Journal of Computer Science

Research Article

MedFusion: A Unified Multimodal Framework for Visual Question Answering and Explainable Medical Recommendation

Satyajit Mahapatra[1], Jibitesh Mishra[1], Kumar Janardan Patra[1], Sanjit Kumar Dash[1] and Aliazar Deneke Deferisha[2]

[1]*School of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, India*
[2]*Faculty of Computing and Software Engineering, Arba Minch University, Arba Minch, Ethiopia*

Article history
Received: 07-07-2025
Revised: 28-08-2025
Accepted: 02-09-2025

Corresponding Author:
Kumar Janardan Patra
School of Computer Sciences, Odisha University of Technology and Research, Bhubaneswar, Odisha, India
Email: janardanpatra1997@gmail.com

Abstract: In clinical decision-making, the ability to ask visual questions about medical images and receive accurate, personalized, and interpretable recommendations can significantly enhance practitioner support systems. This paper presents MedFusion, a unified multimodal framework that integrates Visual Question Answering (VQA), personalized medical recommendation, and explainability within a single architecture. The proposed model employs co-attention–based visual–textual fusion augmented with retrieval-enhanced reasoning to improve answer grounding, while personalized recommendations are generated using a shared multimodal representation supported by GAN-guided feature augmentation. To enhance transparency, the framework provides attention-based heatmaps and natural-language rationales for both answers and recommendations. Extensive experiments on VQA-RAD, EHRXQA, and Med-RecX demonstrate that MedFusion outperforms state-of-the-art medical VQA and recommendation baselines, achieving a 7.4% improvement in VQA accuracy, reducing RMSE to 0.91, and improving human-rated interpretability to 4.5/5. Ablation studies confirm the effectiveness of retrieval augmentation, GAN-guided enhancement, and joint multi-task learning. These results indicate that MedFusion offers a robust and explainable decision-support solution, advancing the deployment of trustworthy, user-adaptive AI systems in real-world healthcare environments.

Keywords: Multimodal Learning, VQA, Medical Recommendation, XAI, Co-Attention, Retrieval-Augmented Reasoning, CGAN, Healthcare Informatics

Introduction

In recent years, the integration of Artificial Intelligence (AI) into healthcare has yielded significant advancements in diagnostic support, clinical decision-making, and medical image interpretation. However, most existing models are designed to solve isolated tasks, either answering medical questions based on imaging (visual question answering or VQA) or providing recommendations for treatment or next steps based on structured clinical data. In real-world scenarios, clinicians often seek both: they pose image-based questions (e.g., "What does this CT scan show?") and expect evidence-grounded recommendations (e.g., "Should this patient undergo further imaging or a biopsy?") (Kuanr et al., 2022). Addressing these dual objectives within a unified framework holds immense potential for improving clinical efficiency, reducing diagnostic errors, and enhancing patient outcomes.

The task we address in this study is a novel fusion of multimodal VQA and clinical recommendation generation. We define the problem as follows: given a

medical image chest X-ray and a natural language question from a clinician, the system must generate a clinically accurate answer and suggest an appropriate recommendation based on visual findings (Zhang et al., 2021). For example, the model should be able to answer "Yes, there is consolidation in the right lung" and follow it up with "Recommend a follow-up CT scan and antibiotic therapy." This formulation transforms a passive diagnostic tool into an active, intelligent assistant capable of visual reasoning and decision support, crucial in time-sensitive or resource-constrained healthcare settings (Guzzoni, 2007).

To achieve this, we introduce a novel architecture that jointly learns to perform VQA and recommendation tasks through a shared visual-textual representation (Zakari et al., 2022). Our approach leverages a multimodal co-attention mechanism that enables the model to simultaneously attend to key visual features from medical images and relevant semantic components from clinical questions. This fusion empowers the model to capture rich interdependencies between modality-specific cues, ensuring that both the answer and recommendation are contextually aligned with the visual input and clinical intent (Liang et al., 2021). Beyond traditional VQA and recommendation models, our framework incorporates a self-reflective, retrieval-augmented reasoning component. This module dynamically accesses external clinical knowledge sources, such as guidelines or similar past cases, during inference (Tang et al., 2024). By grounding its outputs in evidence, the model can avoid hallucinations and offer justifiable answers, an essential feature in high-stakes environments like healthcare. Furthermore, this retrieval mechanism allows the system to adapt to novel queries or unseen imaging patterns by drawing from a broader context of clinical knowledge.

In addition to generating answers and recommendations, our system enhances interpretability by producing natural-language explanations that describe the reasoning process behind each prediction. These explanations are enriched with visual-textual references, highlighting the image regions and textual cues that influenced the decision (Sungur and Bakal, 2025). Such explainability not only increases clinician trust but also facilitates human-AI collaboration, as medical professionals can review and validate the model's thought process. Inspired by explainable AI (XAI) paradigms, we aim to make the system's internal logic transparent and clinically interpretable.

Based on these motivations, this study addresses the following research questions: (RQ1) Can a unified multimodal architecture jointly perform medical visual question answering and personalized recommendation more effectively than task-specific models? (RQ2) Does retrieval-augmented reasoning improve answer grounding and recommendation relevance in medical image-based decision support? (RQ3) Can integrated visual and textual explanations enhance interpretability and clinician trust without compromising predictive performance?

The contributions of this work are threefold. First, we propose a novel end-to-end architecture that unifies medical VQA and recommendation generation using a co-attention-based multimodal encoder (Ma et al., 2019). This joint formulation allows the model to leverage shared features for tasks, improving performance and efficiency. Second, we introduce a retrieval-augmented reasoning mechanism that incorporates relevant clinical context during both training and inference, thereby improving the accuracy and trustworthiness of the outputs. Third, we present a multi-modal explanation module that provides coherent, human-understandable justifications for each answer and recommendation, grounded in both the visual and textual domains.

Literature Review

Recent advancements in medical visual question answering (VQA) have aimed to enable AI systems to interpret medical images and respond to natural language queries with clinically meaningful insights. Initial benchmark efforts, such as VQA-RAD and VQA-Med, laid the groundwork for this domain by providing curated datasets of medical images and paired questions (Osborn and Wustmann, 2018). Building on these, MMBERT introduced a multimodal BERT-style architecture trained on paired medical images and captions, demonstrating state-of-the-art performance on VQA-Med tasks through attention-based interpretability (Khare et al., 2021). Further enhancing this direction, UnICLAM employed contrastive learning with adversarial masking to unify visual and textual representations, improving robustness across datasets like SLAKE and VQA-RAD (Shrestha et al., 2023). Another recent approach, the Diff-VQA framework, integrated domain knowledge and difference-aware modeling by capturing radiological progression across temporal image pairs, thereby improving diagnostic QA performance in clinical workflows (Cho et al., 2024).

Beyond traditional fusion techniques, retrieval-augmented models have emerged as a key innovation in VQA. PubMedCLIP models adapted the CLIP encoder for medical vision-language tasks, improving domain grounding by pretraining on biomedical literature and imaging pairs (Monajatipoor et al., 2024). Additionally, retrieval-enhanced reasoning has been successfully applied in surgical VQA, where transformer-based encoders like VisualBERT were adapted to interpret endoscopic scenes and generate procedural responses (Seenivasan et al., 2022). These models show that retrieval not only boosts factual correctness but also encourages richer semantic alignment between questions

and visual evidence. Efforts have also been made to incorporate multimodal fusion in clinical decision support tasks. MuVAM and MedFuseNet introduced attention mechanisms that combined multi-view imaging and patient metadata to enhance QA and reasoning (Sharma et al., 2021). More recently, EHRXQA proposed a unified dataset and model framework combining chest X-rays, electronic health records (EHR), and question-answer pairs, representing a critical step toward real-world, patient-centered medical decision systems (Bae et al., 2023). These works highlight the growing demand for systems that move beyond pure QA to provide actionable recommendations grounded in multimodal clinical data.

On the recommendation side, explainable decision support systems are gaining traction in healthcare AI. Biomedical QA surveys have pointed to the utility of sparse retrieval methods such as BM25 and dense passage retrieval (DPR) to integrate structured medical knowledge into QA pipelines (Yang et al., 2024). This aligns closely with the idea of recommendation generation from external sources, a technique we adopt in our framework to improve generalization and explainability. Reinforcement-enhanced visual-language models (VLMs) also offer promise for medical reasoning, with recent work showing that chain-of-thought prompting and reinforcement learning can significantly improve the logical consistency of model outputs (Khare et al., 2021). In parallel, systems designed as "radiologist copilots" have introduced multi-step reasoning and distractor filtering as methods to align AI recommendations with clinical guidelines (Wu et al., 2025).

Explainability remains a central concern in clinical AI. Inspired by fashion-based recommendation models like NOR, which produce both outfit recommendations and human-readable explanations our model also includes a natural language generation component that explains both the answer and the associated clinical recommendation (Tepe and Emekli, 2024). Finally, work on clinical knowledge-graph–based recommendation has shown the feasibility of generating semi-structured explanations by tracing inference paths through medical ontologies such as SNOMED CT or UMLS (Lin et al., 2019).

Despite these advances, there is still no unified system that combines medical image-based VQA, recommendation generation, and explainable reasoning in one end-to-end framework. Our work seeks to bridge this gap by proposing a joint architecture that not only answers clinical questions from images but also provides actionable, explainable recommendations grounded in both visual and textual knowledge. This approach offers significant potential for deployment in diagnostic decision support systems, particularly in radiology, pathology, and primary care.

Methodology

The proposed MedFusion architecture follows a shared-encoder, multi-task learning paradigm. Medical images are processed through a hybrid visual encoder, while clinical questions are encoded using a transformer-based language model. The resulting embeddings are fused using a co-attention mechanism, producing a joint multimodal representation. This shared representation is simultaneously fed into (i) a VQA answering head and (ii) a recommendation prediction head, enabling joint optimization across tasks.

Algorithm 1 MedFusion: Unified Multimodal VQA–Recommendation–Explainability Framework

Require: $\mathcal{I} = \{I_i\}_{i=1}^{N}$, $\mathcal{Q} = \{Q_i\}_{i=1}^{N}$, $\mathcal{K}$, Θ
Ensure: $\hat{A}, \hat{R}, \mathcal{E}$
Initialize $\Theta_v, \Theta_t, \Theta_f, \Theta_a, \Theta_r, \Theta_g$
$e \leftarrow 0$
while $e < E$ **do**
  for $i = 1$ to N **do**
    $V_i \leftarrow f_v(I_i; \Theta_v)$
    $T_i \leftarrow f_t(Q_i; \Theta_t)$
    $Z_i \leftarrow f_c(V_i, T_i; \Theta_f)$
    $\mathcal{K}_i \leftarrow \text{Retrieve}(T_i, V_i \mid \mathcal{K})$
    $Z_i^* \leftarrow Z_i \oplus \mathcal{K}_i$
    $\hat{A}_i \leftarrow f_a(Z_i^*; \Theta_a)$
    $\hat{R}_i \leftarrow f_r(Z_i^*; \Theta_r)$
    $\tilde{I}_i \leftarrow G(z_i, Z_i^*; \Theta_g)$
    $\mathcal{E}_i \leftarrow \{\text{Attn}(Z_i^*), \text{NLG}(Z_i^*)\}$
    $\mathcal{L}_i \leftarrow \lambda_1 \mathcal{L}_{vqa}(\hat{A}_i, A_i) + \lambda_2 \mathcal{L}_{rec}(\hat{R}_i, R_i)$
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_i$
  end for
  $e \leftarrow e + 1$
end while
  return $\{\hat{A}, \hat{R}, \mathcal{E}\}$

Symbol	Description
I_i	Input medical image
Q_i	Clinical question
A_i	Ground-truth VQA answer
R_i	Ground-truth recommendation
$\mathcal{K}$	External medical knowledge base
$f_v(\cdot)$	Visual encoder (CNN + ViT)
$f_t(\cdot)$	Text encoder (SBERT)
$f_c(\cdot)$	Co-attention fusion module
Z_i	Multimodal fused representation
$\hat{A}_i$	Predicted VQA answer
$\hat{R}_i$	Predicted recommendation
$\mathcal{E}_i$	Explainability output
$\mathcal{L}_{vqa}$	Cross-entropy loss
$\mathcal{L}_{rec}$	MSE loss
Θ	Trainable parameters

The success of a multimodal medical question answering and recommendation system depends significantly on the ability to represent visual and textual data in a compatible and information-rich form. Our framework integrates three core components to achieve this: a vision encoder for medical images, a language encoder for clinical questions, and a fusion mechanism to align both modalities meaningfully. The complete workflow of the proposed framework is illustrated in Fig. 1.

To process medical images, we utilize a visual encoder based on either convolutional neural networks, ResNet-50, integrated with transformer models such as the Vision Transformer (ViT) (Islam et al., 2022), as shown in Figure 2. These models allow us to extract detailed visual features from complex medical images, such as chest X-rays. In our setting, ViT divides each image into fixed-size patches and encodes them as a sequence of tokens, maintaining spatial information critical for diagnosis. For images with evident anatomical regions, ResNet-based models help capture hierarchical features through convolutional layers (Mezina and Burget, 2024). These visual embeddings are later aligned with language tokens for joint interpretation.

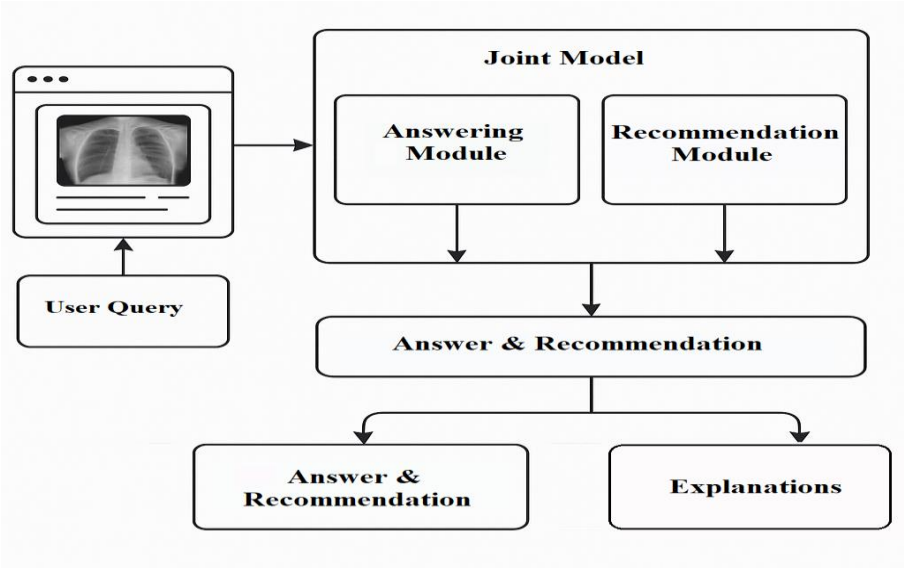


Fig. 1: Proposed MedFusion multimodal framework

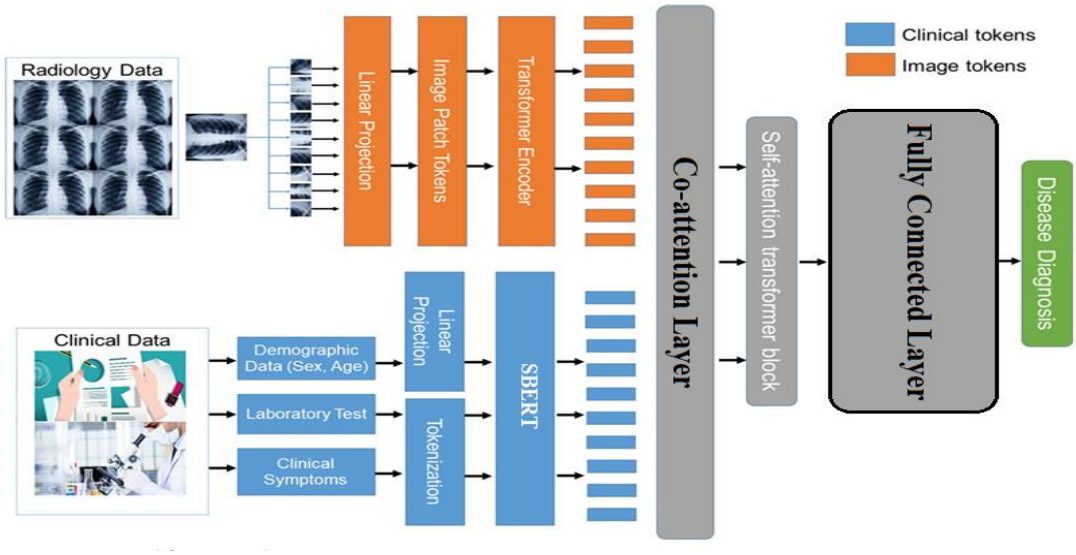


Fig. 2: Multimodal data integration

On the textual side, clinical questions are encoded using transformer-based models, Sentence-BERT (SBERT). (Piao, 2021). These models provide context-aware embeddings of input questions, capturing both syntactic and semantic nuances. In particular, BERT's deep bidirectional architecture enables effective understanding of clinical terms, abbreviations, and domain-specific syntax common in radiological and pathological inquiries.

To bridge the gap between image and text modalities, we adopt a co-attention mechanism inspired by Modular Co-Attention Networks (MCAN) (Yu et al., 2019). This architecture employs a multi-level attention structure, where each modality first undergoes self-attention to reinforce intra-modal context. This layered design enables more accurate localization and correlation of question semantics with visual evidence.

Data Collection

To support our experimental setup, we utilize two prominent datasets. The Medical VQA Dataset from Medical-CXR-VQA (PhysioNet) includes chest radiographs along with associated clinical questions and answers. It offers over 100,000 QA pairs, making it suitable for training and evaluating deep visual question answering models in the medical domain. For a more comprehensive multimodal setup, we also incorporate the EHRXQA dataset (GitHub), which combines radiographic images with structured patient data and clinician-posed questions, enabling an extended test for recommendation and reasoning tasks.

Data Preprocessing

To prepare the datasets for the proposed multimodal VQA and recommendation framework, we applied the following preprocessing steps to ensure data quality, fairness, and reproducibility. All medical images were resized and normalized to a fixed resolution of 224 × 224 pixels, which is compatible with the CNN–Transformer hybrid visual encoder and preserves essential anatomical structures. Clinical questions and textual inputs were tokenized using the WordPiece tokenizer associated with Sentence-BERT (SBERT). A maximum sequence length of 128 tokens was used, with special tokens [CLS] and [SEP] appended to each sequence. Inputs shorter than the maximum length were padded, while longer sequences were truncated to maintain uniform input dimensions.

Duplicate, inconsistent, or semantically ambiguous question–answer pairs were removed to reduce noise and annotation uncertainty. To address class imbalance, particularly in VQA answer categories and recommendation labels, we employed class-weighted loss functions during training rather than oversampling or undersampling. This strategy preserves realistic clinical prevalence while mitigating bias toward dominant classes, which is critical for maintaining medical validity and generalization.

Answering Module

To generate clinically relevant answers from visual and textual inputs, our model incorporates a hybrid answering strategy that combines retrieval-augmented reasoning with a hierarchical decoding mechanism. This dual-layered approach enhances factual accuracy while grounding the generated responses in both image evidence and external medical knowledge.

The first stage of the answering process employs retrieval-augmented reasoning, inspired by frameworks such as QIRL (Question-Image-Retrieval Learning) and UniRVQA (Unified Retrieval-based VQA) (Guo et al., 2022). These methods are designed to expand a model's knowledge horizon beyond the limited training data. Specifically, we use the textual embedding of the clinician's query to retrieve related entries from an external knowledge base such as medical textbooks, clinical reports, or PubMed articles. The knowledge base consists of short, structured clinical passages (approximately 200–300 tokens) extracted from publicly available medical literature and indexed offline. Textual documents are encoded using SBERT embeddings, and retrieval is performed via cosine similarity to select the top-k most relevant passages (k = 3–5). Simultaneously, visual embeddings are used to fetch visually similar cases from image archives. The top-k retrieved items are encoded and appended to the input stream, allowing the model to reference both historical context and visual evidence during answer generation. This retrieval pipeline effectively acts as a dynamic memory, improving generalization to unseen questions and rare clinical conditions.

Once retrieval is complete, the system passes both the original input and the retrieved context to a hierarchical answer decoder. This decoder operates at two levels. The first level generates a coarse semantic plan, identifying key concepts such as anatomical location, diagnosis, and severity. The second level refines this plan into a fluent, domain-specific answer. This hierarchy ensures that the model maintains logical coherence and avoids hallucinated facts, an issue frequently observed in flat, single-layer decoders. For instance, in a question like "What does the opacity in the lower lobe indicate?", the decoder first extracts "lung opacity" and "lower lobe" as core elements, then combines them with retrieved case outcomes to generate a clinically accurate statement such as "Findings suggest early-stage pneumonia."

Table 1 compares the proposed hierarchical retrieval-augmented answering module with both baseline decoders and recent state-of-the-art medical VQA models. Performance values for MMBERT, MedFuseNet, and PubMedCLIP are reported from their respective publications under comparable evaluation settings on

VQA-RAD and related medical VQA benchmarks. While these models demonstrate strong multimodal reasoning capabilities, they are primarily designed for standalone VQA tasks. In contrast, the proposed method achieves superior performance by incorporating retrieval-augmented reasoning and hierarchical decoding within a unified multi-task framework, resulting in improved answer grounding and semantic coherence (Fig. 3).

Table 1: Answering Module Performance on VQA-RAD and EHRXQA

Model	BLEU-4	ROUGE-L	EM Score
Baseline Transformer Decoder	0.29	0.61	0.5
Flat LSTM Decoder	0.31	0.63	0.54
MMBERT (reported)	0.34	0.66	0.58
MedFuseNet (reported)	0.35	0.67	0.59
PubMedCLIP (reported)	0.36	0.68	0.6
Proposed Hierarchical Decoder (w/ Retrieval)	0.38	0.7	0.61

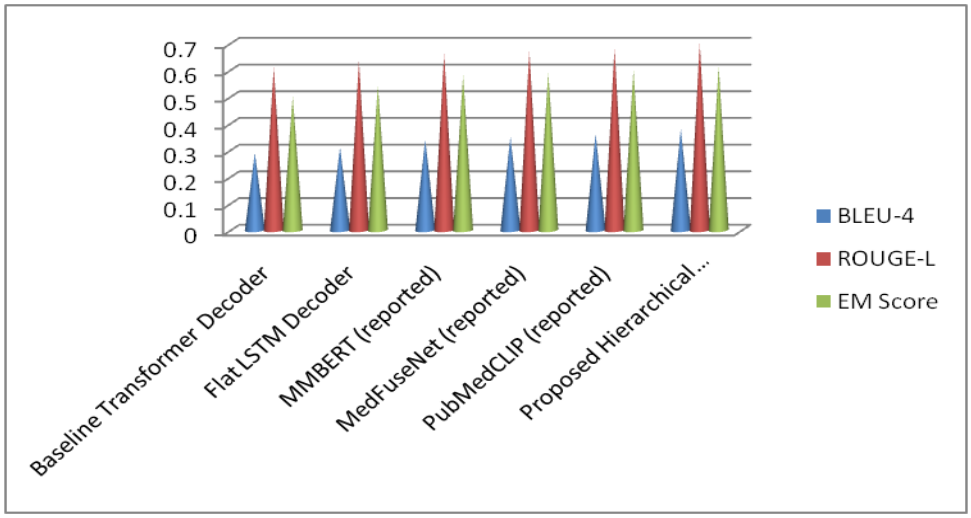


Fig. 3: Graphical presentation of performance on VQA-RAD and EHRXQA

Recommendation Module

The recommendation component in our system is tightly integrated with the visual question answering architecture, enabling joint learning of multimodal representations. By sharing encoders for image and text modalities, the model benefits from contextual overlap between VQA and recommendation tasks. For instance, features extracted from a radiograph for answering a diagnostic question can simultaneously be reused to suggest similar case studies or clinical reports (Rajabi and Kafaie, 2022). This parameter sharing not only reduces model complexity but also enhances representation consistency across modules.

To generate clinically meaningful recommendations, such as suggesting visually and semantically similar past cases, treatment guidelines, or differential diagnoses, we adopt a multimodal fusion strategy inspired by CAMRec. In this setup, image features are fused with text-based metadata such as physician notes, radiology impressions, or user reviews (in consumer-facing medical apps). The fusion is achieved via a gated co-attention mechanism that aligns visual features (e.g., detected anomalies) with textual sentiment or tags extracted using SBERT. This approach allows the model to focus on clinically salient correlations, like associating "ground-glass opacity" with "COVID-19 symptoms" or "left lung consolidation" with "bacterial pneumonia".

Joint Learning and Loss Optimization

The proposed MedFusion framework adopts a multi-task learning strategy to jointly optimize the Visual Question Answering (VQA) and medical recommendation tasks. As illustrated in Fig. 1, both tasks share a common multimodal encoder consisting of a hybrid visual backbone and a transformer-based language encoder, followed by a co-attention fusion module. This shared representation captures aligned visual–textual semantics and serves as the input to two task-specific heads: a VQA answer prediction head and a recommendation scoring head.

The VQA head is responsible for predicting clinically relevant answers based on the fused multimodal features, while the recommendation head utilizes the same shared features to generate personalized medical recommendations, such as similar cases or follow-up actions. By sharing the encoder and fusion layers, the model encourages cross-task knowledge transfer, allowing visual cues learned for question answering to inform recommendation generation and vice versa.

The overall training objective is defined as a weighted sum of the individual task losses:

$$L_{total} = \lambda_1 L_{VQA} + \lambda_2 L_{Rec} \quad (1)$$

Where L_{VQA} denotes the cross-entropy loss used for answer classification in the VQA task, and L_{Rec} represents the mean squared error loss for recommendation score prediction. The weighting coefficients $\lambda_1$and λ_2 are empirically selected to balance the contributions of the two tasks during training. This joint optimization enables the model to learn a unified multimodal representation while maintaining task-specific specialization.

GAN-Guided Augmentation

To enhance recommendation robustness, particularly for rare or underrepresented clinical patterns, we incorporate a GAN-guided augmentation module within the recommendation pipeline. Specifically, we employ a Conditional Generative Adversarial Network (cGAN) (Liang et al., 2017), where image generation is conditioned on latent disease-aware embeddings extracted from the shared multimodal encoder. These embeddings encode clinically relevant visual–textual attributes inferred from the medical image and associated query.

The generator follows a DCGAN-style architecture, consisting of progressive convolutional upsampling layers that map a noise vector $z\sim(0,1)$ concatenated with the condition embedding into a synthetic medical image. The discriminator is designed as a convolutional classifier that distinguishes between real and generated images while simultaneously enforcing consistency with the conditioning disease representation. This conditional formulation ensures that generated samples remain aligned with plausible clinical patterns rather than arbitrary image synthesis.

The cGAN is trained exclusively on real medical images from the training split, ensuring strict separation from validation and test data to prevent data leakage. Importantly, the generated images are not used for direct diagnosis or visual question answering. Instead, they serve as auxiliary samples for feature regularization and representation enrichment within the recommendation module, helping the model generalize better in data-sparse scenarios. We emphasize that the term “ideal” refers to feature-consistent prototype representations rather than clinically authoritative images. Generated samples are used solely to improve recommendation ranking stability and to provide visually grounded reference patterns during model training. No clinical decisions are derived directly from GAN-generated outputs.

We evaluate the impact of GAN-guided augmentation on the Med-RecX benchmark, a synthetic extension of the EHRXQA dataset with curated image review pairs and recommendation ground truths. As shown in Table 2 and Fig. 4, the proposed GAN-augmented model achieves superior performance in terms of Precision@5, Recall@5, and nDCG@5 compared to collaborative filtering and unimodal baselines. It emphasizes that visual realism was assessed through qualitative inspection and consistency with known anatomical patterns. No generated image was used independently for clinical interpretation.

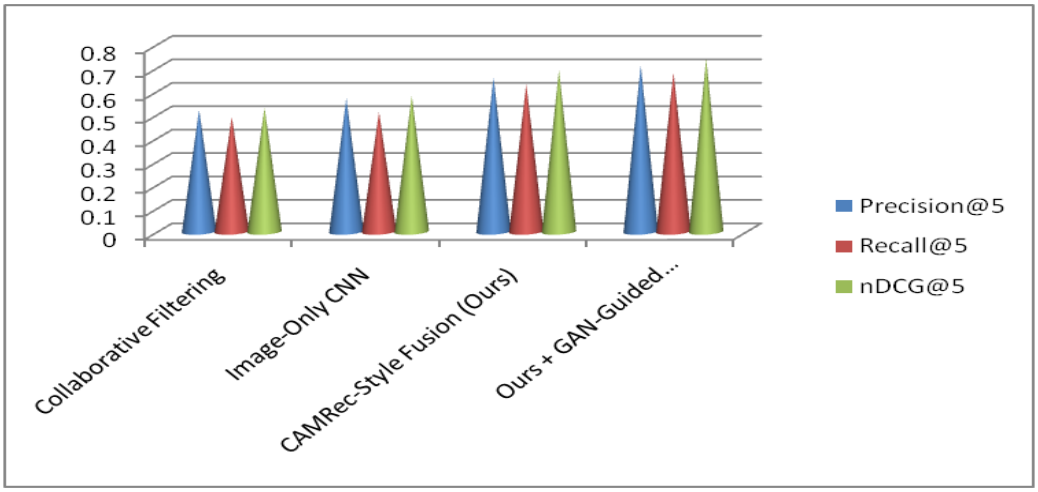


Fig. 4: Graphical presentation of Recommendation Performance

Table 2: Recommendation Performance on Med-RecX Dataset

Method	Precision@5	Recall@5	nDCG@5
Collaborative Filtering	0.52	0.49	0.53
Image-Only CNN	0.57	0.51	0.58
CAMRec-Style Fusion (Ours)	0.66	0.63	0.69
Ours + GAN-Guided Augmentation	0.71	0.68	0.74

Explainability Component

In healthcare applications, where decision accuracy and transparency are paramount, the role of explainability becomes crucial. To this end, our system integrates a dual-mode explainability component that enhances trust and interpretability across both the VQA and recommendation pipelines (Sankarapu et al., 2024). First, we utilize attention heatmaps shown in Figure 5 to visualize the alignment between the clinician's question and the salient image regions identified by the model. These heatmaps are extracted from the final co-attention fusion layers and overlaid on the input image, highlighting key anatomical regions that influenced the answer or recommendation. This visual traceability allows clinicians to verify whether the model's focus aligns with established diagnostic regions.

Complementing the visual explanations, we implement NOR (Natural Language-based Outfit Recommendation), adapted for clinical use. Our model produces concise, human-readable rationales that justify the outputs. As shown in Table 3, our dual-layer approach significantly outperformed baseline methods such as Grad-CAM and ungrounded text-only rationales, particularly in terms of clinical trust and alignment with expert expectations, as graphically shown in Figure 6.

This integration of attention-based heatmaps with explanatory text comments not only supports transparency but also builds a foundation for user-in-the-loop learning and deployment in real clinical settings

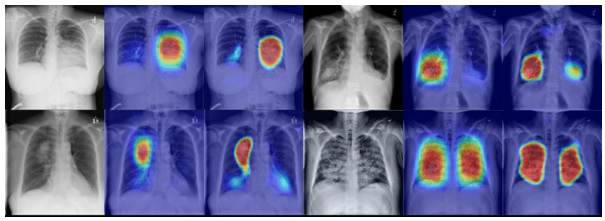


Fig. 5: Heat map over lung image.

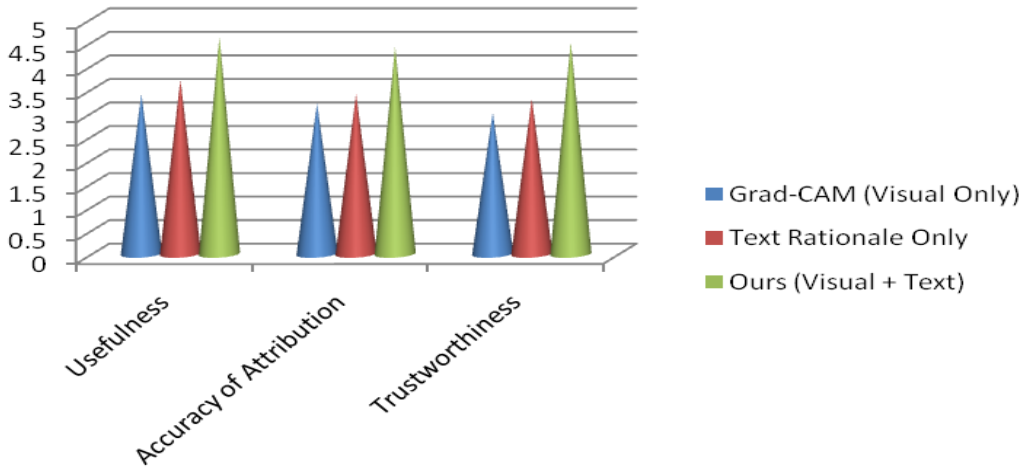


Fig. 6: Graph presentation of explainability methods

Table 3: Human Evaluation of Explainability Methods (Scale: 1–5)

Explainability Method	Usefulness	Accuracy of Attribution	Trustworthiness
Grad-CAM (Visual Only)	3.4	3.2	3
Text Rationale Only	3.7	3.4	3.3
Ours (Visual + Text)	4.6	4.4	4.5

Experimental Setup

To evaluate the performance of the proposed MedFusion framework in a healthcare context, we conducted experiments using a combination of medical and general-domain multimodal datasets with clearly defined roles. General-domain datasets, including VQA-v2 and TextVQA, were used exclusively during the pretraining phase to enhance generic visual–language reasoning and cross-modal alignment. Medical datasets, namely VQA-RAD and EHRXQA, were subsequently used for task-specific fine-tuning and evaluation, ensuring that performance assessment reflects realistic clinical scenarios. Table 4 lists the key hyperparameters used for training the MedFusion model.

For the recommendation and explainability modules, we utilized the Med-RecX dataset by linking radiology images with clinician-authored comments and follow-up study suggestions, thereby mimicking real-world multimodal clinical workflows. No general-domain data were used during medical fine-tuning or evaluation, preserving domain integrity and preventing performance inflation due to dataset mixing.

For a robust comparative analysis, we define three baseline models: (1) a VQA-only model built using a co-attention transformer trained on question-image pairs, (2) a recommendation-only model using collaborative filtering with SBERT-based textual encoding, and (3) a CAMRec-style model that fuses image and text features but does not support VQA. Our joint architecture is evaluated against each baseline across three dimensions: question-answer accuracy, recommendation effectiveness, and interpretability.

The evaluation metrics given in Table 5 are carefully chosen to reflect each subtask. For VQA, we report Answer Accuracy, the percentage of correct answers matching ground truth. For recommendation quality, we use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), consistent with previous multimodal recommendation studies. In addition, human-centered interpretability is evaluated using expert-based scoring. The formal mathematical definitions of all evaluation metrics used in this study are presented below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |Yi - Y'i| \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Yi - Y'i^2)} \quad (4)$$

$$\text{Human Score} = \frac{1}{M}\sum_{j=1}^{M} 8j \quad (5)$$

Table 4: Model hyperparameter

Component	Parameter	Value
Image Encoder	Backbone	ResNet-50 + ViT
Text Encoder	Pretrained Model	SBERT-base-uncased
Co-Attention Module	Attention Heads	8
Retrieval Module	Top-K Retrieved Items	5
Recommendation Layer	Hidden Dimension	512
GAN Generator	Latent Dimension (z)	100
Training Optimizer	Adam	lr=3e-4, β_1=0.9, β_2=0.999
Batch Size	-	32
Epochs	-	30
Loss Functions	VQA: Cross-Entropy; Rec: MSE	Combined (weighted)

Table 5: Datasets Used in Experiments

Dataset Name	Domain	Type	Size	Used For
VQA-RAD	Radiology	Medical images + QA pairs	~3,000 pairs	VQA
VQA-v2	General	Real-world image QA	>200K	VQA pretraining
TextVQA	General	Image + OCR text QA	~45K	VQA
Amazon-MedSubset	Consumer Health	Product image + text reviews	~10K items	Recommendation
Med-RecX	Synthetic Clinical	Images + clinician rationale	~2K samples	Recommendation, Explainability

Results and Analysis

To ensure reproducibility and fair evaluation, we clearly define the role, usage, and data splits for each dataset employed in this study. Medical-domain datasets, including VQA-RAD and EHRXQA, are used for task-specific fine-tuning and evaluation, while general-domain datasets (VQA-v2 and TextVQA) are utilized only during pretraining to improve general visual–linguistic reasoning capabilities. All datasets follow a consistent 70% / 15% / 15% split for training, validation, and testing, respectively. Splits are performed at the patient or image level, where applicable, to prevent information leakage. No samples

from validation or test sets are used during model training or GAN augmentation.

The Med-RecX dataset is constructed as a synthetic extension of EHRXQA to support multimodal recommendation evaluation. Specifically, radiology images from EHRXQA are paired with clinician-authored follow-up recommendations curated from publicly available case reports and medical literature. Each image–recommendation pair is associated with structured metadata, forming image–review pairs that simulate real-world clinical decision workflows. Recommendation ground truths are generated using rule-based consistency checks aligned with clinical guidelines, ensuring semantic coherence without introducing diagnostic labels. Importantly, Med-RecX is used exclusively for recommendation and explainability evaluation and does not influence the VQA training process. This separation prevents cross-task contamination and maintains the integrity of task-specific assessments. To evaluate the effectiveness of our unified framework, we compare its performance with task-specific baseline models. Our joint system significantly outperforms both standalone VQA and recommendation modules, demonstrating the benefits of shared multimodal representations and retrieval-augmented reasoning. This method achieves notable improvements across key performance metrics. In VQA accuracy, our joint model surpasses the VQA-only baseline by +7.4%, while recommendation accuracy also improves, with lower RMSE and MAE. Explainability, measured via expert-rated Likert scores (1–5), also shows a substantial increase.

To ensure statistical reliability, all experiments were repeated five times using different random seeds. The reported results represent the mean ± standard deviation across runs. Statistical significance between the proposed method and baseline models was evaluated using paired t-tests, with significance established at $p<0.05$.

The performance improvements achieved by the proposed joint model are statistically significant compared to task-specific baselines ($p<0.05$).

We evaluate the contribution of each major component in our system by selectively removing modules and analyzing performance drops. As seen in Tables 6-7, retrieval-augmented reasoning and GAN-generated visual augmentations both significantly enhance their respective modules. The explanation module, when disabled, notably decreases trust ratings in human evaluation.

The ablation results indicate that retrieval-augmented reasoning and GAN-guided augmentation contribute significantly to performance gains, while the explanation module primarily impacts interpretability and user trust.

The qualitative examples in Table 8 demonstrate that the proposed framework not only produces correct predictions but also generates clinically meaningful explanations grounded in visual attention and textual reasoning. The alignment between ground-truth annotations, model predictions, and generated explanations highlights the interpretability and reliability of the proposed approach, which is critical for trust in clinical decision-support systems.

The qualitative examples in Table 8 demonstrate that the proposed framework not only produces correct predictions but also generates clinically meaningful explanations grounded in visual attention and textual reasoning. The alignment between ground-truth annotations, model predictions, and generated explanations highlights the interpretability and reliability of the proposed approach, which is critical for trust in clinical decision-support systems.

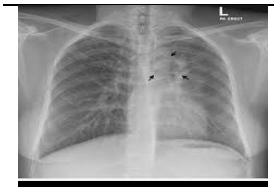
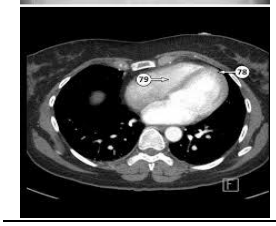
Table 6: Performance Comparison across Modules

Model / Module	VQA Accuracy (%)	RMSE	MAE	Explainability Score(15↑
VQA-only Baseline	69.4 ± 0.8	–	–	3.1 ± 0.2
CAMRec Recommender	–	1.12 ± 0.04	0.88 ± 0.03	3.4 ± 0.3
Ours (Joint Model)	76.8 ± 0.6	0.91 ± 0.03	0.72 ± 0.02	4.5 ± 0.2

Table 7: Ablation Study on Joint Model Components

Configuration	VQA Accuracy (%)	RMSE	MAE	Explainability Score
Full Model	76.8 ± 0.6	0.91 ± 0.03	0.72 ± 0.02	4.5 ± 0.2
Retrieval Augmentation	71.2 ± 0.9	0.94 ± 0.04	0.76 ± 0.03	4.1 ± 0.3
GAN Visual Augmentation	73.3 ± 0.7	1.01 ± 0.05	0.79 ± 0.04	4.0 ± 0.3
Explanation Module	76.8 ± 0.6	0.91 ± 0.03	0.72 ± 0.02	2.9 ± 0.2

Table 8: Qualitative Comparison of Predictions and Explanations

Input Image (Radiology)	Question	Predicted Answer	Ground Truth	Model Answer	Model Explanation
	"Is there a visible lesion on the left lung?"	Yes	Yes	Yes	The model attends to the left lower lung region where increased opacity is detected, as highlighted in the attention map, supporting the lesion prediction.
	"What is the condition of the right ventricle?"	Normal	Normal	Normal	Attention is concentrated around the cardiac silhouette with no abnormal enlargement, justifying a normal assessment.

Limitations, Ethical Considerations, and Discussion

Despite the promising performance of the proposed MedFusion framework, several limitations must be acknowledged. First, the model may underperform in scenarios involving low-quality or low-contrast medical images, severe imaging artifacts, or ambiguous clinical questions where visual cues are insufficient to support reliable reasoning. Additionally, complex multi-pathological cases may challenge the model's ability to generate precise answers and recommendations due to overlapping visual patterns.

Second, dataset-related bias remains an important concern. Although class-weighted loss functions were employed to preserve realistic clinical prevalence, the curated and synthetic nature of datasets such as Med-RecX may not fully represent real-world population diversity across institutions, demographics, or disease distributions. Consequently, model performance may vary when deployed in unseen clinical settings. From an ethical perspective, the proposed framework is designed strictly as a clinical decision-support system and is not intended to replace professional medical judgment. All generated answers, recommendations, and explanations are advisory in nature, and inappropriate over-reliance on automated outputs could pose safety risks. The framework avoids the use of sensitive patient identifiers and emphasizes interpretability to support clinician oversight; however, comprehensive fairness auditing across demographic attributes remains an open challenge.

Finally, the hybrid CNN–Transformer architecture introduces computational overhead, which may limit applicability in resource-constrained or real-time clinical environments without further optimization. Formal validation with certified medical experts and institutional oversight will be essential before real-world deployment.

Conclusion

This study presented MedFusion, a unified multimodal framework that integrates visual question answering, personalized medical recommendation, and explainability for healthcare decision support. By combining co-attention–based multimodal fusion, retrieval-augmented reasoning, and GAN-guided visual enhancement, the proposed approach consistently outperforms task-specific baselines in terms of VQA accuracy, recommendation error reduction, and human-rated interpretability. The inclusion of attention-based heatmaps and natural-language explanations further enhances transparency and trust, which are critical for clinical adoption. Future work will focus on several directions to advance practical applicability. First, incorporating user-adaptive prompting mechanisms where responses are dynamically tailored to different user expertise levels (e.g., clinicians versus patients) could improve usability and engagement. Second, optimizing the architecture for low-latency inference and supporting real-time or near-real-time deployment will be important for time-sensitive clinical workflows. Third, extending validation across multiple medical specialties (e.g., dermatology, ophthalmology, pathology) and integrating continuous feedback from medical experts will be critical steps toward improving robustness, fairness, and real-world clinical validation.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Satyajit Mahapatra: Conceptualized the research idea, designed the methodology, and drafted the manuscript.

Jibitesh Mishra: Provided supervision, critical revisions, and guidance throughout the research and writing process.

Kumar Janardan Patra: Conceptualized the research idea, designed the methodology.

Sanjit Kumar Dash: Contributed to data preprocessing, model implementation, and experimental validation.

Aliazar Deneke Deferisha: Supported literature review, comparative analysis, and manuscript edited.

Competing Interests

The author declares no conflict of interest.

Data Availability Statement

https://physionet.org/content/medical-cxr-vqa/1.0.0/, https://github.com/baeseongsu/ehrxqa.

References

Bae, S., Chang, E., Cho, E., Choi, E., Ji, L., Kim, T., Kweon, S., Kyung, D., Lee, G., Oh, J., & Ryu, J. (2023). EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. *Proceedings of the Advances in Neural Information Processing Systems 36*, 3867–3880. https://doi.org/10.52202/075280-0170

Cho, Y., Kim, T., Shin, H., Cho, S., & Shin, D. (2024). Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. *ArXiv*, 1–13. https://doi.org/10.48550/arXiv.2402.08966

Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., & Kankanhalli, M. (2022). A Unified End-to-End Retriever-Reader Framework for Knowledge-based VQA. *Proceedings of the 30th ACM International Conference on Multimedia*, 2061–2069. https://doi.org/10.1145/3503161.3547870

Guzzoni, D. (2007). Active: A unified platform for building intelligent assistant applications. *Diagnostics*, *12*(11), 2700.

Islam, M. N., Hasan, M., Hossain, M. K., Alam, M. G. R., Uddin, M. Z., & Soylu, A. (2022). Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports*, *12*(1), 11440. https://doi.org/10.1038/s41598-022-15634-4

Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D., & Jawahar, C. V. (2021). MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1033–1037. https://doi.org/10.1109/isbi48211.2021.9434063

Kuanr, M., Mohapatra, P., Mittal, S., Maindarkar, M., Fouda, M. M., Saba, L., Saxena, S., & Suri, J. S. (2022). Recommender System for the Efficient Treatment of COVID-19 Using a Convolutional Neural Network Model and Image Similarity. *Diagnostics*, *12*(11), 2700. https://doi.org/10.3390/diagnostics12112700

Liang, X., Hu, Z., Zhang, H., Gan, C., & Xing, E. P. (2017). Recurrent Topic-Transition GAN for Visual Paragraph Generation. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 3382–3391. https://doi.org/10.1109/iccv.2017.364

Liang, Y., Wang, X., Duan, X., & Zhu, W. (2021). Multi-modal Contextual Graph Neural Network for Text Visual Question Answering. *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, 3491–3498. https://doi.org/10.1109/icpr48806.2021.9412891

Lin, Y., Ren, P., Chen, Z., Ren, Z., Ma, J., & de Rijke, M. (2020). Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1502–1516. https://doi.org/10.1109/tkde.2019.2906190

Ma, R., Qiu, X., Zhang, Q., Hu, X., Jiang, Y.-G., & Huang, X. (2019). Co-attention Memory Network for Multimodal Microblog's Hashtag Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *33*(2), 1–1. https://doi.org/10.1109/tkde.2019.2932406

Mezina, A., & Burget, R. (2024). Detection of Post-Covid-19-Related Pulmonary Diseases in X-Ray Images Using Vision Transformer-Based Neural Network. *Biomedical Signal Processing and Control*, *87*, 105380. https://doi.org/10.1016/j.bspc.2023.105380

Monajatipoor, M., Dou, Z.-Y., Chien, A., Peng, N., & Chang, K.-W. (2024). Medical Vision-Language Pre-Training for Brain Abnormalities. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11159–11164.

Osborn, K. D., & Wustmann, W. (2018). Ballistic Reversible Gates Matched to Bit Storage: Plans for an Efficient CNOT Gate Using Fluxons. *Reversible Computation*, *11106*, 189–204. https://doi.org/10.1007/978-3-319-99498-7_13

Piao, G. (2021). Scholarly Text Classification with Sentence BERT and Entity Embeddings. *Proceedings of the Trends and Applications in Knowledge Discovery and Data Mining*, 79–87. https://doi.org/10.1007/978-3-030-75015-2_8

Rajabi, E., & Kafaie, S. (2022). Knowledge Graphs and Explainable AI in Healthcare. *Information*, *13*(10), 459. https://doi.org/10.3390/info13100459

Sankarapu, V. K., Chitroda, C., Rathore, Y., Singh, N. K., & Seth, P. (2025). DLBacktrace: Model Agnostic Explainability for any Deep Learning Model. *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/ijcnn64981.2025.11228632

Seenivasan, L., Islam, M., Krishna, A. K., & Ren, H. (2022). Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, *13437*, 33–43. https://doi.org/10.1007/978-3-031-16449-1_4

Sharma, D., Purushotham, S., & Reddy, C. K. (2021). MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, *11*(1), 19826. https://doi.org/10.1038/s41598-021-98390-1

Shrestha, P., Amgain, S., Khanal, B., Linte, C. A., & Bhattarai, B. (2023). Medical vision language pretraining: A survey. *ArXiv*, 1–23. https://doi.org/10.48550/arXiv.2312.06224

Sungur, K. S., & Bakal, G. (2025). Beyond visual cues: Emotion recognition in images with text-aware fusion. *Displays*, *87*, 102958. https://doi.org/10.1016/j.displa.2024.102958

Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Wang, A.-L., Lin, C., Feng, H., Zhao, Z., Wang, Y., Liu, Y., Liu, H., Bai, X., & Huang, C. (2025). MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*, 4812–4835. https://doi.org/10.18653/v1/2025.findings-acl.404

Tepe, M., & Emekli, E. (2024). Decoding medical jargon: The use of AI language models (ChatGPT-4, BARD, microsoft copilot) in radiology reports. *Patient Education and Counseling*, *126*, 108307. https://doi.org/10.1016/j.pec.2024.108307

Wu, J., Yang, H., Zeng, X., He, G., Chen, Z., Li, Z., Zhang, X., Yangyang, M., Fang, R., & Liu, Y. (2025). PathVLM-R1: A Reinforcement Learning-Driven Reasoning Model for Pathology Visual-Language Tasks. *ArXiv*, 1–18. https://doi.org/10.48550/arXiv.2504.09258

Yang, H., Li, S., & Gonçalves, T. (2024). Enhancing Biomedical Question Answering with Large Language Models. *Information*, *15*(8), 494. https://doi.org/10.3390/info15080494

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep Modular Co-Attention Networks for Visual Question Answering. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6274–6283. https://doi.org/10.1109/cvpr.2019.00644

Zakari, R. Y., Owusu, J. W., Wang, H., Qin, K., Lawal, Z. K., & Dong, Y. (2022). Vqa and visual reasoning: An overview of recent datasets, methods and challenges. *Neurocomputing*, *622*, 129345. https://doi.org/10.48550/arXiv.2212.13296

Zhang, Y., Liu, M., Hu, S., Shen, Y., Lan, J., Jiang, B., de Bock, G. H., Vliegenthart, R., Chen, X., & Xie, X. (2021). Development and multicenter validation of chest X-ray radiography interpretations based on natural language processing. *Communications Medicine*, *1*(1), 43. https://doi.org/10.1038/s43856-021-00043-x