Medical Visual Question Answering: Improving Answer Quality on Multimodal Datasets with a Generative Output Layer

Naeun Lee
naeunl@andrew.cmu.edu

Virginia Choi-Durham
vdurham@andrew.cmu.edu

Soomin Chung
soominch@andrew.cmu.edu

Michael Chen
mchen5@andrew.cmu.edu

1 Abstract

Healthcare data is inherently multimodal, combining textual (e.g., clinical notes) and visual (e.g., medical images) components. Medical Visual Question Answering (Med-VQA) seeks to address this by enabling AI models to interpret multimodal data and answer clinically relevant questions. However, most Med-VQA models rely on classification-based approaches, which constrain adaptability and clinical applicability due to the use of a fixed answer set.

We propose reframing Med-VQA as a generative task by replacing the classification head in the state-of-the-art M^3AE model with a generative layer. This approach allows the model to produce open-ended, contextually relevant responses, improving its flexibility and practical utility. Experimental results demonstrate significant enhancements in question-answering performance, highlighting the potential of generative Med-VQA models to support diverse and real-world clinical applications effectively.

2 Introduction

Healthcare professionals synthesize diverse data inputs—such as radiological images, patient charts, and lab results—to make critical clinical decisions. This complexity makes healthcare a compelling domain for AI/ML research, as it demands models capable of integrating multimodal inputs effectively.

Vision Language Models (VLMs) have shown strong potential in multimodal healthcare applications, including medical image classification, disease diagnosis, and automated report generation. Med-VQA extends these capabilities by enabling AI models to process question-image pairs and generate clinically relevant answers. However, Med-VQA faces unique challenges, such as the need for precise medical image interpretation and limited data availability due to privacy constraints like HIPAA regulations. Current Med-VQA models primarily treat the task as a classification problem, where answers are selected from a fixed candidate pool. While effective in constrained scenarios, this approach limits adaptability to diverse questions and hinders clinical relevance.

To address these limitations, we propose a generative approach to Med-VQA using the state-of-the-art M^3AE model. M^3AE is a multimodal architecture that processes visual and textual inputs through co-attention fusion layers. It employs a vision transformer (ViT) for image processing and a BERT-based encoder for text inputs. By replacing the classification head with a generative layer, the model can produce open-ended, contextually relevant answers. This design enhances flexibility and better aligns with the complexity of real-world clinical scenarios.

This study evaluates the generative approach, demonstrating its ability to improve question-answering performance compared to traditional classification-based methods. By effectively integrating textual and visual inputs, the proposed model addresses key limitations in existing Med-VQA systems, paving the way for more robust and clinically adaptable AI solutions.

3 Literature Review

Visual Question Answering (VQA) in the medical domain has made significant advancements in recent years, driven by evolving neural network architectures and improved multimodal data integration. The progression of these models reflects a broader trend in deep learning towards more sophisticated, end-to-end trainable architectures that can better capture the complexities of both visual and textual information.

3.1 Architectures

3.1.1 General VQA Architecture Overview

Image Encoder: Often based on Vision Transformers or advanced CNNs, pre-trained on large medical image datasets.

Text/Question Encoder: Usually transformer-based, pre-trained on medical text corpora.

Feature Fusing Layers/Algorithms: Increasingly sophisticated cross-attention mechanisms or joint encoders to facilitate interactions between image and text features.

Answering Component: Most models leverage classification heads, but recently, models have started incorporating generative decoders for open-ended responses.

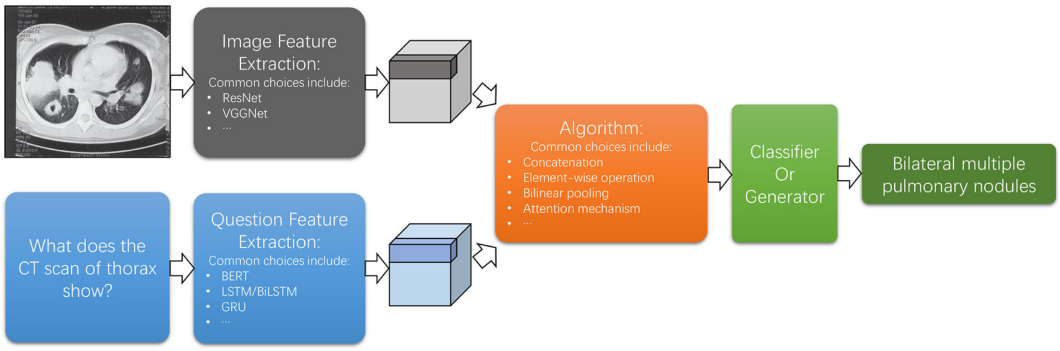

Figure 1: Basic Architecture Components [12]

See the architectures we reviewed in Table 1.

3.1.2 VQA Architecture Changes Over Time [12]

Early Approaches (CNN + LSTM): Initial VQA models combined CNNs for image feature extraction with LSTMs for text processing, using simple fusion methods like concatenation. These models struggled with simplistic text embeddings, difficulty in capturing long-range dependencies, and limited ability to model complex image-text relationships.

Transformers for Text: Transformers revolutionized language modeling with self-attention, enabling models to capture context more effectively than LSTMs. BERT, introduced in 2019, set new standards in NLP by learning bidirectional contextual representations, achieving SOTA performance with minimal fine-tuning. BERT remains the text encoder of choice in medical VQA models like M^3AE and BioMedCLIP due to its efficiency and versatility, despite newer autoregressive architectures excelling in generative tasks.

Table 1: Multimodal Models Reviewed

Model	Visual Encoder	Text Encoder	Fusion Mechanism	Decoder / Classifier
M^3AE[4]	ViT-B/16	$\text{RoBERTa}_{\text{base}}$	Dual Transformer, Co-Attention	Image: Transformer, Text: MLP + Softmax
MISS[3]	ViT-B/16	Transformer-based JTM Encoder containing self-attention, cross-attention, and feed-forward layers		Text-only decoder for generation
DAL[7]	ResNet-152	TSE (Transformer with Sentence Embedding - WordPiece + Sentence-BERT)	DAL Transformer, Self-attention, Guided-attention	MLP + Softmax for classification
MMBERT[9]	ResNet-152	Lightweight BERT	Transformer, Self-attention	MLP + Softmax
Med-Flamingo[13]	Normalizer-Free ResNet		Cross-Attention	Multimodal decoder
BiomedCLIP[19]	ViT-B/16	PubMedBERT	METER (Transformer, Co-attention)	Text-only classifier

Vision Transformers: Early VQA models relied on ResNet CNNs for visual feature extraction. The Vision Transformer (ViT), introduced in 2021, replaced CNNs with patch-based tokenization for images, allowing better scalability on large datasets. While CNNs excel in tasks requiring feature locality, like identifying abnormalities (e.g., VQA-RAD), ViTs outperform in tasks requiring holistic image understanding, such as SLAKE's organ identification. Modern models often combine CNNs and ViTs for improved performance.

Attention-Based Fusion Mechanisms: Advanced fusion techniques like co-attention and cross-attention have enhanced integration of image and text features. Models such as M^3AE and BioViL-T leverage these mechanisms for better multimodal interaction.

Domain-Specific Image and Text Encoders: Recent models employ self-supervised pre-training on large medical datasets for domain-specific image embeddings and use specialized language models like PubMedBERT for medical text to better capture domain-specific nuances.

3.2 Dataset Manipulation and Training

Interleaved (Image and Text) vs. Paired: In medical VQA tasks, interleaved datasets present image and text inputs together (e.g., paired tokens), facilitating contextual multimodal processing. Paired datasets use separate branches for images (e.g., X-rays) and text (e.g., questions) with later fusion, often used in two-stream architectures.

Masking Strategy: Masking selectively hides parts of the input to improve model learning. Text masking hides portions of questions or medical text to enhance contextual understanding, while image masking covers specific regions of medical images (e.g., X-rays) to train the model to focus on clinically relevant areas.

Table 2: Overview of Vision-Language Models and Masking Strategies

Model	Interleaved vs. Pairs	Masking Strategy
M^{3}AE [4]	Interleaved	Image + Text masking
MISS [3]	Pairs	Image + Text masking
DAL [7]	Pairs	No strong masking focus
MMBERT [9]	Interleaved	Text masking
Med-Flamingo [13]	Pairs, Interleaved	None
BiomedCLIP [19]	Pairs	Minimal masking

4 Dataset: VQA-RAD [10]

The VQA-RAD dataset consists of 315 radiology images and 3,515 clinician-generated question-answer pairs, specifically designed for medical Visual Question Answering (VQA) tasks. It is one of the larger Med-VQA datasets, though still smaller compared to datasets from other domains, reflecting the difficulty of obtaining and validating medical data under privacy regulations.

The images cover a range of body regions, including the head, chest, and abdomen, and the questions span various clinical areas, such as abnormalities, anatomical locations, and conditions. This diverse set of question types and images ensures the dataset's broad clinical applicability, making it suitable for training models aimed at real-world medical scenarios.

To make the dataset practically usable, preprocessing steps include normalizing image sizes and tokenizing the question-answer pairs. The data is organized into batches, with instance sampling based on question types to ensure balanced representation across different clinical areas. This method ensures the model is exposed to a diverse range of clinical questions during training, improving its ability to generalize across medical imaging tasks.

5 Baseline Model: M^{3}AE

5.1 Architecture

M^{3}AE combines a ViT-based vision encoder, BERT-based language encoder, and co-attention multimodal fusion module.

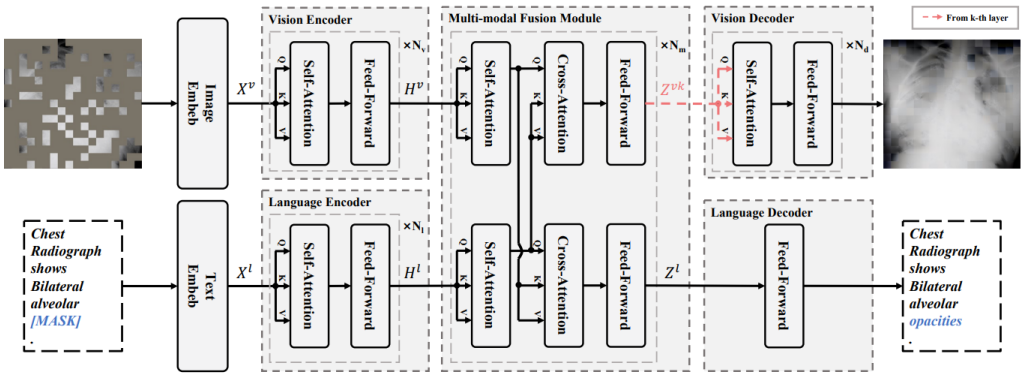

Figure 2: Pre-training M^{3}AE Architecture

5.1.1 Vision Encoder

Following the Vision Transformer (ViT) architecture, the vision encoder first processes the input image by dividing it into non-overlapping patches. Suppose the input image has C channels, and the patch resolution is $P \times P$. Then the image is segmented into patches $p_1, ..., p_N$ with each

$p_i \in \mathbb{R}^{P^2 \times C}$. Let D then be the embedding dimension. To embed the patches, the ViT uses a linear transformation $E^v \in \mathbb{R}^{P^2 C \times D}$, a learnable token embedding $p_I \in \mathbb{R}^D$, and learnable 1D position embeddings $E^v_{pos} \in \mathbb{R}^{N+1} \times D$. The patches are flattened, multiplied by E^v, prepended with p_I, and added to E^v_{pos}:

$$X^v = [p_I; p_1 E^v; p_2 E^v; ...; p_N E^v] + E^v_{pos}$$

X^v is then passed into an N_v-layer Transformer, yielding the contextualized image representations $H^v = [h^v_I; h^v_1; ...; h^v_N]$.

5.1.2 Language Encoder

M^{3}AE uses BERT's WordPiece to tokenize input text into subword tokens $w_1, ..., w_M$, where each $w_i \in \mathbb{R}^v$ is a one-hot encoding with vocabulary size V. To embed the tokens, M^{3}AE uses a projection matrix $E^l \in \mathbb{R}^{V \times D}$. These embeddings are then prepended with a special start-of-sequence token embedding $w_T \in \mathbb{R}^D$ and appended with a boundary token embedding $w_{SEP} \in \mathbb{R}^D$, and finally added to text position embeddings E^l_{pos}:

$$X^l = [w_T; w_1 E^l; ...; w_M E^l; w_{SEP}] + E^l_{pos}$$

X^l is then passed into an N_l-layer Transformer, yielding the contextualized text representations $H^l = [h^l_T; h^1_T; ...; h^l_M; h^l_{SEP}]$.

5.1.3 Multimodal Fusion Module

M^{3}AE uses two Transformer models with cross-attention layers to fuse the image and text representations. Each Transformer layer consists of a self-attention, cross-attention, and feedforward sub-layer. The attention mechanism is defined with $A(Q, K, V) = \mathrm{softmax}(QK^\intercal) \cdot V$.
Self-attention sub-layers keep attention interactions within modalities:

$$H^{vs} = A(H^v, H^v, H^v)$$
$$H^{ls} = A(H^l, H^l, H^l)$$

The self-attention outputs are then used as input in the cross-attention sub-layers, which drive cross-modal attention interactions:

$$H^{vc} = A(H^{vs}, H^{ls}, H^{ls})$$
$$H^{lc} = A(H^{ls}, H^{vs}, H^{vs})$$

The cross-attention outputs are then passed into the feedforward sub-layer, yielding multimodal representations $Z^v = [z^v_I; z^v_1; ...; z^v_N]$ and $Z^l = [z^l_T; z^l_1; ...; z^l_M; z^l_{SEP}]$. Z^l can be passed through an MLP for language decoding, while Z^v is passed through a Transformer for vision decoding. (Reconstructing images requires a more sophisticated decoder as pixel-space outputs have much lower semantic level than language outputs.)

5.2 Training Objectives

M^{3}AE's vision-language pretraining (VLP) uses masked language modeling (MLM) and masked image modeling (MIM) objectives.
The MLM uses Cross Entropy loss between text output logits $o_{L,i}$ and the MLM labels $\ell_{L,i}$:

$$L_L = -\sum_i \ell_{L,i} \log(o_{L,i})$$

The MIM uses pixel-wise MSE loss between image output logits $o_{I,i}$ and the MIM labels $\ell_{I,i}$:

$$L_I = \frac{1}{N} * \sum_i^N (o_{I,i} - \ell_{I,i})^2$$

5.3 Baseline Implementation

After implementing the M3AE model in code with the pre-trained weights, we performed fine-tuning on VQA-RAD for visual question answering, and achieved similar results to the reported test accuracy.

Table 3: Implemented baseline result on VQA-RAD

Model	Test VQA-RAD Score
M^3AE (reported in baseline paper)	77.01
M^3AE	75.89

Table 4: Hyperparameter Combinations

Batch Size	Learning Rate	Max Text Length	LR Multiplier for Heads	LR Multiplier for Multi-Modals
32	1e-5	32	50	5
64	1e-5	32	50	5
64	5e-6	32	50	5
64	1e-6	32	50	5
64	**1e-6**	**16**	**50**	**5**
64	1e-6	64	50	5
128	5e-6	32	50	5

6 Generative Evaluation Metrics

To evaluate the performance of a generative model versus a classification model, we used BLEU and ROUGE metrics to assess the semantic similarity between model predictions and ground-truth answers, as exact-match accuracy was not applicable.

6.1 BLEU [14]

The BLEU (Bilingual Evaluation Understudy) score measures the precision of n-grams in the predicted text compared with those in the target text. To define the BLEU score, we first define its modified n-gram precision. An n-gram is a sequence of n contiguous words in a sentence. To naively calculate n-gram precision between a predicted sentence and target sentence, we could simply count the number of n-grams in the predicted sentence which also exist in the target sentence. This calculation has two important flaws. Firstly, if the predicted sentence repeats an n-gram many times, its naive precision score may be deceivingly high. To solve this, BLEU clips the maximum match count for each predicted n-gram to the number of appearances of that n-gram in the target, then divides the total number of clipped counts by the total number of candidate words. Next, there may be more than one target sentence. To account for this, BLEU considers all candidate target sentences. To calculate the modified n-gram precision p_n over multiple sentences, BLEU adds the clipped n-gram match counts $Count_{clip}(ngram)$ over all candidate target sentences $C \in \{Candidates\}$ and divides by the total number of n-grams in $Candidates$:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{ngram' \in C'} Count(ngram')}$$

The BLEU score also uses a "best match length" for a predicted sentence, which is defined as the closest length to a candidate target sentence, r is defined by summing the best match lengths of all predicted sentences, and c is defined as the total length of the predicted text. Then

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r)/c} & \text{else} \end{cases}$$

Finally, BLEU computes the geometric average of the modified n-gram precisions p_n, with lengths ranging from 1 to N and positive weights w_n for each length summing to 1:

$$\text{BLEU} = BP * \exp \sum_{n=1}^{N} w_n \log p_n$$

6.2 ROUGE [11]

The ROGUE-N metric measures the number of n-gram matches between the predicted text and target text. While BLEU focuses on precision via its clipped n-gram counting, ROGUE-N measures recall:

$$\text{ROUGE}_\text{N} = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count_{match}(ngram)}{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count(ngram)}$$

Here, $Count_{match}(ngram)$ is simply the number of $ngrams$ in the predicted text which also occur in any candidate target text.

7 Extension: Replace Classification Layer with Generative Output Layer for Improved VQA Answer Quality

Currently, the M^{3}AE model's VQA component uses a classification-based approach, which restricts response generation to a fixed set of answers. To address this constraint, we replaced the classification layer with two different generative layers for testing: a pre-trained, encoder-decoder transformer T5-Small and a custom decoder-only transformer.

7.1 Architecture Update: Integrating T5-Small and Decoder-Only Generative Modules

In the updated architecture, we replace the classification-based output layer of M^{3}AE with a generative decoder layer that leverages either a pre-trained T5 model or a single dedicated decoder layer. We choose T5-Small for its relatively small size for a language model (77M parameters); it consists of 6 encoder and 6 decoder layers, each with 8 attention heads. T5 was developed by Google and pre-trained on a large text corpus, reducing our training requirements when re-purposing it for specialized text output.

Table 5: Added Generative Layer Architecture

Name	Encoder params	Decoder params	n_{layer}	d_{model}	d_{ff}	n_{head}
T5-Small	35M	42M	6	512	2048	8
Decoder-Only	0	103M	6	768	3072	8

Model	**Hidden Layers**	**Parameters/Hyperparameters**
T5-Small	Embedding, 6 x {Self-Attention, LayerNorm, Dropout, Linear, Linear, Dropout, LayerNorm, Dropout}, LayerNorm, Dropout, Embedding, {Self-Attention, LayerNorm, Dropout, Cross-Attention, LayerNorm, Dropout, Linear, Linear, Dropout, LayerNorm, Dropout}, LayerNorm, Dropout, Linaer	Embedding dimension (768), Vocabulary size (50,265), Dropout (p=0.1), Learning Rate (.00001), Batch Size (64)
Decoder-Only	Embedding, Dropout, n × {Multi-head self-attention, Multi-head cross-attention, Linear Layer, Residual Connection, LayerNorm, Dropout}, Linear Layer	Embedding dimension (768), Vocabulary size (30,522), Dropout (p=0.1), Layers (4, 6), Learning Rate (.0001, .00001), Batch Size (16, 32)

The new architecture follows these key steps:

1. **Multi-Modal Embedding Processing**:
    - We retain M^{3}AE's multi-modal encoder, which fuses and embeds information from both image and text inputs. The resulting joint embedding, previously used for classification, will now be projected into the appropriate input size for either T5-Small or the Decoder-Only layer.
2. **Generative Module Integration**:
    - A linear projection layer is added to adjust the dimensionality of the multi-modal embedding to match the input size of the chosen generative model. This projected embedding is then used as the conditioning input for either T5 or the Decoder-Only layer.

7.2 Fine-Tuning Process

To adapt our modified M^{3}AE model to the VQA domain, we fine-tune the generative layers on the VQA-RAD dataset. The fine-tuning process includes the following steps:

1. **Dataset Preparation**:
    - We used the VQA-RAD dataset for fine-tuning, as it provides paired question-answer examples relevant to the medical imaging domain.
    - We used an 80-10-10 train-validation-test split for our ablations
2. **Training Procedure**:
    - During fine-tuning, we freeze the parameters of the M^{3}AE encoder and only train the generative component. The training objective is to minimize the cross-entropy loss between the generated answer and the target answer text.
    - When training the single-layer decoder, we unfreeze all its weights, as it is not initialized with any pre-trained weights. When fine-tuning T5, we unfreeze the final e layers of its encoder and the final d layers of its encoder, and experiment with values of e and d from 2 to 5.
    - The output of M^{3}AE consists of image feature encodings, text feature encodings, and a single "cls" vector containing a pooled and compressed representation of both the image and text. We experiment with using only the image and text features, using only the cls vector, and using all the M^{3}AE outputs as inputs to our generative model.

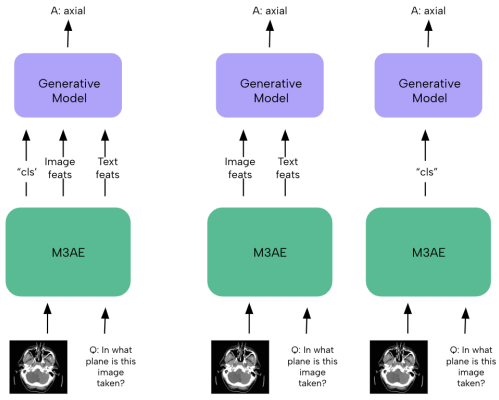

Figure 3: Our generative model inputs methods.

3. **Training Objective**: We use Cross Entropy loss between text output logits $o_{L,i}$ and ground-truth text labels $\ell_{L,i}$:

$$L_L = -\sum_i \ell_{L,i} \log(o_{L,i})$$

8 Results

Our experimental results for training and validation are summarized in Tables 6 and 7, respectively. These experiments compare various generative models applied to VQA task, including the baseline M³AE model, T5-Small, and a decoder-only configuration.

Table 6: Experimental Results on training

Generation Model	M3AE concat img + txt	M3AE pooled img + txt	BLEU Score	ROUGE-1	Loss
M^3AE (baseline)	✓		0.384	0.602	1.061
T5	✓	✓	0.178	0.265	1.598
T5	✓		0.178	0.265	1.057
T5		✓	0.160	0.208	4.495
Decoder only	✓	✓	0.556	0.750	0.891
Decoder only	✓		0.605	0.710	0.974
Decoder only		✓	**0.705**	**0.831**	**0.577**

*The classification loss from the M^3AE baseline is binary cross-entropy loss, while the loss function used for the generative extensions is cross-entropy loss. Because the tasks are different, the loss functions are different, which is why we are using BLEU and ROUGE-1 to compare.

Table 7: Experimental Results on validation

Generation Model	M3AE concat img + txt	M3AE pooled img + txt	BLEU Score	ROUGE-1	Loss
M^3AE (baseline)	✓		0.381	0.520	4.210
T5	✓	✓	0.177	0.275	1.642
T5	✓		0.173	0.265	1.919
T5		✓	0.163	0.200	2.493
Decoder only	✓	✓	0.505	0.644	1.842
Decoder only	✓		0.568	0.615	1.795
Decoder only		✓	**0.755**	**0.697**	**1.825**

*See section 13 for links to WandB graphs for training and validation

8.1 Training Results

The baseline M^3AE model with image and text inputs served as the starting point for comparison. While the T5 configuration underperformed relative to the baseline in BLEU and ROUGE-1 scores, the decoder-only model achieved the best performance overall. Notably, the decoder-only model with pooled image and text inputs attained the highest BLEU score (0.705) and ROUGE-1 score (0.831), alongside a loss of 0.577, highlighting its superior capacity to generate coherent and contextually relevant responses during training.

8.2 Validation Results

The validation results, presented in Table 7, exhibit a consistent trend. The decoder-only model outperformed both the M^3AE baseline and the T5 configurations in BLEU and ROUGE-1 metrics. For example, it achieved a BLEU score of 0.755 and a ROUGE-1 score of 0.697, underscoring its effectiveness in generating meaningful outputs on unseen validation data. These results demonstrate the robustness of the decoder-only approach in handling complex VQA tasks while maintaining strong performance metrics and low loss values.

9 Discussion

Our experiments explored the use of generative models in Medical Visual Question Answering (Med-VQA) by replacing the classification head of the M^3AE model with a generative output layer. This approach sought to address the limitations of fixed-answer classification systems in handling complex and nuanced clinical queries. Below, we discuss key findings, sensitivities, and associated risks.

9.1 Relevance of Key Findings

Generative models demonstrated their ability to produce richer, context-aware answers, addressing the rigidity of classification-based systems. On the VQA-RAD dataset, the classification-based M^3AE model performed better for binary and short-answer questions due to its alignment with the dataset's characteristics. However, it struggled with open-ended and nuanced queries, where generative models displayed significant advantages in flexibility. This is reflected in higher accuracy metrics (the percentage of correct answers) but lower BLEU and ROUGE-1 scores, which weigh token length, meaning short answers contribute less to the total epoch average.

Despite their strengths, the performance of generative models was constrained by the small size and binary-focused nature of VQA-RAD. These findings highlight the need for diverse, large-scale datasets and architectural refinements to bridge the gap between classification-based and generative approaches.

9.2 Analysis

Generative model performance was sensitive to input configurations and dataset characteristics. To integrate M^3AE's output features into the T5 and decoder-only models, a projection layer condensed pooled image and question features into a single token along the sequence length direction. This approach produced suboptimal results for T5—owing to insufficient context and limited impact on pre-trained weights. When the features were concatenated to extend the sequence length, it improved performance slightly, but the T5-Small scores were still worse than baseline overall.

The decoder-only model, being untrained, performed better with M^3AE's multimodal output features because it was able to attend and learn only from the multimodal features. However, its strong performance on VQA-RAD may not generalize well to other datasets, as it only learned VQA-RAD's specific types of images and questions.

Several risks and uncertainties were identified. Overfitting was a significant concern, especially with small datasets like VQA-RAD. Additionally, generative models often faced a trade-off between flexibility and precision, occasionally generating verbose or generic responses that were unsuitable for clinical applications.

10 Future Works

Despite positive results from the decoder-only model, the results, there are still limitations to address to enhance future capabilities.

10.1 Dataset Augmentation

Expanding datasets like VQA-RAD with more examples, particularly those featuring open-ended or long-form answers, would provide a more balanced training set. Data augmentation techniques such as paraphrasing and synthetic question-answer pair generation could further increase diversity.

10.2 Hybrid Model Architectures

Combining classification heads for binary and short-answer questions with generative layers for open-ended tasks could leverage the strengths of both paradigms. Task-specific routing mechanisms could dynamically allocate questions to the appropriate module, optimizing performance.

10.3 Incorporating External Knowledge

Integrating external knowledge bases, such as medical ontologies or structured datasets, could enrich the model's understanding of domain-specific concepts. This would improve handling of specialized terminology and enhance answer accuracy.

By systematically addressing these directions, generative models can achieve improved performance, adaptability, and practical utility in diverse clinical tasks.

11 Conclusion

This study examined the feasibility of generative models for Med-VQA by replacing the classification head of the M^3AE model with a generative layer. Unlike traditional classification systems constrained by predefined answer sets, generative models offer the flexibility to produce nuanced, open-ended responses better suited to real-world clinical scenarios.

Our findings highlight both the strengths and challenges of generative approaches. Generative models demonstrated advantages in handling open-ended queries, as evidenced by the superior performance of the decoder-only model compared to the baseline M^3AE. However, their performance on smaller datasets like VQA-RAD was limited by early saturation and vocabulary constraints. These results underscore the importance of large, diverse datasets and domain-specific pre-training for enhancing performance and adaptability.

Our experiments identified critical areas for improvement, such as dataset augmentation, hybrid architectures, and external knowledge integration. These enhancements hold the potential to unlock greater adaptability and reasoning capabilities in generative Med-VQA systems.

In summary, this work represents a significant step toward flexible and context-aware Med-VQA systems. By addressing the identified limitations, future research can refine generative approaches to better meet the demands of complex clinical applications.

12 Code and WandB Runs

M3AE + Generative Layer Code `https://github.com/vdurham/MM-VQA-Healthcare`

Decoder-Only Generative Layer Training
`https://wandb.ai/soomin-chung-9910-seoul-national-university/VQA-RAD-Decoder-Only?nw=nwuservchoidurham`

T5-Small Generative Layer Training
`https://wandb.ai/soomin-chung-9910-seoul-national-university/VQA-RAD-T5_1209?nw=nwusersoominchung9910`
`https://wandb.ai/soomin-chung-9910-seoul-national-university/VQA-RAD-T5?nw=nwusersoominchung9910`

13 Team Contributions

13.1 Virginia Choi-Durham

Extension Proposal and Generative Architecture Ideas, Writing Support, Dataset Evaluation, Decoder-Only Code and Training, Eval Metric Code

13.2 Naeun Lee

Hyperparameter Experimentation, Writing Support, Code

13.3 Soomin Chung

Model Analysis, Writing Support, Experimentation, Dataset Download and Preparation, Code

13.4 Michael Chen

Reformatted Data for Inference, Writing Support, Evaluation Metric Research, Code

References

[1] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images, 2023.

[2] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023.

[3] Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. Miss: A generative pretraining and finetuning approach for med-vqa, 2024.

[4] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training, 2022.

[5] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest X-ray report generation from longitudinal representations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4891–4904, Singapore, December 2023. Association for Computational Linguistics.

[6] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. Recap: Towards precise radiology report generation via dynamic disease progression reasoning, 2023.

[7] Xiaofei Huang and Hongfang Gong. A dual-attention learning network with word and sentence embedding for medical visual question answering, 2022.

[8] Qiu hui Chen, Qiang Fu, Hao Bai, and Yi Hong. Longformer: Longitudinal transformer for alzheimer's disease classification with structural mris. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3563–3572, 2023.

[9] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa, 2021.

[10] Gayen S. Ben Abacha A. et al. Lau, J. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5, 2018.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[12] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023.

[13] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[15] Santosh Sanjeev, Fadillah Adamsyah Maani, Arsen Abzhanov, Vijay Ram Papineni, Ibrahim Almakky, Bartłomiej W. Papież, and Mohammad Yaqub. Tibix: Leveraging temporal information for bidirectional x-ray and report generation, 2024.

[16] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q O'Neil. Controllable chest x-ray report generation from longitudinal representations, 2023.

[17] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

[18] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 451–466. Springer, 2016.

[19] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.