Vision-Language Model for Multitask Medical Text Generation

Hongkun Sun[1], Lichen Xue[1], and Lushuai Jia[1]

Tomorrow Medical Network Technology Co., Ltd, HangZhou, 310000, China
jialushuai@tomtaw.com.cn

Abstract. Artificial intelligence (AI) has made significant progress in the healthcare domain, where multimodal large models integrating medical imaging and text have garnered considerable attention, yet remain challenging, particularly in generative tasks. This study develops a vision-language model architecture specifically tailored for medical scenarios, based on multimodal medical images (e.g., X-ray, ultrasound, and ophthalmic images) and their corresponding textual descriptions. The model demonstrates remarkable adaptability across diverse imaging modalities and integrates multiple key functionalities, including medical report generation, visual question answering (VQA), and lesion detection in medical images. In the regression task of the MICCAI FLARE 2025 Task 5 challenge, our model achieves state-of-the-art performance with an error of only 13.63 and a detection score of 0.80, classification score of 0.70. It exhibits potential as a unified interface for radiological diagnosis, promising to significantly enhance diagnostic efficiency across various medical imaging applications. Our code have been made publicly available at here.

Keywords: Multimodal · Vision-language · Medical imaging applications.

1 Introduction

Large language models acquire general semantic understanding capabilities through pretraining on massive text corpora, achieving breakthrough advances in natural language processing. Models such as the GPT series and LLaMA demonstrate powerful text generation, question answering, and reasoning abilities. In the medical domain, specialized LLMs like BioBERT and ClinicalBERT, fine-tuned on medical literature and clinical records, significantly enhance the accuracy of entity recognition (e.g., drugs, diseases) and medical question answering. Deep learning-based vision models (e.g., CNNs, Vision Transformers) have become core tools in medical image analysis. Models such as CheXNet (chest X-ray diagnosis) and UNETR (3D medical image segmentation) surpass traditional approaches in tasks like lesion detection and organ segmentation through end-to-end learning. To integrate visual and textual information, multimodal models achieve cross-modal alignment via joint training. General-domain models like CLIP (image-text matching) and BLIP-2 (visual question answering)

demonstrate the feasibility of cross-modal understanding. In healthcare, Med-Flamingo (based on Flamingo architecture) and RadFM (radiology-specific) integrate imaging with diagnostic reports, supporting medical image captioning and visual question answering (VQA).

In the field of multimodal AI, the Qwen-VL series has consistently focused on open-source development and iterative improvements. For instance, Qwen2.5-VL pioneered a dynamic resolution mechanism (spatial dimension) and dynamic frame-rate sampling (temporal dimension), enabling cross-scale analysis from single images to hours-long videos with second-level event localization. It incorporates windowed attention mechanisms to reduce computational complexity and employs SwiGLU activation functions with RMSNorm to enhance vision-language alignment efficiency. LLaVA is designed to improve large language models' understanding of visual content through diverse multimodal instructions. This enhanced comprehension is critical for integrating different types of data inputs. XrayGPT freezes the MedClip visual encoder and Vicuna language model, training only a linear projection layer to achieve cross-modal fusion, thereby reducing data requirements. MedGemma's multimodal variants utilize SigLIP image encoders specifically pretrained on various de-identified medical datasets, including chest X-rays, dermatological images, ophthalmic images, and histopathology slides. Their large language model (LLM) components are trained on diverse medical datasets, encompassing medical texts, medical question-answer pairs, FHIR-based electronic health records (exclusive to the 27B multimodal version), radiology images, histopathology samples, ophthalmic images, and dermatological images.

2 Method

2.1 Model architecture

Our model architecture, as illustrated in the Fig.1, consists of three key components: a vision encoder, a linear projection layer, and a large causal language model.

2.2 Vision Encoder

The vision encoder employs the visual branch of MedSigLIP[1]. MedSigLIP is a variant of SigLIP (Sigmoid Loss for Language-Image Pre-training), specifically designed for cross-modal alignment between medical images and text, mapping medical images and their corresponding textual descriptions into a shared embedding space through contrastive learning. The model supports an input image resolution of 448×448. Throughout the entire training process in this study, the parameters of this visual backbone are kept fixed and frozen to preserve the representational capabilities acquired during pre-training on medical vision-language tasks.

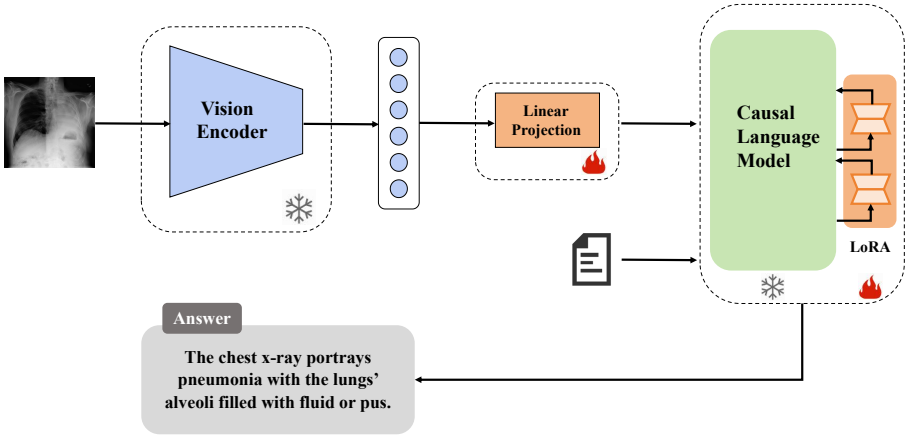

Fig. 1. Overview of our multimodal model architecture.

2.3 Vision Language Alignment

To effectively inject visual information into the language model, this paper employs a learnable linear projection layer to perform cross-modal dimensionality transformation on the visual features. Specifically, four spatially adjacent visual tokens are first aggregated to generate a dimensionality-reduced joint embedding representation, thereby reducing the sequence length and enhancing the semantic consistency of local features. Subsequently, the aggregated visual features are projected into the latent space of the language model through a trainable linear mapping layer, enabling cross-modal semantic alignment. During training, the parameters of this projection layer are continuously updated via end-to-end backpropagation to optimize the fusion of visual and linguistic representations.

2.4 Causal Language Model

In this study, we adopt the open-source causal language model LLaMA2-Chat (7B)[2] as the core linguistic backbone. This large language model has been pre-trained on extensive text corpora and has internalized rich linguistic and domain-specific knowledge, including a broad range of medical expertise. Leveraging its strong capabilities in language understanding and generation, we employ it as a unified interface to handle various medical vision-language tasks, enabling cross-modal semantic reasoning and natural language response generation. To enhance the model's adaptability to specific tasks while preserving the stability of pre-trained knowledge, we apply the Low-Rank Adaptation (LoRA) method for parameter-efficient fine-tuning. Specifically, only the low-rank decomposition matrices are optimized, allowing for effective adaptation and performance improvement under limited computational resources.

2.5 Prompt Template

To enable unified modeling and effective generalization across diverse medical vision-language tasks, this paper proposes a structured prompt template that covers a variety of tasks, including Visual Question Answering (VQA), Report Generation, Object Counting, Disease Detection, and Grounded Image Understanding and Detection. To mitigate potential ambiguities in instruction semantics under multi-task learning scenarios and enhance the model's ability to recognize task intentions, we introduce explicit task-specific identifiers into the training framework to clearly distinguish between different task types. Based on this design, the instruction templates are systematically constructed to ensure clarity of input semantics and consistency in task orientation, thereby improving the model's understanding and reasoning performance in cross-task settings. We present diverse prompt templates in Table 1 to demonstrate how our model effectively deals with the different tasks through task identifiers.

Table 1. Task-specific instruction format. During our model training process, we employ six distinct types of task identifiers to handle diverse tasks (excluding instance detection).

Task	Prompt
Classification	[classification].What is the final diagnostic impression? A) Benign B) Malignant.
Detection	[detection].Capture the lesion coordinates as [x_1, y_1, x_2, y_2].
Multi label Classification	[multi-label classification]. What abnormalities would you report on tooth 11? ...
Report Generation	[report_generation].What are the findings in this chest radiograph?
Counting	[counting].Quantify the number of cells shown in this microscopic capture.
Regression	[regression].Compute the ABR% by analyzing this dental panoramic radiograph.

3 Experiments

3.1 Dataset

The dataset used in the experiments consists of a total of 50,216 image-text pairs, covering both the training set and the public validation set. All images are uniformly resized to a resolution of 448×448 pixels, and no data augmentation techniques are applied. Notably, instance detection is not included as a training task in this study. Based on extensive experimental results, this task exhibits significantly inferior performance under the current multimodal framework, and its optimization process negatively impacts the training of other vision-language tasks. It may introduce gradient interference or attention misalignment during multi-task learning, thereby degrading the overall model's generalization capability. Therefore, to ensure stable convergence and optimal performance on the primary tasks, instance detection is excluded from the training objectives.

3.2 Implementation details

In this experiment, the vision encoder is initialized with the visual branch of MedSigLIP, while the language backbone is initialized based on LLaMA2-7B. Throughout the training process, the parameters of the vision encoder are kept frozen to preserve the representational capabilities acquired during pre-training on medical image-text alignment tasks. To achieve cross-modal feature alignment, a learnable linear projection layer is employed to map visual features into the latent space of the language model, and it is updated in an end-to-end manner. Additionally, Low-Rank Adaptation (LoRA) is applied for parameter-efficient fine-tuning of the LLaMA2 language model, where only the low-rank decomposition matrices are optimized, enabling effective task adaptation while maintaining computational efficiency. Model training is conducted using the standard cross-entropy loss function, optimized with the AdamW optimizer. Training is performed on a single NVIDIA GeForce RTX 3090 GPU for 10 epochs, with a batch size of 2, a maximum learning rate of $1e^{-4}$, and a learning rate warm-up ratio of 0.1. The entire training process lasts approximately 4 days, ensuring sufficient model convergence.

Environment settings The development environments and requirements are presented in Table 2.

Table 2. Development environments and requirements.

System	Linux-5.15.0-139-generic-x86_64-with-glibc2.35
CPU	Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz
GPU (number and type)	Four NVIDIA GeForce RTX 3090 24G
CUDA version	12.4
Programming language	Python 3.9.21
Deep learning framework	torch 2.0.0, torchvision 0.15.1

Training protocols The training protocols are presented in Table 3.

4 Results and discussion

As illustrated in the Fig.2, the changing trend of the loss function during the model's training process is demonstrated. It can be observed that in the early stages of training, despite the base model's strong representational capabilities, the initial loss value remains relatively high due to the lack of established cross-modal alignment. As training progresses, the model gradually achieves semantic alignment between visual and language modalities through the use of a learnable

Table 3. Training protocols.

Network initialization	MedSigLIP(vision encoder), LLaMA2-Chat(7b)
Batch size	2
Patch size	448×448
Total epochs	10
Optimizer	AdamW
Initial learning rate (lr)	$1e^{-4}$
Lr decay schedule	linear_warmup_cosine_lr
Training time	4 days
Loss function	cross-entropy
number of trainable parameters	52432896 [1]

linear projection layer for spatial mapping of visual features and parameter-efficient fine-tuning of the language model via LoRA (Low-Rank Adaptation). Concurrently, the model continuously optimizes its performance across multiple tasks. The loss value exhibits a steady downward trend, indicating that the model progressively adapts to various downstream tasks and eventually reaches a converged state. This demonstrates the effectiveness and stability of the proposed method within the joint training framework.

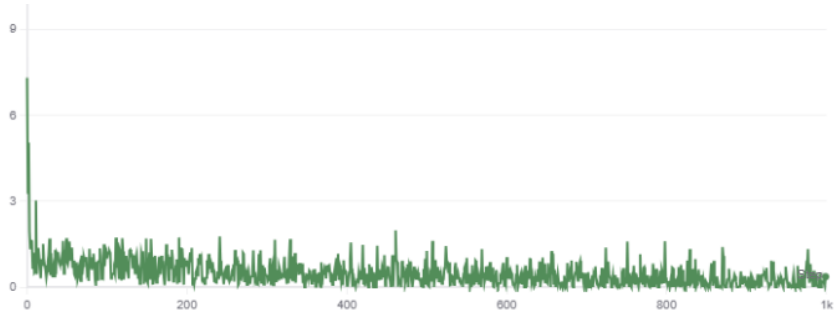

Fig. 2. Loss of training.

4.1 Quantitative results on validation set

We conducted inference and performance evaluation on the hidden validation set, with the inference time for each task presented in Table 4. For comparative analysis, we also performed inference using the competition's officially released

open-source models, Qwen-VL and MedGemma, which exhibited inference durations extending to several hours, indicating relatively low computational efficiency. The performance scores of different models across tasks are summarized in Table 5. Experimental results show that our proposed method achieves the lowest mean absolute error (MAE) of 13.63 on the regression task and a detection score of 0.80, classification score of 0.70, significantly outperforming the baseline models. These results demonstrate the effectiveness and competitiveness of our approach in multi-task medical vision-language understanding scenarios.

Table 4. Inference time for different tasks(batch size is 2).

Task	Inference time(m)
Classification	7.0
Detection	2.5
Multi label Classification	15.0
Regression	2.0

Table 5. Ranking performance on the hidden validation set(top_p is 0.9, temperature is 1).

Participant	Metrics				
	Classification	Multi-label Classification	Detection	Instance Detection	Regression
maiahmed	0.71	0.56	0.82	0.0	18.67
lujiazho	0.68	0.17	0.69	0.0	16.50
phucnlt	0.45	0.54	**0.85**	0.0	22.89
mtyw(ours)	0.70	0.54	0.80	0.0	**13.63**

4.2 Qualitative results on validation set

We randomly selected a set of samples from the validation set for qualitative analysis, with the results presented in Fig.3, to visually illustrate the model's output performance and semantic understanding capabilities across different tasks.

4.3 Limitation and future work

The medical multimodal large model developed in this study still faces challenges due to insufficient training data diversity and the scarcity of high-quality

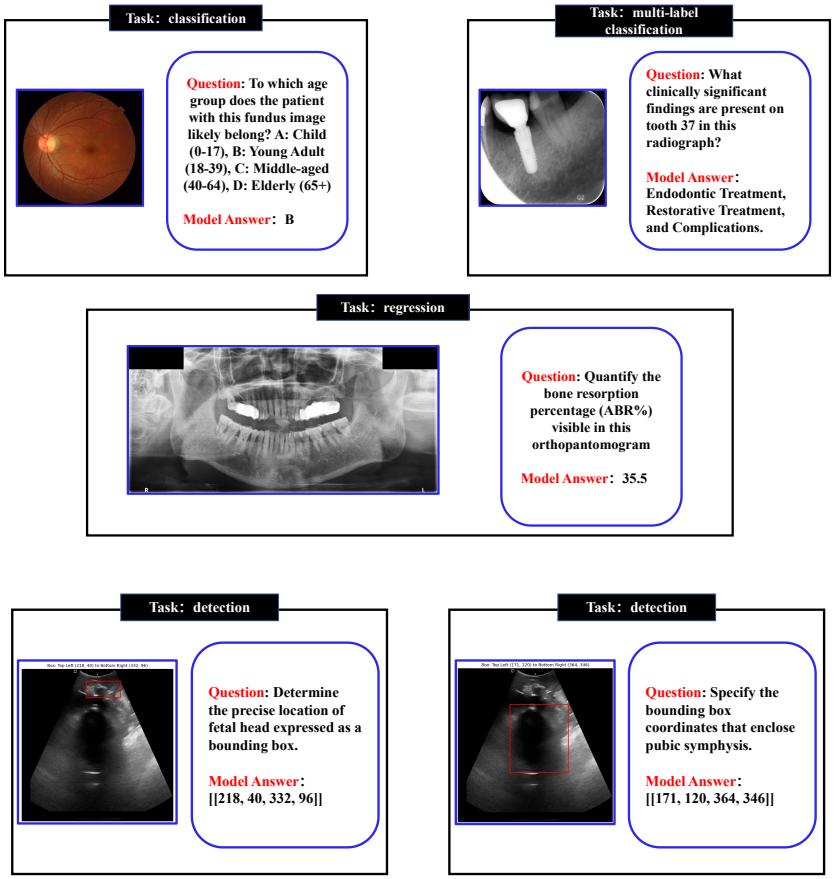

Fig. 3. Examples of our model's multi-task abilities.

annotated datasets, which limits its generalization capability across broad clinical scenarios. To further improve model performance, there is an urgent need to construct larger-scale, more diverse medical vision-language datasets with broader disease coverage and richer modalities. Meanwhile, it is advisable to adopt more advanced vision backbone architectures, particularly medical-specific visual encoders that support higher-resolution inputs, to enhance the representation of subtle lesions and complex anatomical structures. Furthermore, the current model is built upon LLaMA2, whose pre-training corpus has an early cutoff date, leading to limitations in linguistic coverage and depth of medical semantic understanding. Therefore, future work should consider integrating next-generation large language models, such as LLaMA3 or Qwen3 to improve context comprehension, logical reasoning, and domain-specific terminology modeling, thereby comprehensively enhancing the multimodal system's clinical semantic alignment and diagnostic assistance capabilities.

5 Conclusion

In this study, we propose a specialized multimodal large model designed for radiological diagnosis applications, aiming to achieve deep semantic alignment between medical images and natural language. The model is capable of handling a wide range of medical vision-language tasks in a unified framework, including medical report generation, disease classification, lesion detection, numerical regression, and visual question answering. To effectively distinguish task semantics and enhance the robustness of multi-task learning, we introduce explicit task-specific identifiers at the input level, enabling the model to accurately recognize and execute corresponding tasks within a shared parameter architecture. Experimental results demonstrate that the proposed model significantly outperforms existing baseline methods on key tasks such as classification, detection, and regression, exhibiting strong cross-modal understanding capabilities and promising clinical applicability.

Future work will focus on several directions: further integrating more diverse and representative medical imaging datasets to improve generalization across rare diseases and multi-modal scenarios; deepening the understanding of complex medical terminology and clinical expressions to enhance the accuracy and professionalism of language generation; improving model interpretability through techniques such as attention visualization and reasoning chain analysis; strengthening reliability and robustness; and conducting large-scale prospective clinical validation studies to systematically evaluate the model's effectiveness, safety, and integration feasibility in real-world healthcare settings, thereby facilitating its translation into practical clinical decision-support systems.

Acknowledgements The authors of this paper declare that the proposed solution is fully automatic without any manual intervention. We thank all data owners and contributors for making the data publicly available and CodaLab [3] for hosting the challenge platform.

Disclosure of Interests

The authors declare no competing interests.

References

1. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025) 2
2. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 3
3. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 9