Advance Machine Learning (WOA7015)

Alternative Assessment

2025/2026 Semester 1

Group: Crow

Fattah Zul Ikram (24063134)

Nong Haochen (25069080)

1 Introduction

Medical Visual Question Answering (Med-VQA) combines medical artificial intelligence with visual question answering, where a system predicts plausible answers to clinically relevant questions in natural language given medical images. Unlike general VQA systems, Med-VQA is considered more challenging for different factors, including the difficulty in creating large datasets with manually annotated question-answer pairs, the need for sophisticated models that can focus on fine-grained details of medical images, and the necessity to train a model with medical contexts to properly correlate the questions to the images. (Lin et al., 2023) Although the research of Med-VQA is still relatively new (Liu et al., 2021), there are some publicly available datasets that are well-annotated for training, validating, and testing Med-VQA systems, including VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), and SLAKE (Liu et al., 2021).

Traditionally, research on Med-VQA employed discriminative models that focused on image recognition and classification where models were trained to identify specific features within medical images. These architectures commonly utilize CNNs as image encoders and RNN/LSTM as the text encoder. (Bazi et al., 2023) However, these systems typically relied on predefined categories, and while they were effective in closed—set scenarios, they lacked the ability to answer complex, natural language questions that could guide clinical decision-making. (Ben Chaabane & Bal-Ghaoui, 2025)

With the recent advances in **Vision-Language Models (VLM)**, the dynamic has shifted towards fine-tuning powerful foundational models that are pre-trained with large-scale databases of image-text pairs. Models such as BERT, CLIP, BLIP, LLaVA, and others have achieved remarkable results on different VQA datasets. For specific domains, domain-specific knowledge is injected within the models, resulting in domain-specific models like BioMedBLIP (Naseem et al., 2024), BioMedCLIP (Zhang et al., 2025), and so on.

With all these advances, it becomes a necessity to benchmark them to understand the trade-offs between traditional, discriminative models versus modern, pre-trained vision language models. These comparisons can help identify the future direction that can be taken to advance medical VQA systems further.

2 Objective

This study will be conducted on the SLAKE dataset. The objective of this study is to conduct a comparative evaluation of the two approaches to Med-VQA systems:

- **The Baseline—Discriminative Model:** A CNN-LSTM hybrid model will be used as the baseline that will treat the VQA task as a classification task over a fixed vocabulary of answers.
- **The Challenger—Generative Model:** BioMedBLIP, a state-of-the-art VLM that extends the BLIP architectures by integrating medical domain-specific knowledge within the model will be benchmarked against the baseline.

Through this study, it is expected that the following research questions will be addressed.

RQ1. How do discriminative models compare to pre-trained vision-language models in terms of overall accuracy and answer quality on medical VQA tasks on the SLAKE dataset?

RQ2. What are the strengths and weaknesses of the two VQA approaches across different question types (closed-ended vs. open-ended)?

RQ3. How do the models perform on different performance metrics, including accuracy, semantic similarity (F1, BLEU, METEOR, ROUGE), and contextual understanding (BERTScore)?

RQ4. In medical settings, what are the trade-offs between these two approaches in terms of performance, complexity, and resources?

3 Method

3.1 Dataset Description

SLAKE (Semantic Label for Anatomical Knowledge Evaluation)—a publicly avaliable, manually annotated dataset published by Liu et al., 2021—is a benchmark dataset that is specifically designed for evaluating VQA systems on medical images. It is a bilingual dataset consisting of both English and Chinese question-answer pairs. The dataset comprises 642 medical images of three modalities—CT, MRI, and X-ray. These images cover a wide range of anatomical features.

The dataset has 14,028 question-answer pairs in total. Among them, 7,033 pairs are in English, which is the focus of this study. These pairs are categorized into open-ended (4,252) and closed-ended (2,781) depending on the answer type. The closed-ended questions primarily consist of one-word responses with most of them being yes/no answers, with a very few two-word responses. The open-ended questions, on the other hand, consist of more diverse free-form answers, where some of them even surpass 10 words. Most of the questions are based on organs, positions, abnormalities, and modalities.

The entire dataset is split into training (70%), validation (15%), and testing (15%).

3.2 Data Preparation and Preprocessing

There is data in two modalities—images and texts. The text data consists of questions and answers. All of them are prepared differently.

3.2.1 Image Processing

To ensure that consistent input goes to the model, the following preprocessing steps were done.

- **Image Resizing:** All the images were resized to 224*224 pixels, as this is the standard input size for the CNN base in the hybrid model. This resolution is consistent with the input size of the standard ImageNet-pretrained models.
- **Normalization:** The pixel values were normalized using the ImageNet statistics of [0.485, 0.456, 0.406] as the mean and [0.229, 0.224, 0.225] as the standard deviation. For the generative model—BioMedBLIP—this will be changed according to the protocol mentioned there.

- **Grayscale Handling:** As many of the images of the dataset are in grayscale, specifically the X-ray images, the images are converted into 3-channel RGB images for consistency.

3.2.2 Question Preprocessing

The questions are in natural language and contain medical terms as well as filler terms. They were processed in the following steps.

- **Tokenization:** The questions were tokenized to the word level using the 'nltk' module in Python. This effectively converted the questions into sequences of tokens.
- **Vocabulary Construction:** For fair evaluation of the model, only the training dataset was used to construct the vocabulary list. The list included every word that appeared. Appropriate mappings were also created for encoding and decoding vocabularies to and from the tokens. Approximately around 300 vocabularies were available in the list. Special tokens were used for padding and to represent unknown values.
- **Lowercasing:** All the text from the questions was converted into lowercase, which resulted in a smaller vocabulary list but with better generalization.
- **Padding:** In every batch, the question sequences were padded to the max question length of that specific batch. The special padding token was used for this purpose.

3.2.3 Answer Preprocessing

Unlike the questions preprocessing, the answers were not tokenized to the word level. The steps involving answer preprocessing were the following.

- **Answer Vocabulary Construction:** All the unique answers were extracted from the training dataset. This resulted in a vocabulary list of approximately 225 unique answers.
- **Answer Numericalization:** All the answers were mapped to a unique number. This mapping was then used for classification task, making the model a discriminative model.
- **Handling Multi-word Answers:** Answers containing multiple words are considered a single unit. Thus, each multi-word answer got mapped to a unique number.

3.3 Model Architectures

3.3.1 The Baseline: CNN-LSTM

The baseline model used in this study is a discriminative model that focuses on classification task. It has three major components.

Visual Encoder

The model uses a ResNet-34 architecture pre-trained on the ImageNet dataset as the visual encoder. This is chosen over deeper variants, like ResNet-50 or ResNet-101, to balance the capacity of the model with its computational efficiency. It also reduces the risk of overfitting on the relatively small SLAKE dataset. The final fully connected layer is removed from the model to modify it according to the specific classification problem of the study. The RGB images of 3*224*224 dimension are used as the input to this model. The extracted features from the final convolution layer (7*7*512) are then squeezed into a 512-dimensional feature vector.

Question Encoder

For the processing of questions, a bidirectional LSTM is employed with a self-attention mechanism. The tokenized questions are passed to the encoder. The tokens are then mapped to a 512-dimensional word embedding. The embedding dimension is obtained through the hyperparameter tuning stage of the model. The embedded sequences are then processed by a bidirectional LSTM with 512 hidden units in each direction. So, a total of 1024 hidden units are present. A bidirectional LSTM is chosen because it allows the model to understand the medical terminologies better through past and future tokens. This helps the model with understanding complex medical terminologies better. Additionally, to increase the focus on important words in the questions, a self-attention mechanism is integrated into the hidden states. Finally, the question encoder produces a 1024-dimensional feature vector after applying a mean pooling on the resulting output.

Multimodal Fusion and Classification

The visual features (512-dimensional) and the question features (1024-dimensional) are concatenated to create a multimodal representation of 1536 dimensions. This concatenated feature then passes through a series of fully connected layers with dropout regularization. The fusion dimension and fusion dropout values are obtained through hyperparameter tuning.

- FC1: 1536 → 1024, Batch Normalization, ReLU, Dropout (0.27)
- FC2: 1024 → 512, Batch Normalization, ReLU, Dropout (0.27)

Finally, a fully connected layer is used to project the 512-dimensional features to the answer vocabulary space. The answer to a question can be obtained through using a softmax function on the output logits.

Hyperparameter Tuning and Training

The library 'Optuna' is used in this study to optimize the hyperparameters of the model. A 50-trial optimization is conducted in the hyperparameters space with 30 training epochs, a step learning rate scheduler, and early stopping. The best-performing model has the following hyperparameters.

- Embedding dimension: 512
- LSTM hidden layers: 512
- Total LSTM layers: 2
- LSTM dropout: 0.18
- LSTM pooling strategy: Mean pooling
- Attention heads for self-attention: 4
- Fusion dimension: 1024
- Fusion dropout: 0.27

- Batch size: 16
- Learning rate: $7.56e-05$
- Weight decay: $4.62e-06$
- Scheduler step size: 15
- Scheduler gamma: 0.69

The final model is trained on the Adam optimizer with the tuned learning rate and weight decay. A step learning rate scheduler is also used using the obtained step size and gamma. Gradient clipping (max norm = 5.0) is applied to prevent exploding gradients. The model is set to train for 100 epochs with early stopping based on validation accuracy and 15 epoch patience.

Justification for Baseline Selection

The CNN-biLSTM architecture was selected for several factors found through prior literature.

- **CNN-LSTM architecture:** This architecture has been studied a lot in the VQA field. It is also relevant in medical VQA research. (Lin et al., 2023)
- **ResNet-34:** ResNet models have performed outstandingly in medical image analysis tasks, with ResNet-34 being utilized in multiple research studies within this field due to its optimal balance of depth and efficiency. (Gong et al., 2021; Li et al., 2022)
- **Bidirectional LSTM:** TAs contexts from both directions are important for medical terminologies, bidirectional LSTM can understand medical questions better than the vanilla LSTM. Bi-LSTM has also been used in studies and achieved success similar to contemporary VQA models. (Gupta et al., 2020)
- **Self-attention:** Multiple VQA benchmarks have proven the effectiveness of self-attention, as it allows models to focus on question keywords relevant to the input image. (Chen et al., 2023)

3.3.2 The Challenger: BioMedBLIP

BioMedBLIP is a model created for medical image analysis, particularly in the context of visual question answering and image captioning. (Naseem et al., 2024) This model extends on the BLIP (Bootstrapping Language-Image Pre-training) architecture, which is a VLM that is used for unified vision-language tasks, such as VQA. BioMedBLIP extends on this architecture by integrating domain-specific medical knowledge into the BLIP model. This allows BioMedBLIP to work as an excellent choice for medical—related visual-language tasks. For this study, this model will be fine-tuned on the SLAKE dataset. It will then be benchmarked against the baseline on different performance metrics.

Justification for BioMedBLIP Selection

BioMedBLIP is selected as the advanced model for this study for the following factors.

- **Domain Specificity:** Unlike BLIP or CLIP, BioMedBLIP is pre-trained on medical datasets. The domain-specific knowledge lets it outperform general models for the medical domain.

- **Generative Capability:** Unlike the baseline model, BioMedBLIP can generate free-form answers. Thus, it can handle novel question types that were not seen during training. In the case of the baseline, the model can only classify answers that it learns from the training dataset.
- **State-of-the-Art Performance:** BioMedBLIP achieved state-of-the-art (SOTA) performance in 15 of the 20 dataset-task combinations it was evaluated on. (Naseem et al., 2024) This makes this model an excellent choice for comparison.
- **Minimal Modification:** As this model is specifically tailored for tasks like VQA, the model requires minimal modification for this study beyond fine-tuning.

3.4 Evaluation Metrics

To ensure a fair evaluation, both classification and generative metrics will be employed.

- **Accuracy Metrics**
  - **Accuracy:** Measures overall correctness by comparing predicted and ground truth answers.
  - **Exact Match:** Binary metric indicating a perfect string match.
- **F1-Based Metrics**
  - **Macro F1 Score:** Treats all answer classes equally.
  - **Weighted F1 Score:** Accounts for class imbalance by using class weights.
- **Generative Metrics**
  - **BLEU:** It is a standard for NLP tasks. For this assignment, up to 4-grams will be tested. In a study, Lin et al., 2023 noted that BLEU is not suitable for med-VQA, but they still employed it in their study, as it is used in medical report generation tasks.
  - **Rouge:** Assesses how much of the reference text appears in the generated output. This metric is used in studies to evaluate recall. (Xin et al., 2025)
  - **Meteor:** It improves upon BLEU and ROUGE by considering synonyms and stemming, balancing precision and recall. (Xin et al., 2025)
  - **BERTScore** It utilizes contextual embeddings from BERT to compute semantic similarity between text pairs. There are several studies that use this metric for evaluating vision-language tasks. (Yim et al., 2025)
- **Per-Answer-Type Metrics:** Accuracy, Exact Match, and F1 Score will be computed for open- and closed-ended answer types.

4 Preliminary Results

This section presents the preliminary results obtained from training and evaluating the baseline CNN-LSTM model on the SLAKE dataset.

4.1 Model Training

The training of the model converged at epoch 44. After that, the early stop was triggered on epoch 59. The training and validation loss, training and validation accuracy, and learning rates alongside the best epoch with the highest validation accuracy are presented in figure 1.

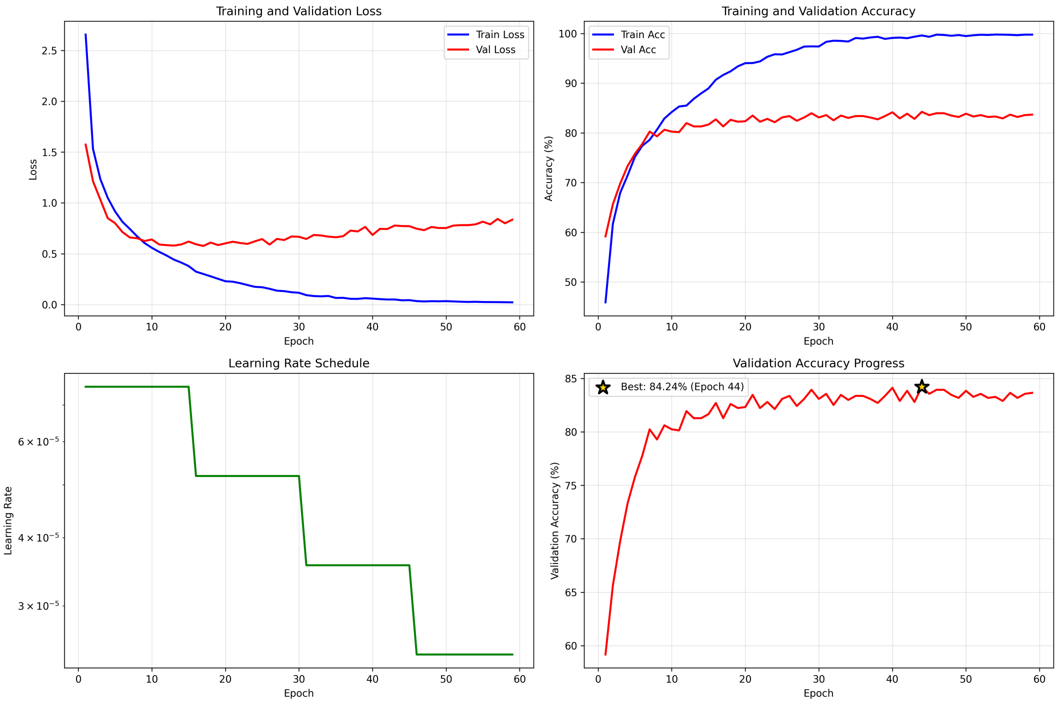

Figure 1: Model Training Curves

4.2 Model Evaluation

After training, the model was evaluated on the test set, and the selected metrics were computed.

Metric	**Value**
Accuracy	81.53%
Exact Match	81.53%

Table 1: Accuracy Metrics

Metric	**Value**
Macro F1 Score	54.53%
Weighted F1 Score	81.46%

Table 2: F1 Scores

Looking at table 1, it can be seen that the model achieved 81.53% accuracy on the overall test dataset. It also achieved 83.65% accuracy on closed and 80.16% accuracy on open answer types, as seen in table 3. From the performances presented in the study of Zhu et al., 2022, it can be seen that the baseline model has achieved close to SOTA performances in all of these metrics. The answer-type-specific metrics are visualized in figure 2.

Question Type	**Metric**	**Value**
Closed	Accuracy	83.65%
	F1 Score	64.75%
	Exact Match	83.65%
Open	Accuracy	80.16%
	F1 Score	53.98%
	Exact Match	80.16%

Table 3: Metrics Per Answer Types

From table 2, it can be seen that there is a 26.93% gap between the weighted and the macro F1 scores. This indicates that there is a class imbalance in the dataset. The low macro F1 score suggests that the model performs well on common answer classes but struggles with rare answer classes. From the answer-type-specific F1 scores, it can be determined that open-ended questions have more rare answer classes.

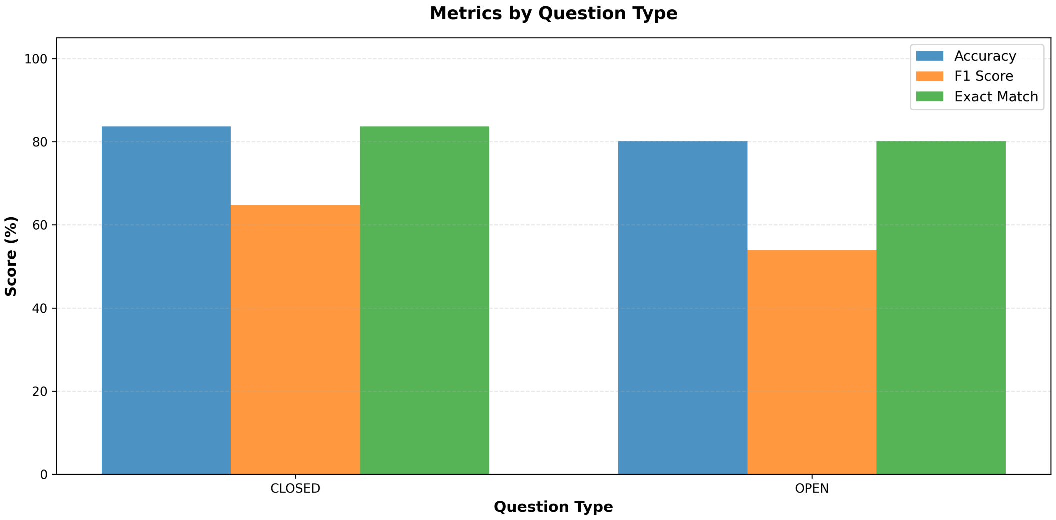

Figure 2: Answer Type Specific Metrics

From the generative metrics presented in table 4, it can be seen that the model has a BLEU-1 score of 83.79%, which drops drastically for BLEU-2, BLEU-3, and BLEU-4. The drop of the score is visualized in figure 3. The reason for this drop is because medical VQA answers are typically 1-2 words (e.g., 'yes', 'liver', 'ct scan'), making higher-order overlaps less common, resulting in lower BLEU-2/3/4 scores. For the same reason, it can be noticed that although the ROUGE-1 score is 84.42%, the ROUGE-2 score plummeted down to 15.89%. Finally, the model achieved a high BERTScore of 96.74%. This indicated that the model gives answers that are semantically very close to the ground truth answers.

Overall, the classification baseline achieves 81.53% accuracy with exceptional semantic understanding (96.74% BERTScore). The model excels at closed-ended questions (83.65%) and common medical terms, with performance consistent across metrics (81.53% exact match = accuracy).

Metric	**Value**
BLEU-1	83.79%
BLEU-2	35.82%
BLEU-3	24.57%
BLEU-4	19.23%
ROUGE-1	84.42%
ROUGE-2	15.89%
ROUGE-L	84.08%
METEOR	48.56%
BERTScore Precision	96.74%
BERTScore Recall	96.75%
BERTScore F1	96.74%

Table 4: Generative Metrics

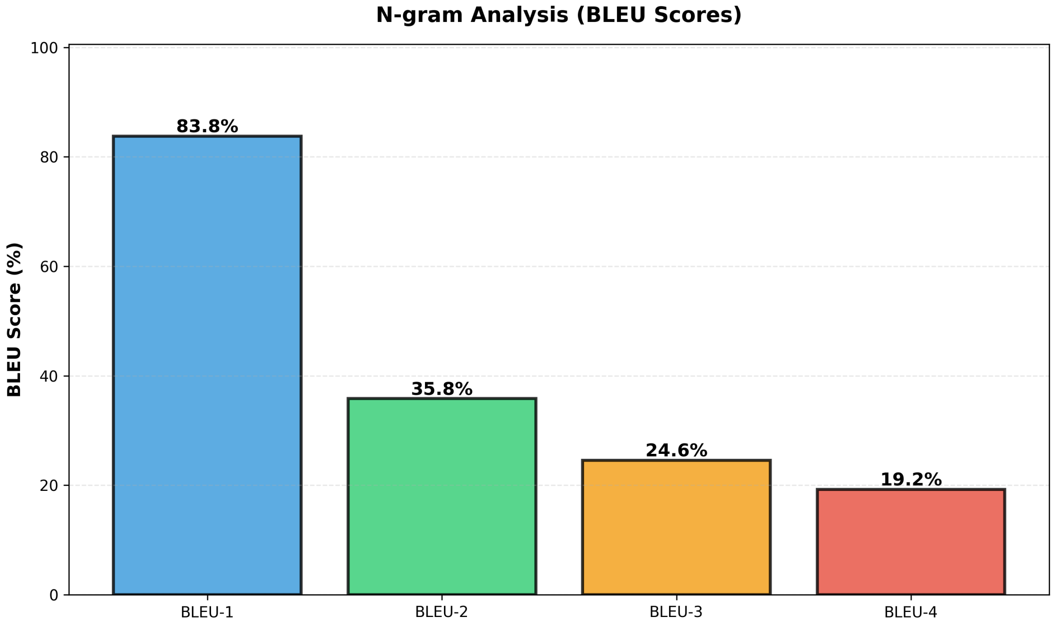

Figure 3: BLEU Scores

5 Summary and Next Step

This report presents a proposal of benchmarking a CNN-LSTM hybrid model against the BioMedBLIP model on the SLAKE dataset. Following this proposal, the baseline model is implemented and evaluated on the dataset. The model has achieved notable success in the classification-based VQA task.

The future steps following the completion of the baseline include:

- Fine-tuning BioMedBLIP on the SLAKE dataset for the VQA task using the same dataset split for fair comparison.
- Conducting similar evaluations using all the selected metrics.
- Answering the research questions presented as the objectives.

References

Bazi, Y., Rahhal, M. M. A., Bashmal, L., & Zuair, M. (2023). Vision–language model for visual question answering in medical imagery. *Bioengineering*, *10*(3). https://doi.org/10.3390/bioengineering10030380

Ben Chaabane, N., & Bal-Ghaoui, M. (2025). Visual question answering for medical diagnosis. *Intelligent Systems with Applications*, *27*, 200545. https://doi.org/https://doi.org/10.1016/j.iswa.2025.200545

Chen, Z., Zou, B., Dai, Y., Zhu, C., Kong, G., & Zhang, W. (2023). Medical visual question answering with symmetric interaction attention and cross-modal gating. *Biomedical Signal Processing and Control*, *85*, 105049. https://doi.org/https://doi.org/10.1016/j.bspc.2023.105049

Gong, H., Chen, G., Liu, S., Yu, Y., & Li, G. (2021). Cross-modal self-attention with multi-task pre-training for medical visual question answering. https://arxiv.org/abs/2105.00136

Gupta, R., Hooda, P., Sanjeev, & Kumar Chikkara, N. (2020). Natural language processing based visual question answering efficient: An efficientdet approach. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 900–904. https://doi.org/10.1109/ICICCS48265.2020.9121068

He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020, March). *Pathvqa: 30000+ questions for medical visual question answering*. https://doi.org/10.48550/arXiv.2003.10286

Lau, 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*, 180251. https://doi.org/10.1038/sdata.2018.251

Li, Y., Long, S., Yang, Z., Weng, H., Zeng, K., Huang, Z., Lee Wang, F., & Hao, T. (2022). A bi-level representation learning model for medical visual question answering. *Journal of Biomedical Informatics*, *134*, 104183. https://doi.org/https://doi.org/10.1016/j.jbi.2022.104183

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, *143*, 102611. https://doi.org/10.1016/j.artmed.2023.102611

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., & Wu, X.-M. (2021). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering [18th IEEE International Symposium on Biomedical Imaging (ISBI), Nice, FRANCE, APR 13-16, 2021]. *2021 IEEE 18TH INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING (ISBI)*, 1650–1654. https://doi.org/10.1109/ISBI48211.2021.9434010

Naseem, U., Thapa, S., & Masood, A. (2024). Biomedblip: Advancing accuracy in multimodal medical tasks through bootstrapped language-image pretraining (preprint). *JMIR Medical Informatics*, *12*. https://doi.org/10.2196/56627

Xin, Y., Ates, G. C., Gong, K., & Shao, W. (2025). Med3dvlm: An efficient vision-language model for 3d medical image analysis. https://arxiv.org/abs/2503.20047

Yim, W.-W., Ben Abacha, A., Doerning, R., Chen, C.-Y., Xu, J., Subbarao, A., Yu, Z., Xia, F., Hall, M., & Yetisgen, M. (2025, January). *Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care*. https://doi.org/10.2139/ssrn.5149002

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., . . . Poon, H. (2025). Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. https://arxiv.org/abs/2303.00915

Zhu, H., He, X., Wang, M., Zhang, M., & Qing, L. (2022). Medical visual question answering via corresponding feature fusion combined with semantic attention. *Mathematical Biosciences and Engineering*, *19*, 10192–10212. https://doi.org/10.3934/mbe.2022478