UNIVERSITY OF MALAYA

Faculty of Computer Science & Information Technology

WOA7015 Advanced Machine Learning
2025/2026 Semester 1

Alternative Assessment

Lecturer
PROFESSOR IR. DR. CHAN CHEE SENG

Prepared By
Group jherng

Team Member	Matric. No.
Hong Jia Herng	U2005313

Abstract

Medical Visual Question Answering (Med-VQA) represents a pivotal intersection of computer vision and natural language processing, promising to democratize medical expertise by enabling automated systems to interpret complex radiological imagery and respond to natural language inquiries. This capability is essential for scaling diagnostic support, reducing clinical workflow burdens, and providing accessible health information. This research report presents an exhaustive comparative analysis between two distinct modeling paradigms: a traditional discriminative architecture utilizing a Convolutional Neural Network (CNN) combined with a Long Short-Term Memory (LSTM) network, and a modern generative approach leveraging a pre-trained Vision-Language Model (VLM), specifically the Bootstrapping Language-Image Pre-training (BLIP) model, fine-tuned via Low-Rank Adaptation (LoRA). Using the Semantically-Labeled Knowledge-Enhanced (SLAKE) dataset, this study rigorously evaluates the efficacy, architectural trade-offs, and clinical applicability of these paradigms.

Our investigation reveals a nuanced landscape where the discriminative CNN-LSTM baseline demonstrates superior accuracy (83.17%) on closed-type answers due to the dataset's structural biases and limited answer vocabulary. However, it fell off a bit on open-ended answers (77.67%) due to its fixed-classifier design. Conversely, the generative BLIP model, while achieving a lower exact match score of 62.02% (Open), 80.53% (Closed), and 69.27% (Overall), exhibits profound semantic understanding and reasoning capabilities, evidenced by a high BERTScore F1 (91.62%) on open-ended questions. This report synthesizes theoretical underpinnings, detailed experimental methodologies, and quantitative results to argue that while discriminative models offer efficiency for routine tasks, the future of Med-VQA lies in generative foundation models, provided challenges regarding hallucination and domain adaptation are addressed.

Keywords: *Medical Visual Question Answering (Med-VQA), Vision-Language Models (VLM), CNN-LSTM, BLIP, Low-Rank Adaptation (LoRA)*

Table of Contents

1. Introduction

1.1. The Clinical Imperative for Automated Image Interpretation

The modern healthcare landscape is characterized by an explosion in medical imaging data. Modalities such as X-ray, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) have become indispensable for diagnosis, yet the human expertise required to interpret these images, i.e., radiologists and pathologists, remains a scarce resource globally. The disconnect between data volume and expert availability leads to diagnostic bottlenecks, increased error rates due to fatigue, and delayed patient care.

Artificial Intelligence (AI) has long been heralded as a solution to this crisis. Initial efforts focused on singular tasks: classification models to detect the presence of pneumonia or segmentation networks to delineate tumor boundaries. While effective, these "narrow AI" solutions lack the interactivity and explanatory power required in a clinical setting. A segmentation mask can show *where* a tumor is, but it cannot explain *why* it appears malignant or answer a follow-up question about its relationship to surrounding organs. This limitation has catalyzed the field of Medical Visual Question Answering (Med-VQA). Med-VQA systems are designed to emulate the consultative role of a radiologist, processing an image alongside a clinical question to produce a relevant, accurate answer. This capability supports diverse applications, from clinical decision support systems (CDSS) for junior doctors to patient-facing portals that explain radiology reports in layman's terms.

1.2. The Paradigm Shift: From Classification to Generation

The technological trajectory of Med-VQA mirrors the broader evolution of deep learning. For the past decade, the field was dominated by discriminative modeling. These architectures, exemplified by the CNN-LSTM framework, treat VQA as a multi-class classification problem. They map the visual and textual inputs to a shared vector space and predict the single most likely answer from a fixed, predefined vocabulary. While computationally efficient and easier to train on small datasets, these models are inherently rigid. They cannot generate answers outside their training distribution, rendering them ineffective for rare diseases or complex, descriptive queries.

We are now witnessing a paradigm shift towards generative modeling, driven by the success of Large Language Models (LLMs) and Vision-Language Models (VLMs). Models like BLIP treat VQA not as classification, but as a text generation task. They utilize Transformer-based decoders to predict answers token-by-token autoregressively, theoretically allowing for infinite flexibility in response formulation. This approach aligns closer to human reasoning, offering the potential for detailed descriptions and handling open-set questions. However, applying these general-domain models to medicine introduces significant challenges, including the "domain gap" between natural and medical images, computational costs, and the risk of "hallucination", i.e., generating plausible but factually incorrect medical advice.

1.3. Objectives

This report documents a comprehensive study benchmarking these two opposing paradigms on the SLAKE dataset. The objectives of this project are:

I. To design, implement, and evaluate a robust classical baseline using a pretrained CNN backbone for visual feature extraction and a LSTM for linguistic encoding, establishing a benchmark for discriminative performance.

II. To adapt a VLM (BLIP) to the medical domain using Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA), demonstrating the feasibility of training large models on SLAKE dataset.

III. To compare the strengths, limitations, and performance of both approaches on the SLAKE dataset.

IV. To identify which modelling approach is more suitable for medical VQA problems at small data scales.

2. Methods

The methodology was designed to ensure a rigorous, reproducible comparison of the two paradigms. The experiment was conducted end-to-end within a Google Colab A100 80 GBs GPU environment, utilizing Python libraries such as PyTorch, Transformers, and PEFT.

2.1. Exploratory Data Analysis

We utilized the SLAKE (Semantically-Labeled Knowledge-Enhanced) dataset, a seminal benchmark for Med-VQA. SLAKE was chosen over alternatives like VQA-RAD due to its richer semantic labels and structured knowledge base, although for this study, we focused on the direct image-question-answer pairs.

2.1.1. SLAKE Dataset

SLAKE contains **642 radiology images** (**CT, MRI, X-ray**) covering diverse body parts including the **Head, Chest, Abdomen, and Pelvic Cavity**. It features **14,028 question-answer pairs** annotated by experienced physicians.

Questions in SLAKE are broadly categorized into:

- Vision-only (requiring only perception, e.g., "Is the image clear?"),
- Knowledge-based (requiring external medical reasoning, e.g., "What is the treatment for this condition?")

This combination makes SLAKE particularly suitable for evaluating both **classical multimodal architectures** (e.g., CNN-LSTM) and **modern Vision–Language Models (VLMs)**

2.1.2. Language Filtering

The original dataset is bilingual (English/Chinese). We filtered the dataset to include **only English** pairs. This decision simplifies tokenization and vocabulary construction, while enabling fair comparison between discriminative and generative models under a unified language setting.

This resulted in a dataset split of:

- **Training:** 4,919 samples.
- **Validation:** 1,053 samples.
- **Test:** 1,061 samples.

2.1.3. Question Distribution

Figure 1 shows the distribution of question lengths (measured in number of words) in the training split. The maximum question length observed in training was **21 words**. To accommodate potential length variations in validation and test splits while maintaining computational efficiency, we set the model input length to:

$$max\_len = 32$$

This choice provides a reasonable buffer and results only in padding overhead, without significantly affecting training cost.

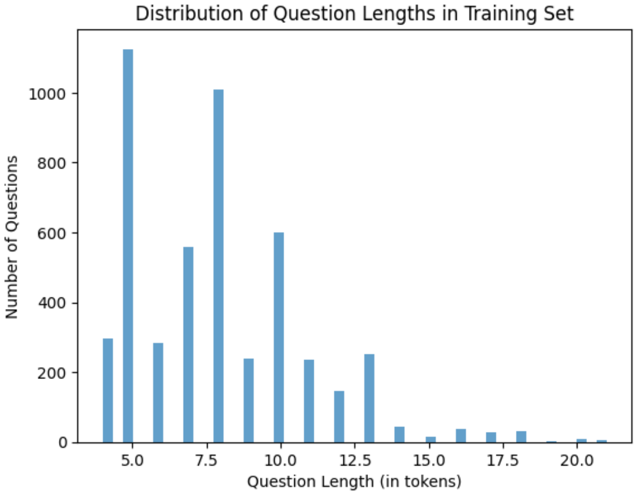


Figure 1: Distribution of the length of questions (by words) in training set.

In addition, the training question vocabulary size was **290 unique tokens**. Therefore, we fixed the vocabulary size to:

max_words = 290

To support robustness to unseen tokens, we included special tokens **<pad>** and **<unk>**, resulting in a final vocabulary size of 292 tokens.

2.1.4. Answer Distribution

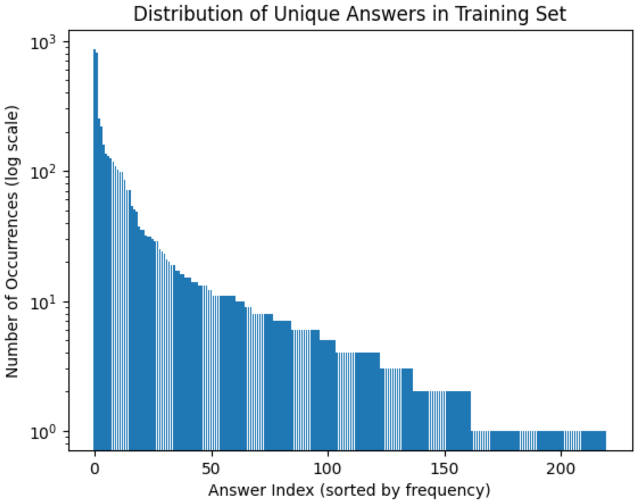


Figure 2: Distribution of the frequency of unique answers in training set (showing long-tailed problem).

Figure 2 illustrates the frequency distribution of unique answers in the training set. The distribution follows a clear **long-tailed pattern**, where a small number of answers occur very frequently while most appear only a few times. This characteristic is known to negatively impact **macro-F1**, as performance on rare classes tends to be disproportionately low.

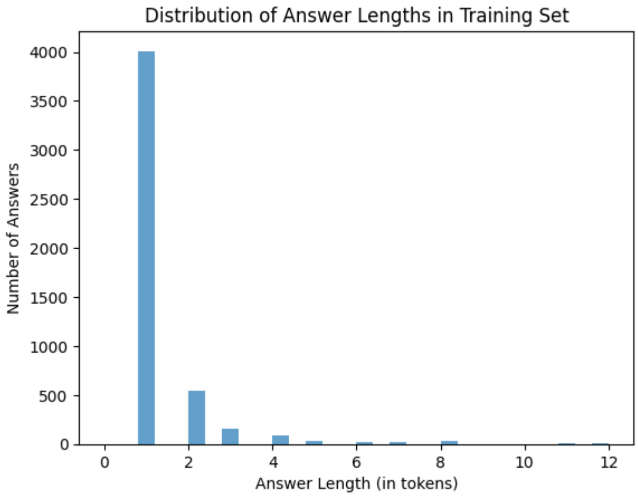


Figure 3: Distribution of the length of answers (by words) in training set.

Figure 3 shows the distribution of answer lengths. The maximum answer length in training was **12 words**, while the average answer length was **1.05 words**, indicating that most answers are extremely short (e.g., *"yes"*, *"no"*, *"lung"*, *"ct"*). This property naturally supports discriminative modelling approaches that treat Med-VQA as classification over a fixed set of answer classes. However, it also motivates exploring generative models that may generalize better to rare or paraphrased answers.

2.1.5. OPEN vs CLOSED Answers Type

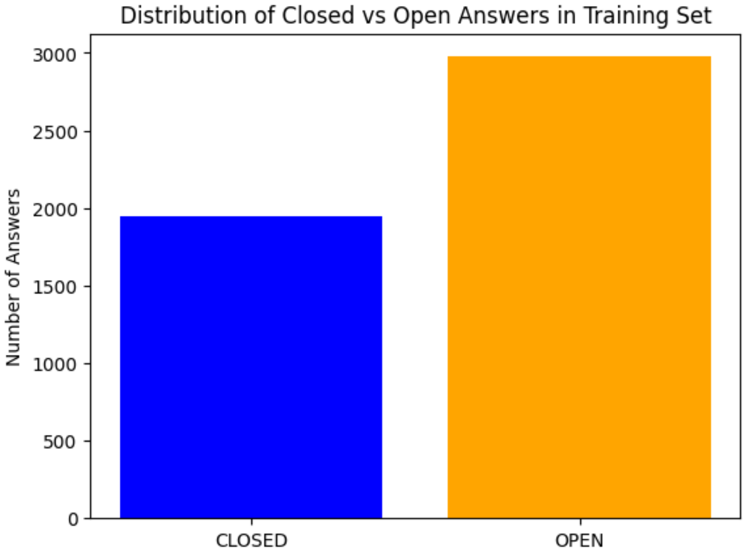


Figure 4: Frequency of CLOSED vs. OPEN answers in training set.

Figure 4 presents the distribution of **OPEN** and **CLOSED** answers in the training set. The analysis revealed a notable imbalance:

- **CLOSED answers** account for approximately **60%** of the dataset (typically yes/no, modality, or short categorical responses).
- **OPEN answers** account for approximately **40%**, requiring more descriptive and flexible responses.

This imbalance is important because CLOSED questions tend to favour classification-based models, while OPEN questions are more aligned with generative VLMs that can produce natural language outputs.

2.1.6. Summary

Overall, SLAKE contains a strong proportion of short-form answers, making classification baselines competitive. However, the presence of OPEN questions and long-tail answers creates a realistic setting where generative VLMs may demonstrate stronger generalization and semantic flexibility.

2.2. Model 1: Discriminative Baseline (CNN-LSTM)

The CNN–LSTM architecture represents a classical discriminative approach in multimodal learning and serves as a strong baseline for medical visual question answering. This model follows the **classification paradigm**, where each question is assumed to have an answer that belongs to a **closed vocabulary** of pre-defined concepts. Formally, the objective is to learn a mapping from the multimodal input pair (I, Q), where I is a medical image and Q is a question sequence, into a joint latent representation that supports answer classification.

This formulation reflects the traditional “extractive” assumption in early VQA research: rather than generating natural language responses, the model predicts one answer category from a fixed candidate set. As a result, it provides a clear benchmark for evaluating how far a conventional supervised architecture can perform under constrained answer spaces.

2.2.1. Model Architecture

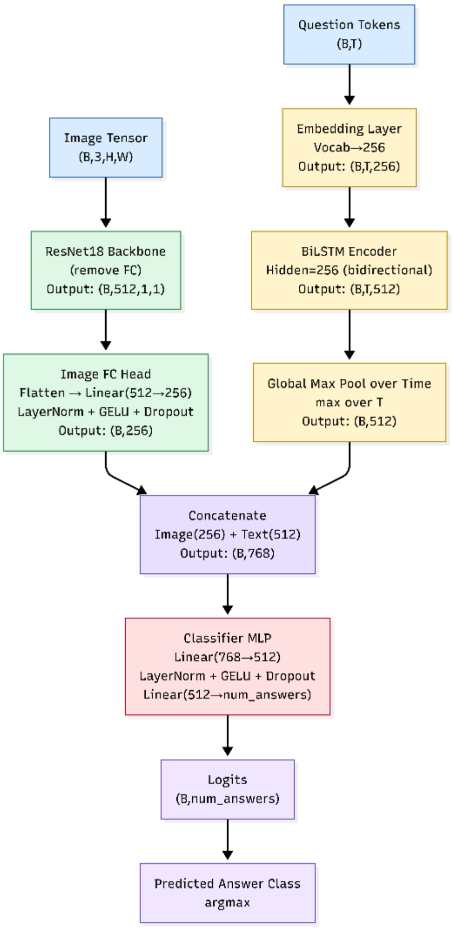


Figure 5: Model Architecture of the CNN-LSTM discriminator.

The CNN–LSTM model consists of three key components: an image encoder, a question encoder, and a fusion-based classifier head.

I. Image Encoder (ResNet-18)

We employed a **ResNet-18** backbone pretrained on ImageNet to encode radiology images into compact visual representations. The final fully connected classification layer was removed, and the network outputs a **512-dimensional feature vector**, which is further projected into a lower-dimensional embedding space through a linear projection block.

To assess the impact of domain adaptation, two training regimes were evaluated:

- **Frozen Backbone:**
  The convolutional layers were frozen, and only the projection layer and classifier head were trained. This setup evaluates how effective generic pretrained features are when transferred to medical images.

- **Fine-tuned Backbone:**
  All layers were unfrozen and jointly optimized, allowing the encoder to adapt its feature extraction to medical imaging characteristics such as grayscale textures, anatomical structures, and modality-specific patterns.

II. Question Encoder (BiLSTM)

Textual questions were tokenized using a word-level vocabulary built from the training set. The resulting vocabulary contained **292 tokens**, including special tokens **<pad>** and **<unk>**. Each question was padded or truncated to a fixed maximum length of **32 tokens**, enabling batch processing.

A **Bidirectional LSTM (BiLSTM)** with hidden size 256 was used to encode the question sequence. The bidirectional formulation enables the model to capture both preceding and subsequent context, which is helpful for medical questions containing modifiers such as “left”, “right”, “upper”, and “lower”.

III. Fusion and Classifier Head

The final image representation and question representation were concatenated to form a fused feature vector. This multimodal embedding was then passed through a Multi-Layer Perceptron (MLP), producing logits over the answer space.

To address the long-tailed answer distribution, the classifier output space was restricted to the **Top-K most frequent answers**, with all other answers mapped to an **<unk>** class. Specifically, we set:

$$K = 220 \Rightarrow \text{num\_answers} = 221 \ (220 + \langle unk \rangle)$$

This approach reduces label sparsity while ensuring unseen answers in validation/test sets remain representable.

2.2.2. Data Preprocessing

The preprocessing steps follow standard multimodal classification pipelines:

- **Images:** resized to 224×224, normalized using ImageNet mean/std
- **Questions:** tokenized into integer IDs, padded to $T = 32$
- **Answers:** mapped into a fixed categorical space of **220 frequent classes + <unk>**

2.2.3. Training Configuration

Both CNN–LSTM variants were trained under the same optimization setup:

- **Optimizer:** Adam, $lr = 1 \times 10^{-3}$
- **Batch size:** 128
- **Epochs:** 50
- **Loss function:** Cross-Entropy Loss
- **Model selection metric:** validation accuracy
- **Training time:** 11 m 23 s (frozen); 12 m 16 s (full fine-tuning)

The parameter counts for both training regimes are reported below:

- **Frozen Backbone:**
    - Total parameters: **12,943,901**
    - Trainable parameters: **1,767,389**
- **Fine-tuned Backbone:**
    - Total parameters: **12,943,901**
    - Trainable parameters: **12,943,901**

2.3. Model 2: Generative Approach (BLIP + LoRA)

To address the limitations of fixed-vocabulary classification models, we implemented a generative Vision–Language Model (VLM) baseline based on **BLIP (Bootstrapping Language-Image Pre-training)**. Unlike discriminative VQA architectures, BLIP formulates Med-VQA as a **conditional text generation task**, where the model generates an answer sequence $\hat{A}$ conditioned on the input image I and question Q. This design enables the model to produce flexible, free-form responses and reduces dependence on a predefined answer vocabulary.

2.3.1. Model Architecture

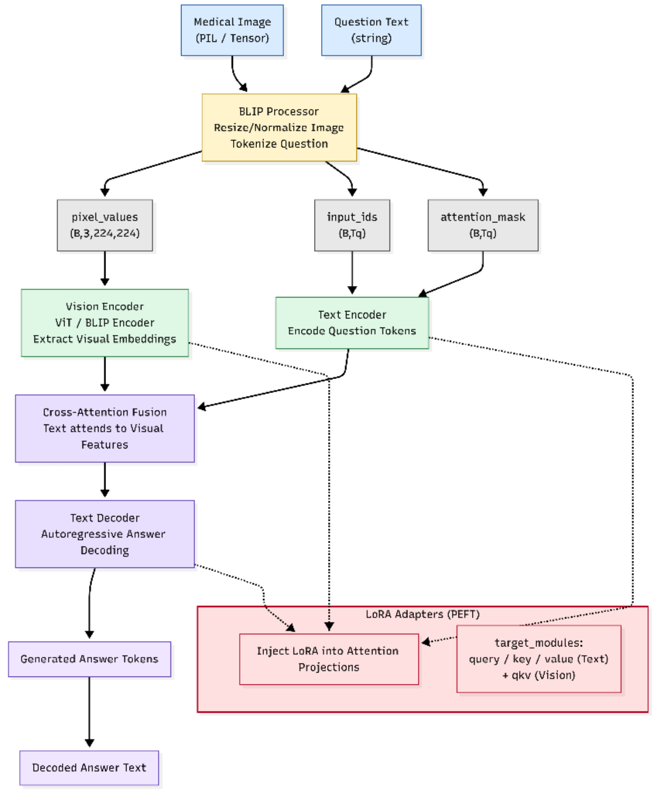

Medical Image (PIL / Tensor)
Question Text (string)
BLIP Processor Resize/Normalize Image Tokenize Question
pixel_values (B,3,224,224)
input_ids (B,Tq)
attention_mask (B,Tq)
Vision Encoder ViT / BLIP Encoder Extract Visual Embeddings
Text Encoder Encode Question Tokens
Cross-Attention Fusion Text attends to Visual Features
Text Decoder Autoregressive Answer Decoding
LoRA Adapters (PEFT)
Inject LoRA into Attention Projections
target_modules: query / key / value (Text) + qkv (Vision)
Generated Answer Tokens
Decoded Answer Text

Figure 6: Model architecture of BLIP + LoRA.

We adopted the pretrained **Salesforce/blip-vqa-base** model as the generative backbone. BLIP follows a transformer-based multimodal encoder–decoder design:

- **Vision Encoder:** A **ViT-Base** transformer encodes the input image into dense visual embeddings.
- **Text Components:** The question is tokenized and combined with visual representations through cross-modal attention, enabling grounded multimodal reasoning.
- **Answer Decoder:** The model generates an answer autoregressively as a sequence of tokens using a transformer decoder.

2.3.2. Parameter-Efficient Fine-tuning (PEFT) with LoRA

Directly fine-tuning all parameters of a pretrained VLM is often computationally expensive and may lead to overfitting in small-data medical settings. To mitigate this, we employed **Low-Rank Adaptation (LoRA)** as a parameter-efficient fine-tuning strategy.

LoRA injects trainable low-rank matrices into attention projections, allowing the model to adapt with minimal updates to the backbone weights. Specifically, LoRA adapters were added to the **query, key, and value projection matrices** within the attention mechanism.

This approach significantly reduces trainable parameters:

- **Total parameters:** 363,294,524
- **Trainable parameters (LoRA):** 2,064,384
- **Trainable fraction:** $\approx 0.57\%$

This demonstrates that BLIP can be adapted to the Med-VQA domain with **extreme parameter efficiency**, enabling practical fine-tuning under limited compute resources.

2.3.3. Training Configuration

The BLIP + LoRA model was fine-tuned using the following settings:

- **Optimizer:** AdamW, $lr = 1 \times 10^{-4}$
    - AdamW was chosen due to its stability for transformer optimization and its decoupled weight decay formulation.
- **Batch size:** 64
- **Epochs:** 20
- **Checkpoint selection metric: Validation Token F1-score**
- **Training time:** 51 m 42 s

Token-F1 was used instead of exact match to prioritize partial correctness and robustness to minor lexical variation (e.g., *"ct"* vs *"computed tomography"*), which is especially important for OPEN-ended medical answers.

2.4. Evaluation Metrics

Evaluating Med-VQA models requires careful metric selection, particularly when comparing **discriminative classifiers** against **generative Vision–Language Models (VLMs)**. While classification models produce discrete answer indices, generative models output free-form text.

Therefore, we employ a combination of classification-based and text-based metrics to ensure both completeness and fairness in evaluation.

2.4.1. Discriminative Metrics (CNN–LSTM)

For the CNN–LSTM baseline, answers are treated as categorical labels. We report:

- **Accuracy (Top-1)**

The proportion of examples where the predicted answer class matches the ground truth class.

- **Top-5 Accuracy**

The proportion of examples where the correct answer appears within the model's top-5 predicted classes. This is useful when multiple answers are plausible or when the model assigns high probability to the correct class but fails to rank it first.

- **Macro-F1 (Overall / OPEN / CLOSED)**

The F1-score averaged equally across answer classes, which makes it sensitive to performance on rare answers. This is particularly important for SLAKE, where the answer distribution is long-tailed.

2.4.2. Generative Metrics (BLIP + LoRA)

For the BLIP generative model, the output is a text sequence. We therefore evaluate answer quality using text-based metrics:

- **Exact Match (EM, normalized)** *(Overall / OPEN / CLOSED)*

  A strict metric that assigns a score of 1 if the predicted answer exactly matches the ground truth after text normalization (e.g., lowercasing and punctuation removal), and 0 otherwise.

  This metric can be interpreted as the text-generation equivalent of classification accuracy.
- **Token-level F1-score (Token-F1)** *(Overall / OPEN / CLOSED)*

  Measures the overlap between predicted tokens and ground truth tokens, providing partial credit for partially correct answers.

  This is especially important for open-ended responses where near-miss answers may still convey clinically relevant information (e.g., predicting *"lung cancer"* instead of *"left lung cancer"*).

For OPEN-ended questions only, we additionally report:

- **BLEU**

  An n-gram precision-based similarity metric, emphasizing exact token matches. BLEU tends to be strict and may under-estimate correctness for short medical answers.
- **ROUGE-L**

  Measures similarity using the longest common subsequence, making it more recall-oriented and less sensitive to minor phrasing differences.
- **BERTScore**

  A semantic similarity metric computed using contextual embeddings and cosine similarity.

  This captures synonymy and paraphrasing (e.g., *“CT”* vs *“computed tomography”*), making it a valuable supplement to Exact Match in medical domains.

2.4.3. Fair Cross-Model Comparison

To enable direct comparison between CNN–LSTM and BLIP despite their different output formats, we compute a unified set of text-level metrics for all models:

- **Exact Match (normalized)**
- **Token-F1**

For CNN–LSTM, class predictions are mapped back into text using the answer vocabulary before computing these metrics. This ensures that discriminative and generative models are compared under the same evaluation protocol, particularly for OPEN-ended answers where lexical variation and partial correctness are common.

3. Results and Analysis

This section presents and analyzes the experimental results on the SLAKE test split, comparing a classical discriminative baseline (CNN–LSTM) against a modern generative Vision–Language Model (BLIP + LoRA). The findings highlight complementary strengths: discriminative models perform reliably when answers fall within a constrained vocabulary, while generative models demonstrate greater flexibility in open-ended answering and semantic alignment.

3.1. Quantitative Performance Comparison

Table 1 summarizes the performance of CNN–LSTM models (frozen and fine-tuned backbones) using classification metrics. Table 2 reports BLIP (LoRA) performance using text-generation metrics. Finally, Table 3 presents a fair cross-model comparison using **Exact Match** and **Token-F1** for all models.

3.1.1. CNN–LSTM (Discriminative) Results

Table 1: CNN–LSTM test performance using classification metrics (Accuracy / Top-5 / Macro-F1).

Model	Subset	Accuracy	Top-5 Accuracy	Macro-F1
CNN–LSTM (Frozen)	Overall	**0.7983**	0.9576	0.4758
CNN–LSTM (Frozen)	OPEN	**0.7767**	0.9318	0.4718
CNN–LSTM (Frozen)	CLOSED	**0.8317**	0.9976	0.6095
CNN–LSTM (Fine-tuned)	Overall	0.7766	0.9576	0.4350
CNN–LSTM (Fine-tuned)	OPEN	0.7473	0.9318	0.4238
CNN–LSTM (Fine-tuned)	CLOSED	0.8221	0.9976	0.6600

Overall, the frozen CNN–LSTM achieved the highest classification accuracy (**0.7983**), outperforming the fine-tuned version (**0.7766**). This suggests that freezing the pretrained visual backbone yields better generalization under small-scale training data. Both CNN variants performed strongest on the CLOSED subset, consistent with the fact that CLOSED questions typically correspond to discrete labels such as modality categories and binary answers.

3.1.2. BLIP + LoRA (Generative) Results

Table 2: BLIP + LoRA test performance using generative metrics.

Model	Subset	Exact Match	Token-F1	BLEU	ROUGE-L	BERTScore-F1
BLIP (LoRA)	Overall	**0.6927**	**0.7364**	–	–	–
BLIP (LoRA)	OPEN	0.6202	0.6919	0.0311	0.7259	**0.9162**
BLIP (LoRA)	CLOSED	0.8053	0.8053	–	–	–

BLIP (LoRA) achieved strong overall performance, with **Token-F1 = 0.7364** and **Exact Match = 0.6927**. Notably, on OPEN questions, the model achieved a high semantic similarity score (**BERTScore-F1 = 0.9162**), indicating that many answers were correct in meaning even when they did not exactly match the ground truth string. On CLOSED questions, BLIP remained competitive (**Exact Match = 0.8053**), suggesting that the generative model can also handle discrete-answer tasks effectively.

3.1.3. Fair Cross-Model Comparison

To compare discriminative and generative models under a unified protocol, we evaluated all models using text-based metrics by mapping CNN predicted classes back into their corresponding answer strings.

Table 3: Fair comparison across all models using Exact Match (normalized) and Token-F1.

Model	Subset	Exact Match	Token-F1
CNN–LSTM (Frozen)	Overall	0.7983	0.8253
CNN–LSTM (Fine-tuned)	Overall	0.7766	0.8131
BLIP (LoRA)	Overall	0.6927	0.7364
CNN–LSTM (Frozen)	OPEN	0.7767	0.8212
CNN–LSTM (Fine-tuned)	OPEN	0.7473	0.8073
BLIP (LoRA)	OPEN	0.6202	0.6919
CNN–LSTM (Frozen)	CLOSED	0.8317	0.8317
CNN–LSTM (Fine-tuned)	CLOSED	0.8221	0.8221
BLIP (LoRA)	CLOSED	0.8053	0.8053

This comparison highlights that CNN–LSTM achieves higher strict correctness on the SLAKE test set when the answer lies within the fixed Top-K vocabulary. However, BLIP demonstrates strong semantic performance, particularly on open-ended questions when evaluated using Token-F1 and BERTScore (Table 2).

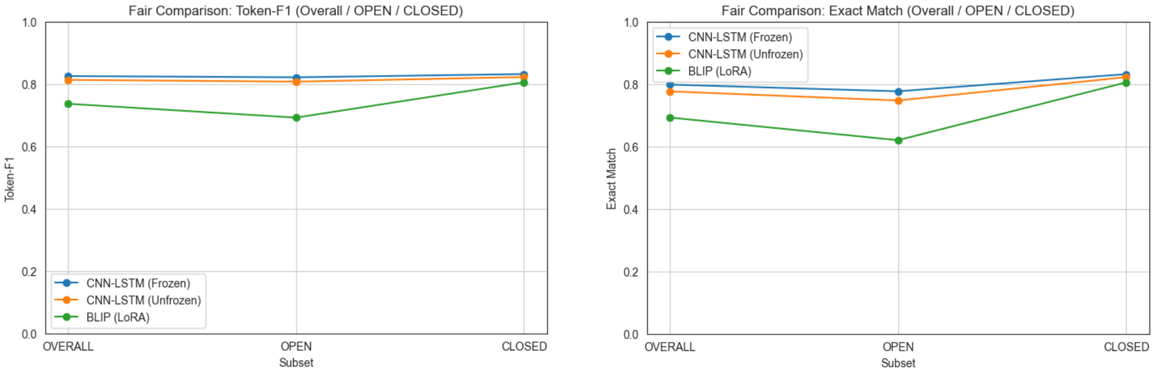


Figure 7: Token F1 / Exact Match across all models and answer types.

3.2. Training Dynamics: Frozen vs Fine-tuned CNN-LSTM

The frozen CNN–LSTM outperformed the fine-tuned variant in overall accuracy, suggesting that fine-tuning the ResNet backbone may lead to overfitting on SLAKE's limited training data. This is consistent with the observation that small medical datasets often lack sufficient diversity to safely update all visual encoder parameters without regularization or additional pretraining.

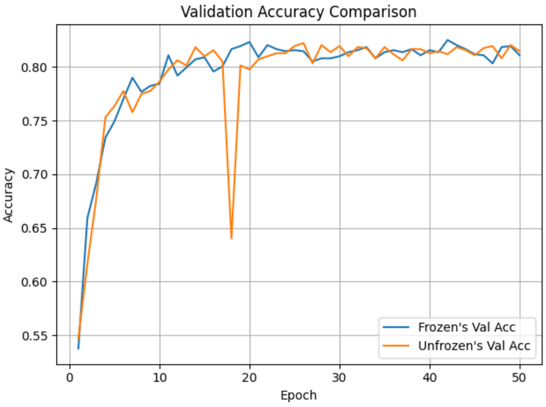


Figure 8: Validation accuracy comparison between frozen and fine-tuned CNN–LSTM across epochs.

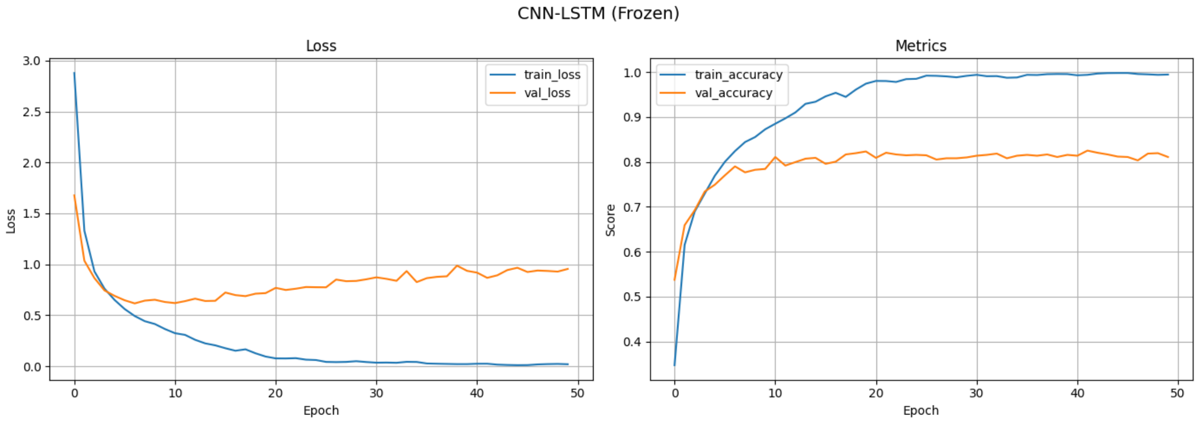


Figure 9: Training history of CNN–LSTM (Frozen Backbone): train/val loss and accuracy curves.

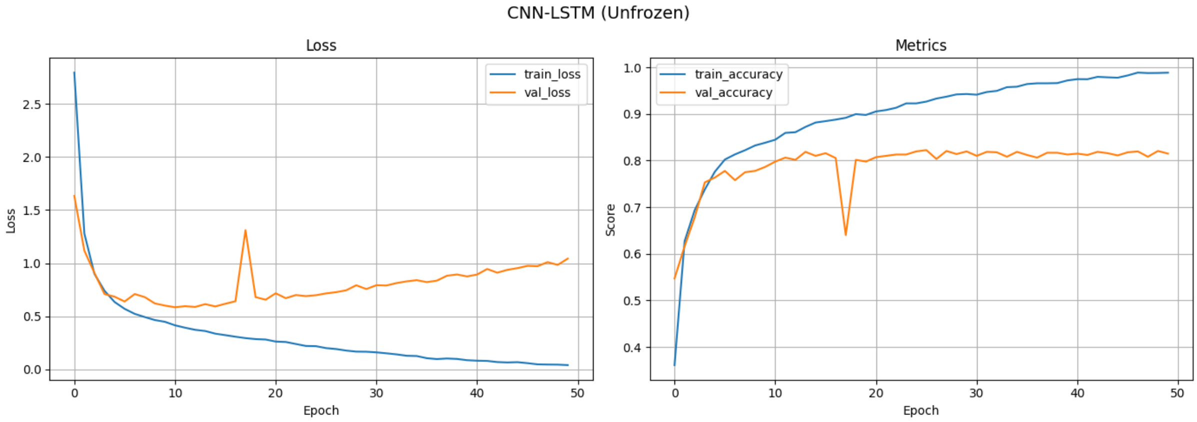


Figure 10: Training history of CNN–LSTM (Fine-tuned Backbone): train/val loss and accuracy curves.

3.3. CNN–LSTM Analysis: Closed-Set Reliability and Top-5 Gap

CNN–LSTM performed particularly well on the CLOSED subset (**0.8317 accuracy** for the frozen variant), which aligns with the classification assumption. Many closed questions correspond to short answers such as modality identification (“CT”, “MRI”) and binary responses (“yes/no”), which are naturally handled by discriminative models.

Additionally, both CNN variants achieve extremely high **Top-5 accuracy (0.9576 overall)**, significantly higher than Top-1 accuracy. This implies that while the model often assigns high probability to the correct answer, it may fail to rank it first. Such behavior is consistent with the long-tail distribution of answers: frequent classes dominate decision boundaries, while semantically similar classes may be confused.

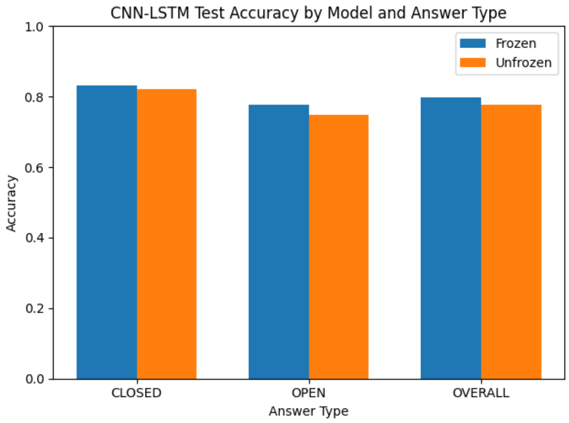


Figure 11: CNN–LSTM test accuracy by answer type (Overall / OPEN / CLOSED).

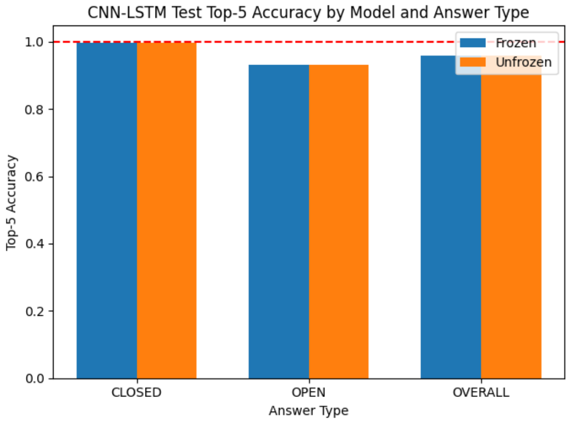


Figure 12: CNN–LSTM test Top-5 accuracy by answer type.

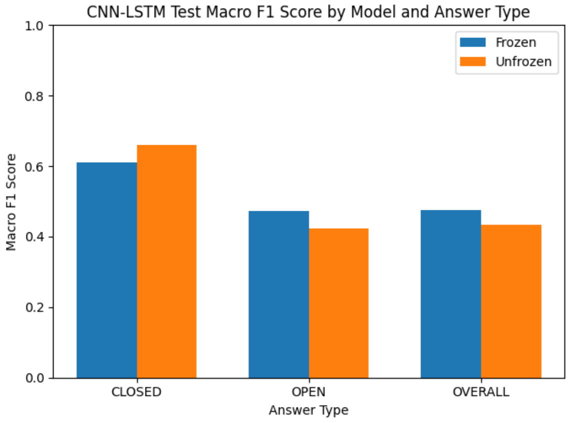


Figure 13: CNN–LSTM test Macro-F1 by answer type.

3.4. BLIP + LoRA Analysis: Semantic Strength and Medical Precision Challenges

Unlike classification models, BLIP produces free-form answers, making strict Exact Match a conservative measure of correctness. On OPEN questions, BLIP achieves **Exact Match = 0.6202**, yet reaches **BERTScore-F1 = 0.9162**, demonstrating that many mismatches arise due to phrasing variations rather than incorrect content. For example, a prediction such as “this is the abdomen” for the ground truth “abdomen” would be scored as incorrect under Exact Match but remains clinically equivalent.

However, qualitative inspection also reveals that BLIP occasionally generates answers that are semantically plausible but clinically imprecise (e.g., confusing left vs right). This highlights a known limitation of generative models: fluent responses can reflect language priors that are not always perfectly grounded in the visual signal.

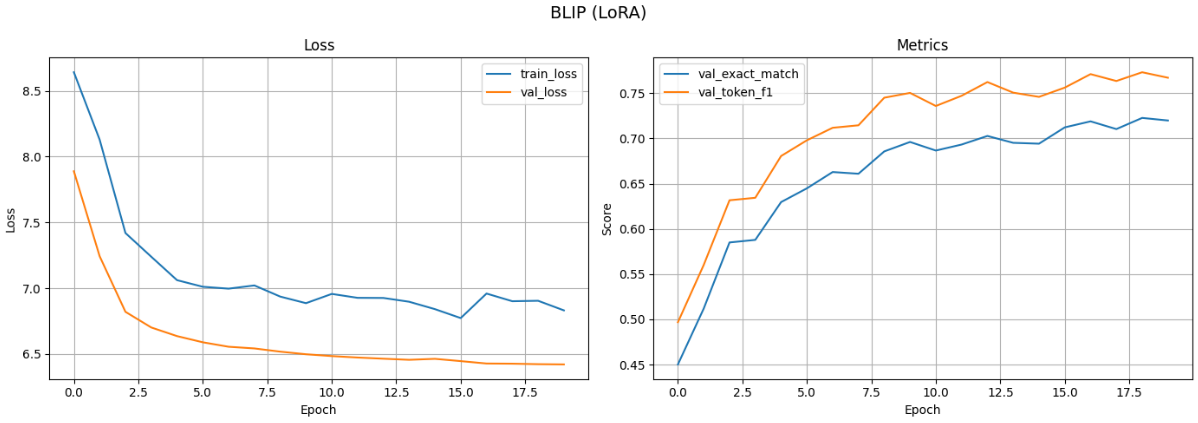


Figure 14: Training history of BLIP + LoRA: train loss and validation Token-F1/Exact Match across epochs.

3.5. Long-Tail Effect and Per-Class Performance Degradation

The SLAKE training set exhibits a long-tail answer distribution, where a small number of answers account for a large proportion of samples while many answers appear infrequently. This affects discriminative models more strongly when evaluated using macro-F1, since rare classes contribute equally to the overall score.

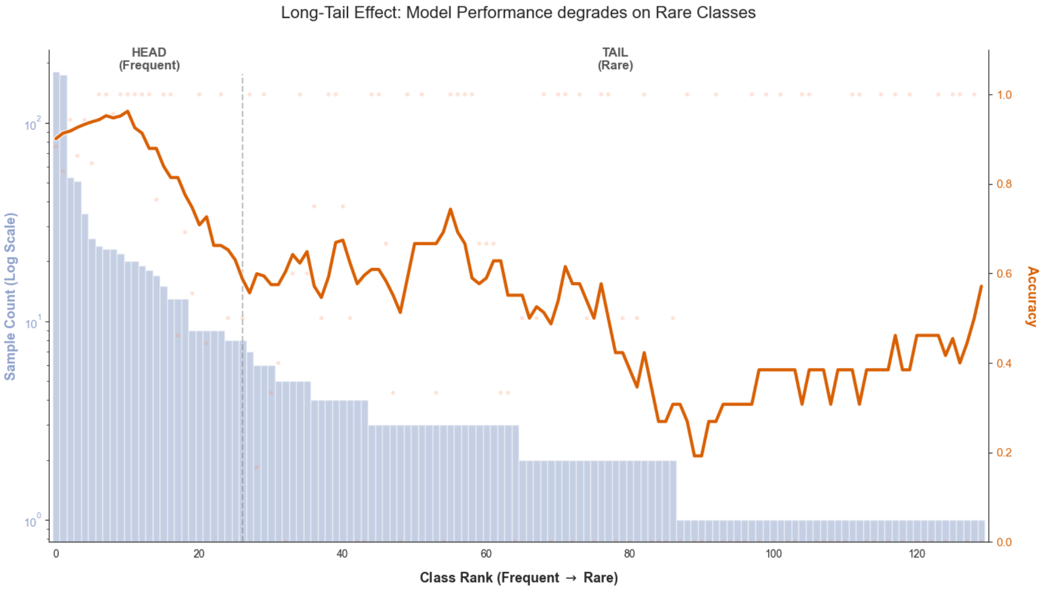


Figure 15: Long-tail effect: per-class accuracy vs class frequency for the best CNN–LSTM model.

Per-class analysis of the best CNN–LSTM model shows that accuracy degrades substantially for rare answer classes, reinforcing the impact of limited supervision for low-frequency labels.

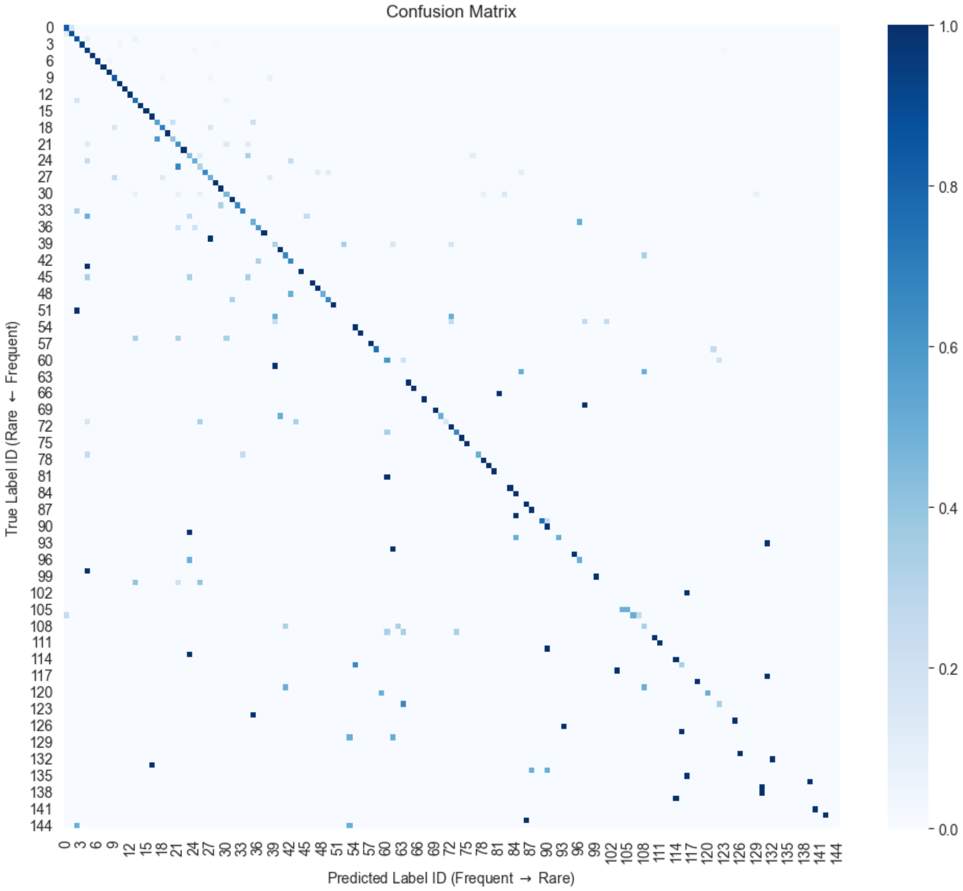
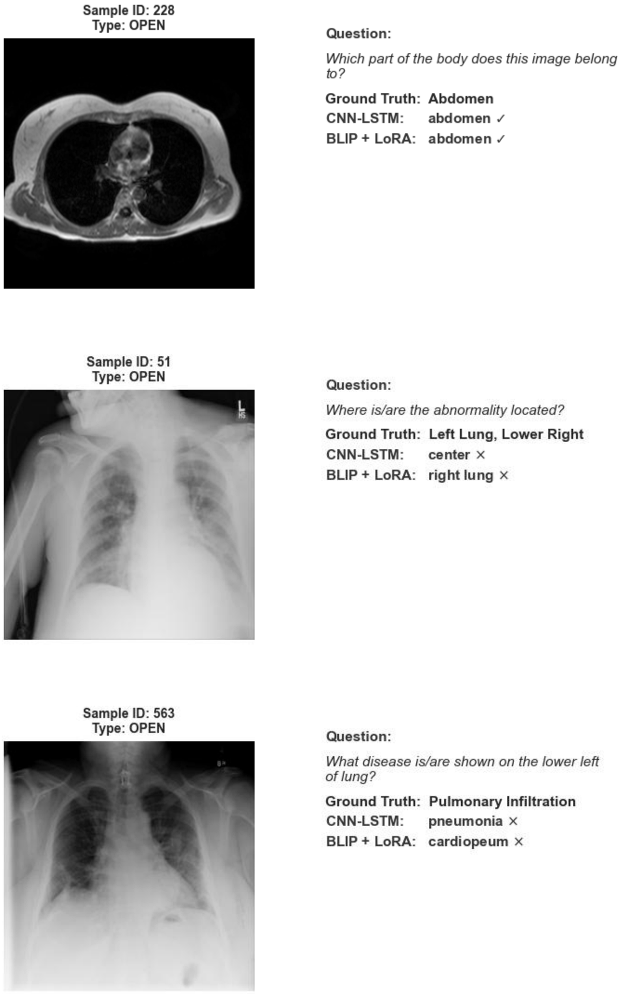


Figure 16: Confusion matrix of best CNN–LSTM model on CLOSED subset answers.

3.6. Qualitative Comparison

Beyond numerical metrics, qualitative examples highlight typical failure patterns. Common errors occur on questions requiring precise location or fine-grained anatomical terminology (e.g., left vs right, specific structures). BLIP often produces semantically correct answers even when the wording differs from the ground truth (e.g., synonyms or paraphrasing). In contrast, CNN–LSTM is constrained to the fixed Top-K answer vocabulary and may output <unk> when the correct answer is rare or unseen in training.



Question:

Which part of the body does this image belong to?

Ground Truth: Abdomen
CNN-LSTM: abdomen ✓
BLIP + LoRA: abdomen ✓

Question:

Where is/are the abnormality located?

Ground Truth: Left Lung, Lower Right
CNN-LSTM: center ×
BLIP + LoRA: right lung ×

Question:

What disease is/are shown on the lower left of lung?

Ground Truth: Pulmonary Infiltration
CNN-LSTM: pneumonia ×
BLIP + LoRA: cardiopeum ×

Figure 17: Qualitative comparison examples (image + question + ground truth + CNN prediction + BLIP prediction).

3.7. Summary of Findings

- The **CNN–LSTM (Frozen)** baseline achieved the strongest classification accuracy (**0.7983**) and performed reliably on CLOSED questions.
- Fine-tuning the ResNet backbone did not improve overall performance, suggesting **overfitting risk** under small-scale training.

- **BLIP + LoRA** achieved strong generative performance with high semantic alignment on OPEN questions (**BERTScore-F1 = 0.9162**), while remaining competitive on CLOSED questions.
- The long-tail distribution in SLAKE affects rare class performance, contributing to the observed gaps between accuracy and macro-F1 in discriminative models.

4. Discussion

4.1. Safety–Reasoning Trade-off in Medical VQA

The central finding of this study is that **model architecture fundamentally determines the balance between clinical safety and diagnostic reasoning capacity**.

Discriminative models as safety-constrained systems.

The CNN–LSTM baseline functions as a tightly constrained decision system. By restricting outputs to a fixed answer vocabulary, it prevents the generation of nonsensical or medically implausible responses. This *safety-by-design* property is advantageous for narrow, closed-set tasks such as modality identification or binary verification (e.g., "Is this a CT scan?"). However, this rigidity comes at a cost: the model cannot generalize beyond its predefined answer space. Rare conditions or unseen concepts are inevitably misclassified or mapped to <unk>, limiting the model's utility for real-world clinical dialogue and rare-disease reasoning.

Generative models as reasoning-oriented systems.

In contrast, BLIP represents a shift toward generative medical reasoning. The model's strong semantic alignment with ground truth answers, evidenced by high BERTScore, suggests an improved ability to interpret medical visual–textual context beyond strict answer selection. This opens the door to explainable and conversational clinical assistants. However, this flexibility introduces the risk of hallucination: generative models may confidently describe abnormalities that are not visually present, driven partially by language priors rather than image evidence.

Overall, these results highlight a fundamental trade-off: **discriminative models prioritize safety through constraint, while generative models prioritize reasoning through flexibility**.

4.2. The Role of Domain-Specific Pre-training

The performance gap between our BLIP model and state-of-the-art Med-VQA systems can be partially attributed to **domain mismatch in pre-training**. BLIP-VQA is pretrained on natural image corpora (e.g., COCO, Visual Genome), which differ substantially from medical imaging in texture, modality, and semantic structure.

Although LoRA enables parameter-efficient adaptation, the underlying vision encoder still perceives images through a natural-image prior. Consequently, subtle radiological cues, such as fine-grained grayscale textures or small anatomical boundaries remain challenging. Future work should explore domain-aligned foundation models (e.g., BioMedBLIP or PMC-CLIP) to reduce this domain gap prior to fine-tuning.

4.3. Limitations of Evaluation Metrics in Medical AI

This study exposes a critical shortcoming in current Medical VQA evaluation practices: **standard NLP metrics do not reflect clinical risk**.

- **Exact Match** is overly rigid and penalizes clinically correct paraphrases.
- **BLEU and ROUGE-L** emphasize n-gram overlap, ignoring diagnostic polarity. For example, “no pneumothorax” versus “pneumothorax” may share high lexical overlap but represent a fatal diagnostic error.
- **BERTScore** better captures semantic similarity but remains agnostic to medical severity and consequence.

These limitations highlight the need for **clinically weighted evaluation metrics**. In medical contexts, hallucinated lesions or missed diagnoses should be penalized far more severely than syntactic variation or synonym choice. Future benchmarks should incorporate **ontology-aware metrics** (e.g., UMLS) or **human-in-the-loop evaluation** to bridge the gap between linguistic correctness and clinical reliability.

4.4. Future Improvements and Extensions

Several improvements could further enhance performance and reliability:

Train BLIP longer with sufficient resources

- BLIP + LoRA was trained for 20 epochs and did not exhibit strong overfitting behaviour, suggesting that extended training or improved optimization schedules may further improve performance.

Explore stronger Vision–Language Models (e.g., BLIP-2).

- While BLIP-VQA already showed promising results under limited training, more advanced VLM architectures such as **BLIP-2** may provide stronger cross-modal grounding and reasoning capacity.

Enhance CNN–LSTM text representations using pretrained embeddings.

- The CNN–LSTM baseline currently learns question embeddings from scratch with a small vocabulary. Incorporating pretrained semantic representations (e.g., **Word2Vec**, **GloVe**, or contextual encoders such as **BERT**) could improve robustness to lexical variation and strengthen language understanding without significantly increasing model complexity.

5. Conclusion

This study compared a classical discriminative CNN–LSTM baseline and a generative Vision–Language Model (BLIP-VQA) fine-tuned with LoRA on the SLAKE Med-VQA dataset. The results show that the CNN–LSTM remains a strong baseline for CLOSED and high-frequency questions, benefiting from a fixed answer space that produces stable predictions under small-data settings. However, its classification formulation limits generalization to rare or unseen answers, which are often mapped to incorrect classes or <unk>.

In contrast, BLIP + LoRA demonstrated stronger semantic flexibility, achieving high Token-F1 and BERTScore on OPEN questions, indicating that generated answers are often meaningfully correct even when phrasing differs from the ground truth. The study also confirms LoRA as an effective parameter-efficient strategy, enabling competitive adaptation while updating only a small fraction of model weights. Overall, discriminative models offer reliability through constrained outputs, while generative VLMs provide greater potential for open-ended medical reasoning.

Author Contribution

This project was completed solely by myself. All design decisions, implementation, experiments, evaluation, and analysis were carried out independently. External resources (e.g., research papers, documentation, and open-source repositories) were consulted for reference and guidance. In addition, AI-assisted tools were used to support productivity, primarily for summarising technical materials and improving writing efficiency.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

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.

Appendix

Code Repository: https://github.com/hongjiaherng/woa7015-medvqa