6 and 7 are absent

gpt-4o-2024-08-06

The two images you uploaded show dental panoramic X-rays.

Observations:

Image 1: The dentition appears mostly intact, with no significant missing teeth visible. The alignment seems generally within a normal range.

Image 2: Several teeth are missing, and dental implants are visible. The implants appear to replace missing teeth on the upper arch, particularly around the central area and towards the back molar regions.

Diagnosis of Missing Tooth Locations:

In Image 2, missing teeth are evident in the upper arch:

The central area has two dental implants, which replace missing teeth likely around the anterior (front) region.

Additional implants or gaps may be present towards the back molars, indicating missing posterior teeth.

If you need a more detailed analysis or specific annotations on these images, please let me know!

Llama-3.1-405B

The missing tooth locations are in the upper front area of the mouth. Two dental implants are visible, indicating that the missing teeth are likely incisors. The implants appear to be positioned in a way that suggests the missing teeth are the upper central incisors, which are the two front teeth on the top row.

Gemini-1.5-Pro

The provided images are panoramic and cross-sectional dental X-rays. They show several missing teeth and dental implants. Let's break down the analysis:

Reasoning:

1. Panoramic X-ray (First Image): This view provides a broad overview of the entire jaw. We can see several missing teeth and the presence of dental implants. The implants appear as bright, dense structures integrated into the jawbone. They are typically topped with abutments, which are not clearly visible in this image, but would support replacement teeth (crowns).

2. Cross-sectional X-ray (Second Image): This view offers a detailed look at a specific section of the jaw. While it shows the existing teeth clearly, it doesn't directly show the missing teeth or implants because it's a slice through a different plane. It helps confirm the overall dental health and bone structure.

3. Missing Tooth Locations: Based on the panoramic X-ray, the following teeth appear to be missing:

o Multiple teeth in the anterior maxilla (upper jaw): The exact number is difficult to determine from this image, but there's a clear gap where multiple teeth should be. The implants in this region suggest that these missing teeth have been or are planned to be replaced with implant-supported restorations.

o Possibly a tooth in the posterior mandible (lower jaw, left side): There's a slight possibility of a missing tooth on the lower left side, but the image isn't completely clear. Further investigation with additional X-rays or clinical examination would be necessary to confirm this.

Diagnosis:

The diagnosis is simply missing teeth at the specified locations. The cause of the missing teeth isn't apparent from the X-rays.

Common reasons for missing teeth include:

- Dental decay: Severe cavities can lead to tooth extraction.
- Gum disease (periodontal disease): This can weaken the supporting structures of the teeth, eventually leading to tooth loss.
- Trauma: Injury to the mouth can knock out or damage teeth beyond repair.
- Congenital absence: Some individuals are born missing certain teeth.

Comparison with those without missing teeth:

Individuals without missing teeth have a complete dentition. Missing teeth can lead to various issues, including:

- Difficulty chewing and speaking: Missing teeth can impair proper function.
- Shifting of remaining teeth: Teeth can drift into the gaps left by missing teeth, leading to misalignment.
- Bone loss in the jaw: The stimulation provided by teeth is essential for maintaining bone density. Missing teeth can result in bone resorption.
- Aesthetic concerns: Gaps in the smile can affect appearance and self-confidence.

Important Note: This analysis is based on the provided images and is for informational purposes only. A definitive diagnosis and treatment plan can only be provided by a qualified dental professional after a thorough clinical examination.