UNIVERSITI MALAYA

Faculty of Computer Science and Information Technology

WOA7015 – Advanced Machine Learning

Group Assignment

A Comparative Study of CNN Baseline Classifier and LLM-Based Model for Medical Visual Question Answering

Submit By

NAME	**MATIRX NO.**
ZHONG JUN PEI	24214748
ZOU TING	24201617

Lecturer: Dr. SAW SHIER NEE (OCC1)

Submission Date: 19 / 12 / 2025

Academic Session: 2025/2026

1. Background

In the context of the rapid development of clinical informatization, modern patients are increasingly obtaining health data through hospital portals or personal health platforms. These data include medical images such as X-rays,CT, MRI, and pathological sections[1].This trend has given rise to an urgent need for patients to have a deeper understanding of their own health conditions, especially in the interpretation of medical images. Meanwhile, In recent decades, with significant progress made in data analysis, machine learning and deep learning, data-driven models are expected to bring breakthroughs in image understanding and clinical decision support [2]. Medical Visual Question Answering (Med-VQA), as a combined task of "image + natural language", is capable of answering natural language questions based on medical images [3]. Within a reasonable safety boundary, Med-VQA can serve as a communication aid and educational tool for patients, facilitating information preparation before visits and efficient communication during visits, thereby supporting the decision-making process.

2. Research problem and research objectives

This project intends to construct two types of Med-VQA methods: one is a lightweight baseline to ensure the stability and controllability of closed-loop questions, and the other is a VLM-based scheme to enhance the expressive ability and knowledge coverage of open-ended responses. Our research problem is how to quantitatively evaluate the performance of closed-ended and open-ended questions, thereby responding to patients' fundamental demands for "understanding medical images", while ensuring clinical availability and safety boundaries.

Based on the research problem, we have proposed research questions and objectives in two aspects: closed-end performance comparison, open-end expression and safety.

- **RQ1**: Which model, baseline or VLM, performs better in answering closed-ended questions about radiology images? How big is the gap between the two models?

- **RO1**: Construct a unified evaluation framework, calculate closed classification indicators (Overall Acc, Macro-F1), calculate the gap, and evaluate and output the classification results under the two routes respectively.

- **RQ2**: How is the expression quality of large visual language models(VLM) in medical open-ended question answering? What is the risk that the content it generates may cause hallucinations?

- **RO2**: Use language similarity metrics (BLEU, ROUGE-L, METEOR, etc.) and human assistance to evaluate the expression quality of open-ended questions. Conduct sensitivity analysis on repetitive questions, systematically quantify the expression quality and illusion risk of answers generated by vlm, and formulate specific operation guidelines to reduce risks.

3. Analysis of Datasets

3.1. Comparison of Med-VQA datasets

Table 1 Comparison of Med-VQA datasets

	# Images	# QA Pairs	Question Type	Language	Modality
VQA-RAD[3]	315	3515	Vision-only	EN	2D (X-ray, CT)
SLAKE[5]	642	14028	Knowledge-ba	EN & ZH	Multimodal

			sed & Vision-only		imaging (CT, MRI, X-ray, etc.)
PathVQA[6]	4998	32799	-	EN	Pathology

According to Table 1, we observe that PathVQA is the largest dataset, thus enabling the training of more complex and more generalized models that cover a wider range of cases. VQA-RAD focuses on "pure visual" problems, where the answers are completely dependent on the content of the image itself, while SLAKE encompasses both "knowledge-based" and "pure visual" problems. This means that answering certain questions requires not only looking at pictures but also external medical knowledge (such as the causes of diseases, treatment plans, anatomical connections, etc.). This makes SLAKE's task more challenging and closer to the actual reasoning process of clinicians, which is also of the highest clinical value. Meanwhile, SLAKE supports both English and Chinese (EN & ZH), which enables it to evaluate the performance of models in different languages and promote the application of medical AI in non-English environments, thus having a broader potential application scope.

From the perspective of data sources and professional fields, VQA-RAD is based on Radiology images (such as X-rays and CT). SLAKE covers a wider range of medical imaging modalities (such as CT, MRI, X-rays, etc.) and includes carefully labeled anatomical structures and disease labels. PathVQA is specifically designed for Pathology images, such as tissue sections under a microscope.

3.2. Reasons to choose VQA-RAD

(1) A clean "pure visual" benchmark

VQA-RAD prioritizes questions answerable solely through image content. This design provides a robust baseline for assessing visual understanding, ensuring that model failures reflect an inability to process visual features rather than a deficit in medical knowledge.

(2) High Efficiency & Accessibility

The extremely small amount of data (315 images, 3,515 QA pairs) become the key reason we chose it for exercise, and the dataset has the main advantages:

Rapid Training: Training and fine-tuning cycles are highly efficient, often completing within minutes or a few hours.

Low Hardware Demands: The model requires minimal resources, remaining compatible with standard consumer-grade GPUs or even CPUs.

Agile Prototyping: This lightweight nature facilitates rapid iteration, allowing us to quickly validate the effectiveness of new architectural concepts in medical imaging.

3.3. Analysis of VQA-RAD

3.3.1. Description and analysis of images

We downloaded the latest VQA-RAD dataset from kaggle, which contains 315 images and 2,247 Q&A pairs.After analyzing all the pictures, we obtain Figure *1*.

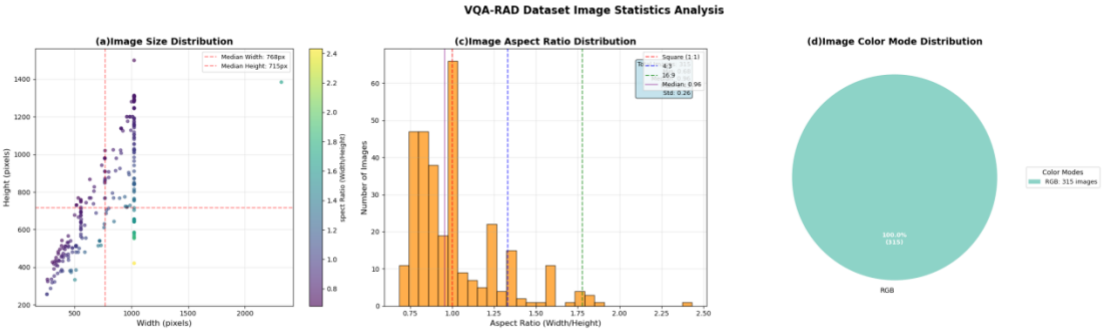

Figure 1 VQA-RAD dataset image statistic analysis (a)Most of the length and width pixels of both train and test do not exceed 1500, indicating that the image resolution is not too large and no corresponding processing is required. (b) A few images have extreme aspect ratios

(aspect_ratio > 2.0). When tuning the VLM model, it is necessary to letterbox it to a reasonable size. (c)The image mode is 100% RGB,with no bad images,no additional pretreatment is required.

3.3.2. Description and analysis of Q-A pairs

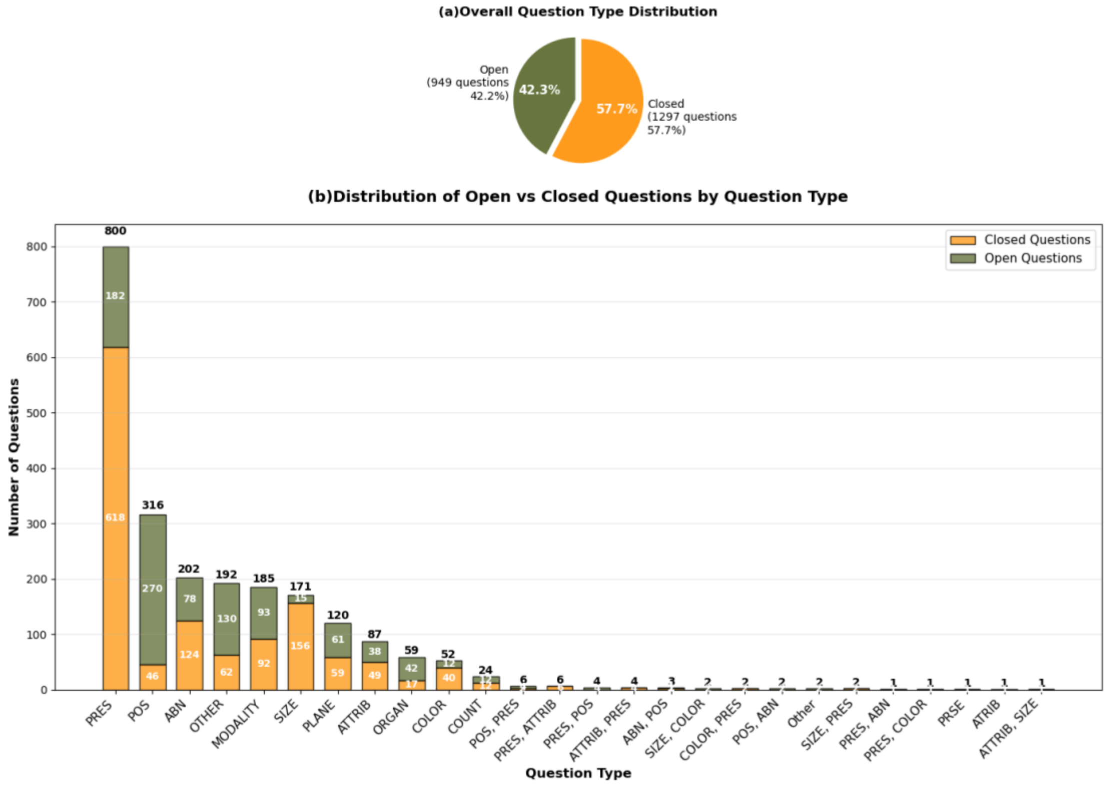

Figure 2 VQA-RAD dataset question type distribution (a) In the current dataset, there are 1,297 closed-end questions, accounting for 57.7%, indicating that this dataset is mainly composed of closed-end questions. (b) Analysis of different types of closed-ended and open-ended free-form questions shows that the majority of question types are PRES, POS, ABN, etc., and some types of questions are more likely to be open-ended: POS, OTHER.

3.3.3. Duplicates question checking

The same question occurs repeatedly, the model may rely on text patterns rather than images to guess the answer (question-only bias), So we counted the duplicate data of questions and answers in the dataset as shown in Figure 3.

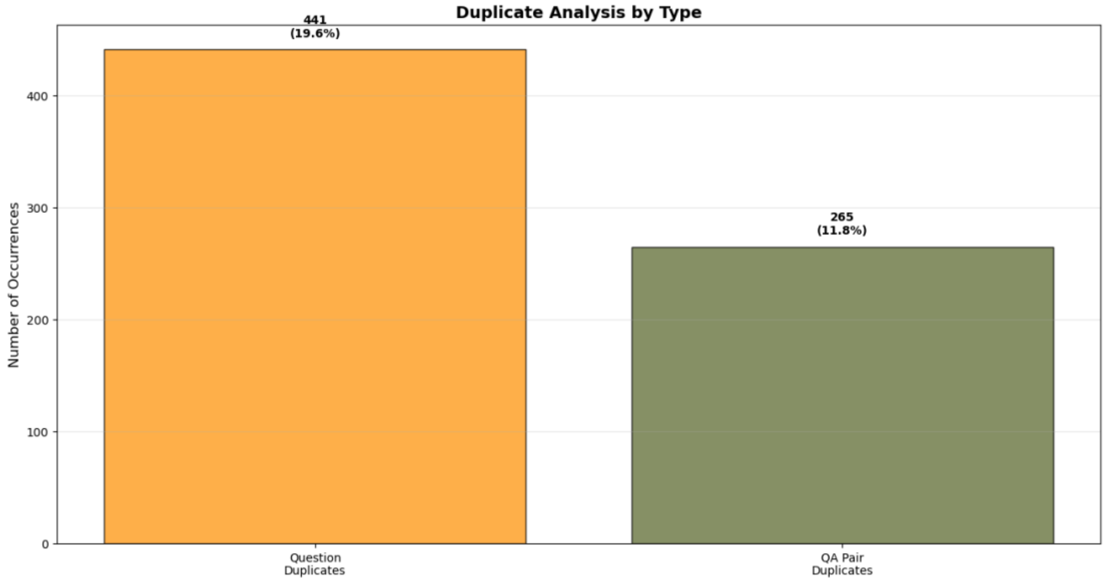

Figure 3 "question duplicate" refers to the number of duplicate questions alone, with 441 cases, accounting for 19.6% of the entire dataset. "QA pair duplicate" refers to the number of duplicate questions and answers alone, with 265 cases, accounting for 11.8% of the entire dataset.

The current repetition rate level needs to be disclosed in the report. Moreover, in the test normalized questions, how many of them appear in the train (text overlap rate), the corresponding analysis results should also be disclosed in the report and a sensitivity analysis should be conducted.

3.3.4. Data preprocessing

3.3.4.1. For baseline model

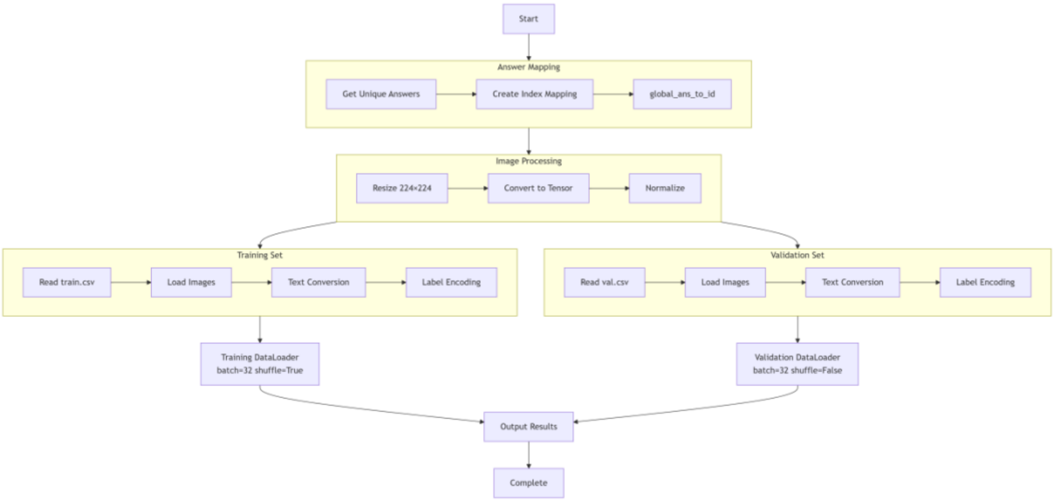

Figure 4 This flowchart presents the complete processing flow of the VQA-RAD medical visual question answering dataset from the original format to the model-readable format.

(1) Answer mapping establishment

The system first analyzes all the answers in the dataset and creates a digital index for each unique medical diagnosis description. This step converts the text-based answer (such as "nodules in the lungs") into numerical labels, facilitating subsequent classification task processing.

(2) Image standardization processing

All medical images are uniformly processed through a standardized preprocessing pipeline. The image was adjusted to the standard size of 224×224, converted to tensor format, and normalized according to the statistical characteristics of medical images to ensure the consistency of the input data.

(3) Dataset construction

Create the training set and validation set respectively.

The system reads data from the CSV file, loads the corresponding medical images, converts the question text into a numerical sequence, and maps the answers to digital labels. This step generates two dataset objects with the same structure but different data.

(4) Batch loading of data

In the final stage, the dataset is encapsulated into an efficient data loader.

The training data loader is set to 32 samples per batch and randomly shuffled in order to enhance the model's generalization ability.

Verify that the data loaders maintain the same batch size without shuffling the order to ensure the comparability of the evaluation results.

3.3.4.2. For LLM

The data preprocessing work before the LLM model runs is divided into four steps: data preparation stage, image processing process, text processing process, and final output format, refer to Figure 5.

(1) Data preparation stage

The system finds the corresponding data based on the index number and simultaneously reads medical images and answers to questions. Images are loaded from files, and the answers to questions are obtained from the data list. Both are processed in parallel to improve efficiency.

(2) Image processing process

First, open the pictures and uniformly convert them into the RGB color format, and then use a dedicated processor for standardization processing. The processor will adjust the image to the size and format required by the model, and finally convert it into a digital tensor to provide input for the visual model.

(3) Text processing procedure

The text of the question and answer is concatenated into a complete sentence and organized in a fixed format of "question: content - Answer: content". This complete sentence is converted into a sequence of numbers by a tokenizer and uniformly filled or truncated to a fixed length.

(4) Final output format

The processed image tensor and the sequence of text numbers are combined into a standard data packet. In generative training, text sequences serve both as input and training targets, forming a three-element output structure that can be directly applied to the training of multimodal models.

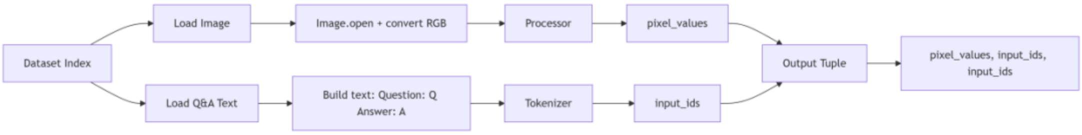

Figure 5 The processing procedure of a single data sample for the design of the generative visual question-answering model.

4. Comparison of Popular Models

4.1. Comparison of Model Architecture

We selected five representative models for architectural comparison, with their detailed configurations summarized in Table 2. The comparison primarily focuses on their Image Encoders, Text Encoders, Modality Fusion Modules, and Answer Generators. Based on this comparative analysis, we will subsequently select two models for further experimentation and explain the specific reasons of our selection.

Table 2 Comparison of five popular model

Models	Image Encoder and Text Encoder	Modality Fusion Module	Answer Generator
CNN baseline[7]	ResNet+ LSTM	MLP	Classifier
Med-MoE[13]	CLIP-ViT + Text tokenizer of the LLM	MLP projection	MoE (Router+LLMs)
Sonsbeek et.al.[8]	CLIP-ViT + Text Tokenizer of the LLM	MLP projector	GPT2-XL
VGG-Seq2Seq [8][11]	VGG+LSTM	concatenation	LSTM
LLaVa-Med[12]	CLIP-ViT + Text Tokenizer of the LLM	MLP projector	LLaMA

The submission by PwC US-Advisory at CLEF2019 utilized two distinct methodologies: one based on a classifier model[7], and the other using a VGG-Seq2Seq[8][11] approach. The first approach primarily focuses on providing answers from a fixed pool of predefined answer categories. The second approach involves generating answers based on anomalies observed within the images. Their strict accuracy reached 48%, while the BLEU score achieved 53%. VGG-Seq2Seq was utilized as a secondary approach to complement the CNN baseline. While both methods involve relatively low training costs and achieve similar accuracy, with VGG-Seq2Seq showing slightly higher accuracy and a better BLEU score than classification-based methods, VGG-Seq2Seq serves primarily as a supplementary solution specifically for generating diagnostics for abnormal images.

However, VGG-Seq2Seq exhibits significant limitations: its language component is overly simplistic, relying on a basic RNN/LSTM Decoder. This results in limited generative capacity and a susceptibility to grammatical errors that can obscure diagnostic clarity. Furthermore, it requires training from scratch or extensive fine-tuning. Consequently, VGG-Seq2Seq is not considered an optimal model for open-ended questions. The CNN baseline was retained to establish a foundational benchmark based on fundamental VQA models, aiming to facilitate the development of low-cost Med-VQA systems.

The work by van Sonsbeek et al.[8] developed a network architecture that maps extracted visual features onto a set of learnable tokens. These tokens are subsequently used alongside text-based questions to directly prompt a Large Language Model (LLM). The evaluation methodology primarily relies on standard VQA metrics to assess the model's performance.

LLaVA-Med[12] is a pioneering model that adapts general multimodal capabilities into specialized biomedical expertise. It employs a "two-stage fine-tuning" strategy: first aligning visual-language features using large-scale biomedical image-text pairs, and then performing instruction tuning via clinical dialogue data generated by GPT-4. This approach enables the model to transition from simple visual recognition to complex clinical reasoning, making it a cornerstone for Med-VQA.

Med-MoE[13] addresses the significant challenge of "modal heterogeneity" within medical imaging. Unlike traditional dense models, it introduces a Mixture-of-Experts architecture featuring multiple specialized modules tailored for distinct medical domains, such as radiology, pathology, or ultrasound. A gating router dynamically selects the most relevant experts for each input, significantly enhancing the model's precision in handling rare cases and cross-departmental data without a proportional increase in computational cost.

4.2. Evaluation Metrics

Prior to model selection and preliminary experimentation, it is essential to define the metrics for evaluating model performance. These metrics are categorized into evaluation standards for both closed-ended and open-ended questions, all of which are derived from established benchmarks commonly used in previous research.

4.2.1. Closed-Ended Evaluation Metrics

Table 3 Evaluation metrics for closed-ended questions in Med-VQA

Evaluation Metric	**Used By**
Accuracy	All Works
AUC-ROC,AUC-PRC	[11]

Accuracy: Accuracy is the most common metric for closed-ended MedVQA tasks, measuring the proportion of correctly predicted answers over the total number of questions. It provides a simple and interpretable assessment of model performance in classificationbased settings.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): AUC-ROC evaluates the model’s ability to distinguish between positive and negative cases by plotting the true positive rate against the false positive rate. A higher AUC-ROC indicates better discriminatory power, making it particularly useful for binary classification tasks such as disease detection.

AUC-PRC (Area Under the Precision–Recall Curve): AUC-PRC is well-suited for imbalanced datasets, where positive cases are rare. It measures the trade-off between precision and recall, capturing the model’s performance in correctly identifying relevant answers while minimizing false positives.

4.2.2. Open-Ended Evaluation Metrics

Table 4 Evaluation metrics for open-ended questions in Med-VQA

Evaluation Metric	Used By
Accuracy	[8]
Recall	[12]
BLEU	[8]
BERTScore	[8]

Accuracy: While traditional accuracy is difficult to apply directly to open-ended responses, some studies use exact-match accuracy, where a response is considered correct only if it matches the ground truth exactly. However, this metric is often too rigid for natural language generation tasks.

Recall: Recall measures how many ground-truth tokens appear in the generated response. This metric is particularly relevant for MedVQA tasks where completeness of information is crucial, such as in diagnostic explanations.

BLEU (Bilingual Evaluation Understudy): BLEU is a widely used metric that measures the overlap of n-grams between generated responses and reference answers. It assigns higher scores to outputs that closely match human-written references.

BERTScore (BERT-Sim): BERTScore computes similarity between generated and reference answers using contextualized embeddings from BERT. It assesses token-wise similarity in a way that accounts for synonymy and paraphrasing, making it a more robust metric for evaluating MedVQA models.

4.3. Performance of Models

Compared to earlier discriminative or smaller-scale generative MedVQA approaches, LLM/MLLM-based models generally offer higher performance and greater flexibility in handling open-ended queries. In Table 4,we examine the performance of the currently selected five models on open-ended questions and the size of the parameters they use. It can be seen that the accuracy of the early methods is inferior to that of LLM-based models.

The earlier model like the CNN baseline model and VGG-Seq2Seq model achieves an strict accuracy of 48% around, while larger models like LLaVa-Med exhibit higher performance at 64.4%. Notably, the architecture developed by Sonsbeek et al. demonstrates a superior accuracy of 84.3% on the Slake dataset.

Table 5 Evaluation metrics for open-ended questions in Med-VQA.

Models	**Parameters**	**Generative**	**LLM/MLLM-Based**	**Accuracy of Open-Ended Questions for VQA-RAD(%)**
CNN baseline	Not Reported	No	No	Not reported but strict accuracy is 48.4
Med-MoE	3.6B	Yes	Yes	58.6
Sonsbeek et.al.	1.5B	Yes	Yes	Not reported but 84.3 for

				Slake
VGG-Seq2Seq	Not Reported	Yes	No	Not reported but strict accuracy is 48.8
LLaVa-Med	13B	Yes	Yes	64.4

4.4. Model Selection

As shown in the model performance of Section 4.3,considering both baseline stability and state-of-the-art potential, we have chosen to focus on the CNN-baseline and the model architecture created by Sonsbeek et al. for my research.

In the CLEF 2019 work notes [7], the VGG-Seq2Seq model in their proposed scheme only provides generative diagnostic results for abnormal images, serving as a supplementary solution. Its weaknesses lie in the relative simplicity of the language model, relying solely on an RNN/LSTM Decoder, and its limited generative capacity, which is prone to grammatical errors that can impede the understanding of diagnostic results. Furthermore, its training process requires starting from scratch or extensive fine-tuning.

Due to these limitations, we do not consider VGG-Seq2Seq a robust model for open-ended questions. In contrast, we chose the CNN baseline as it helps us establish a foundational benchmark based on VQA fundamental models, aiming to explore the possibility of constructing low-cost Med-VQA systems.

At the same time, generative models have undergone continuous improvement and development. In 2023, the approach proposed by Sonsbeek et al.[8] emerged as a highly representative model. Their team proposed leveraging the generative capabilities of Pre-trained Language Models (LMs), enabling the model to generate variable-length, natural language answers similar to a chatbot.

Since medical datasets are typically small, this method effectively avoids overfitting by freezing the weights of the pre-trained LM and only fine-tuning the mapping layers and prefix parameters. We selected this model for our experiments because it outperformed the contemporary state-of-the-art (SOTA) classification-based models across multiple mainstream MedVQA benchmarks in terms of BLEU and BERTScore.

4.5. Model Architecture

4.5.1. CNN Baseline Classifier Model

This model adopts a "two-tower" structure, consisting of a visual encoder, a text encoder and a multimodal fusion classifier.

Vision Encoder: It utilizes ResNet50 as the visual backbone to extract high-level semantic features from medical images. The final fully-connected layer is replaced with nn.Identity() to output a 2048-dimensional feature vector.

Text Encoder: A standard LSTM (Long Short-Term Memory) network is employed for text encoding. The words are first mapped to 512-D dense embeddings. The LSTM then processes the sequence, and the hidden state of the final time step (h_n[-1]) is extracted as a 512-D summary of the question.

Multimodal Fusion & Classifier: The visual and textual features are integrated via feature concatenation, resulting in a 2560-D joint vector. This combined representation is fed into an MLP classifier, which includes a 1024-D hidden layer with ReLU activation, a Dropout (0.3) layer for regularization, and a final linear layer that outputs probabilities across all possible answer classes.

4.5.2. Van Sonsbeek Model

This model adopts a parameter-efficient multimodal alignment architecture designed to inject powerful pre-trained visual features into a generative language model. Its core consists of three key components:

Vision Encoder: Utilizes CLIP (ViT-B/32). This component is completely frozen and is responsible for extracting high-dimensional semantic features from medical images.

Mapping Network (Projector): This is the only trainable part of the model. It consists of a Multi-Layer Perceptron (MLP) responsible for mapping visual features into the language model's embedding space, generating prefix_length virtual token vectors.

Language Model: Utilizes a pre-trained GPT-2. This component is also frozen and generates answers in an autoregressive manner by receiving the concatenated visual prefixes and text embeddings.

5. Preliminary Results

5.1. CNN Baseline Classifier Model

5.1.1. Core Parameters

The preliminary training phase utilized the VQA-RAD dataset, comprising a total of 315 images and 2247 pairs of Q&A for the preliminary training and testing purposes. The raw data was partitioned into an 80% training set and a 20% validation set to facilitate model development and evaluation.

Table 6 Core training parameters of CNN baseline model

Parameter	**Value**	**Description**

Epochs	30	Total number of training cycles.
Batch Size	32	Number of samples processed in each batch.
Optimizer	Adam	Optimization algorithm for weight updates.
Learning Rate	10^(-4)	Learning rate controlling the step size of updates.
Loss Function	CrossEntropy	Cross-entropy loss for multi-class classification.
Dropout	0.3	Dropout rate used to prevent overfitting.

5.1.2. Preliminary Performance Evalution of Model

During the preliminary training process, we recorded the loss history and the trend chart of accuracy changes, as shown in Figure 1. The training loss consistently decreased from approximately 4.8 to below 0.5. The smooth curve without significant oscillations indicates that the learning rate (10^(-4)) is appropriately configured, and the model successfully captured the associative features between images and text from the VQA-RAD dataset. But the significant gap between training and test accuracy indicates overfitting. This is primarily due to the limited sample size of VQA-RAD (only 315 images) and the simplicity of the baseline's concatenation-based architecture, which struggles with complex medical generalization.

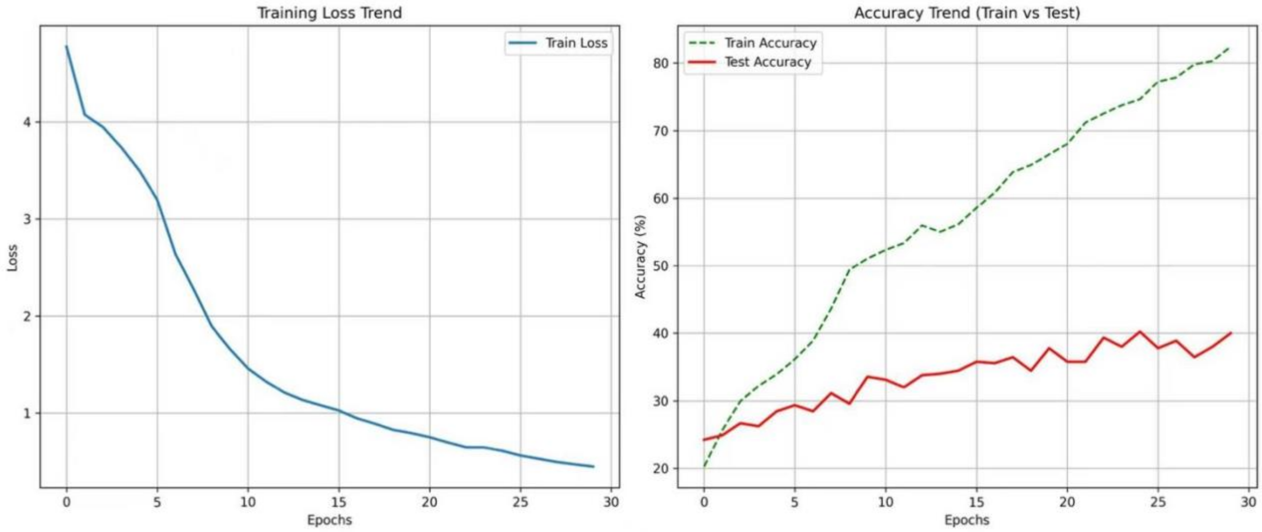

Figure 6 Loss Curve and Accuracy Comprarison of Baseline Classifier

By distinguishing between CLOSED and OPEN questions, the model clearly reveals performance discrepancies across different task types, helping to identify specific challenges in medical QA. And the comparison in last training epoch be shown in Figure 7.

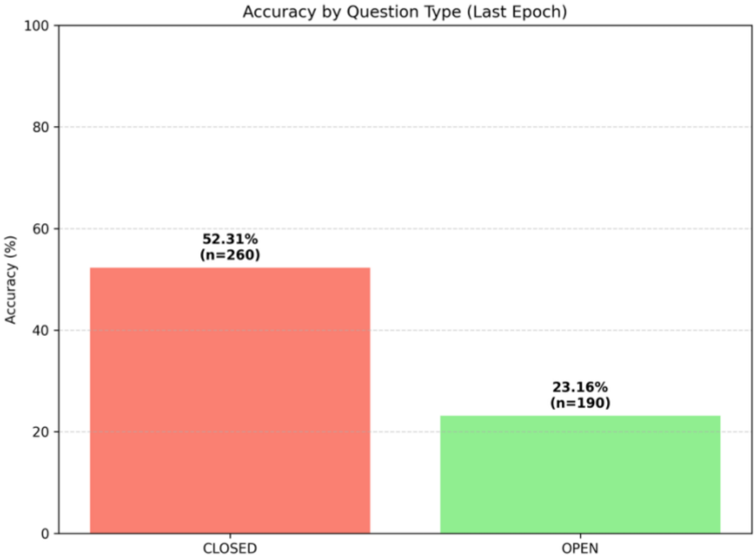

Figure 8 Comparison of the Accuracy of Different type questions for Baseline Classifier

5.2 Van Sonsbeek Model

5.2.1 Core Parameters

Our training and hyperparameters are configured as shown in the table 6. We froze clip_model and gpt2 in the code using param.requires_grad = False. Consequently, backpropagation only needs to compute the gradients for the tiny mapping_network, which significantly saves VRAM (Video RAM)

Table 7 Core Parameters for Van Sonsbeek Model

Parameter	Value	Description
prefix_length	10	Prefix Length: Number of virtual tokens the image features are converted into.
clip_model_name	openai/clip-vit-base-patch32	Vision Encoder: CLIP model used to extract high-dimensional semantic feature vectors from images.
clip_dim	512	Vision Feature Dimension: The length of the raw vector output by the CLIP model.
gpt_dim	768	LM Embedding Dimension: The length of the internal token vectors in the GPT-2 model.
mapping_network	nn.Sequential	Mapping Network: The only trainable part, mapping vision space to language space.
inputs_embeds	torch.cat(...)	Combined Embeddings: The matrix formed by concatenating visual prefixes and question text embeddings.

labels (Training)	input_ids	Training Labels: The target sequence the model learns to predict, containing questions and answers.
-100	ignore_index	Ignore Index: Tells the model to ignore the prefix tokens when calculating loss.
lr (Learning Rate)	1e-4	Learning Rate: Controls the step size of parameter updates for the mapping network.
max_new_tokens	10	Generation Length: The maximum number of tokens the model generates during inference.

5.2.2 Loss Curve Analysis

As table x shown, the loss curve exhibits a rapid initial descent, followed by a plateau.Within the first 20 steps of training, the loss value shows a rapid decline, dropping from approximately 13 to below 3. This indicates that the mapping network is quickly learning to align visual features with GPT-2's semantic space.

From step 40 onwards, the loss curve flattens out, eventually stabilizing around 0.5. This smooth convergence suggests that the learning rate is appropriately set and the model shows no gradient oscillation.

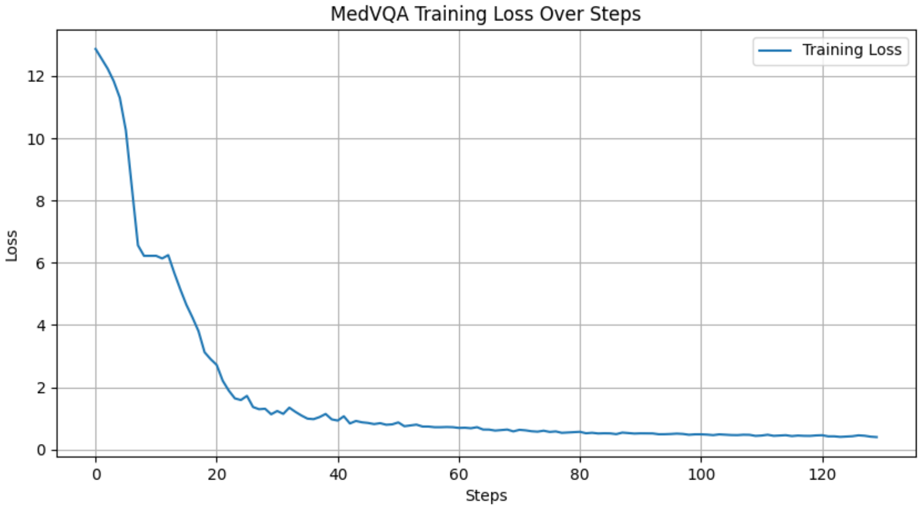

Figure 9 The Loss Curve of Van Sonsbeek Architecture in emulate excecise

Compared to the earlier VGG-Seq2Seq model which relies on a basic RNN/LSTM decoder with limited generative capacity, this model fully leverages the prior knowledge of pre-trained language models via Prefix-Tuning. This approach effectively avoids common overfitting issues in small-scale medical datasets, providing more accurate and natural responses for open-ended medical diagnostics.

Reference:

[1] Tapuria, A., Porat, T., Kalra, D., Dsouza, G., Xiaohui, S., & Curcin, V. (2021). Impact of patient access to their electronic health record: systematic review. Informatics for Health and Social Care, 46(2), 194-206.

[2] Thapa, S., & Adhikari, S. (2023). ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering*, *51*(12), 2647-2651.

[3] Zhan, L. M., Liu, B., Fan, L., Chen, J., & Wu, X. M. (2020, October). Medical visual

question answering via conditional reasoning. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2345-2354).

[4] Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5(1), 1-10.

[5] Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021, April). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th international symposium on biomedical imaging (ISBI) (pp. 1650-1654). IEEE.

[6] He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286.

[7] Bansal, M., Gadgil, T., Shah, R., & Verma, P. (2019, September). Medical Visual Question Answering at Image CLEF 2019-VQA Med. In *CLEF (working notes)*.[CrossRef]

[8] Van Sonsbeek, T.; Derakhshani, M.M.; Najdenkoska, I.; Snoek, C.G.; Worring, M. Open-ended medical visual question answering through prefix tuning of language models. In Proceedings of the International Conference on Medical Image Computing and ComputerAssisted Intervention; Springer: Berlin/Heidelberg, Germany, 2023; pp. 726–736.

[9] Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. Neural Inf. Process. Syst. 2024, 36, 28541–28564.

[10] Talafha, B.; Al-Ayyoub, M. JUST at VQA-Med: A VGG-Seq2Seq Model. In Proceedings of the CLEF (Working Notes), Avignon, France, 10–14 September 2018.

[11] Sharma, D.; Purushotham, S.; Reddy, C.K. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. 2021, 11, 19826.[CrossRef]

[12] Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. Neural Inf. Process. Syst. 2024, 36, 28541–28564.

[13] Jiang, S.; Zheng, T.; Zhang, Y.; Jin, Y.; Yuan, L.; Liu, Z. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 3843–3860.