Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MedCapVQA: A Two-Stage Explainable Framework for Medical Visual Question Answering

Nadeesha Perera

nadeesha.perera@tuni.fi

Tampere University

Jyrki Nummenmaa

Tampere University

Kostas Stefanidis

Tampere University

Article

Keywords: Medical VQA, Visual Question Answering, Caption-Based Reasoning, Medical Image Captioning, Vision-Language Models, BioBERT, Parameter-Efficient Fine-Tuning, Multimodal Learning

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11622355/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MedCapVQA: A Two-Stage Explainable Framework for Medical Visual Question Answering

Nadeesha Perera, Jyrki Nummenmaa and Kostas Stefanidis

Data Science Research Centre, Faculty of Information and Communication Technology, Tampere University, Korkeakoulunkatu, Tampere, 33720, Finland.

*Corresponding author(s). E-mail(s): nadeesha.perera@tuni.fi;
Contributing authors: jyrki.nummenmaa@tuni.fi;
konstantinos.stefanidis@tuni.fi;

Abstract

Medical Visual Question Answering (Med-VQA) aims to answer natural language questions about clinical images and has significant potential for diagnostic support, medical education, and improved access to expert interpretation. Most Med-VQA systems fuse image features and question embeddings directly, a mapping that is hard to learn from the few thousand question–answer pairs that medical benchmarks provide and that is usually compensated for with large-scale domain pre-training. We propose MedCapVQA, a two-stage, caption-driven framework that describes the image first and answers the question second. A MedGemma captioner, adapted with parameter-efficient fine-tuning on ROCOv2, converts each medical image into a clinically meaningful caption; a BioBERT-based reasoner with dedicated caption and question encoders and question-to-caption cross-attention then predicts the answer from text alone. We evaluate MedCapVQA on the closed-ended questions of the VQA-Med 2019 benchmark, training on the official training set (3,200 medical images, 12,792 question–answer pairs) and evaluating on the official validation set (500 images, 2,000 question–answer pairs), against six baselines that represent the main prior design families: a convolutional–recurrent fusion model, BioBERT and Med3D domain-adapted variants, a self-attention fusion network, a Vision Transformer model with cross-attention, and an end-to-end fine-tuned vision-language model. All models are trained under the same protocol and across training-set proportions from 5% to 100%, with each data-size result averaged over ten random training subsets. MedCapVQA achieves the highest validation accuracy on both tasks: 81.9% on binary (yes/no) questions, 11.6 percentage points above the strongest baseline, and 73.9% on 71-category multiclass questions, 5.4 points

above the fine-tuned vision-language model and 13.5 points above the Vision Transformer model. The vision-language model attains a higher multiclass training accuracy but lower validation accuracy compared to MedCapVQA and none of the baseline models are strong on both tasks. The MedCapVQA reasoner trains on a single GPU in ∼5 hours. These results show that converting medical images into textual captions before reasoning is an effective and resource-efficient alternative to direct multimodal fusion, and they highlight the value of intermediate semantic representations for the design of next-generation Med-VQA systems.

Keywords: Medical VQA, Visual Question Answering, Caption-Based Reasoning, Medical Image Captioning, Vision-Language Models, BioBERT, Parameter-Efficient Fine-Tuning, Multimodal Learning

1 Introduction

Visual Question Answering (VQA) emerged to develop systems capable of providing accurate answers to natural language questions about the content of an image. Early VQA models typically involved a combination of Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks [1, 2] for processing questions. Some of these initial approaches later began to incorporate attention mechanisms to allow models to focus on specific parts of the image relevant to the question, significantly improving the model's ability to provide contextually relevant answers [3–5]. Later work moved to using pre-trained networks such as VGG or ResNet for images and pre-trained word embeddings for text to improve feature extraction capabilities and efficiency [6]. The advent of transformer-based architectures [7] marked a further paradigm shift: models such as VisualBERT [8], VL-BERT [9], and LXMERT [10] demonstrated that joint vision-language pre-training yields powerful representations that transfer effectively to downstream VQA tasks. More recently, instruction-tuned Multimodal Large Language Models (MLLMs) such as LLaVA [11] and InstructBLIP [12] have pushed the state of the art further by enabling open-ended, conversational image understanding.

Visual question-answering models can be categorized along two principal axes. First, based on the *answer type*, VQA is formulated either as a closed-set classification problem—treating answers as labels from a fixed vocabulary with a softmax classifier and cross-entropy loss—or as an open-ended generative problem that produces free-form textual answers [6]. Second, based on *architecture*, VQA systems range from two-step early fusion models that separately encode image and text before combining them, to end-to-end models that jointly learn all representations, to modular networks that decompose the task into specialized sub-components. Attention-based extensions of these paradigms, including hierarchical and co-attention mechanisms [5], and knowledge-augmented models that incorporate external domain information, have further enriched the architectural landscape.

Medical Visual Question Answering (Med-VQA) applies these techniques to clinical imaging data—such as X-rays, MRI scans, CT scans, and ultrasounds—to help

healthcare professionals interpret images, identify anomalies, and support medical education [13]. A Med-VQA system must jointly understand visual content (anatomical structures, pathological findings) and natural language queries ranging from simple modality identification to complex diagnostic reasoning, before generating either a free-text answer or a classification. The field holds significant promise for diagnostic assistance, second-opinion support, and improving access to expert-level image interpretation in resource-constrained settings [14].

Despite rapid progress, Med-VQA faces substantial challenges: high-quality annotated datasets are scarce due to privacy constraints and the cost of expert labelling; medical images vary greatly across modalities and institutions, limiting generalisability; clinical questions are often ambiguous and domain-specific; and the *black box* nature of deep learning models creates barriers to clinical adoption where explainability and trust are paramount [15, 16]. Ethical and legal considerations surrounding patient data and automated decision-making add further complexity [17].

Most existing Med-VQA systems respond to these challenges in the same way: they learn a joint representation directly from image features and question embeddings, and they scale up domain-specific pre-training to compensate for the scarcity of labelled question–answer pairs. This strategy has two practical drawbacks. First, a fusion module trained on a few thousand question–answer pairs has to learn, at the same time, *what is visible in the image* and *how that evidence relates to the question*, across two very different feature spaces. Second, the pre-training that makes direct fusion work—hundreds of thousands to millions of image–text pairs and multi-GPU budgets—is out of reach for many research groups and clinical sites.

In this paper we take a different route. We propose **MedCapVQA**, a two-stage, caption-driven framework that *describes first and answers second*. In the first stage, a medical vision-language model (MedGemma, fine-tuned on ROCOv2 with QLoRA) converts each image into a clinically meaningful caption. In the second stage, a BioBERT-based reasoner with dedicated caption and question encoders and question-to-caption cross-attention predicts the answer from text alone. The caption acts as an intermediate semantic representation: it expresses the visual content in the same biomedical vocabulary as the question, it can be computed once and cached, and it can be read and checked by a clinician.

To establish that the gain comes from the caption-driven design rather than from a favourable experimental setting, we compare MedCapVQA with six baseline architectures that represent the main design families of prior work—convolutional–recurrent fusion, text-side and image-side domain adaptation, self-attention fusion, Vision Transformer encoding with cross-attention, and an end-to-end fine-tuned vision-language model (VLM)—all trained and evaluated under the same conditions on VQA-Med 2019. MedCapVQA achieves the best validation accuracy on both closed-ended tasks: 81.9% on binary yes/no questions, 11.6 percentage points above the strongest baseline, and 73.9% on 71-category multiclass questions, 5.4 points above the fine-tuned VLM and 13.5 points above the Vision Transformer model. No other baseline shows such competitive results on both tasks.

The contributions of this paper are as follows:

- We propose MedCapVQA, a two-stage explainable caption-driven Med-VQA framework that recasts multimodal question answering as text-only reasoning over a generated medical caption and the clinical question (Section 3).
- We describe a decoupled, parameter-efficient scheme for adapting the captioner—offline visual encoding, a lightweight projector producing eight virtual tokens, and QLoRA adaptation of the language decoder—that trains only about 0.5% of the captioner's parameters.
- We compare MedCapVQA with six representative baseline architectures under unified training and evaluation conditions, on binary (yes/no) and multiclass (modality, imaging plane, and organ system) questions and across training-set proportions from 5% to 100% (Sections 4 and 5).
- We analyse the training dynamics of MedCapVQA, relate its accuracy and resource requirements to published Med-VQA systems, and discuss practical implications and limitations (Section 6).

The remainder of this paper is organised as follows. Section 2 reviews related work on Med-VQA architectures, fusion methods, datasets, pretrained models, and caption-mediated VQA. Section 3 presents MedCapVQA. Section 4 describes the dataset, the evaluation protocol, and the baseline models. Section 5 reports the results, Section 6 discusses them, and Section 7 concludes with directions for future research.

2 Related Work

Medical Visual Question Answering (Med-VQA) is a sub-area of VQA research that aims to develop models capable of analysing clinical images such as radiographs, CT scans, and MRI scans and answering natural language questions about them. Driven by advances in computer vision and NLP, the field has grown substantially over the past decade. This section reviews the principal architectural components of Med-VQA systems (Section 2.1), key datasets (Section 2.2), and the main families of approaches: multimodal fusion methods (Section 2.3), attention mechanisms (Section 2.4), knowledge graph-based models (Section 2.5), transformer-based architectures (Section 2.6), dedicated pretrained Med-VQA models (Section 2.7), and caption-mediated VQA (Section 2.8), the line of work on which MedCapVQA builds and positions our contribution with respect to this literature.

2.1 Components of a Med-VQA System

A Med-VQA system consists of the following principal modules, which work together to map a medical image and a natural language question to an answer (Figure 1).

Image Processing Module. This component analyses the input image using deep learning techniques such as CNNs or Vision Transformers (ViTs) to extract hierarchical visual features capturing anatomical structures, pathological findings, and imaging artefacts. It must handle diverse medical imaging modalities including X-rays, MRIs, CT scans, and ultrasound images.

Natural Language Processing (NLP) Module. The NLP module interprets the user's query, mapping it from raw text to a dense feature representation. Techniques employed range from RNNs and LSTMs to transformer-based encoders such as BERT [18], the latter of which have become dominant due to their ability to capture long-range linguistic dependencies and their compatibility with joint vision-language pre-training.

Fusion Module. The fusion module is arguably the most critical component, integrating the visual and textual representations into a unified joint embedding from which an answer can be decoded. Fusion can be broadly categorised as *early fusion* (combining feature vectors before classification), *late fusion* (combining outputs of independently trained modality-specific models), or *joint fusion* (interleaving visual and textual processing throughout the network, as in cross-attention transformer architectures). The choice of fusion strategy has a direct impact on the model's ability to align fine-grained visual regions with specific question semantics, and it is the component that MedCapVQA rethinks: instead of fusing visual and textual features, we move the image into the text domain before fusion.

Answer Generation Module. This module decodes the fused representation into an answer, either by classifying it against a fixed answer vocabulary (closed-set paradigm) or by auto-regressively generating free-form text (generative paradigm). The appropriate choice depends on the question type and the clinical application.

Knowledge Base (Optional). Some Med-VQA systems augment the above pipeline with an external knowledge component, such as a clinical ontology (e.g., UMLS, SNOMED-CT) or a knowledge graph, to support reasoning over established medical relationships that may not be inferable from image-text pairs alone [19].

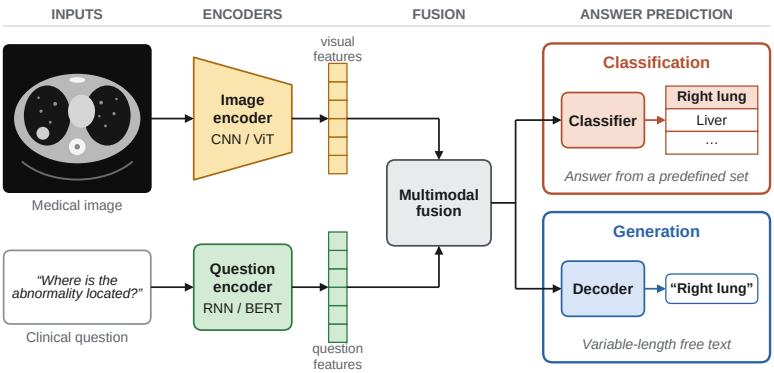


Fig. 1 General architecture of a conventional Medical VQA system

2.2 Medical VQA Datasets

Benchmark datasets are critical infrastructure for Med-VQA research. Several domain-specific datasets have been established, each with different imaging modalities,

question types, and annotation styles. Table 1 summarises their size and access conditions; the paragraphs below describe how each divides into closed-ended, open-ended, and multiple-choice questions.

The VQA-Med series. The ImageCLEF medical VQA challenges produced a series of radiology benchmarks covering X-ray, CT, MRI, and ultrasound images, all distributed through the ImageCLEF challenge platform. VQA-Med 2018 [20] contains 2,866 images with 6,413 question–answer pairs, all with free-form answers; the release does not divide them into closed- and open-ended subsets. **VQA-Med 2019** [21], among the most widely used Med-VQA benchmarks, comprises 4,200 radiology images and 15,292 free-form question–answer pairs. Its clinically motivated questions span four categories—abnormality detection, modality identification, organ recognition, and plane classification—which makes it well suited for category-level performance analysis. The official release defines no closed/open split, but the modality, plane, and organ questions have a small fixed answer vocabulary and behave as closed-ended classification, while abnormality questions are mostly open-ended. In the 12,792 training pairs used in this study, 1,294 are yes/no questions, 8,404 other closed-ended questions, and 3,094 open-ended questions (Table 2). The 2020 [22] and 2021 [23] editions narrow the task to abnormality questions, with one question per image: 5,000 question–answer pairs over 5,000 images and 5,500 pairs over 5,500 images, respectively. Their answers name the abnormality in free text, and no closed/open split is reported.

VQA-RAD [24] provides a smaller but carefully expert-curated, publicly available set of 3,515 question–answer pairs over 315 radiology images covering head CT, chest X-ray, and abdominal CT. Its questions are formulated by radiologists, lending high clinical authenticity, and are almost evenly divided between closed-ended questions (yes/no or a limited set of options; 1,797, 51.1%) and open-ended questions (1,718, 48.9%); there is no multiple-choice format. It has become a standard benchmark for closed-ended (yes/no, modality, organ) question evaluation.

PathVQA [25] extends Med-VQA to histopathology, presenting open-ended and binary yes/no questions over microscopy slides. Its 32,799 question–answer pairs over 4,998 images, derived from pathology textbooks and online educational resources, are split almost equally between closed-ended yes/no questions (16,334; 49.8%) and open-ended questions in six categories (16,465; 50.2%), with no multiple-choice format, and test a model's ability to reason over fine-grained cellular and tissue-level visual features. The dataset is publicly available.

SLAKE [26] is a publicly available, semantically labelled, bilingual (English/Chinese) dataset of 14,028 question–answer pairs over 642 CT, MRI, and X-ray images of multiple organs; the dataset contains both closed-ended and open-ended questions, but the original paper does not report how many of each. It includes structured knowledge graph annotations alongside the image–question–answer triples, and its rich semantic labelling makes it particularly valuable for evaluating knowledge-augmented Med-VQA systems.

Medical-Diff-VQA (also MIMIC-Diff-VQA) [27] introduced difference-based VQA over chest X-rays: answering questions about the difference between a main image and a reference image. Derived from MIMIC-CXR, it contains 700,703 question–answer pairs over 164,324 image pairs, spanning seven question types (abnormality,

presence, view, location, type, level, and difference). The paper reports no closed/open or multiple-choice split, and, like MIMIC-CXR, it is distributed through PhysioNet under credentialed access.

PMC-VQA [28] is a large-scale, multi-modality, publicly available benchmark of 226,946 question–answer pairs over 149,075 images. All 226,946 questions are released in two formats: as multiple-choice questions with four candidate answers and as open-ended questions without options, so the same pairs serve both a closed-ended and an open-ended benchmark. About 80% of the images are radiological, and the remainder include pathology and microscopy, so that it surpasses earlier datasets in both size and diversity.

OmniMedVQA [29] is among the largest real-image medical VQA benchmarks, with 127,995 question–answer pairs over 118,010 images, and is publicly available. All of its questions are multiple-choice, with two to four candidate answers, and it has no open-ended subset. It is collected from 73 different medical datasets, covers 12 imaging modalities and more than 20 anatomical regions, and draws all of its images from authentic clinical scenarios, which makes it well suited to evaluating large vision-language models.

RadImageNet-VQA [30] is a large-scale CT and MRI VQA dataset built from the expert-annotated CT and MRI subsets of RadImageNet [31]. It covers eight anatomical regions (abdomen, ankle/foot, brain, hip, knee, lung, shoulder, and spine) and 97 pathology categories, with questions on three tasks: anatomy recognition, abnormality detection, and pathology identification. Its training split pairs 750,009 images with about 6.75 million question–answer pairs (nine per image), plus one image–caption pair per image for vision–language alignment; a validation split adds 83,668 images and about 753,000 pairs, and the benchmark split has 9,000 pairs over 1,000 images. The questions come in three formats: open-ended, closed-ended yes/no (with positive and negative variants), and multiple-choice with four options, and the benchmark is balanced across tasks and formats. The dataset is distributed through Hugging Face as a gated dataset: users must accept a research-use agreement that restricts it to non-commercial research.

Together, these datasets expose several recurring question categories that form the basis for evaluation in this work: binary yes/no questions, modality classification, imaging plane identification, and organ system recognition.

2.3 Multimodal Fusion Approaches

Multimodal fusion is the process of integrating visual and textual representations to support joint reasoning. The field has evolved considerably, progressing from simple concatenation strategies to sophisticated architectures capable of modelling complex cross-modal dependencies [32]. In the medical domain, these approaches must additionally contend with the high-dimensional nature of clinical imaging data, the scarcity of annotated training examples, and the need for clinically interpretable outputs [33].

Concatenation and Early Fusion. The CNN-LSTM family of models [6] represents one of the earliest fusion paradigms, extracting image features with a CNN and question representations with an LSTM before combining them through concatenation or element-wise operations. This established many architectural conventions

Table 1 Summary of the medical VQA datasets described in Section 2.2. Access: *Open*, freely downloadable; *ImageCLEF*, distributed through the ImageCLEF challenge platform; *Credentialed*, PhysioNet credentialed access with data-use training; *Gated*, download after accepting a research-use agreement. Medical-Diff-VQA images are counted as main–reference pairs; RadImageNet-VQA counts combine its training, validation, and benchmark splits.

Dataset	Images	QA pairs	Access
VQA-Med 2018 [20]	2,866	6,413	ImageCLEF
VQA-Med 2019 [21]	4,200	15,292	ImageCLEF
VQA-Med 2020 [22]	5,000	5,000	ImageCLEF
VQA-Med 2021 [23]	5,500	5,500	ImageCLEF
VQA-RAD [24]	315	3,515	Open
PathVQA [25]	4,998	32,799	Open
SLAKE [26]	642	14,028	Open
Medical-Diff-VQA [27]	164,324 pairs	700,703	Credentialed
PMC-VQA [28]	149,075	226,946	Open
OmniMedVQA [29]	118,010	127,995	Open
RadImageNet-VQA [30]	834,677	≈7.5M	Gated

that subsequent work built upon. Joint Embedding Space models [34] extend this by projecting both modalities into a shared vector space where alignment scores can be computed directly. In medical VQA, concatenating ResNet-derived image features with GRU-encoded question embeddings has demonstrated competitive performance on VQA-RAD [24] and VQA-Med [21].

Bilinear Pooling Methods. Multimodal Compact Bilinear (MCB) pooling [35] computes the outer product of visual and textual feature vectors to capture higher-order cross-modal interactions, and has been applied to medical image question answering where subtle visual cues must be precisely associated with specific question semantics. The MUTAN framework [36] improves upon MCB by employing Tucker decomposition to model multimodal interactions efficiently, offering greater expressiveness over concatenation while remaining computationally tractable.

Late and Joint Fusion. Late fusion combines the outputs of independently trained modality-specific models at decision time, providing modularity at the cost of reduced cross-modal interaction during training. Joint fusion approaches—most prominently transformer-based cross-attention architectures—interleave visual and textual processing throughout the network, enabling richer bidirectional alignment between image regions and question tokens.

2.4 Attention Mechanisms

Attention mechanisms allow models to weight image regions or question tokens selectively words based on their relevance to the current inference step, substantially improving both performance and interpretability. Hierarchical Attention Networks [5] apply attention at both image region and question word levels, enabling fine-grained cross-modal alignment. Co-attention models [37] implement joint attention over both modalities simultaneously, allowing bidirectional reasoning between visual and linguistic representations.

Stacked Attention Networks (SANs) [38] apply multiple rounds of attention to progressively refine image region selection based on the query, a strategy that has been adapted for chest X-ray VQA to iteratively localise findings such as effusions and pneumothoraces. In the medical domain, Zhan et al. [39] proposed a conditional reasoning framework that employs question-conditioned attention over anatomical image regions, achieving strong performance on the VQA-Med 2019 benchmark. Cross-modal attention, as implemented in CMSA [40], jointly aligns specific question words with corresponding image regions, producing more clinically coherent answers. Moon et al. [41] further demonstrated that self-attention within transformer architectures can implicitly learn anatomically meaningful attention patterns when fine-tuned on chest radiograph VQA data.

2.5 Knowledge Graph-Based Models

Knowledge graph-based models augment standard vision-language pipelines with structured external information, enabling reasoning over clinical relationships that may not be apparent from image-text pairs alone. Scene graphs encoding spatial and anatomical relationships between image regions [42] provide explicit structural priors about how organs and pathological findings co-occur in clinical images. Medical knowledge graphs built from clinical ontologies such as UMLS and SNOMED-CT [19] allow models to chain clinical facts across both visual and textual evidence.

Li et al. [43] proposed a relation-aware graph attention network that constructs patient-specific knowledge graphs from clinical notes and imaging reports, dynamically updating node representations based on the input question. Zhang et al. [44] introduced a knowledge-aware multimodal pre-training strategy that incorporates entity-level medical knowledge from structured databases into the vision-language pre-training objective, improving performance on open-ended medical VQA. The integration of anatomical ontologies such as the Foundational Model of Anatomy (FMA) [45] has further enabled precise localisation of anatomical structures referenced in clinical questions.

2.6 Transformer-Based Architectures

Transformer-based architectures have become the dominant paradigm in Med-VQA, owing to their ability to model long-range dependencies and their compatibility with large-scale pre-training. VisualBERT [8] adapts BERT for multimodal tasks by integrating visual embeddings from a pre-trained object detector directly into the BERT input sequence, allowing self-attention to model cross-modal interactions implicitly; this approach proved effective in general VQA and served as an influential blueprint for subsequent medical adaptations. LXMERT [10] introduced a cross-modality encoder pre-trained with a diverse set of objectives including masked language modelling, masked object prediction, and cross-modal matching, demonstrating strong generalisation to downstream VQA tasks. OSCAR [46] leverages detected object tags as anchor points to align visual and textual representations during pre-training, reducing the semantic gap between modalities, while VinVL [47] improves upon this paradigm with a stronger visual backbone pre-trained on a larger collection of image-region-label

data, yielding richer visual representations that benefit downstream medical VQA when combined with domain-specific fine-tuning.

Zhou et al. [34] presented a Unified Vision-Language Pre-training (VLP) framework supporting both understanding and generation within a single model, adapted for medical VQA through fine-tuning on domain-specific image-text corpora. In the medical domain specifically, Chen et al. [48] proposed a multi-scale transformer that processes medical images at multiple resolutions simultaneously, capturing both global anatomical context and fine-grained pathological detail within a unified transformer framework. ClinicalBERT [49], while primarily a unimodal language model pre-trained on clinical notes from the MIMIC-III database, has been incorporated as the text encoder component in multimodal transformer pipelines, providing domain-adapted linguistic representations that improve the model's ability to parse complex clinical questions. Most recently, generalist medical multimodal models such as Med-PaLM M [50] have demonstrated that scaling transformer-based vision-language models to billions of parameters and pre-training on diverse biomedical data sources can yield strong zero-shot and few-shot performance across a wide range of medical VQA benchmarks, suggesting that large-scale pre-training may be a viable path toward generalisable clinical AI systems.

2.7 Pretrained Models for Medical VQA

A growing number of models have been pre-trained specifically for medical VQA, ranging from domain-adapted transformers to contrastive vision-language models and instruction-tuned MLLMs. Representative models are discussed below, together with their backbones and the results reported in their original papers.

MedViLL [41] employs a unified transformer pre-trained on large-scale chest X-ray data from MIMIC-CXR using combined masked language modelling and image-text matching objectives, capturing intricate relationships between radiological findings and textual descriptions. It combines a ResNet-50 visual extractor with a BERT-base joint encoder and is pre-trained on 91,685 frontal chest X-ray–report pairs. On VQA-RAD it reports 77.7% accuracy on closed-ended and 59.5% on open-ended questions. The model generalises beyond VQA to related tasks such as medical image-report retrieval.

MedQA [51], though primarily a text-based question answering framework derived from medical licensing examinations such as the USMLE, has been integrated into multimodal Med-VQA pipelines to support complex clinical reasoning where textual inference plays a central role.

Med3D [52] extends Med-VQA to three-dimensional volumetric imaging by pre-training 3D CNNs on eight medical segmentation datasets covering multiple organs and pathologies. Its 3D ResNet backbones are evaluated by transfer to segmentation and classification tasks rather than on VQA benchmarks, but they provide volumetric features that enable spatial reasoning in three dimensions, which is critical for CT and MRI-based VQA tasks.

BioViL [53] applies contrastive learning inspired by CLIP [54] to align chest X-ray images with their radiology reports from MIMIC-CXR, combining a ResNet image encoder with a biomedically-adapted BERT text encoder (CXR-BERT) to produce

representations sensitive to clinically relevant features. It is not evaluated on VQA; its reported strengths are phrase grounding on MS-CXR and zero-shot pneumonia classification (AUROC 0.831 on RSNA Pneumonia), which make it a strong visual encoder for chest X-ray VQA pipelines.

PubMedCLIP [55] fine-tunes CLIP on the ROCO dataset of over 80,000 image–caption pairs from PubMed articles, covering diverse modalities including radiology, histopathology, and dermatology. The semantic richness of PubMed captions—which often describe subtle pathological findings—makes it particularly effective for fine-grained medical VQA. Used as the visual encoder of an existing Med-VQA model, it reaches 72.1% overall accuracy on VQA-RAD [24] (80.0% closed, 60.1% open) and 80.1% on SLAKE [26] (82.5% closed, 78.4% open).

M3AE [56] adopts a masked autoencoder pre-training strategy, jointly reconstructing masked image patches and text tokens from paired medical image-text data sourced from PubMed and MIMIC-CXR. Its ViT image encoder and BERT text encoder are unified through a multimodal fusion transformer, enabling strong downstream performance with limited labelled data. It reports 83.5% closed and 67.2% open accuracy on VQA-RAD, 87.8% closed and 80.3% open on SLAKE, and 79.9% overall accuracy on VQA-Med 2019.

LLaVA-Med [57] represents the current frontier of instruction-tuned Med-VQA, adapting the general-purpose LLaVA model [11] to the medical domain through a two-stage process: first aligning biomedical visual features with the language model's embedding space using PubMed image-text pairs, then fine-tuning on medical VQA instruction data generated with GPT-4 assistance. LLaVA-Med achieves strong performance on VQA-RAD, PathVQA [25], and SLAKE [26], and supports open-ended medical dialogues grounded in visual evidence. Its best 7-billion-parameter variant reports closed-ended accuracy of 83.1% on VQA-RAD, 86.8% on SLAKE, and 91.1% on PathVQA, with open-ended recall of 64.8%, 87.1%, and 39.6%, respectively.

Across these models, a consistent trend is evident: combining domain-specific pre-training with strong backbone architectures yields substantial gains over general-domain transfer learning. This progress comes at a price, however. The strongest systems depend on hundreds of thousands to millions of image–text pairs and, at the frontier, on models with billions of parameters and resources that are unavailable in many research and clinical settings. This motivates approaches, such as the one proposed in this paper, that reach strong accuracy from a small amount of task-specific supervision.

2.8 Caption-Mediated Visual Question Answering

A complementary line of work avoids feature-level fusion altogether by translating the image into text and letting a language model reason over the description. PICa [58] converts images into captions and tags and prompts GPT-3 to answer knowledge-based questions in a few-shot manner. Plug-and-Play VQA [59] generates question-guided captions that are passed to a pre-trained question-answering model without any additional training, and Img2LLM [60] derives synthetic question–answer exemplars from captions to prompt a frozen large language model. These studies, conducted on natural-image benchmarks, show that language is an effective interface between

perception and reasoning. In the medical domain, captioning and report generation have mostly been studied as tasks in their own right, on datasets such as ROCO [61], IU X-Ray [62], and MIMIC-CXR [63], rather than as an intermediate representation for supervised Med-VQA.

MedCapVQA brings the caption-mediated paradigm to supervised Med-VQA and differs from the work above in three respects. First, the captioner is not used zero-shot: it is adapted to medical imagery with parameter-efficient fine-tuning, so that the captions carry the attributes that clinical questions ask about. Second, the reasoner is not a prompted general-purpose large language model but a compact, trainable biomedical encoder with explicit question-to-caption attention, which can be trained on a few thousand question–answer pairs on a single GPU. Third, we do not evaluate the approach in isolation: it is compared directly with direct-fusion architectures of increasing sophistication under one experimental protocol, which makes it possible to attribute the observed gains to the caption-driven design.

3 The MedCapVQA Framework

We propose **MedCapVQA**, a caption-based Medical Visual Question Answering framework that replaces direct image feature extraction with generated medical image captions. Unlike conventional Med-VQA systems, which rely on visual encoders such as convolutional neural networks [64] or vision transformers [65] to supply image representations to a fusion module, MedCapVQA uses a two-stage pipeline: (1) offline caption generation from medical images using a vision-language model, and (2) text-only reasoning over the generated caption and the clinical question using a transformer-based architecture (Figure 2). This design converts the multimodal VQA task into a purely textual reasoning problem, which allows powerful pre-trained biomedical language models to be applied directly and means that the reasoning stage never processes pixels.

3.1 Overview and Design Rationale

Let I denote a medical image, q a clinical question, and $\mathcal{A}$ the set of candidate answers. A direct-fusion model learns a single mapping $(I, q) \mapsto a$. MedCapVQA factorises this mapping into a perception step and a reasoning step:

$$c = g_\phi(I), \qquad \hat{a} = \arg\max_{a \in \mathcal{A}} \; p_\theta(a \mid c, q), \tag{1}$$

where g_ϕ is the captioner (Section 3.2), c is the generated caption, and p_θ is the caption-grounded reasoner (Section 3.3). Four considerations motivate this factorisation.

A shared representation space. The caption and the question are both biomedical text. Aligning them is a text-to-text matching problem that a biomedical language model is already well prepared for, whereas aligning convolutional or patch features with word embeddings has to be learned from the question–answer pairs themselves.

Perception and reasoning learn from different data. The captioner is adapted on image–caption pairs, which are far more plentiful than expert-written

question–answer pairs. The reasoner then only needs the small Med-VQA training set to learn how questions relate to descriptions.

An inspectable intermediate result. The caption is human-readable. A clinician or developer can see what the system "saw" before it answered, and an incorrect answer can be traced either to a caption that omits the relevant finding or to the reasoning step.

Low computational cost. Captions are generated once and cached. Training and running the reasoner involve only short text sequences, so the question-answering model can be trained, retrained, or extended to new question sets on modest hardware.

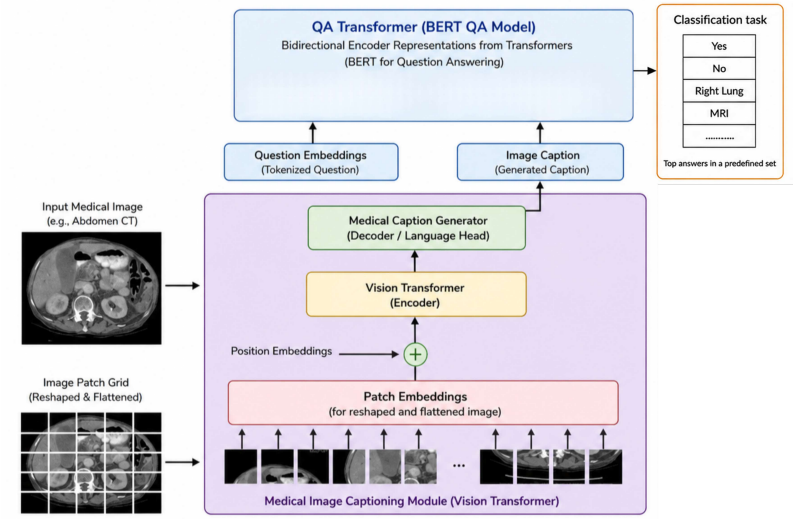


Fig. 2 Architecture of the proposed MedCapVQA framework. Stage 1 converts the medical image into a caption; Stage 2 encodes the caption and the question with BioBERT and dedicated transformer encoders, fuses them with question-to-caption cross-attention, and classifies the answer.

3.2 Stage 1: Medical Caption Generation

The first stage of our pipeline generates descriptive medical captions for input images using a fine-tuned MedGemma model [66]. Rather than relying on zero-shot captioning, we fine-tune MedGemma-4B-IT on the ROCOv2 dataset [67]—a large-scale collection of over 79,000 medical images paired with captions sourced from PubMed Central—to produce domain-adapted captions that capture clinically relevant visual features.

We employ a decoupled two-phase approach that separates visual encoding from language model training:

Phase 1: Offline Visual Encoding.

All ROCOv2 images are passed through MedGemma's frozen, 4-bit NF4-quantized [68] vision encoder (based on SigLIP-2 [69]). MedGemma employs a Gemma 3 processor [70] that jointly processes text–image pairs; accordingly, each image is wrapped in a chat-templated prompt before encoding. For each image I_i, we extract the final-layer hidden states corresponding to the N_v visual tokens and compute a pooled representation:

$$\mathbf{v}_i = \frac{1}{N_v} \sum_{j=1}^{N_v} \mathbf{h}_{i,j} \ \in \ \mathbb{R}^{d_v} \tag{2}$$

where d_v is the vision encoder's hidden dimension. When a multi-modal projector is available, the hidden states are additionally projected into the language model's embedding space before pooling. The resulting embeddings are serialized to disk, eliminating the need to load the vision encoder during fine-tuning.

Phase 2: Language Model Fine-Tuning.

A lightweight embedding projector maps the pre-extracted visual embedding $\mathbf{v}_i$ to K=8 virtual tokens in the language model's input space via a two-layer MLP with GELU activation [71] and layer normalization [72]:

$$[\mathbf{t}_1, \ldots, \mathbf{t}_K] = \mathrm{Reshape}\big(\mathrm{LN}\big(W_2\, \sigma\big(\mathrm{LN}(W_1\, \mathbf{v}_i + b_1)\big) + b_2\big),\ [K, d_{\mathrm{lm}}]\big) \tag{3}$$

where d_{lm} is the language model's hidden size, σ denotes GELU, and LN is layer normalization. Projector weights are initialized with Xavier uniform scaling [73] (gain = 0.01) for training stability. The Gemma 3-based language model decoder [70] is adapted using QLoRA [68, 74]—4-bit NF4 quantization with double quantization, combined with low-rank adapters (r=16, α=32, dropout = 0.05) applied to all attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and feed-forward layers (`gate_proj`, `up_proj`, `down_proj`)—yielding only ∼0.5% trainable parameters.

The model is trained with the standard autoregressive cross-entropy loss over caption tokens, with virtual-token positions masked:

$$\mathcal{L}_{\mathrm{cap}} = -\sum_{j=1}^{T} \log P_\theta\big(w_j \mid \mathbf{t}_1, \ldots, \mathbf{t}_K, w_1, \ldots, w_{j-1}\big) \tag{4}$$

where $\{\mathbf{t}_k\}_{k=1}^{K}$ are the projected virtual tokens and $\{w_j\}_{j=1}^{T}$ are the ground-truth caption tokens. Training employs 8-bit Paged AdamW [75, 76] with separate learning rates for the projector (1×10^{-3}) and LoRA parameters (2×10^{-4}), cosine learning rate scheduling with 3% warmup [77], and gradient checkpointing [78]. Early stopping with patience of 5 epochs monitors validation loss.

At inference, a new medical image is encoded by the quantized vision encoder, projected into virtual tokens, and decoded by the fine-tuned language model using beam search with repetition penalty and n-gram blocking. The resulting caption c_i serves as the visual proxy for the downstream VQA model, converting the multimodal task into a text-only reasoning problem.

3.3 Stage 2: Caption-Grounded Answer Prediction

The second stage receives the generated caption and the clinical question and predicts the answer. It consists of a shared BioBERT encoder, two dedicated transformer encoders, a question-to-caption fusion module, and a classification head.

3.3.1 Text Encoding with BioBERT

Both the generated captions and the clinical questions are encoded using BioBERT [79], a domain-specific language model pre-trained on large-scale biomedical corpora including PubMed abstracts and PubMed Central full-text articles. BioBERT is initialized from BERT-Base [18] and further pre-trained on biomedical text, enabling it to capture domain-specific terminology and semantic relationships that are critical for medical VQA.

Given a caption c and a question q, each is tokenized and passed through the shared BioBERT encoder to obtain contextualized token-level representations:

$$\mathbf{H}_c = \text{BioBERT}(c) \in \mathbb{R}^{L_c \times d}, \quad \mathbf{H}_q = \text{BioBERT}(q) \in \mathbb{R}^{L_q \times d} \tag{5}$$

where L_c and L_q denote the sequence lengths of the caption and question respectively, and $d = 768$ is the hidden dimension of BioBERT.

To balance training efficiency and representation quality, we employ a progressive unfreezing strategy for BioBERT [80]. The embedding layer and the first N transformer layers (default $N = 6$) are frozen at the start of training, while the upper layers remain trainable. This prevents catastrophic forgetting of pre-trained biomedical knowledge while allowing task-specific adaptation in the higher layers.

3.3.2 Dedicated Caption and Question Encoders

The BioBERT outputs are passed through two dedicated transformer-based text encoders—one for captions and one for questions. Each encoder consists of a linear projection layer that maps the BioBERT representations from the original hidden dimension d to a fusion dimension $d_f = 512$, followed by layer normalization and a stack of transformer encoder layers [7]:

$$\mathbf{Z}_c = \text{TransformerEncoder}_c(\text{LayerNorm}(\mathbf{W}_c\mathbf{H}_c + \mathbf{b}_c)) \tag{6}$$

$$\mathbf{Z}_q = \text{TransformerEncoder}_q(\text{LayerNorm}(\mathbf{W}_q\mathbf{H}_q + \mathbf{b}_q)) \tag{7}$$

where $\mathbf{W}_c, \mathbf{W}_q \in \mathbb{R}^{d_f \times d}$ are learnable projection matrices. The caption encoder uses 3 transformer layers and the question encoder uses 2 transformer layers, both employing 8 attention heads and GELU activation functions [71]. Each encoder produces a pooled representation via a learned attention pooling mechanism, which computes a weighted sum of the token representations based on learned attention scores:

$$\alpha_i = \frac{\exp(f_{\text{att}}(\mathbf{z}_i))}{\sum_{j=1}^{L} \exp(f_{\text{att}}(\mathbf{z}_j))}, \quad \mathbf{r} = \sum_{i=1}^{L} \alpha_i \mathbf{z}_i \tag{8}$$

where f_{att} is a two-layer feed-forward network with a Tanh activation.

3.3.3 Question-to-Caption Fusion

To enable the model to identify caption regions relevant to the posed question, we employ a multi-stage attention-based fusion mechanism. First, self-attention [7] is applied independently to the projected caption and question sequences with residual connections:

$$\hat{\mathbf{Z}}_c = \mathbf{Z}_c + \text{SelfAttn}(\mathbf{Z}_c, \mathbf{Z}_c, \mathbf{Z}_c) \tag{9}$$

$$\hat{\mathbf{Z}}_q = \mathbf{Z}_q + \text{SelfAttn}(\mathbf{Z}_q, \mathbf{Z}_q, \mathbf{Z}_q) \tag{10}$$

Subsequently, cross-attention is applied where the question tokens attend to the caption tokens, allowing the model to focus on caption content that is most relevant to answering the question:

$$\mathbf{A} = \text{CrossAttn}(\text{query} = \hat{\mathbf{Z}}_q, \text{key} = \hat{\mathbf{Z}}_c, \text{value} = \hat{\mathbf{Z}}_c) \tag{11}$$

The attended question representation is obtained via masked mean pooling over the cross-attention output. This is concatenated with the attention-pooled caption representation and passed through layer normalization:

$$\mathbf{f} = \text{LayerNorm}([\mathbf{r}_c \parallel \text{MeanPool}(\mathbf{A})]) \in \mathbb{R}^{2d_f} \tag{12}$$

3.3.4 Classification Head

The fused representation $\mathbf{f}$ is passed through a multi-layer classification head consisting of two fully connected layers with GELU activations [71], layer normalization, and dropout ($p = 0.1$):

$$\hat{y} = \mathbf{W}_3(\text{Dropout}(\text{GELU}(\mathbf{W}_2(\text{Dropout}(\text{GELU}(\mathbf{W}_1\mathbf{f})))))) \tag{13}$$

where the hidden dimensions are 512 and 256 respectively.

For binary classification (e.g., yes/no questions), a sigmoid activation is applied to produce a probability estimate, and binary cross-entropy loss is used for training. For multiclass classification, the raw logits are used with cross-entropy loss with optional label smoothing ($\epsilon = 0.1$) [81].

3.4 Training Procedure

The reasoner is trained with the AdamW optimizer [76] with a weight decay of 10^{-4}, and gradient clipping with a maximum norm of 1.0 is applied to stabilise training. The learning rate follows a warm-up and cosine decay schedule, with a separate, lower rate for the BioBERT backbone: the backbone is held at a negligible learning rate for the first two epochs and is then unfrozen, so that the newly initialised encoders, fusion layers, and classification head can produce meaningful gradients before the pre-trained representations are adapted (Section 5.4). Early stopping monitors validation accuracy.

4 Experimental Setup

4.1 Dataset

We use the VQA-Med 2019 dataset [21] with its official splits. The training set, which contains 3,200 medical images with 12,792 question–answer (QA) pairs (Table 2), is used only for training. All evaluations are carried out on the official validation set of 500 images with 2,000 QA pairs, none of which appear in the training set; no image, and therefore no generated caption, is shared between training and evaluation. In this study we focus on closed-ended questions in order to establish the performance of the models under a well-defined classification setting; open-ended questions are left for future work.

Table 2 Statistics of the VQA-Med 2019 training set used in this study

	Images	**Questions**			
Modality	**Count**	**Yes/No**	**Other closed**	**Open**	**Total**
Radiology	801	176	2,272	755	3,203
CT-Scan	824	485	1,996	814	3,295
MRI	1,358	620	3,497	1,312	5,429
Ultrasound	196	10	576	196	782
Nuclear Medicine	21	3	63	17	83
Total	3,200	1,294	8,404	3,094	12,792

4.2 Tasks and Evaluation Protocol

Two closed-ended tasks are derived from the dataset.

Binary classification. The yes/no questions form a two-class task: 1,294 in the training set (615 “yes”, 679 “no”) and 202 in the validation set (80 “yes”, 122 “no”).

Multiclass classification. The remaining closed-ended questions, which ask about the imaging modality, the imaging plane, and the organ system, form a classification task over 71 answer categories, with 8,404 questions in the training set and about 1,320 in the validation set.

For each task, models are trained on the corresponding questions of the official training set and evaluated on those of the official validation set. To study the effect of the amount of training data, every model is additionally trained on random subsets ranging from 5% to 100% of the training data. To reduce the bias of any single draw, ten random training subsets are sampled at each proportion, and the reported accuracy is the average over the ten runs; every run is evaluated on the full validation set. We report classification accuracy on the training and validation sets.

Evaluation metric. Both tasks are evaluated with classification accuracy, the fraction of questions whose predicted answer exactly matches the reference answer:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\hat{y}_i = y_i\right], \tag{14}$$

where N is the number of evaluated questions, y_i is the reference answer of question i, and $\hat{y}_i$ is the model's prediction. In the binary task, $\hat{y}_i \in \{\text{yes}, \text{no}\}$ is the predicted class label. In the multiclass task, $\hat{y}_i = \arg\max_k z_{i,k}$ is the top-1 prediction over the logits $z_{i,k}$ of the 71 answer categories. No partial credit is given, so the metric corresponds to the strict accuracy used in the VQA-Med 2019 challenge [21], restricted to closed-ended questions. The same function is applied to every model, on the training set to obtain training accuracy and on the validation set to obtain validation accuracy.

4.3 Baseline Models

We compare MedCapVQA with six baseline architectures (Table 3). They were chosen to cover the main design families of prior Med-VQA work and to form a ladder in which one design factor changes at a time: the text encoder (Model A to Model B), the image encoder (B to C), the fusion mechanism (C to D), the visual backbone together with cross-attention (E), and finally a pre-trained vision-language model fine-tuned end-to-end (G). The proposed model is labelled Model F in the figures. Figure 3 and Figure 4 illustrate the baseline architectures.

Baselines A–D derive from AquaNet [82], a simple multimodal question-answering architecture originally proposed for audio. We replace its audio branch with an image branch built on pre-trained image embeddings (Figure 3A); features from the two branches are fused, passed through two dense layers, and classified.

4.3.1 Baseline A: General-Domain Base Architecture

The first baseline employs general-domain pre-trained representations for both modalities (Figure 3A). Visual features are extracted from a VGG-19 [83] network pre-trained on ImageNet [84], yielding a 512-dimensional image feature vector per input. Textual representations are obtained using FastText [85], a subword-level word embedding model that maps each token in the medical question to a 300-dimensional vector. The image and question representations are independently encoded through separate 2-layer bidirectional LSTMs [86] (hidden size 512), producing 1024-dimensional encoded representations for each modality. The final hidden states of both LSTMs are concatenated to form a 2048-dimensional joint embedding, which is passed through a two-layer classification head ($1024 \rightarrow 1024$ units) with ReLU activations, followed by a linear projection over the answer categories. This architecture serves as a baseline to assess performance when neither the visual nor the textual encoder is adapted to the biomedical domain.

Table 3 Models compared in this study. Models A–E and G are baselines that operate on image features; MedCapVQA (Model F) operates on a generated caption.

	Model	Image representation	Question representation	Fusion
A	Base architecture	VGG-19 (ImageNet) features	FastText + Bi-LSTM	Concatenation
B	Partially domain adapted	VGG-19 (ImageNet) features	BioBERT + Bi-LSTM	Concatenation
C	Fully domain adapted	Med3D ResNet features	BioBERT + Bi-LSTM	Concatenation
D	Self-attention fusion	Med3D ResNet features	BioBERT + Bi-LSTM	Multi-head self-attention over the joint token sequence
E	Vision Transformer	ViT patch features	BioBERT + Bi-LSTM	Modality-specific self-attention and bidirectional cross-attention
G	End-to-end VLM	Qwen3.5-VL-2B with LoRA adapters (interleaved image and text tokens)		Native joint transformer
F	**MedCapVQA (ours)**	**MedGemma-generated caption (text)**	**BioBERT + transformer encoder**	**Question-to-caption cross-attention**

4.3.2 Baseline B: Partially Domain Adapted Base Architecture

The second architecture (Figure 3B) replaces the general-domain FastText embeddings with domain-adapted textual representations from BioBERT [79], a BERT-based [18] language model pre-trained on large-scale biomedical corpora including PubMed abstracts and PMC full-text articles. Each medical question is tokenized and passed through the frozen BioBERT encoder to obtain contextualized token-level embeddings of dimension 768, which capture domain-specific semantic relationships absent in general-purpose word embeddings. The visual pathway remains identical to Baseline A, utilizing VGG-19 pre-computed image features. Both modalities are encoded through separate 2-layer bidirectional LSTMs and fused via concatenation, as in the first architecture. By isolating the change to the textual encoder, this configuration enables direct evaluation of the impact of domain-adapted language representations on medical VQA performance while holding the visual representations constant.

4.3.3 Baseline C: Fully Domain Adapted Base Architecture

The third architecture (Figure 3C) extends the domain adaptation to the visual modality by replacing the ImageNet pre-trained VGG-19 features with representations extracted from a 3D ResNet [64] pre-trained on diverse medical imaging datasets through the Med3D framework [52]. Med3D provides transfer learning for 3D medical image analysis by pre-training on a large-scale aggregation of segmentation datasets spanning multiple imaging modalities and anatomical regions, enabling the learned representations to capture three-dimensional spatial relationships and domain-specific visual characteristics that are absent in natural image features. Each medical image is passed through the frozen Med3D network, and the penultimate layer activations are

extracted as a 512-dimensional feature vector ($\mathbf{v}_i = f_{\text{Med3D}}(\mathbf{I}_i) \in \mathbb{R}^{512}$). The textual pathway employs BioBERT embeddings identical to Baseline B. Both modalities are encoded through separate 2-layer bidirectional LSTMs and fused via concatenation, following the same concatenation-based fusion strategy. This configuration evaluates the benefit of domain-adapted representations in *both* modalities, where the visual encoder captures medically relevant spatial patterns and the textual encoder leverages biomedical linguistic knowledge.

4.3.4 Baseline D: Domain Adapted Self-Attention Fusion Architecture

The fourth baseline (Figure 4D) retains the domain-adapted encoders from Baseline C—Med3D for visual features and BioBERT for textual embeddings—but replaces the simple concatenation-based fusion with an attention-driven joint-modal interaction mechanism. After encoding both modalities through their respective bidirectional LSTMs, the image features are linearly projected to match the textual feature dimension, ensuring dimensional compatibility. The projected image feature sequence and the text feature sequence are then concatenated along the sequence dimension to form a unified multimodal token sequence. A multi-head self-attention module [7] with 8 attention heads is then applied, where the combined multimodal sequence serves as queries, keys, and values simultaneously. This mechanism enables every image and text token to attend to all other tokens in the joint sequence, allowing the model to capture intra-modal and inter-modal dependencies in a unified representational space, rather than conditioning attention strictly from one modality to the other. The attended joint representations are aggregated via global average pooling across the full sequence dimension to produce a single fused vector, which is then classified through a two-layer head identical to the preceding architectures. Compared to the concatenation-based baselines, this architecture enables holistic cross-modal reasoning by jointly modelling the relationships between all visual and textual tokens, which is particularly beneficial for medical VQA tasks that require integrating distributed visual evidence with the semantic content of the question to localise specific anatomical structures or pathological findings.

Baselines A–D are trained using the Adam optimizer [87] with a learning rate of 1×10^{-3} and early stopping based on training accuracy with a patience of 10 epochs to mitigate overfitting.

4.3.5 Baseline E: Vision Transformer with BioBERT Architecture

This baseline (Figure 4E) employs a multi-modal architecture that integrates visual and textual representations through cross-modal attention mechanisms. For image processing, the model supports both pre-computed feature vectors and raw image inputs; when raw images are provided, a pre-trained Vision Transformer (ViT) [65] extracts patch-level visual representations, while pre-computed features are projected into a sequence format to maintain architectural consistency. Textual inputs, comprising medical questions, are first encoded using BioBERT [79], a domain-specific

language model pre-trained on biomedical corpora, yielding contextualized token-level embeddings of dimension 768. These embeddings are subsequently processed by a bidirectional LSTM with 2 layers and a hidden dimension of 512, producing a 1024-dimensional text representation that captures sequential dependencies within the question. Both image and text features are then projected into a shared embedding space of dimension 512 via learned linear projections with ReLU activation and dropout regularization ($p = 0.1$). Within this shared space, modality-specific self-attention is first applied independently to the image and text representations using multi-head attention with 8 heads, followed by residual connections to preserve the original representations. Cross-modal interaction is then achieved through bidirectional cross-attention, where text features attend to image features and vice versa, enabling the model to ground textual queries in visual evidence and contextualize visual content with respect to the posed question. The attended representations from both modalities are pooled via global average pooling across their respective sequence dimensions and concatenated to form a 1024-dimensional fused representation. This fused vector is passed through a classification head comprising two fully-connected layers ($1024 \rightarrow 1024$ units) with ReLU activations, layer normalization, and dropout regularization ($p = 0.3$), followed by a final linear layer that produces logits over the 71 answer categories. The model is trained end-to-end using the AdamW optimizer with a learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a step-based learning rate schedule that decays the rate by a factor of 0.8 every 10 epochs. Early stopping based on training accuracy with a patience of 10 epochs is employed to prevent overfitting, and gradient clipping (max norm = 1.0) is applied for training stability.

4.3.6 Baseline G: End-to-End Vision-Language Model (VLM)

Model G (Figure 4G) is the strongest direct baseline in terms of capacity and pre-training: a general-purpose vision-language model that receives the raw image and the question and is fine-tuned for the same classification tasks. It tests whether a modern VLM, adapted end-to-end, makes an explicit captioning stage unnecessary.

The baseline employs Qwen3.5-VL-2B [88], a 2-billion parameter vision-language model that natively processes interleaved image and text tokens within a unified transformer architecture, eliminating the need for separate visual and textual encoders or explicit cross-modal fusion mechanisms. Unlike conventional multi-modal pipelines that extract features independently from each modality before combining them, Qwen3.5-VL jointly encodes visual and linguistic information from the outset.

To enable parameter-efficient fine-tuning the pretrained Qwen3.5-VL-2B weights are loaded in 4-bit NF4 quantization with double quantization enabled [68]. Low-Rank Adaptation (LoRA) [74] modules with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate $p = 0.1$ are injected into all seven linear projection layers within each transformer block—namely `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`—enabling adaptation of the full attention and feed-forward sub-layers while keeping the vast majority of parameters frozen. The native language modeling head of the VLM is also frozen, as the model is repurposed for classification rather than text generation.

For classification, no answer tokens are generated; instead, the hidden states from the final transformer layer of the language decoder $\mathbf{H} \in \mathbb{R}^{B \times L \times d}$ are extracted, where B is the batch size, L is the variable-length sequence comprising both visual and textual tokens, and d is the hidden dimension automatically detected from the model configuration. A masked mean pooling operation is applied over the sequence dimension using the attention mask $\mathbf{m} \in \{0, 1\}^{B \times L}$ to aggregate information across all valid tokens into a single fixed-dimensional representation:

$$\mathbf{z} = \frac{\sum_{i=1}^{L} m_i \cdot \mathbf{h}_i}{\sum_{i=1}^{L} m_i} \tag{15}$$

where $\mathbf{h}_i$ denotes the hidden state of the i-th token. The resulting pooled vector $\mathbf{z} \in \mathbb{R}^d$ is passed through a classification head comprising two fully-connected layers ($d \to 512 \to 256$ units) with ReLU activations and dropout regularization ($p = 0.2$), followed by a final linear layer that produces either a single logit passed through a sigmoid activation for binary classification, or logits over the 71 answer categories for multiclass classification. In the binary setting, the model is trained with binary cross-entropy loss, while the multiclass setting employs standard cross-entropy loss. The model is trained end-to-end (across all LoRA-adapted and classification head parameters) using the AdamW optimizer [76] with a learning rate of 2×10^{-5}, weight decay of 0.01, and a linear warmup schedule over 6% of total training steps followed by linear decay. Training employs gradient accumulation over 2 steps (yielding an effective batch size of 32) and gradient clipping with a maximum norm of 1.0 for stability. Early stopping based on validation accuracy with a patience of 5 epochs is applied to prevent overfitting.

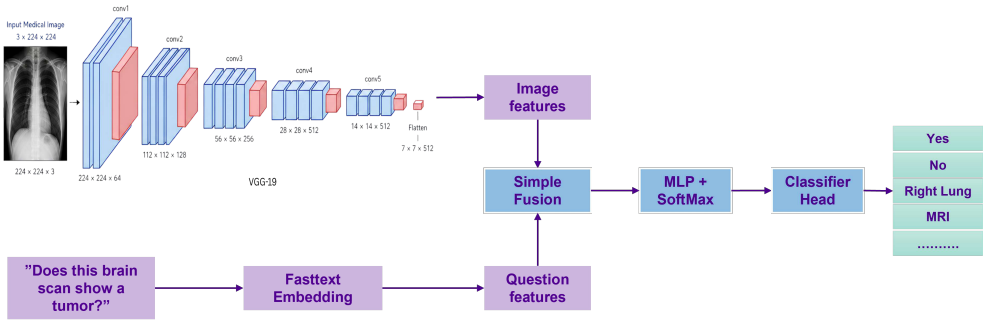
5 Results

Table 4 reports the training and validation accuracy of MedCapVQA and the six baselines when trained on the full training data, and Figure 5 shows how accuracy evolves with the proportion of training data.

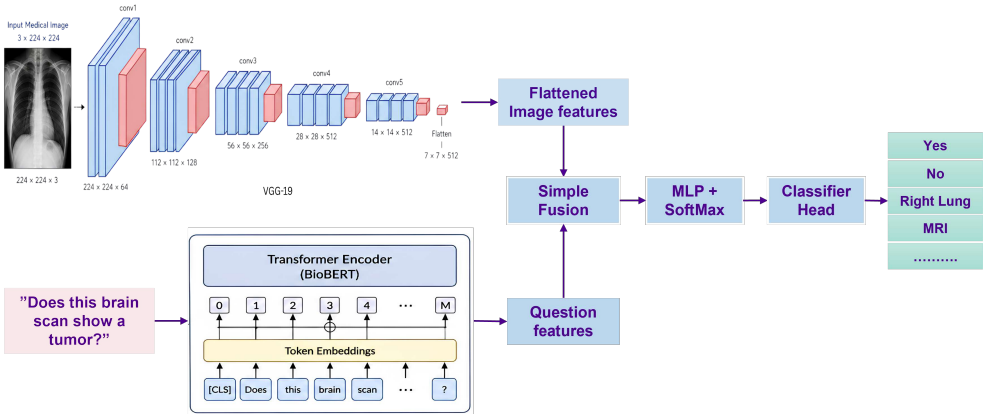
5.1 Overall Comparison

MedCapVQA achieves the highest validation accuracy on both tasks. On binary questions it reaches 81.9%, compared with 70.3% for the best baseline (Model A), a margin of 11.6 percentage points; the remaining baselines lie between 39.0% and 57.4%. On the 71-class task it reaches 73.9%, compared with 68.5% for the end-to-end VLM (Model G, +5.4 points), 60.4% for the Vision Transformer model (Model E, +13.5 points), and 43.2%–50.2% for the convolutional–recurrent baselines (Models A–D).

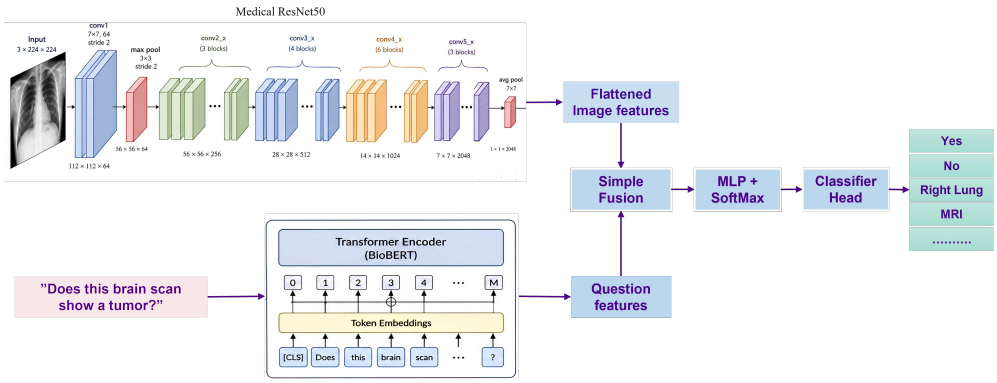
Two further observations follow from Table 4. First, MedCapVQA is the only model that is strong on both tasks. Model A ranks second on binary questions but fifth on multiclass questions, whereas Models E and G rank third and second on multiclass questions but reach only 39.0% and 57.4% on binary questions. Second, the model with the highest multiclass *training* accuracy is not the one that generalises best: Model G reaches 77.4% on the training set but 68.5% on the validation set, a gap of 8.9 points, whereas the gap for MedCapVQA is 1.6 points (75.5% versus 73.9%).



Model A: Base Architecture

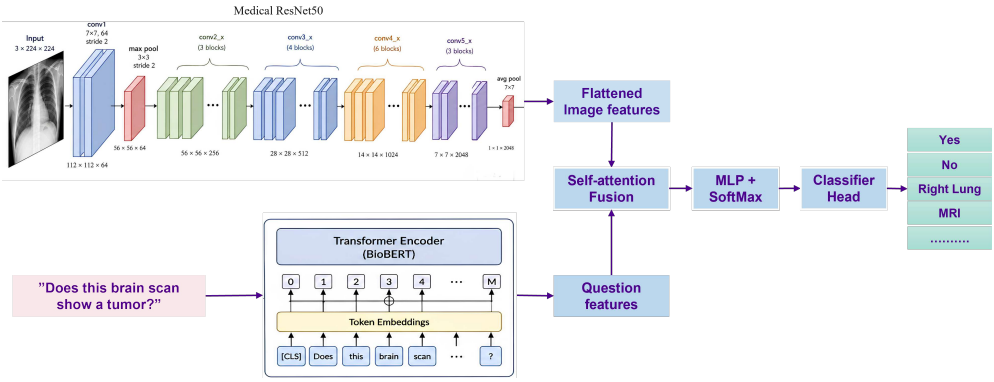


Model B: Partial Domain Adaptation

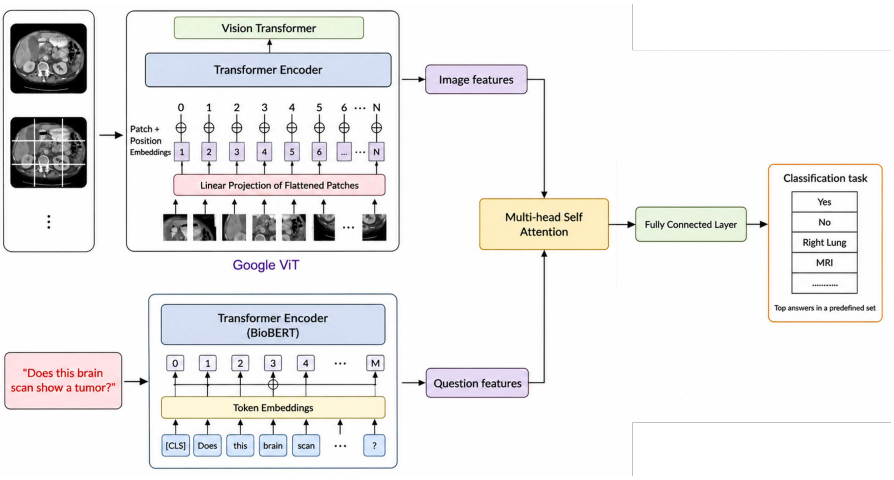


Model C: Full Domain Adaptation

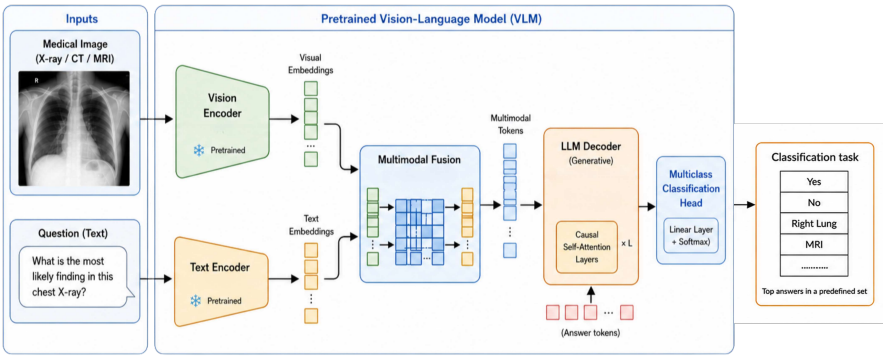
Fig. 3 Baseline architectures (Part 1): Base architecture (Model A), partial domain adaptation (Model B), full domain adaptation (Model C).



Model D: Self-Attention Fusion



Model E: Vision Transformer Model



Model G: Vision Language Model

Fig. 4 Baseline architectures (Part 2): Self-attention fusion (Model D), Vision Transformer-based architecture (Model E), end-to-end Vision-Language Model (Model G).

Table 4 Training and validation accuracy on the binary and multiclass tasks at 100% of the training data, with validation accuracy measured on the official VQA-Med 2019 validation set. Values are means over ten runs; *Train* is the training accuracy with dropout active at the early-stopping epoch. The best value in each column is in bold. Δ is the difference in validation accuracy between MedCapVQA and the best baseline, in percentage points.

Model	Binary Accuracy		Multiclass Accuracy	
	Train	**Val**	**Train**	**Val**
Baselines				
A. Base Architecture	0.7396	0.7030	0.4871	0.4587
B. Partially Domain Adapted Base	0.5313	0.5573	0.5157	0.5020
C. Fully Domain Adapted Base	0.5547	0.5469	0.4767	0.4321
D. Domain Adapted Self-Attention	0.5258	0.5729	0.4705	0.4341
E. Vision Transformer	0.5450	0.3900	0.6258	0.6043
G. End-to-End VLM	0.6258	0.5740	**0.7739**	0.6851
Proposed				
F. MedCapVQA (ours)	**0.9353**	**0.8194**	0.7549	**0.7390**
Δ vs. best baseline (Val)		+11.6		+5.4

5.2 Binary Classification Results

Figures 5a and 5b present the training and validation accuracy of all seven models on the binary (yes/no) task, evaluated across increasing data percentages from 5% to 100%; each point is the average over ten random training subsets.

MedCapVQA (Model F) clearly dominates both training and validation performance. Its training accuracy rises sharply beyond 40% of the data, reaching approximately 0.90 at 45% and stabilising around 0.93–0.96 from 70% onward (Figure 5a). Correspondingly, its validation accuracy climbs steadily, reaching approximately 0.82–0.85 at 75%–85% of the data (Figure 5b), which indicates strong generalisation. The consistent gap between the training and validation curves suggests a moderate degree of overfitting, although validation accuracy remains high throughout.

Among the baselines, **Model A (Base Architecture)** is the strongest on this task, with training accuracy fluctuating between 0.52 and 0.78 at lower data regimes, with dips at 25%, 50%, and 75% of the data, before stabilising around 0.72–0.75 at higher data percentages (Figure 5a). Its validation accuracy tracks relatively closely, reaching approximately 0.70 at full data (Figure 5b), and its behaviour across both curves is more consistent than that of the other baselines.

Model G (End-to-End VLM) shows moderate binary performance, with training accuracy around 0.62–0.67 across most data regimes and validation accuracy around 0.55–0.60. Although it is the largest model in the comparison and processes the image and the question jointly, its binary accuracy remains far below that of MedCapVQA, which indicates that end-to-end fine-tuning of a general-purpose VLM on the small binary subset does not match the caption-driven approach.

Model B (Partially Domain Adapted) shows noticeable fluctuations in both training (0.52–0.77, highest at the smallest subsets) and validation (0.43–0.63) accuracy, suggesting sensitivity to data sampling and limited robustness at lower data regimes. **Model C (Fully Domain Adapted)** exhibits erratic behaviour, particularly in validation, where its accuracy oscillates between approximately 0.44 and 0.63 across data percentages (Figure 5b) without a clear upward trend. This instability suggests that full domain adaptation introduces training difficulties when data is scarce.

Model E (Vision Transformer) performs modestly in the binary setting, with training accuracy clustering around 0.52–0.60 across most data percentages, broadly comparable to Models B, C, and D. Notably, its validation accuracy drops sharply to approximately 0.39 at 40%, 90%, and 100% of the data (Figure 5b), the last being the lowest validation accuracy of any model at 100% data. This counter-intuitive degradation at high data availability may indicate overfitting to training-specific patterns or sensitivity to the class distribution of the full binary subset.

Model D (Domain Adapted Self-Attention) consistently underperforms most other models in binary training (0.52–0.58 from 25% of the data onward) while its validation accuracy varies widely, from approximately 0.38 at 20% to 0.63 at its best, which suggests that its self-attention fusion mechanism provides limited discriminative power for simple yes/no classification.

Overall, the binary results show that the caption-driven approach of MedCapVQA yields a substantial advantage over every direct-fusion baseline. Model A remains a stable reference point, while the domain-adaptation strategies (Models B, C, and D) introduce instability, particularly at lower data availability, and Model E's collapse in validation accuracy at full data warrants further investigation.

5.3 Multiclass Classification Results

Figures 5c and 5d present the training and validation accuracy of all seven models on the multiclass task (71 classes), evaluated across increasing data percentages from 5% to 100%, again averaged over ten random training subsets per proportion.

MedCapVQA (Model F) achieves the highest validation accuracy of all models, reaching approximately 0.74 at full data (Figure 5d). Its learning curve has a distinctive shape: it starts with the lowest training accuracy of all models at 5% of the data (approximately 0.22), rises sharply to overtake Models A–E by approximately 15% of the data, and plateaus around 0.71–0.76 from 40% onward (Figure 5c). The reasoner therefore needs a minimum amount of supervision to learn the 71-way mapping, but once that amount is available it makes better use of the data than any direct-fusion baseline. Its training and validation curves stay close together, with a gap of 1.6 points at full data (Table 4).

Model G (End-to-End VLM) is the strongest baseline. It achieves the highest training accuracy of all models, rising steeply from approximately 0.30 at 5% to a plateau around 0.75–0.77 from 40% onward (Figure 5c). Its validation accuracy, however, reaches only approximately 0.69 at full data (Figure 5d), which places it second and leaves a markedly larger gap between its training and validation curves than for

MedCapVQA. The VLM has the capacity to fit the training set but generalises less well than the caption-driven model.

Model E (Vision Transformer) is the clear third-best performer in the multiclass setting. Its training accuracy increases steadily from approximately 0.48 at 5% to 0.64 at full data (Figure 5c), while its validation accuracy follows a similar trajectory, rising from approximately 0.42 to 0.60 (Figure 5d). The small and consistent gap between training and validation indicates that Model E generalises reliably in the multiclass regime—a marked contrast to its instability in the binary task.

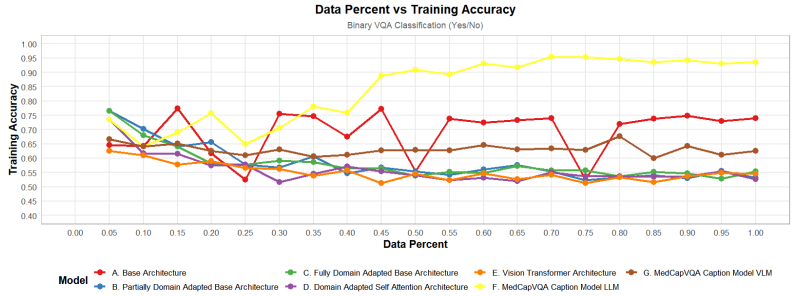
Model B (Partially Domain Adapted) ranks fourth, with training accuracy stabilising around 0.51–0.54 and validation accuracy around 0.49–0.50 at higher data percentages. While it does not approach Models F, G, or E, its performance is notably more stable in the multiclass setting than the erratic behaviour observed in the binary task (Figures 5a and 5b).

Models A (Base Architecture), **C (Fully Domain Adapted)**, and **D (Domain Adapted Self-Attention)** cluster closely together, with training accuracies converging around 0.47–0.49 and validation accuracies around 0.43–0.46 at full data (Figures 5c and 5d). Model A begins with the second-lowest training accuracy ($\sim$0.35) at 5% of the data but converges with Models C and D by approximately 15%–20% of the data. Model D, which performed weakest in binary training, is comparable to Models A and C here, suggesting that its self-attention fusion mechanism provides marginally more useful representations when the output space is larger, although it still does not surpass the simpler baselines.

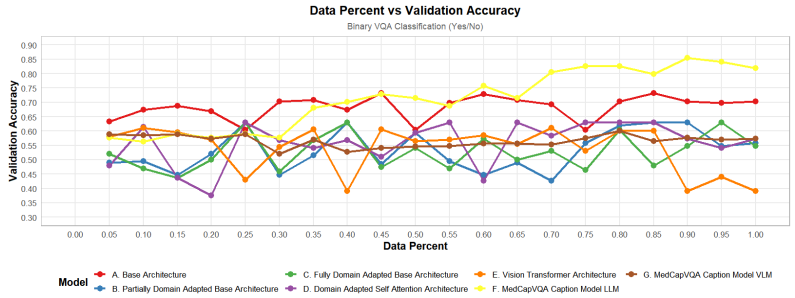
Comparing the multiclass results with the binary results, several observations emerge. First, accuracy values are expectedly lower across all models because of the much larger number of target classes (71 versus 2). Second, the ranking of the baselines changes considerably between tasks: Model E, which struggles in binary classification (particularly in validation), recovers to become the third-best model in the multiclass setting, while Model A, which is clearly second-best in binary classification, falls to the lower cluster. MedCapVQA is the exception, ranking first in both. Third, Model G leads in multiclass training accuracy while MedCapVQA leads in validation accuracy, which underscores the stronger generalisation of the caption-driven approach. Fourth, the models exhibit smoother and more monotonically increasing learning curves in the multiclass task, with fewer of the erratic fluctuations seen in the binary setting, particularly for Models B and C; this suggests that the multiclass objective provides a more stable training signal across data percentages. Finally, the ability of MedCapVQA to reach approximately 0.74 validation accuracy across 71 classes underscores the advantage of using captions as an intermediate representation in the VQA pipeline.

5.4 Training Dynamics of MedCapVQA

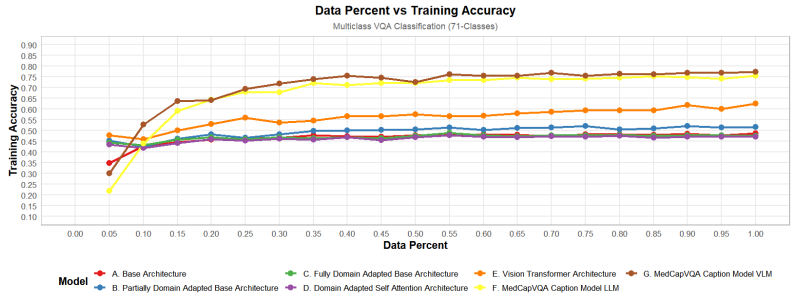
Figure 6 presents detailed training metrics for MedCapVQA on the multiclass classification task across 20 epochs, comprising four diagnostic panels: loss curves, accuracy curves, overfitting analysis, and learning rate schedules.



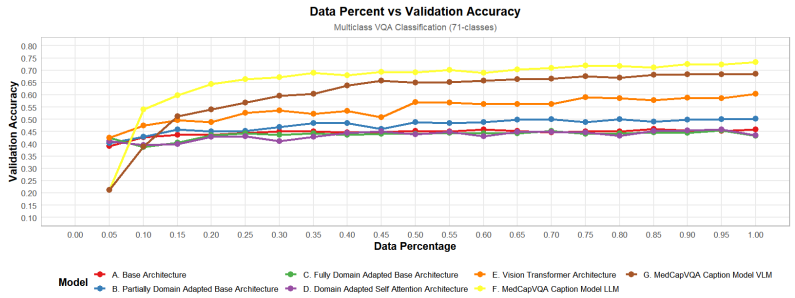
(a)



(b)



(c)



(d)

Fig. 5 Binary classification training and validation accuracy (a, b) and multiclass classification training and validation accuracy (c, d) for the six baselines (Models A–E and G) and the proposed MedCapVQA (Model F) across data percentages from 5% to 100%. Each point is the mean over ten random training subsets; validation accuracy is measured on the official VQA-Med 2019 validation set.

5.4.1 Loss Curves

The top-left panel shows the training loss (with dropout), clean training loss (without dropout), and validation loss over the training period. All three loss curves exhibit a steep initial decline during the first three epochs, reflecting rapid early learning. The training loss (with dropout) starts at approximately 3.3 and decreases sharply, remaining slightly above the clean training loss throughout, as expected when dropout is active during training. The clean training loss and validation loss converge closely in the early epochs and then gradually diverge, with the validation loss reaching its minimum of **1.4770 at epoch 8** before slowly increasing. This divergence beyond epoch 8 signals the onset of overfitting, where the model continues to improve on the training set but begins to lose generalisation performance on the validation set.

5.4.2 Accuracy Curves

The top-right panel displays the corresponding accuracy metrics. The training accuracy (with dropout) climbs from approximately 25% at epoch 1 to nearly 98% by epoch 20, while the clean training accuracy rises from approximately 65% to over 98%. The validation accuracy improves more gradually, reaching its best value of **73.90% at epoch 13**, after which it plateaus around 72–74%. The widening gap between clean training accuracy and validation accuracy beyond epoch 10 corroborates the overfitting trend identified in the loss curves. Nevertheless, the validation accuracy remains stable rather than declining sharply, indicating that the model does not catastrophically overfit.

5.4.3 Overfitting Analysis

The bottom-left panel visualises overfitting dynamics through two gap metrics: the *Dropout Gap* (Clean Train Accuracy − Train Accuracy with dropout) and the *Overfit Gap* (Clean Train Accuracy − Validation Accuracy). The Dropout Gap starts high at approximately 21% in the first epoch due to the strong regularisation effect of dropout on an undertrained model, but decreases rapidly, fluctuates between about 3% and 7.5% up to epoch 12, and stays below 5% from epoch 13 onward, indicating that the model becomes increasingly robust to dropout noise as training progresses. In contrast, the Overfit Gap begins near 0% and steadily increases from epoch 6 onward, reaching approximately 26% by epoch 18. Training therefore falls into two phases: an early phase (epochs 1–8) in which the overfit gap stays small and the model learns generalisable features, and a later phase (epochs 9–20) in which the gap widens and the model increasingly memorises training-specific patterns. This analysis suggests that **early stopping around epochs 8–13** would be optimal, balancing minimal validation loss (epoch 8) and peak validation accuracy (epoch 13).

5.4.4 Learning Rate Schedule

The bottom-right panel shows the learning rate schedules for both the BERT backbone and the classification head. The head learning rate follows a warm-up and cosine decay pattern, peaking at approximately 4.5×10^{-5} around epochs 2–4 before gradually declining to approximately 5×10^{-7} by epoch 20. The BERT learning rate remains

at a very low value ($\sim 10^{-7}$) for the first two epochs, after which the BERT layers are unfrozen and the learning rate rises to approximately 2×10^{-5}, following a similar cosine decay thereafter. This staged unfreezing strategy allows the classification head to develop meaningful gradient signals before fine-tuning the pre-trained BERT representations, preventing catastrophic forgetting of the language model's learned features.

5.4.5 Summary

Overall, the training diagnostics reveal that MedCapVQA learns efficiently within the first 8–13 epochs, achieving strong convergence on both loss and accuracy. The dropout regularisation is effective during the early training phase, while the staged learning rate schedule with BERT layer unfreezing enables careful adaptation of the pre-trained language model to the medical VQA domain. However, the growing overfit gap beyond epoch 10 indicates that additional regularisation strategies or early stopping should be considered to further improve generalisation in future work.

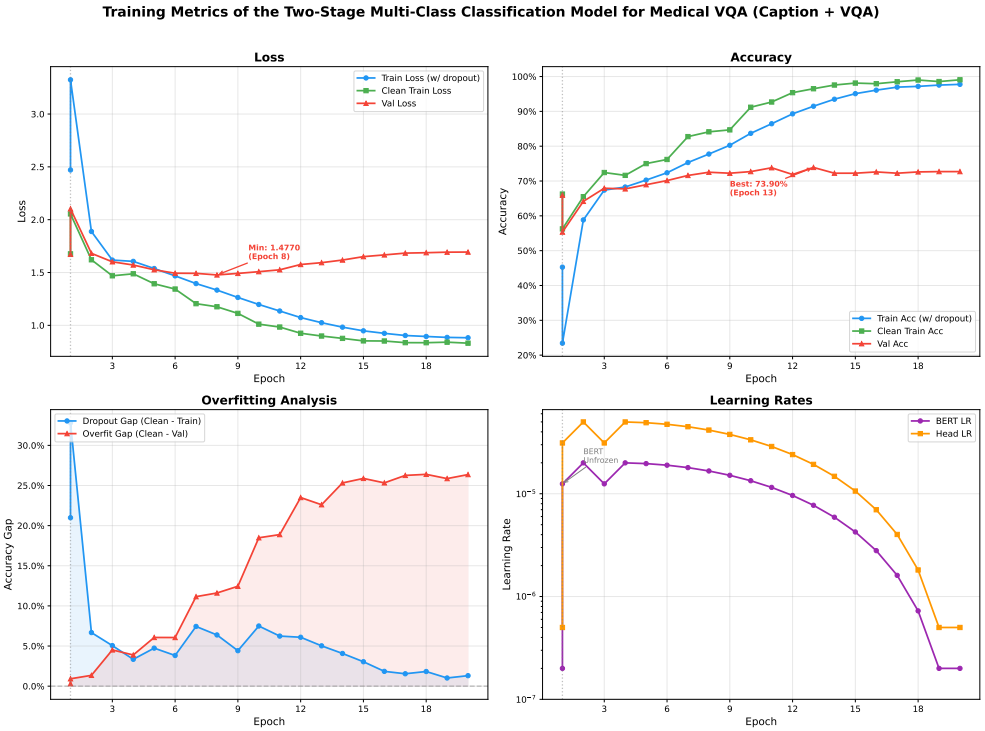


Fig. 6 Training and validation metrics for MedCapVQA on the 71-class multiclass classification task over 20 epochs, for a single representative run. Panels show (top-left) loss curves, (top-right) accuracy curves, (bottom-left) overfitting analysis via dropout and overfit gap metrics, and (bottom-right) learning rate schedules for the BioBERT backbone and classification head.

6 Discussion

6.1 Why Does Describing First Help?

The controlled comparison suggests an explanation for the advantage of MedCapVQA that goes beyond the capacity of its components.

Better encoders alone do not solve the problem. Replacing FastText with BioBERT (Model B) and VGG-19 with Med3D (Model C) changes multiclass validation accuracy by only a few points in either direction and reduces binary validation accuracy relative to Model A. The encoders used by the baselines are therefore not the limiting factor; the difficulty lies in relating image features to question semantics with little supervision.

More expressive fusion helps only partly. Self-attention fusion (Model D) brings no material improvement over concatenation (Model C), and the Vision Transformer model with cross-attention (Model E) improves multiclass accuracy substantially but is the weakest model on binary questions. Richer fusion mechanisms add parameters that must themselves be learned from the same small set of question–answer pairs.

Capacity without the right interface overfits. The end-to-end VLM (Model G) fits the multiclass training set better than any other model but shows the largest gap between multiclass training and validation accuracy, and its binary validation accuracy stays below 60%.

Captions move the hard part to where the data is. In MedCapVQA, visual understanding is delegated to a captioner adapted on image–caption pairs, and the reasoner is left with a text-matching problem between two pieces of biomedical text. We attribute the combination of high accuracy and a small generalisation gap to this division of labour. The multiclass questions in this study concern modality, imaging plane, and organ system, which are attributes that a good medical caption tends to state explicitly; this is consistent with the large margin of MedCapVQA over the feature-fusion baselines on that task.

6.2 Comparison with Published Med-VQA Systems

Table 5 places MedCapVQA alongside established Med-VQA models in terms of training data requirements, computational resources, and reported accuracy. The published figures were obtained with different data splits, question subsets, and answer formulations than ours, so this comparison is indicative rather than like-for-like; the controlled comparison in Section 5 remains the primary evidence.

With these caveats, four observations can be made.

(1) Competitive accuracy with far less task-specific data. MedCapVQA achieves 73.90% validation accuracy on a 71-class task with a reasoner trained on only **8,404 multiclass question–answer pairs**. The published systems in Table 5 draw on substantially larger corpora: 91,685 image–report pairs (∼11× more), over 80,000 image–caption pairs (∼10× more), and about 300,000 image–caption pairs (∼36× more). Despite this difference, MedCapVQA is within about six points of the overall accuracy reported for M3AE on VQA-Med 2019.

Table 5 MedCapVQA alongside published Med-VQA models: pre-training data and reported accuracy. The result for MedCapVQA is accuracy on the official VQA-Med 2019 validation set for the 71-class multiclass task and is not directly comparable with the published figures, which come from different benchmarks and splits.

Model	Pre-training / training data	Reported results
MedViLL [41]	MIMIC-CXR (91,685 image–report pairs)	VQA-RAD: 77.7% (Closed), 59.5% (Open)
PubMedCLIP [55]	ROCO (>80K image–caption pairs)	VQA-RAD: 72.1% overall (80.0% Closed, 60.1% Open); SLAKE: 80.1% overall
M3AE [56]	ROCO + MedICaT (∼300K image–caption pairs)	VQA-Med 2019: 79.9% overall; VQA-RAD: 77.0% overall
MedCapVQA (ours)	**8,404 multiclass VQA-Med 2019 training pairs for the reasoner; trained on 1× V100**	**VQA-Med 2019: 73.90% (71-class multiclass, validation)**

(2) A large answer space. Closed-set classification over 71 answers is harder than yes/no classification and easier than free-form answer generation. For reference, the published systems in Table 5 report closed-ended accuracies of 77.7–80.0% and open-ended accuracies of 59.5–60.1% on VQA-RAD, so the 73.90% obtained by MedCapVQA on a 71-way task lies between the two. A direct comparison on open-ended questions requires a generative answer module and is left for future work.

(3) Single-GPU training. The published systems rely on pre-training with substantially larger corpora; MedViLL, for example, was pre-trained on eight GPUs. The MedCapVQA reasoner was trained on a **single NVIDIA V100 GPU in approximately 5 hours**. This is a significant practical advantage for researchers and clinicians working in resource-constrained environments.

(4) No large-scale Med-VQA pre-training. Models such as MedViLL, PubMedCLIP, and M3AE derive much of their performance from extensive domain-specific vision-language pre-training on large corpora (MIMIC-CXR, PubMed). MedCapVQA needs no such pre-training of its question-answering model. The visual knowledge enters through the captioner, a publicly available medical vision-language model that is adapted once on ROCOv2 by updating about 0.5% of its parameters; the captioning stage distils the visual information into natural language, and the downstream BioBERT-based classifier operates in a purely textual domain.

6.3 Practical Implications

The results suggest three guidelines for the design of Med-VQA systems when annotated question–answer pairs are scarce. First, effort is better spent on the quality of an intermediate description than on increasingly elaborate feature-fusion modules: in our comparison, moving from concatenation to self-attention or cross-attention fusion brought smaller and less consistent gains than moving to captions. Second, separating perception from reasoning makes a system easier to maintain: captions are computed once per image, and the lightweight reasoner can be retrained for a new set of questions

or answer categories without touching the vision model. Third, a textual intermediate representation gives clinicians something to inspect, which addresses part of the "black box" concern raised in the Introduction.

6.4 Limitations

Several limitations should be kept in mind. (i) The evaluation covers closed-ended questions only; abnormality and other open-ended questions are not addressed. (ii) Accuracy is reported on the official VQA-Med 2019 validation set, which was also used for early stopping, rather than on the official test set or on additional benchmarks; and the binary validation set is small (202 questions), so differences of a few points on the binary task should be interpreted with caution. (iii) MedCapVQA can only answer from what the caption contains: a finding that the captioner omits or describes incorrectly cannot be recovered by the reasoner, and we have not yet measured caption quality or its effect on answer accuracy separately. (iv) The training analysis shows that the reasoner begins to overfit after roughly ten epochs, so the results depend on early stopping. (v) The captioner is a 4-billion-parameter model; although it is run once per image, it adds an inference cost that the feature-based baselines do not have.

7 Conclusion

We presented MedCapVQA, a two-stage, caption-driven framework for medical visual question answering that describes the image first and answers the question second. A parameter-efficiently adapted MedGemma captioner turns each image into a clinically meaningful caption, and a BioBERT-based reasoner with dedicated caption and question encoders and question-to-caption cross-attention predicts the answer from text alone.

In a controlled comparison on the closed-ended questions of VQA-Med 2019, MedCapVQA outperformed six baselines spanning convolutional–recurrent fusion, domain-adapted encoders, self-attention fusion, a Vision Transformer model with cross-attention, and an end-to-end fine-tuned vision-language model. It reached 81.9% validation accuracy on binary questions and 73.9% on 71-category multiclass questions, 11.6 and 5.4 percentage points above the strongest baseline on each task, and no baseline was competitive on both. The comparison also showed that domain-adapted encoders and more expressive fusion mechanisms bring limited or inconsistent gains on their own, and that a fine-tuned vision-language model fits the training data well but generalises less well than the caption-driven model. The MedCapVQA reasoner achieves these results with a few thousand question–answer pairs and a single GPU.

These findings indicate that converting medical images into textual captions before reasoning is a strong and resource-efficient alternative to direct multimodal fusion, and that the choice of intermediate representation deserves as much attention as the choice of encoders. Future work will extend MedCapVQA to open-ended and abnormality questions with a generative answer module, evaluate it on the official VQA-Med test set and on further benchmarks such as VQA-RAD, SLAKE, and PathVQA, quantify the contribution of caption quality through ablations of the captioner, investigate

question-conditioned captioning so that the description focuses on what is being asked, and explore stronger regularisation and the integration of external medical knowledge.

Declarations

- **Funding:** The research leading to these results has received partial funding from Emil Aaltonen foundation funded project "Using Language to Interpret Nonstructured Data".
- **Conflict of interest/Competing interests:** Not Applicable
- **Ethics approval and consent to participate:** Not Applicable
- **Consent for publication:** Provided
- **Data availability:** The VQA-Med 2019 (https://zenodo.org/records/10499039) [21] and ROCOv2 (https://huggingface.co/datasets/eltorio/ROCOv2-radiology) [67] datasets used in this study are publicly available, on mentioned repositories.
- **Materials availability:** Not Applicable
- **Code availability:** The code used for the experimentation is provided upon request. The code for the MedCapVQA will be made available on GitHub OA if the manuscript is accepted for publication.

References

[1] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)

[2] Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–9 (2015)

[3] Kim, J.-H., Jun, J., Zhang, B.-T.: Bilinear attention networks. Advances in Neural Information Processing Systems **31** (2018)

[4] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015). PMLR

[5] Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in Neural Information Processing Systems **29** (2016)

[6] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural

Information Processing Systems, vol. 30 (2017)

[8] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W.: VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

[9] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)

[10] Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 5100–5111 (2019)

[11] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)

[12] Dai, W., Li, J., Li, D., et al.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)

[13] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)

[14] Shen, X., Wu, L., Zhang, F., et al.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)

[15] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Towards medical XAI. IEEE Transactions on Neural Networks and Learning Systems **32**(11), 4793–4813 (2020)

[16] Topol, E.J.: High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine **25**, 44–56 (2019)

[17] Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature Medicine **25**, 37–43 (2019)

[18] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)

[19] Naseem, U., Khushi, M., Dunn, A.G., Kim, J.: K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering. IEEE Journal of Biomedical and Health Informatics (2023)

[20] Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.: Overview of ImageCLEF 2018 medical domain visual question answering task. Proceedings of CLEF 2018 Working Notes (2018)

[21] Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In: CLEF 2019 Working Notes (2019). CEUR Workshop Proceedings

[22] Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the VQA-Med task at ImageCLEF 2020: Visual question answering and generation in the medical domain. In: Proceedings of the CLEF 2020 Working Notes (2020)

[23] Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the VQA-Med task at ImageCLEF 2021: Visual question answering and generation in the medical domain. In: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum — Working Notes (2021). 21–24 September 2021

[24] Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 180251 (2018)

[25] He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. In: Machine Learning for Health Workshop, NeurIPS (2020)

[26] Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Piscataway, NJ, pp. 1650–1654 (2021). IEEE

[27] Hu, Y., et al.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. arXiv preprint arXiv:2307.11986 (2023)

[28] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)

[29] Hu, Y., *et al.*: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In: arXiv Preprint arXiv:2402.09181 (2024)

[30] Butsanets, L., Corbière, C., Khlaut, J., Manceron, P., Dancette, C.: RadImageNet-VQA: A large-scale CT and MRI dataset for radiologic visual question answering. In: Medical Imaging with Deep Learning (MIDL). Proceedings of Machine Learning Research, vol. 280 (2026). arXiv:2512.17396

[31] Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A.,

Yang, Y.: RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), 210315 (2022)

[32] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018)

[33] Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)

[34] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13041–13049 (2020)

[35] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 457–468 (2016)

[36] Ben-younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2612–2620 (2017)

[37] Teney, D., Anderson, P., He, X., Van Den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4223–4232 (2018)

[38] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)

[39] Zhan, L.-M., Liu, B., Fan, L., Chen, J., Wu, X.-M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, pp. 2345–2354 (2020). ACM

[40] Yan, A., He, Z., Li, X., Zhang, T., Hall, A., Yamada, A., Naeim, A., Yang, Y., *et al.*: Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)

[41] Moon, S.-J., Lee, H., Choi, M., Kang, J.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. In: IEEE Journal of Biomedical and Health Informatics. IEEE, Piscataway, NJ (2022)

[42] Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: Generation and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1–26 (2021)

[43] Li, G., Zhu, L., Liu, Z., Tang, Y.: Relation-aware graph attention network for visual question answering in medical domain. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, pp. 177–187 (2021). Springer

[44] Zhang, S., Chen, Y., Shen, Z., *et al.*: Multi-modal graph learning for disease prediction. In: IEEE Transactions on Medical Imaging, vol. 41, pp. 2207–2216. IEEE, Piscataway, NJ (2021)

[45] Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 1439–1449 (2020)

[46] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., *et al.*: OSCAR: Object semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision, Cham, pp. 121–137 (2020). Springer

[47] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2021)

[48] Chen, G., Gong, H., *et al.*: Multi-scale cross-modality attention for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham (2022). Springer

[49] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78 (2019)

[50] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical AI. arXiv preprint arXiv:2307.14334 (2023)

[51] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)

[52] Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. In: arXiv Preprint arXiv:1904.00625 (2019)

[53] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking,

B., Sharma, H., Bouzid, K., Thieme, A., *et al.*: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15016–15027 (2023)

[54] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763 (2021). PMLR

[55] Eslami, S., Melo, G., Meinel, C.: Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? In: arXiv Preprint arXiv:2112.13906 (2021)

[56] Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.-H.: Multi-modal masked autoencoders for medical vision and language. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham, pp. 679–689 (2022). Springer

[57] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems (2023)

[58] Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of GPT-3 for few-shot knowledge-based VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36 (2022)

[59] Tiong, A.M.H., Li, J., Li, B., Savarese, S., Hoi, S.C.H.: Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In: Findings of the Association for Computational Linguistics: EMNLP 2022 (2022)

[60] Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

[61] Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in COntext (ROCO): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 180–189. Springer, Cham (2018)

[62] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)

[63] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)

[64] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[65] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR) (2021)

[66] Google DeepMind: MedGemma: Medical Image Understanding with Gemma. https://ai.google.dev/gemma/docs/medgemma. Accessed: 2025 (2025)

[67] Rückert, J., Pelka, O., Friedrich, C.M., *et al.*: ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset. In: arXiv Preprint arXiv:2402.02498 (2024)

[68] Dettmers, T., Pagnoni, A., Farhadi, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems (2023)

[69] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)

[70] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

[71] Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). In: arXiv Preprint arXiv:1606.08415 (2016)

[72] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. In: arXiv Preprint arXiv:1607.06450 (2016)

[73] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010). JMLR Workshop and Conference Proceedings

[74] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen,

W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)

[75] Dettmers, T., Lewis, M., Shleifer, S., Zettlemoyer, L.: 8-bit optimizers via block-wise quantization. International Conference on Learning Representations (ICLR) (2022)

[76] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

[77] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[78] Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)

[79] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[80] Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 328–339 (2018)

[81] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

[82] Lipping, S., Sudarsanam, P., Drossos, K., Virtanen, T.: Clotho-AQA: A crowdsourced dataset for audio question answering. In: 2022 30th European Signal Processing Conference (EUSIPCO), Piscataway, NJ, pp. 1140–1144 (2022). IEEE

[83] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015)

[84] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, pp. 248–255 (2009). IEEE

[85] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. In: Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146 (2017)

[86] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[87] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)

[88] Yang, A., Yang, B., Zhang, B., Hui, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)