ChestVQATrans: A Transformer-Based Approach for Medical VQA in Radiology

Rao Faizan
Department of Computer Science and Engineering
Kyung Hee University (Global Campus)
Yongin-si, Republic of Korea
rao.faizan@khu.ac.kr

Sungyoung Lee
Department of Computer Science and Engineering
Kyung Hee University (Global Campus)
Yongin-si, Republic of Korea
sylee0301@gmail.com

Seong Tae Kim
Department of Computer Science and Engineering
Kyung Hee University (Global Campus)
Yongin-si, Republic of Korea
st.kim@khu.ac.kr

Abstract

Visual Question Answering (VQA) in the medical domain is a challenging yet critical task that combines image analysis and natural language understanding to address clinical queries effectively. It requires models to comprehend complex medical imagery, such as chest X-rays, while interpreting diverse textual questions that often demand domain-specific reasoning. This paper investigates the potential of fine-tuning the Vision-and-Language Transformer (ViLT) model for medical VQA tasks by leveraging a unified dataset combining MIMIC-CXR and MIMIC VQA. By adopting a full fine-tuning approach, our model achieves a competitive accuracy of 71.57%, underscoring the advantages of tailoring pre-trained transformers to domain specific data. Our study meticulously outlines the steps for preparing the unified dataset, including the integration of radiological images and question-answer pairs. We explore effective strategies for optimizing the ViLT model for medical VQA, emphasizing the importance of domain alignment in both training and evaluation. The results demonstrate that full fine-tuning enables the model to capture nuanced visual and textual correlations, setting a robust benchmark for chest X-ray VQA tasks. Additionally, we analyze the challenges posed by medical VQA, such as question complexity, variability in image quality, and the need for clinically accurate reasoning, providing valuable insights for future research in this domain.

CCS Concepts

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision tasks; Scene understanding.

Keywords

Visual Question Answering (VQA), Vision-and-Language Transformer (ViLT), Chest X-ray Interpretation

ACM Reference Format:
Rao Faizan, Sungyoung Lee, and Seong Tae Kim. 2025. ChestVQATrans: A Transformer-Based Approach for Medical VQA in Radiology. In *2025 8th Artificial Intelligence and Cloud Computing Conference (AICCC 2025), December 20–22, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3789982.3790004

AICCC 2025, Tokyo, Japan

ACM ISBN 979-8-4007-1889-2/2025/12
https://doi.org/10.1145/3789982.3790004

1 INTRODUCTION

Visual Question Answering (VQA) represents a frontier at the intersection of computer vision and natural language processing. It poses a unique challenge by requiring systems to comprehend visual content and interpret textual questions to provide accurate answers. In the medical domain, this challenge is further amplified by the complexity and criticality of medical images and domain-specific knowledge required to interpret them. Unlike general purpose VQA, where models are trained on datasets with everyday objects and scenes, medical VQA demands systems to reason over specialized datasets, such as chest X rays, to answer clinically relevant questions. With the rise of large multimodal models, the need to align vision-language representation with domain-specific semantics has become increasingly vital. In this context, VQA in radiology represents a practical use-case where AI tools can directly support clinical workflows. As hospitals transition toward AI-assisted diagnostics, integrating explainable, image-grounded question answering systems becomes not only desirable but essential for patient safety and transparency. The importance of VQA in healthcare is immense. Chest X-rays, being one of the most common imaging modalities in medicine, are extensively used for diagnosing a wide range of conditions, including pneumonia, heart failure, and lung cancer. Automating the process of answering clinical questions related to chest X-rays can significantly aid radiologists, enhancing diagnostic workflows, reducing errors, and ensuring timely decision-making. For instance, a VQA system could answer questions such as "Is there evidence of pleural effusion?" or "What are the findings in the right lung?" with explanations derived from the visual content of the X-ray and the associated textual information in medical reports. The contributions of this paper are threefold:

- **Dataset Integration and Preparation:** We present a unified dataset that combines MIMIC-CXR [1] and MIMIC-VQA [2], detailing the preprocessing steps, question categorization, and answer encoding strategies.
- **Model Fine-Tuning and Optimization:** We describe the methodology for full fine-tuning of the ViLT model [9], including architectural adaptations, hyper parameter tuning, and training protocols.

- **Evaluation and Analysis:** We provide a thorough evaluation of the model's performance, highlighting its strengths and limitations through quantitative metric.

2 RELATED WORK

Visual Question Answering (VQA) in the medical domain has gained significant attention in recent years as an intersection of computer vision, natural language processing, and medical informatics. This section provides an overview of the foundational efforts and recent advancements in medical VQA, with a specific emphasis on chest X-ray datasets and performance evaluation.

2.1 General VQA Frameworks

The development of VQA systems has been spearheaded by efforts in general-purpose datasets like VQA 1.0 [3], VQA 2.0 [4], and CLEVR [5], which introduced challenges requiring deep understanding of visual and textual data. Early approaches primarily used convolutional neural networks (CNNs) to extract image features, which were then incorporated into question embeddings. [6]. [7] introduced a visual attention–based encoder–decoder framework that dynamically focuses on salient image regions during text generation, demonstrating the effectiveness of attention mechanisms for aligning visual representations with natural language descriptions in image understanding tasks. The success of these methods in generic domains paved the way for domain-specific adaptations, including medical imaging. However, these models often lack fine-grained alignment at the level of pixel-wise attention to text queries. Few works explicitly address the synthesis of structured question answering pipelines within chest X-ray interpretation, leaving a gap that our study seeks to address through a ViLT-centric fine-tuning approach on a unified dataset. Additionally, recent studies have explored knowledge graph completion methods [10] as a complementary strategy to enrich the knowledge base for VQA systems.

3 DATASET FORMULATION

The unified dataset was constructed by integrating question–answer pairs from MIMIC-CXR and MIMIC-VQA to create a consistent chest-X-ray visual question answering corpus. Duplicate instances were identified and removed using patient, study, and image identifiers, ensuring each record represents a unique image–question pair. Question text was normalized through lower-casing and stop-word removal, and ambiguous or incomplete entries were excluded. To mitigate potential dataset bias such as the predominance of yes/no questions and uneven label distribution stratified sampling was applied during training, preserving the diversity of question types and anatomical focus. The final unified dataset comprises 142,778 samples. This unified formulation provides a standardized foundation for reproducible medical VQA research and ensures fair evaluation across both visual and textual modalities.

3.1 Data Acquisition

The dataset was sourced from publicly available MIMIC-CXR and MIMIC-VQA repositories, providing a rich blend of chest X-ray images and corresponding question-answer pairs.

Workflow for dataset formulation as shown in Figure 1. The dataset metadata included essential fields such as:

- **Subject ID, Study ID, and Image ID:** Unique identifiers for each patient and associated image.
- **Question:** Natural language queries related to the image, focusing on clinical insights.
- **Answer:** The corresponding response to the question, derived from radiological annotations or text reports.
- **Split:** Indicates whether the sample belongs to the training, validation, or test set.

The unified dataset was split into training, validation, and test sets using an **80:10:10** ratio to ensure balanced model development and evaluation. Table 1 summarizes the distribution of images, questions, and answers across each split. While every sample includes both an image and a question, the answer count is slightly lower due to quality filtering.

During dataset unification, overlapping or duplicate entries between MIMIC-CXR and MIMIC-VQA were carefully resolved by aligning on patient ID and study ID. Class imbalance (e.g., dominance of yes/no questions) was mitigated using stratified sampling during training.

4 PROPOSED METHODOLOGY

The proposed methodology as shown in the Figure 2 involves developing a VQA model based on the Vision-and-Language Transformer (ViLT). This section outlines the model architecture, training procedure, and evaluation strategy.

4.1 Model Architecture

The framework accepts two inputs: a chest X-ray image and a corresponding natural language question. The image is processed through a vision transformer branch that generates rich visual embeddings, while the question is encoded using a text transformer to obtain contextual word features. These visual and textual representations are fused through a multimodal interaction module employing cross-attention to learn semantic alignment between image regions and question tokens. The resulting joint representation is then passed to a classification head that predicts the most probable answer from a predefined answer space. This unified architecture enables end-to-end reasoning over medical images and textual queries, facilitating accurate and interpretable decision-making in radiology VQA tasks.

The ViLT model was selected for its capability to integrate visual and textual information effectively. The model was initialized with a pre-trained checkpoint, dandelin/vilt-b32-mlm, and further fine-tuned for the VQA task. The model configuration was updated to accommodate the number of unique answers in the dataset by setting the num.labels parameter in the configuration. We also introduced a class-balancing sampler during training to mitigate the skewed distribution of frequent versus rare answers. This step improved the model's generalization on minority classes, particularly on medical conditions that are less prevalent but clinically significant. The loss function was enhanced with label smoothing ($\epsilon = 0.1$), which reduced overconfidence in model predictions.

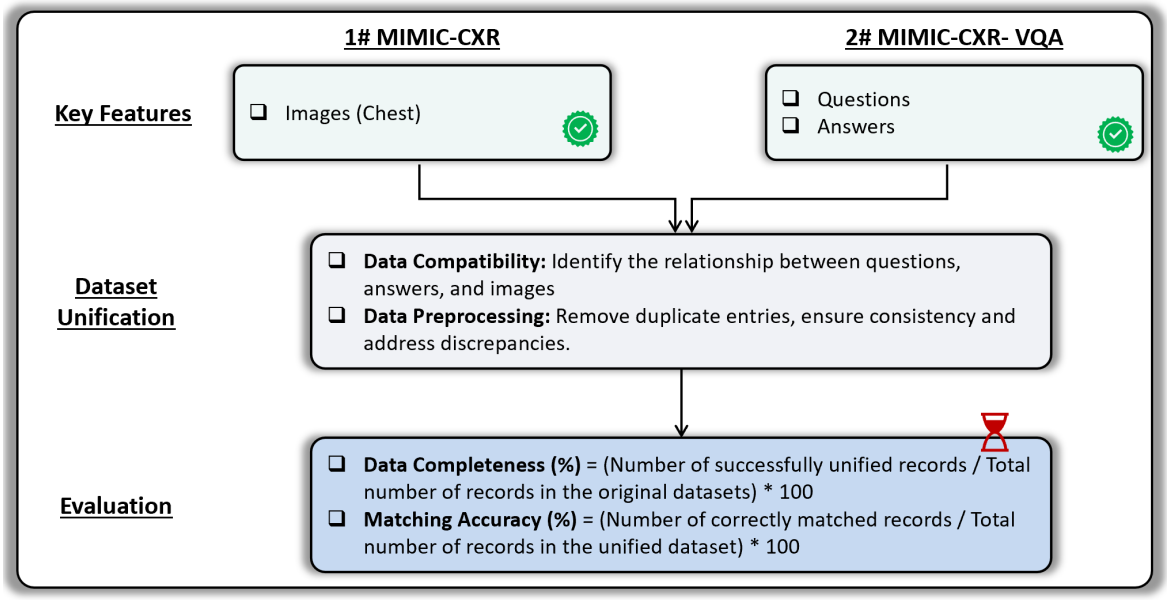

Figure 1: Workflow for unified medical dataset generation: Key features from MIMIC-CXR (chest images) and MIMIC-CXR-VQA (questions, answers) are systematically merged through compatibility assessment and preprocessing. Evaluation includes measures of data complete.

Table 1: Summary of unified dataset statistics: Distribution of image, question, and answer counts for training, validation, and test splits, highlighting answer coverage and sample allocation across each subset.

	Train	Valid	Test	Total # of Samples
Images	114,222	14,277	14,277	142,778
Questions	134,796	16,849	16,849	168,496
Answers	127,440	15,930	15,930	159,300

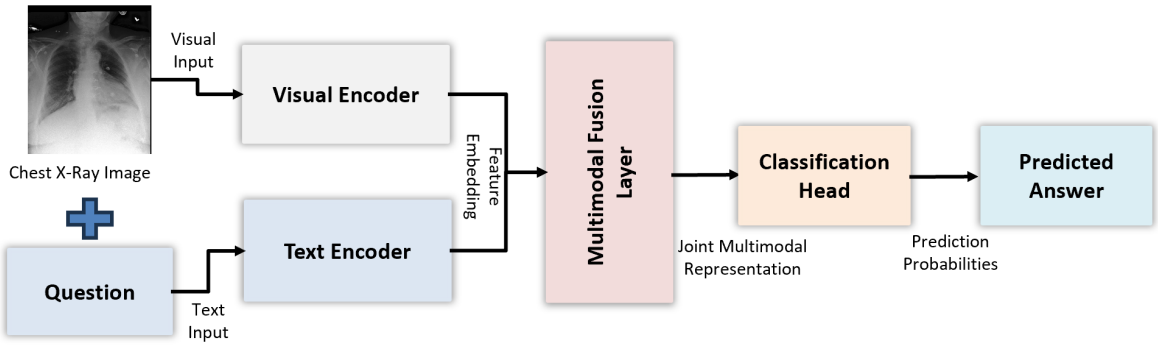

Figure 2: Architecture of the proposed unified ViLT-based chest X-ray VQA framework. The model jointly encodes visual and textual inputs through transformer-based encoders, fuses their embeddings via a multimodal attention mechanism, and predicts the final answer through a classification layer. The output block represents the model's predicted response to the input question based on the given chest X-ray image.

Table 2: Performance overview of VQA models on chest X-ray datasets. Our model is evaluated on the unified dataset. Note that datasets differ in composition, thus, results serve to contextualize the proposed approach rather than establish direct comparison.

Model	Dataset	Accuracy
PubMedCLIP [8]	MIMIC-CXR-VQA	54.9%
MedViLL [8]	MIMIC-CXR-VQA	63.6%
Transformer-based VQA [8]	MIMIC-CXR-VQA	68.8%
M3AE [8]	MIMIC-CXR-VQA	69.2%
Ours (ViLT-FT)	Unified (CXR+VQA)	71.57%

4.1.1 Label Mapping. To enable supervised learning, each unique answer in the dataset was mapped to a numeric identifier. This mapping was achieved by constructing label2id and id2label dictionaries, where:

- label2id maps each answer to a unique integer.
- id2label maps each integer back to its corresponding answer.

These mappings were integrated into the ViLT model configuration, allowing the model to predict discrete answer classes. With over 100 unique answers, this multi class classification task required careful handling of label distribution to ensure balanced learning.

4.1.2 Image Preprocessing. Given the high resolution and varying sizes of chest X ray images, preprocessing was essential to standardize inputs. Images were resized to a fixed resolution of 384×384 pixels using the Pillow library.

- Each image was converted to RGB format to match the ViLT model's input requirements.
- Invalid or inaccessible images were identified and excluded, ensuring only high-quality data was used.

4.1.3 Dataset Splitting. To facilitate training and evaluation, the dataset was divided into three subsets:

- **Training Set:** Used to optimize the model's parameters, containing approximately 80% of the data.
- **Validation Set:** Used to monitor performance during training and prevent overfitting, accounting for 10% of the data.
- **Test Set:** Reserved for final performance evaluation, comprising 10% of the data.

5 EXPERIMENT & RESULTS

5.1 Experimental Setup

The model was trained using the AdamW optimizer with a learning rate of $\mathbf{5 \times 10^{-5}}$. Training was performed for 60 epochs with a batch size of 32. The loss function used was the cross-entropy loss, calculated between the predicted logits and the ground truth labels.

- **Training Loop:** During each epoch, the model's parameters were updated using backpropagation. The training accuracy and loss were computed after processing each batch.
- **Validation Loop:** After each epoch, the model was evaluated on the validation dataset. This evaluation provided insights into the model's generalization performance.

5.2 Implementation Details

The training and validation pipelines were implemented using PyTorch. The training process was conducted on a GPU to leverage accelerated computation. The model was saved after each epoch, allowing for resumption in case of interruption.

5.3 Evaluation Metric and Performance Benchmarks

The proposed ViLT-based model was evaluated against existing chest X-ray VQA baselines. Table 2 summarizes accuracy results alongside published baselines for contextualization. It is important to note that direct numerical comparison across rows should be interpreted with caution, as the models were trained and evaluated on datasets that differ in scale, composition, and question distribution. Accuracy is a standard metric for VQA tasks, measuring the proportion of correctly predicted answers [3]. The baseline results PubMedCLIP (54.9%), MedViLL (63.6%), M3AE (69.2%), and Transformer-based VQA (68.8%) are reported from [8] on the MIMIC-CXR-VQA dataset. Our ViLT-based model, fine-tuned on the unified dataset combining MIMIC-CXR and MIMIC-VQA (168,496 question-answer pairs), achieves 71.57% accuracy. This result demonstrates the effectiveness of full fine-tuning of pre-trained vision-language transformers on a comprehensive medical VQA corpus.

5.4 Challenges and Future Directions

Despite recent advancements, medical VQA faces persistent challenges. The lack of large-scale annotated datasets limits the generalizability of trained models. Additionally, the interpretability of VQA models is crucial for clinical adoption. Efforts to integrate explain ability modules into transformer architectures are ongoing.

5.5 Hyper-parameter Optimization and Ablation Studies

The meticulous selection of optimal hyperparameters was crucial for maximizing the performance of our fine-tuned ViLT model. We systematically explored a range of learning rates, batch sizes, and optimization algorithms, ultimately settling on the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 32, determined through iterative experimentation and validation performance monitoring. The unified dataset design, combining MIMIC-CXR and MIMIC-VQA, was motivated by the hypothesis that a larger and more diverse training corpus would improve the model's

ability to generalize across different question types and image conditions. Future work will include formal ablation studies to quantify the individual contributions of each dataset component and the impact of specific design choices on model performance.

6 DISCUSSION

The experimental outcomes affirm the viability of leveraging pre-trained vision-language transformers like ViLT in specialized domains such as medical imaging. Notably, the full fine-tuning approach significantly outperformed previous methods, suggesting that domain adaptation at all transformer layers is critical for nuanced visual-textual reasoning. However, the model occasionally struggled with ambiguous or compound questions, often defaulting to dominant answer patterns from training data. This behavior underscores the need for more diverse annotations and balanced question distributions. Another limitation lies in the reliance on single-frame chest X-rays without temporal or 3D contextual information, which may hinder performance on time-sensitive diagnoses. Future enhancements could explore self-supervised pre-training using masked language modeling on radiology reports or integrating multi-view imaging data. Additionally, incorporating clinician feedback into the model's reasoning loop may foster explainability and trustworthiness, facilitating safer clinical integration.

7 CONCLUSION

In this study, we have successfully full fine-tune Vision and-Language Transformer (ViLT) model, focusing on a unified dataset comprising MIMIC-CXR and MIMIC VQA. Our approach aimed to enhance the interpretability and accuracy of VQA tasks in the healthcare domain, specifically leveraging multimodal data sources for improved clinical insights. The dataset preparation phase involved rigorous preprocessing techniques, including the removal of outliers, standardization of image dimensions, and the creation of structured mappings for labels. By ensuring a balanced and representative dataset split for training, validation, and testing, we achieved a robust setup that mitigated overfitting and enhanced the generalization capabilities of our model. The integration of image and question modalities through ViLT's multi-modal transformer layers enabled us to leverage contextual relationships between visual and textual data effectively. The proposed system establishes a strong foundation for advancing transformer-based medical VQA research and its potential integration into real-world radiology workflows.

Acknowledgments

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) (Explainable Logical Reasoning for Medical Knowledge Generation) under Grant IITP-2022-0-00078; in part by the Information Technology Research Center (ITRC) Support Program Supervised by the IITP under Grant RS-2023-00259004; in part by the Ministry of Science and Information and Communications Technology (MSIT), South Korea, under the Grant Information Technology Research Center Support Program under Grant IITP-2024-2020-0-01489; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) under Grant 2020-0-00004.

References

[1] Alistair EW Johnson, Tom J Pollard, Samuel J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chuanhong Deng, Roger G Mark, and StevenHorng. Mimic-cxr: A large publicly available database of labeled chest radiographs. Nature Scientific Data, 6(1):1–12, 2019.

[2] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, *et al.* Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2425–2433, 2015.

[4] Yash Goyal, Tejas Khot, Daniel Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6904–6913, 2017.

[5] Johnson, Justin, *et al.* "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander Smola. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 21–29, 2016.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning (ICML), pages 2048–2057, 2015.

[8] Bae, Seongsu, *et al.* "Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images." Advances in Neural Information Processing Systems 36 (2023): 3867-3880.

[9] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision and-language transformer without convolution or region supervision. Proceedings of the International Conference on Machine Learning (ICML), pages 5583–5594, 2021.

[10] Nguyen, T. D. T., Rehman, U. U., Hussain, M., Rao, F., Hussain, J., Bae, S.-H., Kim, J. U., Kim, S. T., & Lee, S. (2025). Unified link prediction modeling for enhanced knowledge graph completion task. Expert Systems with Applications.