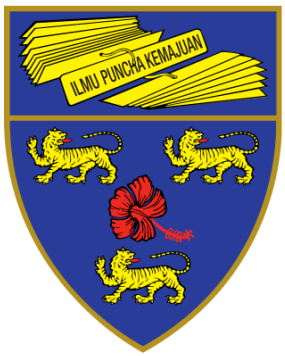

UNIVERSITY OF MALAYA

Faculty of Computer Science & Information Technology

WOA7015 Advanced Machine Learning
2022/2023 Semester 1

Alternative Assessment – Preliminary Report

Lecturer
PROFESSOR IR. DR. CHAN CHEE SENG

Prepared By
Group jherng

Team Member	Matric. No.
Hong Jia Herng	U2005313

Table of Contents

1 – Background

Medical images such as X-rays, CT scans, and MRIs contain rich anatomical and pathological information that is difficult to interpret without clinical training. Deep learning has achieved strong performance in tasks such as image classification and segmentation, but these models lack the ability to understand questions about an image. Medical Visual Question Answering (Med-VQA) aims to close this gap by enabling models to answer natural-language questions grounded in medical images.

This capability is clinically valuable because it supports interactive querying, clarifies radiological findings, and may provide decision support for clinicians.

However, Med-VQA remains challenging due to:

1. limited dataset sizes compared to general-domain VQA,
2. high variability of medical imaging modalities,
3. the need for models to process both visual and linguistic information jointly, and
4. the requirement to generate medically precise answers.

To address these challenges, recent research has explored both traditional multimodal deep learning models and large-scale Vision–Language Models (VLMs), which are pretrained on billions of image–text pairs. This project investigates the differences between these two classes of models on the SLAKE dataset (Liu et al., 2021).

2 – Objectives

The objectives of this project are:

1. To build and evaluate a classical multimodal baseline using a CNN for image encoding and an LSTM for question encoding.
2. To benchmark a modern pretrained Vision–Language Model (BLIP-VQA) on the same dataset.
3. To compare the strengths, limitations, and performance of both approaches on the SLAKE Med-VQA dataset.
4. To identify which modelling approach is more suitable for medical VQA problems at small data scales.

This preliminary report focuses on describing the dataset, explaining the modelling approaches, and presenting early results from a zero-shot BLIP-VQA baseline and a partially trained CNN-LSTM model baseline.

3 – Method

3.1 - Dataset Description, Preparation, & Processing

Dataset

This project uses the SLAKE (Semantically-Labeled Knowledge-Enhanced) dataset, a multilingual medical VQA dataset containing radiology images paired with clinical questions and short factual answers.

SLAKE includes:

- Medical images (e.g., CT, MRI, X-ray, etc.)
- Both closed-ended questions (Yes/No, modality type, organ identification)
- And open-ended factual questions (e.g., "What organ is enlarged?")

The dataset provides a diverse mixture of visual concepts and medical terminology, making it suitable for evaluating both classical and pretrained VQA models. In this project, we scoped the question-answering pairs down to just English (skipping Chinese), effectively reducing the number of training examples from 9835 to 4919.

Answer Space Construction

Although SLAKE contains free-text answers, most answers are short (< 20 words) and drawn from a limited vocabulary.

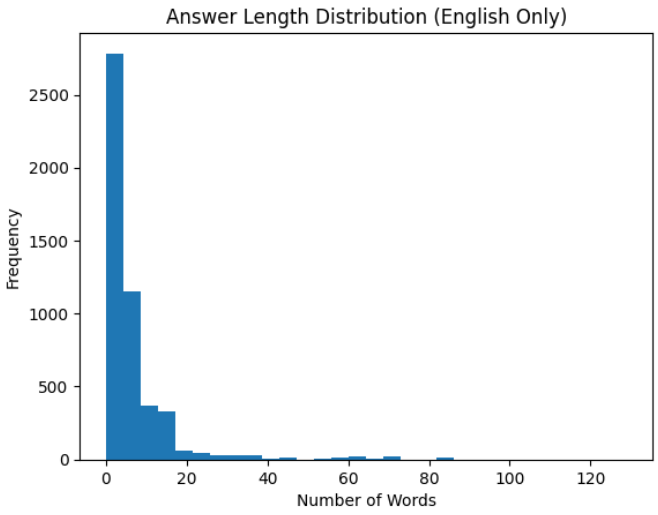

Figure 1: Answer Length Distribution (English Only)

With such insight, we can treat the baseline model as multi-class classification problem.

Steps:

1. Extract all answers from the training set.
2. Lowercase and normalise text (remove punctuation).
3. Build an answer vocabulary of the top K most frequent answers (e.g., K=222 in our experiment, <UNK> class included).
4. Map all rare answers to an <UNK> token.

The BLIP-VQA model outputs text generatively, but for comparison, predictions will still be evaluated using exact-match accuracy against the ground-truth SLAKE answers.

Image Preprocessing

- Images resized to 224×224
- Convert to RGB to fit ResNet input
- Normalised using ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225)

Question Preprocessing (Baseline only)

- Lowercase
- Tokenisation with NLTK
- Convert each token to a hash value modulo 5000, so they'll span within 0-4999
- Take the ID (0 – 4999) as input to the word embedding layer
- Word embedding via learnable embedding matrix

3.2 - Model Architectures

A. Baseline Model: ResNet + LSTM

The baseline model follows the traditional two-branch VQA architecture that independently encodes the image and question, then fuses both modalities for answer prediction. The current model has 13,418,622 trainable parameters.

Image Encoder

A **ResNet-18** backbone pretrained on ImageNet (torchvision ResNet18_Weights.IMAGENET1K_V1) is used to extract global image features:

- The final classification layer is removed,
- producing a **512-dimensional** feature vector for each image.

This encoder provides a strong, lightweight visual representation, suitable for small medical datasets like SLAKE.

Question Encoder

The question branch consists of:

- A **learnable embedding layer** (Embedding(5000, 300)) that maps discrete tokens into 300-dimensional embeddings.
- A **single-layer LSTM** with hidden size 256 that processes the token sequence.

Only the final hidden state is used as the **textual representation**, yielding a compact summary of the question semantics.

Fusion Mechanism

The model applies **early multimodal fusion** by concatenating the image and text embeddings. This fused representation is passed through a fully connected classification layer to predict an answer from the vocabulary.

Output Layer and Training Objective

The output head is a linear layer, which produces a probability distribution over the K answer categories via softmax (k=222).

The model is optimized using cross-entropy loss, treating the Med-VQA task as a multiclass classification problem over a fixed answer vocabulary.

B. Vision–Language Model: BLIP-VQA

BLIP (Bootstrapping Language Image Pretraining) is a pretrained Vision–Language Model designed for VQA, captioning, and image–text matching. The version that we're using "blip-vqa-base" has a total of 361,230,140 trainable parameters.

BLIP-VQA consists of:

- A ViT-based image encoder
- A Transformer-based text encoder
- A multimodal fusion encoder
- A generation-based decoder that produces the answer token by token

Unlike the baseline, BLIP-VQA:

- Is pretrained on 129M image–text pairs
- Has strong zero-shot VQA ability

- Generates answers in natural language form, not from a fixed vocabulary

For this Week-9 report, BLIP-VQA is used zero-shot (no fine-tuning).

In the final report, fine-tuning or feature-extraction-only training will be considered depending on compute availability.

4 - Preliminary Results

Baseline

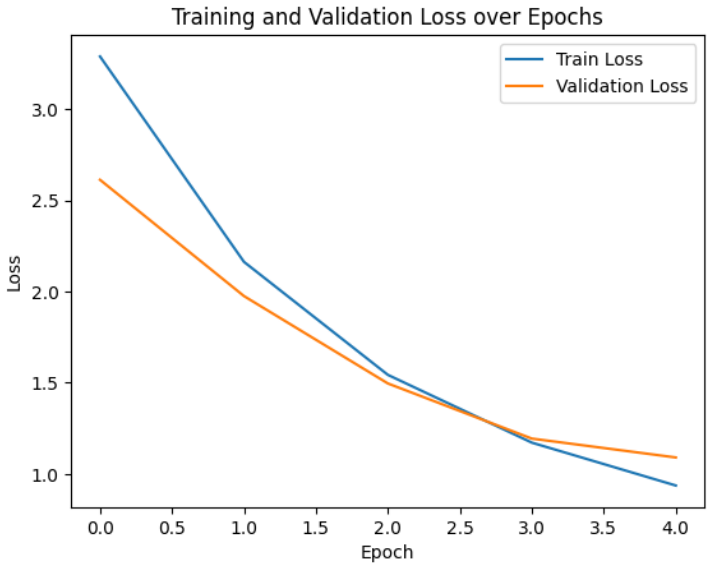

Figure 2: Learning Curve of Baseline Model (5 epochs)

The baseline model was trained for **5 epochs** using the Adam optimizer (learning rate = 1e-4). Figure 2 shows the learning curve, where the training and validation loss decrease steadily, indicating stable optimization with a bit of overfitting starting at epoch 4.

To better understand model behavior, we evaluated performance on the test set and reported accuracy separately for **open-ended** and **closed-ended** questions.

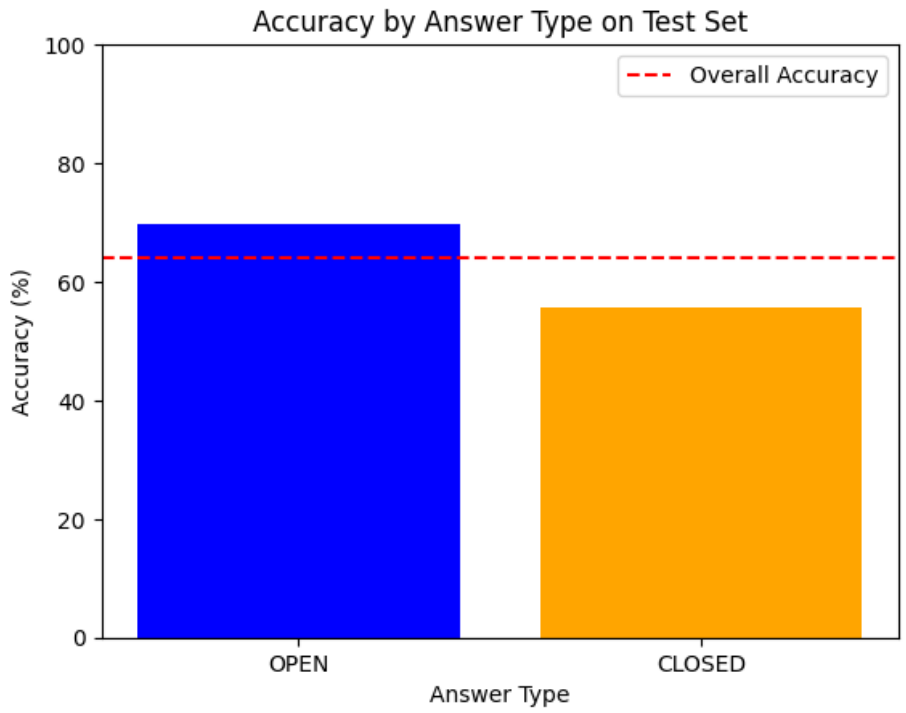

Figure 3: Accuracy by Answer Type on Test Set (Baseline)

Metric	Accuracy
Overall	64.10%
OPEN	69.73%
CLOSED	55.69%

Interpretation

- **Closed-ended performance (55.69%)** indicates that the baseline is performing above trivial guessing, suggesting successful learning of core visual–textual alignment.
- **Open-ended performance (69.73%)** is higher, partly because answers in this category include more frequently occurring factual terms that the model learns more easily.
- The gap between OPEN and CLOSED accuracy suggests that the baseline benefits from predictable answer distributions, but struggles with binary or categorical reasoning.

These results confirm that a classical architecture can learn meaningful image–question relationships even without large-scale multimodal pretraining.

Zero-shot BLIP-VQA

Zero-shot inference was conducted using **BLIP-VQA (base)** with no training or fine-tuning on SLAKE.

Because BLIP is pretrained on general-domain image–text pairs (not medical images), zero-shot performance is expected to be limited.

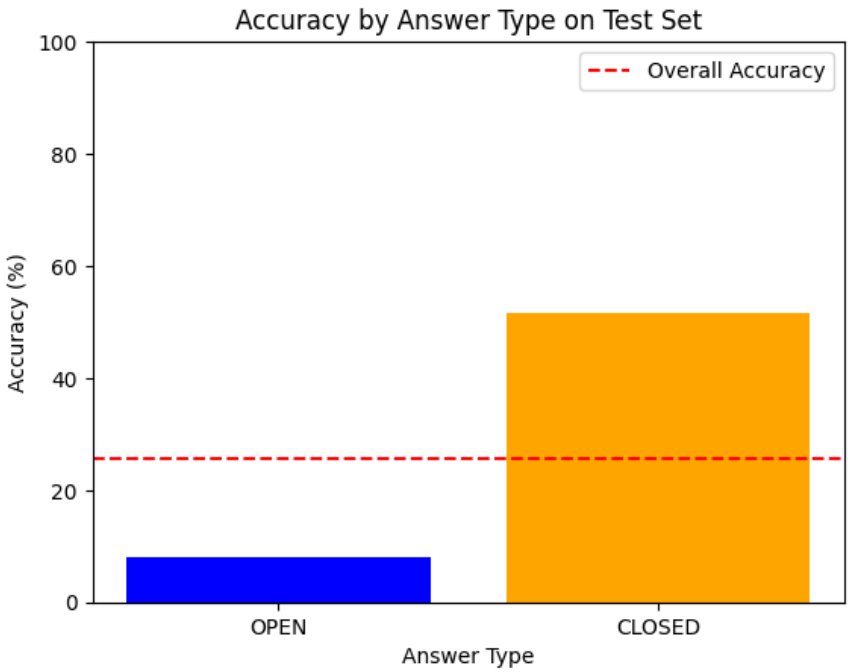

Figure 4: Accuracy by Answer Type on Test Set (Zero-shot BLIP-VQA)

Metric	Accuracy
Overall	25.55%
OPEN	8.08%
CLOSED	51.66%

Interpretation

- **Closed-ended performance (51.66%)** is close to guess-level behaviour, suggesting the model does not possess reliable radiology-specific reasoning out-of-the-box.
- **Open-ended accuracy (8.08%)** is extremely low. This reflects:
  - the **domain mismatch** between medical images and BLIP's pretraining data, and
  - the mismatch between **free-form generation** and SLAKE's answer vocabulary.
- These results strongly indicate that **domain-adaptive fine-tuning** or **medical-specific VLMs** (e.g., MedBLIP variants) would be required for reliable zero-shot VQA performance

Side-by-side Comparison

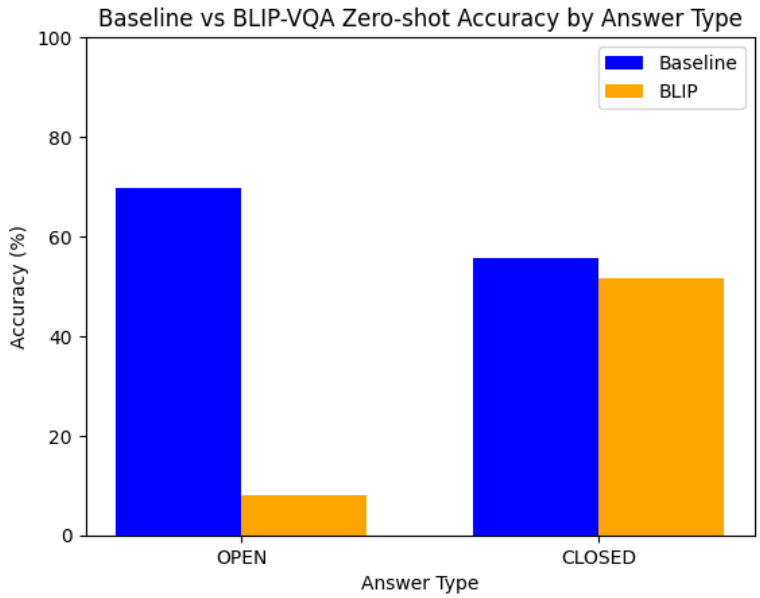

Figure 5: Baseline vs. BLIP-VQA Zero-shot Accuracy by Answer Type

Key Observations

- The **baseline outperforms zero-shot BLIP-VQA** on both OPEN and CLOSED questions, despite being a relatively simple architecture.
- BLIP's strength lies in **broad natural-image reasoning**, but it lacks exposure to medical imaging conventions and terminology.

This contrast highlights the importance of dataset-specific learning even for powerful pretrained multimodal models.

Inspecting the Samples

To further understand model behaviour, we examine several samples.

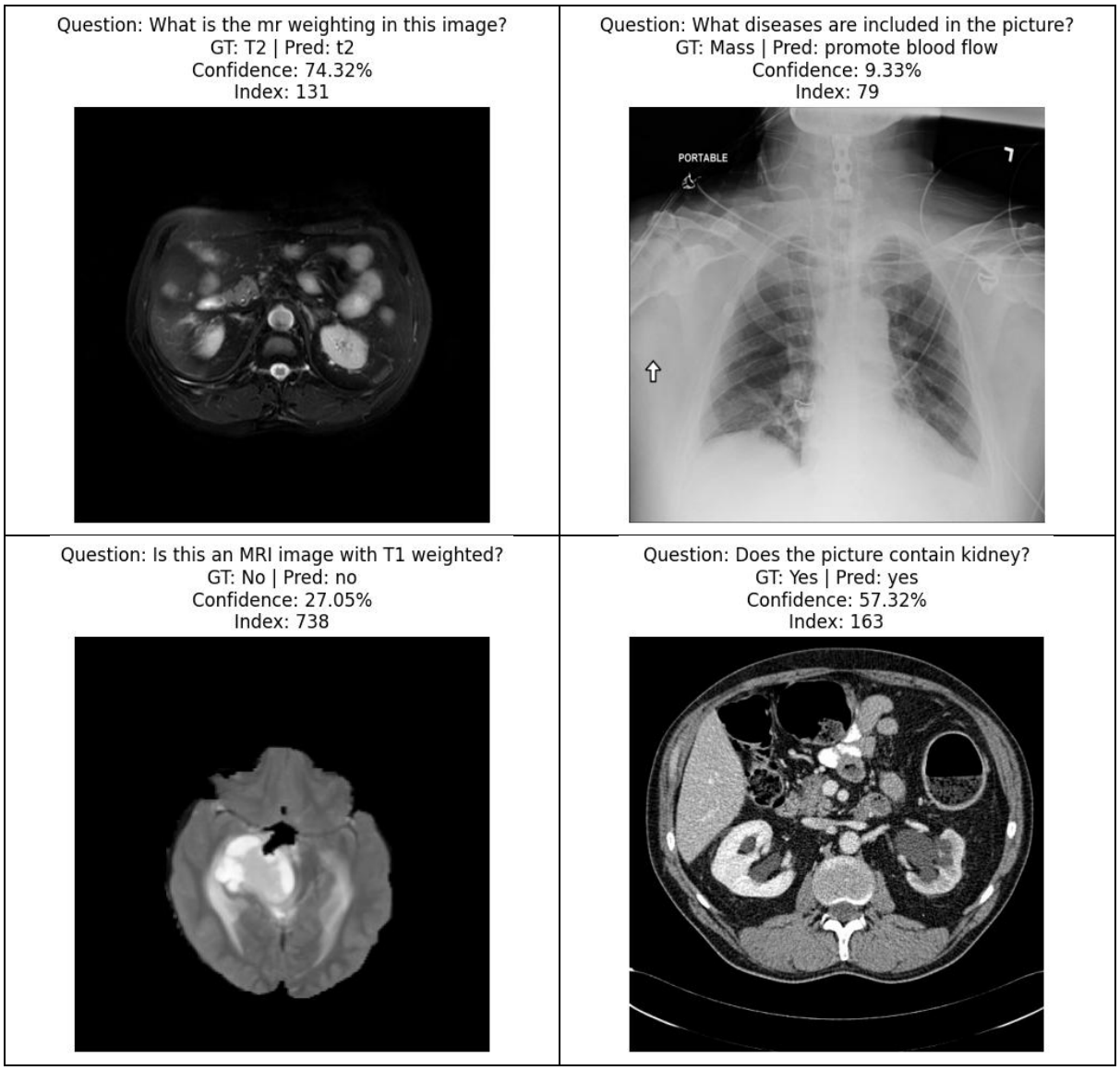

Figure 6: Inspecting a few samples with baseline prediction

Looking at Figure 6, The baseline generally produces reasonable predictions for modality, anatomical position, or common findings. However, errors remain. For instance, in one sample ("What diseases are included in the picture?"), the ground truth answer *"Mass"* is misclassified as *"Promote blood flow"*, reflecting the difficulty of rare classes and context-dependent visual reasoning.

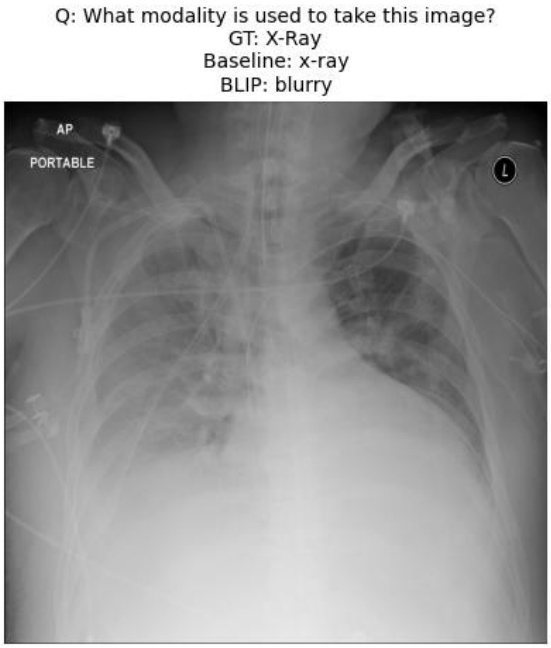

Figure 7: Prediction of Baseline Model vs. Zero-shot BLIP-VQA

Based on Figure 7, this illustrates BLIP's domain gap: although strong on natural images, BLIP tends to hallucinate or produce semantically irrelevant text when applied zero-shot to radiology.

References

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. arXiv preprint arXiv:2102.09542.