applied sciences

MDPI

Review

Generative Models in Medical Visual Question Answering: A Survey

Wenjie Dong [1], Shuhao Shen [1], Yuqiang Han [1], Tao Tan [2], Jian Wu [1] and Hongxia Xu [1,*]

[1] College of Computer Science & Technology and Institute of Wenzhou, Zhejiang University, Hangzhou 310012, China; wenjiedong@zju.edu.cn (W.D.); 3210103717@zju.edu.cn (S.S.); wujian2000@zju.edu.cn (J.W.)
[2] Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China; taotan@mpu.edu.mo
* Correspondence: einstein@zju.edu.cn

Abstract: Medical Visual Question Answering (MedVQA) is a crucial intersection of artificial intelligence and healthcare. It enables systems to interpret medical images—such as X-rays, MRIs, and pathology slides—and respond to clinical queries. Early approaches primarily relied on discriminative models, which select answers from predefined candidates. However, these methods struggle to effectively address open-ended, domain-specific, or complex queries. Recent advancements have shifted the focus toward generative models, leveraging autoregressive decoders, large language models (LLMs), and multimodal large language models (MLLMs) to generate more nuanced and free-form answers. This review comprehensively examines the paradigm shift from discriminative to generative systems, examining generative MedVQA works on their model architectures and training process, summarizing evaluation benchmarks and metrics, highlighting key advances and techniques that propels the development of generative MedVQA, such as concept alignment, instruction tuning, and parameter-efficient fine-tuning (PEFT), alongside strategies for data augmentation and automated dataset creation. Finally, we propose future directions to enhance clinical reasoning and intepretability, build robust evaluation benchmarks and metrics, and employ scalable training strategies and deployment solutions. By analyzing the strengths and limitations of existing generative MedVQA approaches, we aim to provide valuable insights for researchers and practitioners working in this domain.

Keywords: medical visual question answering; multimodal representation learning; vision–language pretraining; large language models; multimodal large language models

Academic Editors: Paolino Di Felice, Chilukuri K. Mohan, Aleksander Mendyk and Luis Javier García Villalba

Received: 10 February 2025
Revised: 28 February 2025
Accepted: 5 March 2025
Published: 10 March 2025

Citation: Dong, W.; Shen, S.; Han, Y.; Tan, T.; Wu, J.; Xu, H. Generative Models in Medical Visual Question Answering: A Survey. *Appl. Sci.* **2025**, *15*, 2983. https://doi.org/10.3390/app15062983

1. Introduction

Medical Visual Question Answering (MedVQA) involves responding to questions related to medical images [1]. This emerging field lies at the intersection of artificial intelligence (AI) and healthcare, where machine learning models are developed to interpret and analyze medical imagery, such as X-rays, MRIs, CT scans, and pathology slides [2]. MedVQA aims to enhance clinical decision-making by providing accurate and contextually relevant answers to questions posed by clinicians, thereby alleviating medical professionals' burden, reducing diagnostic errors, and ultimately improving patient outcomes [3–5]. A fully developed MedVQA system can also benefit patients by integrating into online consultation platforms. It provides reliable answers in scenarios where medical professionals are unavailable, such as remote or automated health examinations. This capability helps reduce the risk of misinformation that can arise from online searches [4–6]. As a

significant advancement in multimodal learning, MedVQA addresses societal challenges by expanding access to high-quality healthcare and fostering informed decision-making.

A typical MedVQA framework consists of four key components: an image feature extractor, a question feature extractor, a feature fusion module, and an answering component [4]. The ImageClef 2018 Challenge first introduced MedVQA task and received five model submissions, with most participants employing deep learning techniques, leveraging sequence-to-sequence learning, encoder–decoder frameworks, CNNs for image encoding, RNNs for question processing, and stacked attention networks (SAN), multimodal compact bilinear (MCB) pooling or multimodal factorized bilinear (MFB) pooling for modality fusion [1]. To address data scarcity, meta-learning was introduced to enhance model training with limited labeled data [7]. Conditional reasoning was later explored to improve modality fusion [8]. Efforts to refine visual and textual representations led to the integration of contrastive learning in PubMedCLIP [9] and masked language modeling (MLM) pretraining tasks in MMBERT [5]. Further advancements leveraged multi-task learning for image feature pretraining, enhancing image comprehension [10]. MedViLL introduced a novel vision–language pretraining (VLP) strategy, incorporating MLM, image-report matching (IRM), and a bi-directional auto-regressive self-attention mask to improve vision–language understanding [11].

While substantial research efforts have been directed towards advancing MedVQA, the predominant focus of current endeavors lies in discriminative approaches. These models predict the most appropriate answer from a predefined set of candidates from the training set, which limits their ability to generate free-form responses, particularly when addressing out-of-domain or open-ended questions [3,12]. Consequently, such models struggle to harness the full potential of MedVQA, especially in complex and less structured scenarios in real-world clinics, calling for more solutions to the challenge [13–15]. On the other hand, recent advancements in LLMs, particularly transformer-based architectures, have catalyzed a significant paradigm shift in MedVQA. These models, renowned for their ability to handle complex natural language tasks, have driven the transition from traditional discriminative approaches to generative frameworks, enabling MedVQA systems to produce open-ended, contextually nuanced answers beyond predefined categories.

Generative MedVQA, in comparison with the discriminative MedVQA framework, uses a generator instead of a classifier as the answer production module, as shown in Figure 1. In as early as 2020, CGMVQA first added a generator branch to deal with open-ended questions, producing answers with beam search algorithm, while keeping the classifier for closed-ended questions [6]. In the following two years, MedFuseNet [16] and MED-GPVS [17] also tried out different answer generators. However, it was not until 2023 that generative MedVQA approaches emerged buoyantly, making use of transformer decoders as well as LLMs and MLLMs. This development aligns with the broader trend of incorporating LLMs and MLLMs into medical AI, reflecting a transformative approach to addressing complex, real-world clinical scenario.

Given the recent rise of generative approaches in MedVQA, it is timely and valuable to pause and critically examine the paradigm shift from discriminative to generative models. Existing surveys and reviews on MedVQA have primarily focused on discriminative methods or broader aspects of medical VQA [4]. To address this gap, this review aims to summarize and analyze the key developments in the integration of generative models into MedVQA, providing an overview of their advantages. Although there have been a few survey papers on the subject of medical VQA [4,18,19], this work is, to our knowledge, the first to review it with an emphasis on generative models and underscore the paradigm shift from the classificative to generative models.

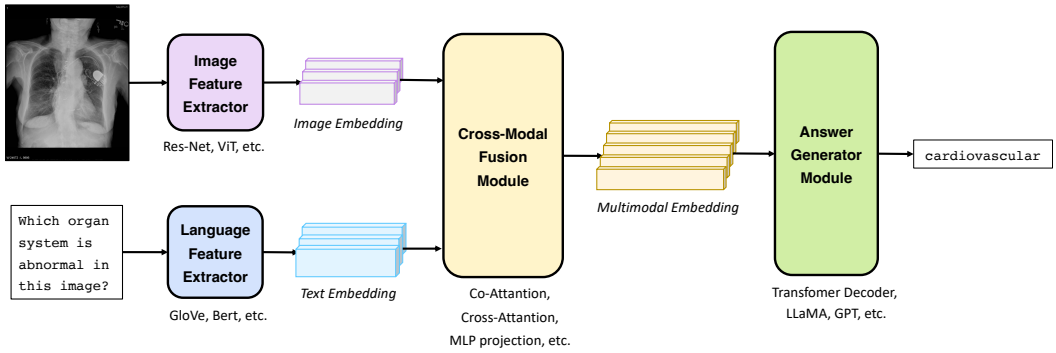

Figure 1. A representative illustration of generative MedVQA systems. Similar to classificative MedVQA systems, generative MedVQA systems typically consists of an image encoder which encodes image input, a text encoder which encodes the question input, a modality fusion module which aligns different modalities, and an answer generator that produces the answer.

By exploring the evolving landscape of MedVQA, this paper seeks to highlight the significant paradigm shift from discriminative to generative models and review techniques adopted in generative MedVQA methods. We conclude by presenting available pretraining datasets and MedVQA benchmarks as well as introducing relevant automated dataset building and data augmentation techniques.We detail the techniques employed in the vision–language pretraining and the direction to build one unified model for multiple multimodal and unimodal tasks. Then, we elaborate the techniques related to adopting large language models (LLMs) and multimodal large language models (MLLMs), namely instruction tuning, domain adaptations, and parameter-efficient fine-tuning. Last but not least, we discuss the remaining challenges and future directions of generative MedVQA from the aspects of high-quality dataset, proper evaluation metrics, clinical reasoning, efficiency, LLM hallucination and application in real-world clinical settings. In doing so, this review will offer a comprehensive understanding of the motivations behind the shift to generative models and the future directions of generative MedVQA.

2. The Evolution of Generative MedVQA

2.1. The Paradigm Shift: From Discriminative to Generative Models

The evolution of MedVQA has seen a fundamental shift from discriminative-based approaches to generative models. Early MedVQA systems relied on classification architectures that selected answers from predefined sets, limiting their ability to address complex, open-ended medical queries that require detailed, context-sensitive responses [3,12]. While these models demonstrate success in structured tasks such as binary and multiple-choice question answering, they struggle with flexibility and lacked the capability to generate nuanced, patient-specific explanations [13,14]. As clinical demands grew, it becomes evident that a more expressive and adaptable approach is necessary to enhance AI-driven medical reasoning [15].

This shift has been driven by advancements in generative architectures, allowing models to move beyond rigid classification frameworks toward decoder-based approaches capable of producing free-form, context-sensitive answers. Early generative models, such as MedFuseNet [16] and CGMVQA [6], combined classification and generative techniques, using LSTM-based decoders with beam search, offering a transition between fixed answer selection and open-ended text generation. The introduction of transformer-based decoders, including BART [20] and T5 [21], further refined generative MedVQA by improving long-range dependencies and sequence-to-sequence learning [22,23].

More recently, the integration of LLMs and MLLMs has significantly expanded the capabilities of MedVQA, becoming a popular direction as shown in Table 1. These models

leverage large-scale pretraining [24–26], instruction tuning [14,22,24,25], and multimodal alignment [27,28] techniques to produce more contextually relevant and clinically useful responses. The development of parameter-efficient fine-tuning (PEFT) methods, such as prefix tuning, LoRA, and visual instruction tuning, has further enabled LLMs to be adapted for medical reasoning with reduced computational overhead [12,13,29].

The diversity of generative MedVQA model architectures, fusion techniques, and training strategies employed in generative MedVQA is systematically summarized in Table 2, providing a comparative overview of key works in the field following the chronological order. Over the years, MedVQA has evolved from early encoder–decoder approaches that employed CNN-based vision encoders (e.g., VGG, ResNet) and LSTM text decoders to increasingly sophisticated transformer-based architectures. Early works typically fused image and text features through simple concatenation or shallow attention mechanisms, as seen in TUA1 and CGMVQA. More recent models, including MedVInT, Uni-Med, and LLaVA-Med, integrate pretrained vision encoders (e.g., CLIP-ViT, EVA) with large language models (e.g., GPT-2, LLaMA) via cross-attention or projection modules, reflecting the broader AI trend toward multimodal alignment and instruction tuning. The emergence of frameworks like Med-Flamingo, MLeVLM, and MEDIFICS further underscores the shift toward MLLMs, emphasizing task-specific adapters and parameter-efficient fine-tuning methods (e.g., LoRA, QLoRA) to handle domain complexity without incurring prohibitive training costs.

This transformation represents a significant leap in AI-driven medical question answering, enabling models to generate richer, more interpretable, and adaptable answers that better align with real-world clinical needs. Figure 2 provides a timeline of generative MedVQA works alongside selected discriminative models, offering an overview of the field's evolution. From a chronological perspective, the burst of generative frameworks takes place in 2023, two years after the MedVQA surge, largely driven by the trend of LLMs and MLLMs. The number of working employing LLMs and MLLMs has grown to 13, surpassing that of those without. As listed in Table 1, works adopting LLMs/MLLMs become the mainstream in 2024. The following subsections delve into the evolution of generative MedVQA with a detailed explanation of the model architecture and training processes.

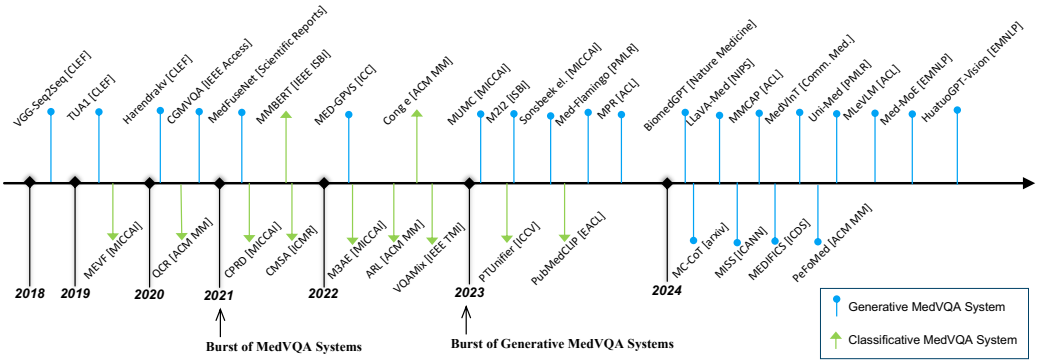

Figure 2. A timeline of generative MedVQA systems along with milestone discriminative MedVQA systems.

Table 1. Chronological list of generative MedVQA works.

Category	2018	2019	2020	2021	2022	2023	2024
Works without LLMs/MLLMs	[30]	[31]	[6,32]	[16]	[17]	[23,33,34]	[15,22]
Works with LLMs/MLLMs	-	-	-	-	-	[13,25]	[12,14,24,26–29,35–38]

2.2. Early Explorations of Generative MedVQA

The early exploration of generative models in MedVQA marked a shift from traditional discriminative-based systems to fully generative systems, with an intermediate phase featuring hybrid approaches that combined both classification and generation. Some early models [30,32] were fully generative systems that used decoders to generate answers for both open-ended and closed-ended questions. These models were pioneering efforts in exploring how generative approaches could address the complex and variable nature of medical question answering.

VGG-Seq2Seq [30], proposed for the ImageCLEF 2018 task [1], adopts an encoder–decoder architecture. The model combines a pretrained VGG network with an LSTM for encoding image and question inputs, followed by a simple concatenation of the visual and textual embeddings. The decoder, also LSTM-based, generates the answer sequentially.

The VQA-Med 2020 task introduced a sequential model by Harendra K. Verma and Sindhu Ramachandran S. [32]. This model uses VGG16 for image feature extraction and BERT for question encoding. It combines these features using Multimodal Factorized Bilinear Pooling (MFB) and refines them with self-attention. The decoder, based on LSTM, generates answers word-by-word, utilizing Bahdanau attention and GLOVE embeddings.

MED-GPVS [17], a model for both MedVQA and object detection, uses ResNet50 and BERT for image and text feature extraction, respectively. The image feature is processed through DEtection TRansformer (DETR) to extract area descriptors. Vision and language features are fused in ViLBERT's co-attention layers, followed by a text decoder that generates answers for MedVQA tasks.

In contrast, some models took a hybrid approach, using both classification and generation to handle different types of questions [6,16,31]. These models demonstrated the flexibility of combining classification and generation but still had to balance the strengths and weaknesses of both approaches.

TUA1 [31] is a hybrid system that combines classification and generation models designed for the ImageCLEF VQA-Med 2019 task [39]. The model leverages Inception-ResNet-V2 for image feature extraction and BERT for question feature processing. The model concatenates the image and question embeddings and then uses classifiers for modality, plane, and organ system questions, while a sequence-to-sequence generator handles "abnormality" questions using LSTM with beam search.

CGMVQA [6] is another hybrid model that performs both classification and answer generation tasks in a unified framework. It utilizes ResNet-152 for image feature extraction and simple concatenation for image–text fusion. The model employs a lightweight transformer-based architecture and is trained with a combination of cross-entropy loss for classification and mask-based loss for generative tasks. A beam search algorithm is incorporated for sequential answer generation.

Lastly, another hybrid framework MedFuseNet [16] is a multimodal attention-based model. It uses ResNet-152 for image feature extraction and BERT for question encoding, with Multimodal Factorized Bilinear Pooling (MFB) for fusion. The answer generator employs an LSTM-based decoder with attention mechanisms and beam search for generating answers to open-ended "abnormality" questions in datasets like MED-VQA-2019 and PathVQA.

These early generative models paved the way for the more advanced, large-scale models seen today. While each model introduced unique innovations, the common challenge was to balance the flexibility of generative systems with the need for structured and interpretable answers in the medical domain.

2.3. Integrations of Vision–Language Pretraining with Various Answer Generators

The advancement of generative models in the general domain has led to several MedVQA systems exploring different answer generator architectures, such as BART [20,22] and T5 [21,23]. Meanwhile, starting in 2021, vision–language pretraining (VLP) became a dominant strategy in building MedVQA systems. These generative models not only focused on answer generation but also incorporated VLP to improve image–text alignment, enabling more efficient handling of multimodal data.

One notable model, M2I2 [34], integrates multimodal learning by unifying pretraining objectives across image and text features. The model, pretrained with tasks like Masked Image Modeling (MIM), Masked Language Modeling (MLM), Image–Text Matching (ITM), and Image–Text Contrastive Learning (ITC) on the ImageCLEF2022 dataset [40], consists of a 12-layer Vision Transformer (ViT) as the image encoder and a 6-layer transformer initialized with BERT for text encoding. It further integrates a 6-layer transformer with cross-attention for multimodal fusion and a 6-layer transformer-based decoder for answer generation.

Another approach, MUMC [33], introduces Unimodal and Multimodal Contrastive Losses (MUMC) to complement pretraining tasks like MLM and ITM. The Unimodal Contrastive Loss (UCL) differentiates examples within the same modality (either image or text), while the Multimodal Contrastive Loss (MCL) aligns image and text embeddings in a shared space [33]. The architecture mirrors M2I2, and the model is pretrained on multiple datasets, including ROCO [41], MedICaT [42], and ImageCLEF2022 [40].

The Multimodal Prompt Retrieval (MPR) framework [23] leverages a T5 transformer [21] for auto-regressive answer generation. The framework integrates a retrieval module that selects relevant multimodal prompts from a pre-indexed database of medical image–text pairs. These prompts are then used as additional context during inference. The vision–language encoder, based on CLIP-ViT, extracts visual embeddings aligned with retrieved text embeddings for seamless multimodal fusion. A generative decoder, implemented using a pretrained GPT model, generates open-ended answers conditioned on the question, retrieved prompts, and encoded visual features. The pretraining follows a two-stage strategy: the first stage trains the model with self-supervised contrastive learning tasks, aligning vision and language features using a synthetic dataset of medical images with associated captions and metadata from ROCO; while the second stage fine-tunes the model on MedVQA datasets, such as Slake and PathVQA, using a retrieval-augmented generative learning objective.

BiomedGPT [22] is a unified encoder–decoder model designed to handle vision, language, and multimodal tasks. Based on the BART architecture [20], it integrates a BERT-style encoder for bidirectional understanding and a GPT-style autoregressive decoder for text generation. Pretraining tasks include MIM, MLM, image captioning, VQA, and object detection (OD). The model is pretrained on a wide range of datasets, including CheXpert [43], CytoImageNet [44], Slake, PathVQA, Peir Gross [45], IU X-ray [46], PubMed Abstracts [47], MIMIC-III Clinical Notes [48,49], and NCBI BioNLP Corpus, among others.

The MISS (Multi-task Self-Supervised-learning-based) model [15] adopts a Joint Text-Multimodal (JTM) encoder, which unifies text and multimodal feature extraction to replace traditional dual-tower architectures. The model incorporates a ViT-based image encoder for visual feature extraction and a transformer-based text decoder for generating free-form answers. To address the challenge of limited multimodal medical datasets, MISS introduces Transfer-and-Caption (TransCap), a technique that leverages LLMs to generate structured captions for unimodal medical images, effectively expanding pretraining data. The model is pretrained with Image–Text Contrastive Learning (ITC), Image–Text Matching (ITM),

and Masked Language Modeling (MLM) to enhance image–text representation alignment, followed by fine-tuning on VQA-RAD and Slake.

These advancements in generative MedVQA systems, particularly with the integration of VLP and diverse answer generation architectures, have significantly enhanced the ability of models to handle complex medical question answering. The shift toward unifying vision and text features within multimodal encoders, combined with the use of generative decoders, has led to more flexible, interpretable, and context-aware models. Moving forward, the combination of VLP with generative models will continue to improve the precision and applicability of MedVQA systems in clinical settings, especially as they adapt to new, large-scale medical datasets.

2.4. Explorations of Employing LLMs

The integration of LLMs into MedVQA has significantly enhanced the ability of models to generate flexible, context-aware responses. Many works integrate frozen LLMs with vision encoders via a simple MLP projector, requiring efficient fine-tuning techniques to bridge the gap between modalities. Recent works employ LoRA, prefix tuning, and instruction tuning to adapt LLMs without excessive computational overhead. The other challenge is medical knowledge integration, where models incorporate domain-specific knowledge, instruction datasets, and staged training pipelines to improve contextual understanding. This subsection reviews various methods for employing LLMs in MedVQA, examining how researchers optimize model architectures and training strategies to improve performance while maintaining efficiency.

Van Sonsbeek et al. [13] employ prefix tuning to adapt LLMs to medical VQA tasks. Their model follows a two-stream encoder–decoder framework, where a vision encoder (e.g., CLIP with a ViT backbone) extracts image features that are mapped into learnable visual tokens via an MLP-based mapping network. These tokens are combined with tokenized question embeddings in the language encoder before passing into an LLM decoder (GPT-2 XL, BioGPT, or BioMedLM) for autoregressive answer generation. To overcome the limitation of small medical datasets, parameter-efficient fine-tuning (PEFT) methods such as prefix tuning, prompt tuning, and LoRA are used to efficiently adapt the model without extensive full-model fine-tuning. The pretraining process leverages medical text corpora and optimizes for MLM, ITM, and contrastive learning to enhance vision–language alignment.

Ha et al. [12] propose a transformer-based fusion framework inspired by BLIP-2 [50], integrating a biomedical-adapted vision encoder with an LLM specialized for radiology. The model employs either BiomedCLIP-ViT or PMC-CLIP ResNet50 as the vision encoder, while RadBloomz-7B serves as the LLM, with an MLP projector aligning their embeddings. The training process first focuses on medical concept alignment via image-caption pretraining on PMC-OA. The model then undergoes general MedVQA pretraining on PMC-VQA to enhance domain-specific reasoning. The final stage involves fine-tuning on radiology datasets like VQA-RAD and Slake to improve medical image interpretation. LoRA is applied to fine-tune the LLM while keeping the vision encoder frozen, ensuring efficient adaptation.

PeFoMed [29] adapts LLMs with PEFT techniques for both MedVQA and medical report generation (MRG). The model consists of a frozen ViT-based EVA encoder, LLaMA2-7B for text generation, and a lightweight linear projection layer to align image and text representations. The training process includes image captioning pretraining using ROCO, CLEF2022, MEDICAT, and MIMIC-CXR for vision–language alignment, followed by task-specific fine-tuning on MedVQA datasets (VQA-RAD, Slake, PathVQA) and MRG datasets

(IU-Xray). By fine-tuning only the projection layer and LoRA adapters, the model significantly reduces computational costs while preserving pretrained knowledge.

MedVInT (Medical Visual Instruction Tuning) [14] introduces a visual instruction-tuning framework to bridge a pretrained vision encoder with an LLM. The visual encoder (PMC-CLIP-ResNet) maps extracted image features into a unified embedding space via a trainable projection module. The language encoder (PMC-LLaMA) encodes the input question with a structured prompt to guide response generation. The multimodal decoder has two variants: MedVInT-TE (encoder-based transformer), which reformulates answer generation as MLM, and MedVInT-TD (decoder-based transformer), which directly generates free-form text answers. The pretraining consists of PMC-VQA alignment, followed by fine-tuning on benchmarks like VQA-RAD and Slake.

MMCAP (Multimodal Concept Alignment Pretraining) [28] integrates medical knowledge graphs (KGs) with vision–language pretraining to improve knowledge-intensive MedVQA. Its architecture consists of a CLIP-based ResNet vision encoder, a BioMedBERT text encoder, and a Graph Attention Network (GAT) that encodes UMLS-based KGs for aligning medical concepts with images. A Knowledge Adapter maps visual features into the language model's embedding space using learnable vision tokens and transformer decoders. The training consists of two stages: pretraining on ROCO and MedICaT datasets using contrastive learning and masked modeling, followed by fine-tuning on VQA-RAD and Slake with Type Conditional Prompting (TCP) for structuring responses based on question type.

MLeVLM [27] introduces a multi-level feature alignment (MLFA) module to improve the interaction between visual and textual representations. The model employs EVA-G as the image encoder, Q-Former as the text encoder, and Vicuna-7B as the answer generator. The training process begins with medical modality alignment, pretraining the MLFA module on large multimodal datasets to align image–text representations. It then undergoes medical instruction-tuning, using LLaVA-Med instruction dataset [24] to improve domain-specific understanding. The final stage involves level instruction-tuning, optimizing on MLe-VQA-60K to enhance hierarchical reasoning. LoRA-based fine-tuning is applied to efficiently adapt the model, ensuring robust recognition, diagnosis, and medical reasoning while maintaining computational efficiency.

Med-MoE [37] is a lightweight multimodal model that introduces a Mixture-of-Experts (MoE) architecture with domain-specific experts. The model consists of a CLIP-ViT vision encoder, a Joint Text-Multimodal (JTM) encoder, and a meta-expert that remains active to provide global context. The training process focuses on multimodal medical alignment, where image and text features are jointly trained, followed by instruction tuning, where the model learns task-specific prompts. The final stage involves domain-specific MoE tuning, where experts are fine-tuned for specialized medical tasks while the router remains fixed. By selectively activating experts during inference, Med-MoE reduces computational overhead, achieving competitive performance with fewer parameters than traditional large models like LLaVA-Med [24].

The integration of LLMs into MedVQA has demonstrated the effectiveness of PEFT in adapting text-based models to multimodal tasks. A recurring theme is the use of projection layers to align vision–language representations, along with structured training pipelines to refine medical reasoning. The shift towards instruction tuning and knowledge-enhanced MedVQA highlights the growing importance of domain-specific adaptations for improving model interpretability and performance. Future research is expected to refine multimodal fusion techniques, domain-specific adaptation, and efficient inference strategies, paving the way for more accurate and interpretable MedVQA systems.

2.5. Explorations of Working with MLLMs

MLLMs provide a natural extension of LLMs into the multimodal space, allowing models to process images and text natively. Unlike LLM-based MedVQA approaches that require explicit vision–language alignment, MLLMs already incorporate cross-modal understanding as part of their pretraining. The focus in MedVQA research has shifted towards optimizing these models through instruction tuning, modular architectures, and knowledge augmentation. By refining existing MLLMs such as Flamingo, LLaVA, and IDEFICS, researchers have enhanced zero-shot generalization, reasoning capabilities, and task adaptability. The key difference between the approaches discussed here and those in the previous section is that these works leverage existing MLLMs as the foundation, modifying them for better domain alignment and interpretability. The primary strategies, however, overlaps with those in the previous subsection, involving instruction tuning, knowledge augmentation, modular reasoning frameworks, and PEFT techniques to optimize these MLLMs for medical applications. This subsection reviews how MLLMs are being leveraged for MedVQA, examining novel training methodologies and fine-tuning strategies that maximize performance across medical datasets.

Med-Flamingo [25] builds on the OpenFlamingo-9B framework, adapting it specifically for the medical domain. The architecture integrates a frozen LLaMA-7B language model and CLIP ViT/L-14 vision encoder through gated cross-attention layers and perceiver layers, specializing in few-shot settings. The model consists of 1.3 billion trainable parameters focused on cross-attention and perceiver layers, while 7 billion parameters remain frozen within the decoder and vision encoder, totaling 8.3 billion parameters. Med-Flamingo undergoes continued pretraining using paired and interleaved image–text data from PMC-OA [51] and the constructed Medical Textbook Dataset (MTB), aligning visual and textual representations for improved medical reasoning. The model is trained with memory-efficient optimizations, ensuring computational efficiency and scalability.

LLaVA-Med [24] extends LLaVA from the general domain to medical applications. It employs a linear projection layer to map visual features extracted by a vision encoder (e.g., CLIP-ViT) into the embedding space of an LLM, functioning similarly to visual prefix tuning. The model is trained in two stages using curriculum learning. The first stage aligns biomedical concepts by training on 600K image–caption pairs sampled from PMC-15M, updating the projector to align visual and textual embeddings. The second stage involves instruction tuning with a dataset of 60K GPT-4-generated instruction-following samples, enabling the model to handle open-ended, conversational MedVQA tasks.

MC-CoT [35] introduces a Modular Collaborative Chain-of-Thought (CoT) framework designed to improve zero-shot MedVQA by leveraging LLMs for reasoning and MLLMs for image interpretation. The architecture comprises three specialized medical modules: a radiology module for imaging modality and lesion localization, an anatomy module for structural identification, and a pathology module for clinical significance assessment. The workflow begins with an LLM-guided task assignment, where the LLM receives the question and a caption of the image to provide strategic guidance. The MLLM then extracts image-specific features before the LLM synthesizes the final answer. Unlike traditional fine-tuned MedVQA models, MC-CoT does not require task-specific pretraining, making it highly adaptable for zero-shot settings.

HuatuoGPT-Vision [26] is a medical MLLM built on Yi-VL-34B [52], integrating medical visual knowledge into its architecture. To enhance domain-specific knowledge alignment, it is pretrained with structured knowledge-injection approach on the constructed PubMedVision dataset containing 1.3 million medical image–text pairs refined using GPT-4V-assisted data curation. The training process begins with PubMedVision Alignment VQA for image–text understanding, followed by PubMedVision Instruction-Tuning VQA

for task-specific optimization. This structured knowledge-injection approach enables state-of-the-art performance on VQA-RAD, Slake, and PMC-VQA, significantly improving zero-shot generalization and medical diagnostic reasoning.

MEDIFICS [38] introduces a model-calling mechanism to dynamically select specialized models for different medical tasks. Built on IDEFICS-9B-Instruct [53], the architecture includes an OpenCLIP vision encoder and a LLaMA-based language model, aligned using a Perceiver Resampler for vision–text fusion. The external model-calling system allows MEDIFICS to invoke domain-specific models such as ConvNeXt for classification and DINOV2 for self-supervised learning, improving accuracy across tasks. The model is fine-tuned using QLoRA, where only attention layers are updated, optimizing memory efficiency. To expand training data, synthetic doctor-patient conversations generated by GPT-3.5-Turbo are incorporated, further improving generalization in clinical scenarios.

The Uni-Med framework [36] introduces a Universal Instruction-to-Answer Navigator (IAI) to improve interpretability and zero-shot capabilities in MedVQA. Its architecture consists of three key modules: the Instruct-to-Answer Clues Interpreter (IAI), which generates visual explanations and refines user query understanding; the Task-guided Token-Level Cut-Mix (TC-Mix), which enhances token-level feature alignment by associating visual explanations with image regions to improve answer traceability; and the Intention-guided Attention (IGA), which dynamically reduces attention weights for non-core instructions, allowing the LLM to focus on essential aspects of the question. The pretraining process involves constructing the IAI-Med VQA dataset, where the model is fine-tuned on annotated medical images with structured instruction-to-answer mappings. The framework employs a Universal-Navigator Prompt (UNP) to guide MLLMs in generating step-by-step explanations alongside their answers, ensuring structured reasoning and improved model interpretability.

A key theme among MedVQA systems involving MLLMs is progressive multimodal alignment, where models undergo stage-wise training to refine visual-textual interactions before task-specific fine-tuning. Additionally, zero-shot capabilities and efficient fine-tuning strategies are recurring priorities, as seen in MC-CoT, Uni-Med, and MEDIFICS, which explore methods to reduce reliance on extensive labeled medical datasets. Moving forward, the combination of adaptive task routing, hierarchical instruction tuning, and model-calling mechanisms is likely to further enhance the flexibility and efficiency of MLLMs in MedVQA.

Table 2. A summary of generative MedVQA works.

Model	Image Encoder and Text Encoder	Modality Fusion Module	Answer Generator	Pretraining Tasks and Datasets
VGG-Seq2Seq [30]	VGG + LSTM	Concatenation	LSTM	NA
TUA1 [31]	Inception-ResNet-V2 + BERT	Concatenation	Classifier + LSTM	NA
Harendrakv [32]	VGG16 + BERT	Attention + MFB	GLOVE + LSTM + Attention	NA
CGMVQA [6]	ResNet-152 + BERT	Concatenation	Classifier + LSTM	NA
MedFuseNet [16]	ResNet-152 + BERT	Image Attention + MFB + Co-Attention	Classifier + (LSTM + Attention)	NA
MED-GPVS [17]	(ResNet-50 + DETR) + (BERT + ViLBERT)	Co-attention	Text Decoder	NA
MUMC [33]	ViT + BERT	Cross-Attention	Text Decoder	ROCO, MIMIC-CXR, MedICaT
M2I2 [34]	ViT + BERT	Cross-Attention	Text Decoder	MIMIC-IV, PathVQA; self-supervised masked learning
MPR [23]	CLIP-ViT-B/32 + T5 encoder	Multimodal embedding retrieval	T5-small	Retrieval set augmentation on ROCO-sythesized VQA data
Sonsbeek et al. [13]	CLIP-ViT + Text tokenizer of the LLM	MLP projector	GPT2-XL, BioGPT and BioMedLM	NA
Med-Flamingo [25]	ViT/L-14 + Text tokenizer of the LLM	Gated cross attention layers and perceiver layers	LLaMA-7B	Pretrained on MTB and PMC-OA, initialized with Open-Flamingo-9B
LLaVA-Med [24]	CLIP-ViT + Text tokenizer of the LLM	MLP projector	LLaMA	Biomedical concept feature alignment on filtered PMC-15M Instruction tuning on GPT-4 generated data from PMC-15M
BiomedGPT [22]	ResNet + BPE and Token Embedding	Transformer encoder	GPT-style autoregressive decoder	Image infilling on CheXpert, CytoImageNet, ISIC, Retinal Fundus OD on DeepLesion, OIA-DDR MLM on PubMed Abstracts, NCBI BioNLP, MIMIC-III Clinic Notes Image captioning on MedICat, IU X-ray, Peir Gross VQA on Slake, PathVQA
MedVInT [14]	PMC-CLIP-ResNet + Text tokenizer of the LLM	MLP/transformer-based projection	PMC-LLaMA	Instruction-tuned on PMC-VQA dataset
Uni-Med [36]	PMC-CLIP-ViT + instruction-based LLM	Task-Guided Embedding Tokenizer (TET) + Task-Guided Token-CutMix + Intention-Guided Attention + Projection	PMC-LLaMA	Pretrained on PMC-VQA
Ha et al. [12]	BiomedCLIP ViT + Text tokenizer of the LLM	Linear projector	RadBloomz-7b	Medical Concept Alignment on PMC-OA Fine-tuning on on the PMC-VQA
PeFoMed [29]	EVA (ViT)+ Text tokenizer of the LLM	Linear projector	LLM	Fine-tuning on ROCO, CLEF2022, MEDICAT and MIMICCXR
Med-MoE [37]	CLIP-ViT + Text tokenizer of the LLM	MLP projection,	MoE (Router+LLMs)	Multimodal Medical Alignment on LLaVA-Med alignment dataset Instruction Tuning and Routing on LLaVA-Med instruction-tuning dataset Domain-Specific MoE Tuning on VQA-RAD, Slake, PathVQA
MLeVLM [27]	EVA-G (ViT) + Q-former	Multi-Level Feature Alignment (MLFA) module: Attention-based Token Selector + Context Merger	Vicuna-7B	Medical modality alignment on MedMINIST, Medicat, PMCVQA Medical instruction-tuning on LLaVA-Med instruction-tuning dataset Level instruction-tuning on MLe-VQA
MISS [15]	Image + JTM encoder	Casual Attention + Cross Attention	Text decoder	ITC, ITM, MLM, momentum model MoCo update
MEDIFICS [38]	OpenCLIP + Text tokenizer of the VLM	Perceiver Resampler, cross-attention	IDEFICS-9B-Instruct (based on Flamingo)	QLoRA, model calling fine-tuning on ISIC, ROCO, MURA
MMCAP [28]	PMC-CLIP-ResNet + Text tokenizer of the VLM	Knowledge Adapter (Transformer Decoder)	GPT-2	Multimodal Concept Alignment Pretraining
HuatuoGPT-Vision [26]	CLIP-Large + Text tokenizer of LLM	MLP/transformer-based projection	LLaMA	Pretrainied on PubMedVision PubMedVision Alignment VQA Instruction tuning on PubMedVision Instruction-Tuning VQA
MC-CoT [35]	MLLM	MC-CoT framework: Task Distributor + Anatomy, Radiology, and Pathology Modules	MLLM	NA

3. Dataset and Evaluation

3.1. MedVQA Benchmark Datasets

There are several key benchmark datasets in MedVQA research, namely VQA-RAD [2], PathVQA [54], and Slake [55]. Recently, new MedVQA datasets are crafted using semi-automatic or automatic methodologies with more data volume, scope, and modalities to better match the real-world clinical settings. This subsection briefly introduces the benchmark datasets including the data sources, creation methods, as well as dataset sizes and contents, presenting a summary table of MedVQA benchmark datasets in Table 3.

VQA-RAD [2] contains 3515 QA pairs curated manually by radiologists, derived from the MedPix® Radiology Database. It primarily focuses on X-rays and CT scans, ensuring clinical relevance and high-quality annotations for radiology-based queries. It is the earliest manually-curated MedVQA dataset and the most used MedVQA benchmark despite its relatively small size.

Slake [55] is a semantically labeled and knowledge-enhanced medical VQA dataset with 642 radiology images and 14,028 QA pairs in both English and Chinese. It includes segmentation masks, bounding boxes, and a structural knowledge graph of 2603 English and 2629 Chinese triplets, enabling diverse and complex queries like organ functions and disease prevention. Images and questions are carefully annotated and reviewed by experienced physicians, making Slake a highly comprehensive and clinically relevant resource.

PathVQA [54] focuses on pathology images, containing 32,800 QA pairs manually curated by domain specialists. With 5000 images, this dataset ensures high relevance to pathology diagnostics and supports reasoning-intensive QA tasks. It is the most recognized benchmark for pathological MedVQA.

VQA-Med-2019 dataset from ImageCLEF 2019 VQA-Med competition [39] focuses on answering clinically relevant questions based on radiology images. It includes 4200 images and 15,292 QA pairs divided into four categories: modality, plane, organ system, and abnormality. The dataset was semi-automatically created using question patterns inspired by medical students, ensuring natural phrasing, and was validated by experts for the test set. This dataset supports the development of MedVQA systems with high-quality, clinically grounded QA pairs.

PMC-VQA [14] leverages automatic synthetic QA generation through NLP models, resulting in 227,000 QA pairs across 149,000 images. Covering radiology, pathology, and microscopy, the dataset relies on automated filtering, leading to moderate annotation quality.

IAI-Med VQA [36] integrates AI-driven QA generation with expert validation, producing 14,000 QA pairs from 642 images. Covering X-rays, CT, and MRI modalities, the dataset combines automation and manual validation for high-quality outputs.

MIMIC-CXR-VQA [56] integrates structured EHR data from MIMIC-IV with chest X-ray images from MIMIC-CXR. Offering 46,152 QA pairs from 377,110 images, the dataset is created semi-automatically using expert-reviewed templates and machine annotations, making it ideal for multimodal medical reasoning.

P-VQA [57] is designed to address patient-oriented Medical Visual Question Answering, incorporating real hospital cases. It features 2169 images spanning multiple modalities, including X-rays, CT scans, MRI, and ultrasound, alongside 24,800 QA pairs. The dataset was semi-automatically generated using templates curated by physicians and enriched with a medical knowledge graph, ensuring clinically relevant and high-quality annotations.

Table 3. MedVQA benchmark datasets.

Dataset	Data Source	Creation Method	Size	Modalities
VQA-RAD [2]	MedPix® Radiology Database	Manual: Curated by radiologists for clinical relevance	315 images, 3515 QA pairs	X-ray, CT
Slake [55]	Medical Segmentation Decathlon, NIH Chest X-ray, CHAOS	Manual: Curated by medical professionals with semantic labeling	642 images, 14,028 QA pairs	X-ray, CT, MRI
PathVQA [54]	Electronic pathology textbooks, PEIR Digital Library	Manual: Curated by domain specialists for pathology-focused QA	4998 images, 32,799 QA pairs	Pathology
VQA-Med-2019 [39]	MedPix Database, Radiology Images	Semi-Automatic: Patterns from medical students, validated by experts	4200 images, 15,292 QA pairs	X-ray, CT, MRI, Ultrasound
PMC-VQA [14]	PubMed Central	Automatic: Synthetic QA generation using NLP models and filtering	149 K images, 227 K QA pairs	Radiology, Pathology, Microscopy
IAI-Med VQA [36]	Augmented Slake and VQA-RAD	Semi-Automatic: AI-generated QA pairs with expert validation	642 images, 14 K QA pairs	X-ray, CT, MRI
MIMIC-CXR-VQA [56]	MIMIC-IV, MIMIC-CXR	Semi-Automatic: MIMIC-CXR integration with expert-reviewed templates	142,797 images, 377,391 QA pairs (16,366 image-related)	X-ray, structured EHR data
P-VQA [57]	Deidentified Hospital Cases	Semi-Automatic: Generated with KGs and templates curated by physicians	2169 images, 24,800 QA pairs	X-ray, CT, MRI, Ultrasound
O-VQA [58]	Clinically Generated Data	Semi-Automatic: Based on clinical scenarios with expert-verified QA pairs	2001 images, 19,020 QA pairs	X-ray, CT
MLe-VQA [27]	ROCO, Medicat, MIMIC-CXR, PMC-VQA,	Automatic: generated with GPT-4 with five levels	5352 images, 59,969 QA pairs	X-ray, CT, MRI, Ultrasound, PET, Angiogram, Pathology, Endoscopy, Ophthalmic Imaging, and others
MLe-Bench [27]	PathVQA, VQA-RAD, VQA-Med, Slake	Automatic: generated with GPT-4 with five levels	1492 QA pairs	X-ray, CT, MRI, Pathology

O-VQA [58] is generated using electronic medical records (EMRs) from orthopedic hospitals. It contains 2001 images from X-rays and CT modalities, along with 19,020 QA pairs created semi-automatically using templates derived from frequently asked questions (FAQs). Experienced orthopedic surgeons validated the dataset, ensuring clinical relevance and annotation quality.

MLe-VQA [27] is constructed with GPT-4 with data from ROCO, Medicat, MIMIC-CXR and PMC-VQA, aiming to evaluate not only image and question comprehension but also the reasoning capability of the model. It contains 59,969 QA pairs with five levels from easy to difficult: Recognition, (e.g., identifying anatomical structures), Details (e.g., describing abnormalities), Diagnosis (e.g., identifying diseases), Knowledge Application (e.g., discussing treatment), and Reasoning, with the last two levels testing the reasoning capability.

MLe-Bench [27] is a evaluation benchmark generated with selected QA pairs from PathVQA, VQA-RAD, VQA-Med [59] and Slake. It contains 1492 QA pairs that are categorized into five levels—the same as MLe-VQA.

3.2. Evaluation Metrics

Evaluating generative MedVQA models requires specialized metrics that assess both the accuracy and quality of generated responses. Unlike classification-based models, which rely on straightforward accuracy measurements, generative models must be evaluated based on fluency, relevance, and clinical correctness. This section outlines key evaluation metrics used in closed-ended and open-ended MedVQA tasks.

3.2.1. Closed-Ended Evaluation Metrics

Closed-ended MedVQA questions involve selecting an answer from a predefined set, making traditional classification metrics applicable. Table 4 shows a summary of evaluation metris for closed-ended questions.

Table 4. Evaluation metrics for closed-ended questions in MedVQA.

Evaluation Metric	Used by
Accuracy	All Works
F1	[13]
AUC-ROC, AUC-PRC	[16]

Accuracy: Accuracy is the most common metric for closed-ended MedVQA tasks, measuring the proportion of correctly predicted answers over the total number of questions. It provides a simple and interpretable assessment of model performance in classification-based settings.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): AUC-ROC evaluates the model's ability to distinguish between positive and negative cases by plotting the true positive rate (sensitivity) against the false positive rate. A higher AUC-ROC indicates better discriminatory power, making it particularly useful for binary classification tasks such as disease detection.

AUC-PRC (Area Under the Precision–Recall Curve): AUC-PRC is well-suited for imbalanced datasets, where positive cases are rare. It measures the trade-off between precision and recall, capturing the model's performance in correctly identifying relevant answers while minimizing false positives.

3.2.2. Open-Ended Evaluation Metrics

Open-ended MedVQA questions involve generating free-form text responses, requiring more sophisticated evaluation techniques beyond accuracy. Table 5 presents adopted evaluation metrics for open-ended questions.

Table 5. Evaluation metrics for open-ended questions in MedVQA.

Evaluation Metric	Used by
Accuracy, Exact-match	[13–15,17,22,25,29,31,33,34]
BLEU	[6,13,14,16,23,27,31]
WBSS	[6]
CBSS	[30]
F1 score	[13,16]
BERTScore	[13,23,25,27]
ROUGE-L	[27]
Recall	[24,35,37]
Clinical Evaluation	[25,27,29]
GPT-4-Based Evaluation	[24,27,29]
Deepseek-Based Evalutation	[35]
VLM-Based Evaluation, not specified	[38]

Accuracy: While traditional accuracy is difficult to apply directly to open-ended responses, some studies use exact-match accuracy, where a response is considered correct only if it matches the ground truth exactly. However, this metric is often too rigid for natural language generation tasks.

Recall: Recall measures how many ground-truth tokens appear in the generated response. This metric is particularly relevant for MedVQA tasks where completeness of information is crucial, such as in diagnostic explanations.

F1 score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure of both false positives and false negatives.

BLEU (Bilingual Evaluation Understudy): BLEU is a widely used metric that measures the overlap of n-grams between generated responses and reference answers. It assigns higher scores to outputs that closely match human-written references.

WBSS (Word-Based Semantic Similarity): WBSS evaluates how semantically similar a generated response is to the reference answer by considering the relationships between words. Unlike BLEU, it captures meaning rather than exact word matches, making it more suitable for clinical applications where different phrasing can still convey the correct diagnosis or interpretation.

CBSS (Concept-Based Semantic Similarity): CBSS evaluates the similarity between two texts based on extracted biomedical concepts rather than exact word overlap with the help of MetaMap.

BERTScore (BERT-Sim): BERTScore computes similarity between generated and reference answers using contextualized embeddings from BERT. It assesses token-wise similarity in a way that accounts for synonymy and paraphrasing, making it a more robust metric for evaluating MedVQA models.

ROUGE-L: ROUGE-L evaluates the longest common subsequence (LCS) between the generated and reference texts, offering a measure of lexical similarity that accounts for

word order while being more flexible than n-gram-based metrics, making it suitable for tasks like summarization and report generation.

Clinician Evaluation: Some studies ask clinician to score the model answers by giving quality scores on correctness in a 10-point [29] or 5-point range [25]. Some also score from different aspects such as relevance, completeness, coherence, and explainability [27].

LLM-Based or MLLM-Based Evaluation: Recent studies leverage GPT-4, Deepseek, and some other LLMs and VLMs as an evaluator, scoring generated responses based on accuracy.

By employing a combination of these metrics, researchers can obtain a more comprehensive assessment of MedVQA models, balancing precision, semantic similarity, and clinical utility.

4. Advances in Generative MedVQA

4.1. Adopting LLMs and MLLMs

The adoption of LLMs and MLLMs has marked a significant evolution in the capabilities of MedVQA systems. These models, with their ability to integrate vast amounts of textual and visual information, have enabled MedVQA models to handle complex, open-ended queries with greater accuracy and contextual awareness and profoundly enhanced generative reasoning and conversational capabilities. This section highlights key techniques in the application of LLMs and MLLMs to MedVQA, focusing on instruction tuning, domain adaptation, and parameter-efficient fine-tuning. These methodologies exemplify how LLMs and MLLMs are tailored to meet the specialized needs of the medical field.

Instruction Tuning. Adapting large (multimodal) language models (LLMs/MLLMs) to MedVQA via instruction tuning has proven effective for aligning these models with clinical requirements. Rather than a purely general-purpose setup, instruction-tuned frameworks leverage carefully designed prompts and task-specific directives—often sourced from domain experts or synthetic medical datasets—to guide model outputs. For instance, LLaVA-Med [24], MedVInT [14], and Uni-Med [36] each illustrate how refining LLMs with medical task instructions improves answer relevance, reduces hallucination, and bolsters clinical usability. The key advantage lies in steering the model's generative behavior with precise, context-sensitive cues, ultimately enhancing both accuracy and interpretability.

Domain Adaptation. Beyond instruction tuning, domain adaptation techniques tailor LLM-based vision encoders and textual components to medical data. By combining pretrained components (e.g., CLIP-ViT) with fine-tuning strategies that incorporate radiology images and clinical text, these methods hone representational power on specialized domains. Such approaches, demonstrated by Haetal. [12], emphasize lightweight modifications (e.g., LoRA-based layers) that localize model parameters to the intricacies of radiological imagery. This targeted adjustment balances strong performance on MedVQA tasks (e.g., Slake, VQA-RAD) with reduced overhead, ensuring feasibility in clinical environments where computational resources may be constrained.

Parameter-Efficient Fine-Tuning (PEFT). As LLMs and MLLMs grow in complexity, cost-effective strategies are paramount for broader adoption. PEFT methods, such as the LoRA-based fine-tuning used in PeFoMed [29], tackle the resource bottleneck by updating only a small fraction of parameters while keeping the model core intact. This selective approach not only lowers training and storage demands but also increases adaptability: new medical data or tasks can be incorporated without overhauling the entire model. In practice, PEFT techniques pave the way for MedVQA systems deployable on edge devices and in low-resource settings—a critical factor for scaling advanced AI support in global healthcare.

Adopting LLMs/MLLMs for MedVQA increasingly hinges on three complementary strategies—instruction tuning to align outputs with clinical demands, domain adaptation to capture medical subtleties, and parameter-efficient fine-tuning to maintain feasibility across diverse healthcare contexts. Collectively, these techniques enable models to bridge the gap between general-purpose language understanding and the specialized rigor of real-world clinical applications, driving innovation and expanding access to high-quality AI solutions in medicine.

4.2. Automated Dataset Building and Data Augmentation

The scarcity of annotated datasets presents a significant challenge in advancing Med-VQA systems. Medical data are often difficult to obtain, requiring expert annotations that are both time-consuming and costly. To overcome this limitation, scalable data generation methods, such as synthetic dataset creation and data augmentation, have emerged as vital tools. Automated dataset building not only addresses the scarcity of labeled data but also enhances diversity and robustness in model training, ensuring that generative MedVQA systems generalize better to unseen scenarios. This subsection explores key techniques in data augmentation and automated dataset building, including synthetic dataset generation, data augmentation on existing datasets, and retrieval-based enrichment to expand training data effectively.

Synthetic Dataset Generation leverages automated methods to create large-scale datasets tailored for pretraining and fine-tuning MedVQA models. Frameworks such as Med-Flamingo [25], LLaVA-Med [24], and PMC-VQA [14] exemplify this approach. These models generate synthetic QA pairs using LLMs, MLLMs or domain-specific algorithms.

The PMC-15M dataset [60] comprises 15 million biomedical image–text pairs sourced from PubMed Central. The dataset creation involves two distinct subsets: biomedical concept alignment data and instruction-following data. The concept alignment data include basic image–caption pairs curated to provide foundational visual–textual alignment. Meanwhile, the instruction-following data are generated using GPT-4, which produces conversational instruction–answer pairs based on image captions and their surrounding textual context. For example, the pipeline prompts GPT-4 to simulate questions and answers as if it could view the images, enriching the data with high-quality multi-turn conversations. This dataset covers a variety of imaging modalities and clinical contexts, ensuring robust pretraining for a broad spectrum of biomedical tasks. Together, these datasets provide the necessary diversity and detail to train a versatile and capable multimodal assistant.

The MTB (Medical Textbook Benchmark) dataset [25] was constructed to evaluate and fine-tune the performance of Med-Flamingo in medical contexts by leveraging structured knowledge from medical textbooks. This dataset consists of image–text pairs extracted and curated from authoritative medical literature, including diagnostic atlases and foundational medical textbooks. Each data point is designed to simulate real-world clinical scenarios, comprising visual inputs like medical images (e.g., X-rays, CT scans, histopathology slides) paired with text that includes captions, explanations, and contextual information. The dataset captures a wide range of medical domains, from anatomy and radiology to pathology and surgery, ensuring a comprehensive scope. To maintain quality and relevance, medical experts reviewed and annotated the dataset, focusing on clarity, accuracy, and clinical applicability. By providing high-quality multimodal data, the MTB dataset facilitates pretraining and evaluation of models like Med-Flamingo, ensuring they can generalize across diverse medical imaging tasks and deliver clinically relevant insights.

The Visual USMLE dataset [25] draws inspiration from the format and structure of the United States Medical Licensing Examination (USMLE) and is curated to assess the model's ability to handle real-world, complex medical reasoning tasks, encompassing open-

ended questions that require advanced understanding of medical concepts. The dataset includes a diverse array of question types, such as diagnosis-based queries, treatment recommendations, and pathology explanations, each paired with supporting images like X-rays, CT scans, or MRIs. The dataset contains 618 USMLE-style questions modified from the real USMLE questions from the Amboss platform. The original multiple-choice questions are turned into open-ended questions to increase the difficulty.

Uni-Med framework [36] goes a step further by generating datasets with built-in explanations. These explanations provide additional layers of interpretability, allowing the model to learn not only to generate answers but also to justify them. For example, when tasked with identifying abnormalities in a chest X-ray, the generated dataset includes explanations detailing why specific regions are flagged, fostering more transparent and interpretable MedVQA outputs.

PMC-VQA [14] is built using PMC-OA, a biomedical corpus containing 1.6 million image–text pairs extracted from PubMed Central's Open Access subset. The dataset construction process involves automated question-answer generation, leveraging ChatGPT to synthesize five diverse question–answer pairs per image caption, covering different medical domains such as radiology, pathology, and microscopy. To ensure high-quality data, automated filtering methods were applied, including a text-only QA filtering model trained on LLaMA-7B to remove questions that could be answered without visual input. Additional manual verification was conducted to curate a high-quality test set (PMC-VQA-test-clean), ensuring relevance and diversity. The final dataset contains 227k VQA pairs across 149k images, significantly surpassing existing MedVQA datasets in size and modality diversity.

The MLe-VQA dataset [27] is a multi-level instruction dataset designed to enhance progressive reasoning capabilities in MedVQA. It consists of 5352 medical images collected from public biomedical datasets like MIMIC-CXR, ROCO, MedICaT, and PMC-VQA, covering X-ray, CT, MRI, ultrasound, angiograms, pathology, and ophthalmic imaging. Images were carefully selected based on caption richness to ensure meaningful question-answer generation. The dataset includes 59,969 multi-level QA pairs, generated using GPT-4, following a structured approach to simulate the diagnostic reasoning process of a doctor. The questions are categorized into five progressive levels: Recognition, Details, Diagnosis, Knowledge Application, and Reasoning Process, which integrates a step-by-step logical explanation. GPT-4 was guided with structured prompts to ensure medical relevance, and expert manual validation was applied to enhance the dataset's accuracy. This dataset enables comprehensive training and benchmarking for MLLMs in MedVQA, helping them understand medical images, reason step-by-step, and provide clinically relevant responses.

The dataset generation process in MEDIFICS involves synthetic data augmentation using GPT-3.5-Turbo to create realistic medical dialogues [38]. This process ensures the model is trained on diverse and clinically relevant question–answer pairs, improving its performance in Medical Visual Question Answering (MedVQA). The dataset is built by extracting metadata from medical image datasets such as ISIC (skin cancer), ROCO (radiology), and MURA (bone X-rays). Using few-shot prompting, GPT-3.5-Turbo generates structured doctor–patient conversations, simulating real-world clinical interactions. This enhances the model's ability to handle complex medical queries by providing context-rich training samples. The dataset also includes annotated image–text pairs, aligning radiology reports with corresponding images, ensuring multimodal learning. By leveraging synthetic data, MEDIFICS addresses the lack of large-scale MedVQA datasets, improving generalization while maintaining domain-specific accuracy.

PubMedVision is a large-scale medical multimodal dataset with 1.3 million medical VQA samples [26]. The dataset is built by extracting medical image–text pairs from PubMed, refining them using GPT-4V, and generating structured question-answer pairs to enhance medical multimodal learning. Unlike previous datasets that rely solely on text-based LLM reformulation, PubMedVision employs "unblinded" MLLMs, allowing GPT-4V to process both images and their contextual text to generate highly aligned, clinically relevant descriptions. The synthetic dataset is generated through a multi-stage pipeline: first, 914,960 medical images are extracted from PubMed, filtering out non-medical images using semantic retrieval and classification models. Next, GPT-4V reformats image–text pairs into VQA-style structured data, producing two types of synthetic data: Alignment VQA, where predefined medical questions are paired with AI-generated answers, and Instruction-Tuning VQA, where GPT-4V generates diverse clinical interactions in 10 different medical scenarios (e.g., doctor–patient discussions, intern–specialist dialogues). This approach ensures rich and diverse synthetic medical knowledge, improving model alignment for real-world medical tasks. The synthetic dataset is further validated through medical expert reviews, confirming higher accuracy, completeness, and relevance than previous reformatted datasets.

Data Augmentation enhances existing datasets by introducing variations that improve model robustness. The CGMVQA framework [6] employs augmentation techniques on both images and texts to diversify the training data. For visual data, methods such as rotation, scaling, and cropping are applied to simulate real-world variations in imaging. For textual data, tokenization and paraphrasing are used to create diverse representations of clinical questions and annotations. These augmentation strategies not only improve the model's ability to generalize across different scenarios but also reduce its susceptibility to overfitting.

Retrieval-Based Enrichment focuses on integrating multimodal prompts to provide contextually rich training data. The MPR framework [23] uses retrieval methods to curate relevant text and image pairs from large medical corpora. By augmenting the training process with these contextually enriched prompts, the model gains a deeper understanding of complex medical queries. For instance, when answering a question about organ abnormalities, the retrieval-based approach supplements the query with similar cases, relevant anatomical details, and diagnostic guidelines, enabling the model to generate more accurate and informed responses.

In summary, data augmentation and automated dataset building are critical for scaling MedVQA systems and improving their performance. Techniques such as synthetic dataset generation, augmentation on existing datasets, and retrieval-based enrichment address the limitations of annotated data scarcity while enhancing model robustness and generalization. As these methods continue to evolve, they will play a pivotal role in advancing the capabilities and clinical applicability of MedVQA technologies.

4.3. Comparison of Generative MedVQA with Discriminative MedVQA

Compared to discriminative MedVQA systems, generative MedVQA systems tend to perform better on open-ended questions. Table 6 provides a concise overview of recent MedVQA models with respect to generative vs. classic discriminative formulations, approximate parameter counts, and accuracy on open-ended questions of two prominent benchmarks: Slake and VQA-RAD. In terms of accuracy, Uni-Med achieves the highest reported accuracy on Slake (85.3%), closely followed by Med-MoE (85.1%), and is built upon a LLM/MLLM backbone. These results suggest that leveraging large-scale generative architectures—and thus more extensive parameter counts—can confer robust language generation and reasoning abilities. As for the smaller VQA-RAD benchmark, although

Uni-Med also excels with an accuracy of 74.2%, MPR reports a higher accuracy (77.5%). Overall, generative MedVQA models suit open-ended questions well and often attains strong accuracy on Slake and VQA-RAD. However, they tend to require substantial computational resources—Uni-Med is 7B parameters, LLaVA-Med is 13B—suggesting potentially higher memory usage and longer training times. On the other hand, discriminative models, such as BiomedCLIP and ARL, generally use fewer parameters (e.g., 422M for BiomedCLIP and 362M for ARL) and often exhibit more moderate accuracy levels (82.5% and 81.9% on Slake, respectively), yet they remain competitive. However, while parameter size often correlates with improved performance, targeted architectures and domain-specific design choices can sometimes outperform straightforward scaling.

Compared to earlier discriminative or smaller-scale generative MedVQA approaches, LLM/MLLM-based models generally offer higher performance and greater flexibility in handling open-ended queries. Their large-scale pretraining on diverse textual and/or multimodal datasets can yield stronger generalization, as demonstrated by improvements on standard benchmarks like VQA-RAD and Slake in Table 6. However, this increased capacity often comes at the expense of both interpretability and computational efficiency. Due to the black-box nature of many large-scale transformer architectures, interpretability is more challenging to achieve, though emerging strategies—such as chain-of-thought prompting, rationale generation, and instruction tuning—are beginning to address this shortfall. In terms of computational efficiency, LLM/MLLM-based solutions typically require greater GPU/TPU resources and may pose deployment challenges in resource-constrained clinical environments. Recent advances in parameter-efficient fine-tuning (e.g., LoRA, QLoRA) and lightweight adaptation techniques (e.g., adapters) help mitigate some of these costs, but the trade-off between performance gains and operational feasibility remains a key consideration for real-world adoption.

Table 6. Performance of generative MedVQA and classic discriminative MedVQA models.

Model	#Parameters	Generative	LLM/MLLM-Based	Accuracy on Open-Ended Questions	
				Slake	VQA-RAD
Uni-Med	7B	Yes	Yes	**85.3**	**74.2**
Med-MoE	3.6B	Yes	Yes	**85.1**	58.6
MedVInT	7B	Yes	Yes	84.5	73.7
LLaVA-Med	13B	Yes	Yes	84.5	64.4
Ha, fusion	Not Reported	Yes	Yes	84.5	57.5
van Sonsbeek	1.5B	Yes	Yes	84.3	-
PeFoMed	57M	Yes	Yes	77.8	62.6
BiomedGPT	182M	Yes	No	84.3	60.9
M2I2	252M	Yes	No	74.7	61.8
MUMC	211M	Yes	No	71.5	-
MPR	Not Reported	Yes	No	62.6	**77.5**
BiomedCLIP	422M	No	No	82.5	67.6
ARL	362M	No	No	81.9	67.6
M3AE	Not Reported	No	No	80.3	67.2
MMBERT	Not Reported	No	No	79.5	55.3
VQAMIX	Not Reported	No	No	-	56.6

5. Challenges and Future Directions in Generative MedVQA

5.1. Dataset Limitations

While existing datasets have advanced MedVQA research, their limitations must be critically addressed to guide future improvements.

Data Bias. Both manual and automated datasets exhibit data bias through two primary mechanisms, originating from data sources and annotation processes. *(1) Case selection bias.* MedVQA encompasses a wide range of cases while diseases follow extreme long-tail patterns, which makes it challenging for individual datasets to represent the full spectrum of clinical scenarios. For instance, MIMIC-CXR-VQA [56] predominantly focuses on chest radiographs, while PathVQA [54] extracts pathology images and captions largely from textbooks, resulting in a focus on commonly documented cases with standardized professional language. *(2) Modality bias.* Medical images contain a variety of modalities, including radiology, pathology, microscopy, etc. Most of the existing datasets exhibit bias toward certain modalities, which potentially over-represent common modalities (e.g., X-ray) while under-representing rare emerging imaging techniques. PMC-VQA [61] covers various modalities and diseases in 149 K images though, 80% of them are radiological images obtained from PubMed Central.

Insufficient Annotation. Many of the existing MedVQA datasets face critical limitations in annotation sufficiency and diversity. While Slake [55] provides bilingual (English and Chinese) annotations, most datasets like VQA-RAD [2] and PathVQA [54] rely solely on English annotations. This restricts cross-lingual generalization and creates barriers for non-English medical contexts. Although image-caption datasets (e.g., ROCO) are generally used and adequate for model training/fine-tuning to some extent, other forms of annotations are equally crucial for achieving improved performance on MedVQA models. For instance, both Slake [55] and P-VQA [57] provide knowledge graphs to enhance models' diagnostic and analytical capabilities, offering additional critical information beyond pure imaging data. Slake further provides labeled segmentations of organs and tumors or bounding boxes on objects, enabling precise region-of-interest (ROI) localization for visual models. Additionally, MIMIC-CXR-VQA [56] connects medical images with structured electronic health records (EHRs) of corresponding patients, which contain multi-layered information for question answering. These supplementary annotations can effectively improve models' performance and interpretability through structured clinical correlation.

Risks in Automatic Generation. Stemming from the complexities of annotation, an increasing number of datasets are prompting LLMs to create QA pairs either directly or indirectly. Current AI-assisted approaches primarily manifest in two forms. *(1) QA data augmentation*. For instance, MIMIC-CXR-VQA [56] makes efforts to generate questions by prompting GPT-4 to augment the origin expressions with an average of 16.5 paraphrases per template, thereby enhancing linguistic diversity while preserving semantic integrity. This measure may enhance the robustness of MedVQA models at the expense of inducing unexpected semantic deviations from standard terminology. *(2) Caption-based QA generation.* MLe-VQA [27] collects public datasets that provide images and captions, using the captions of the images to prompt GPT-4 for generating QA pairs. Such approach of generation poses a high risk for introducing ambiguity and inaccuracies to QA pairs, resulting in MedVQA models being prone to hallucinations and deviating from clinical validity.

Balancing expert oversight with efficient annotation processes remains a major challenge in constructing high-quality MedVQA datasets. Annotation by clinicians requires domain expertise, resulting in small-scale pure-human datasets (VQA-RAD [2] and Slake [55]). While semi-automated approaches extract textbook figures (PathVQA [54]) or open source databases (MIMIC-CXR-VQA [56]), they risk inheriting source-specific biases and oversimplifying clinical scenarios. Moreover, the annotation of MedVQA datasets requires establishing well-designed and diverse templates to ensure the question diversity, as well as accuracy and interpretability. This dual requirement adds another layer of complexity to creating clinically relevant yet efficient annotation frameworks.

5.2. Evaluation Benchmarks and Metrics

Recent benchmarks in MedVQA indicate key trends shaping the field. One major trend is the integration of more data modality, where datasets like MIMIC-CXR-VQA [56] combine chest X-ray images with structured EHR data to enhance the model's ability to perform multimodal medical reasoning. This approach allows MedVQA systems to understand both visual and textual aspects of medical records, improving their real-world applicability. As more structured clinical data become available, future benchmarks are likely to expand the range of medical knowledge that models can leverage.

Another key development is the emphasis on reasoning and clinical context. Datasets such as MLe-VQA [27] introduce multi-level difficulty settings, testing models beyond simple recognition tasks and evaluating their ability to diagnose and reason through medical scenarios. Visual USMLE [25] adopts questions from USMLE that requires a higher-level of clinical understanding and reasoning ability. On the other hand, IAI-Med VQA [36] integrates user instructions to into model inference to align better with clinician's intent. This shift towards higher-order reasoning suggests that future benchmarks will place greater demands on models to interpret medical findings, integrate prior knowledge, and provide clinically meaningful responses rather than just factual retrieval.

As the field matures, domain-specific benchmarks are becoming more prevalent, with datasets like O-VQA [58] focusing specifically on orthopedic medical records. Similarly, P-VQA [57] captures patient-oriented questions across multiple imaging modalities, making the dataset more relevant for real hospital settings. This specialization ensures that MedVQA models are fine-tuned for specific clinical applications, which could lead to AI systems that better support subspecialties in medicine such as cardiology, neurology, and dermatology.

A notable shift is the growing use of automation in dataset creation, combined with expert validation to maintain data quality. IAI-Med VQA [36] demonstrates this approach by using AI-driven QA generation alongside manual expert validation to ensure the reliability of its 14,000 QA pairs. Similarly, MIMIC-CXR-VQA [56] leverages semi-automated template-based question generation reviewed by medical experts. This hybrid approach balances scalability and accuracy, making it a practical method for producing large-scale, clinically relevant datasets.

Overall, these trends highlight the increasing complexity of MedVQA tasks and the evolving expectations for AI systems in clinical settings. Future benchmarks will likely continue expanding in scale, reasoning capability, and domain specialization, pushing AI towards more advanced clinical decision support applications.

Evaluation metrics for MedVQA are evolving beyond traditional accuracy-based measures that assess the exact match between generated answers and ground truth. While BLEU remains a common choice for evaluating n-gram overlap, it fails to account for synonymy and semantic similarity, which are crucial in medical contexts. This limitation has led to the adoption of metrics such as BERTScore, which leverages contextualized embeddings from BERT to measure the similarity between generated and reference answers. Unlike BLEU, BERTScore can capture semantic equivalence and paraphrasing, making it a more suitable choice for medical applications.

The reasonability of traditional NLP metrics (BLEU, ROUGE, BERTScore) that widely adopted in MedVQA can be further discussed, in terms that they fall short in capturing clinical correctness due to the following reasons. *(1) Surface-level alignment does not ensure clinical validity.* Metrics such as BLEU and ROUGE prioritize lexical or statistical overlap but ignore the contextual relevance and clinical appropriateness of the generated answers. A generated answer may achieve high BLEU scores yet contain life-threatening errors (e.g., misclassifying "benign" as "malignant"). *(2) Lack of reasoning-specific evaluation.* Medical

reasoning requires multi-step inference, but existing metrics have not taken into account the rationality of the model's logical chain, which means that a causal inversion answer caused by hallucination may be overlooked. The clinical relevance and key limitations of these metrics are highlighted in Table 7.

Table 7. Limitations of traditional NLP metrics in MedVQA task.

Metric	Clinical Relevance	Key Limitations
BLEU	Measures n-gram overlap Useful for detecting terminology consistency	Fails to penalize clinically invalid paraphrases (e.g., "GGO" vs. "ground-glass opacity")
ROUGE	Evaluates content recall Identifies missing critical information	Ignores logical order errors (e.g., reversing cause–effect chains in diagnosis)
BERTScore	Captures semantic similarity better than BLEU/ROUGE	Struggles with domain-specific abbreviations and rare diseases

To address the shortcomings discussed above, some studies have introduced clinician evaluation to their MedVQA models. Expert-driven judgment is a gold standard for clinical validity, which assesses diagnostic logic coherence and contextual appropriateness. Med-Flamingo [25] conducts clinical evaluations by experts, revealing that a model with the lowest BERT-sim score may actually achieve the best performance on clinical eval score. Especially when dealing with ambiguous or high-risk scenarios such as tumor detection, expert review is essential.

A more recent trend is the adoption of LLMs and MLLMs, such as GPT-4 and DeepSeek, for evaluation [27,29,35]. These models offer a more nuanced assessment by considering different valid expressions of the same medical concept, addressing the limitations of traditional lexical-based metrics. By incorporating LLMs and MLLMs into evaluation frameworks, researchers can better assess a model's ability to generate clinically meaningful and contextually appropriate responses, rather than focusing solely on surface-level text similarity.

Looking ahead, the development of more robust and clinically-aligned evaluation metrics will be critical in advancing MedVQA. In addition to measuring accuracy and linguistic fidelity, additional standards are needed to gauge both the thoroughness of reasoning and the system's reliability in real-world clinical contexts. Future metrics may integrate domain-specific knowledge graphs, expert-in-the-loop validation, or human–AI collaborative assessment to ensure clinical relevance and trustworthy outputs. Incorporating real-world case studies and empirical evidence from actual medical settings will be essential in validating these metrics, identifying data limitations, and addressing issues such as hallucinations. As MedVQA continues to evolve, the refinement of evaluation methods will play a key role in bridging the gap between AI-generated responses and practical, evidence-based medical reasoning.

5.3. Reasoning Ability and Interpretability

Reasoning and interpretability are fundamental to the advancement of clinical AI systems, particularly in generative MedVQA. Accurate reasoning ensures that AI models can address complex medical queries in a way that aligns with clinical logic and domain-specific requirements. Meanwhile, interpretability fosters trust and transparency, allowing clinicians to understand the rationale behind a model's responses. Together, these elements are critical for the safe and reliable integration of MedVQA systems into real-world medical workflows, where decision-making carries significant consequences.

Recent advancements in MedVQA have introduced novel reasoning techniques tailored to handle medical complexity and enhance interpretability. MC-CoT [35] employs chain-of-thought (CoT) reasoning, where the model incrementally constructs a response

by integrating intermediate outputs from modular components. By breaking down a query into logically connected sub-problems, MC-CoT ensures that the reasoning process remains transparent and clinically sound, improving the model's reliability in medical decision-making.

Another significant advancement is rationale generation, which enhances interpretability by providing explanatory justifications alongside model outputs. The Med-Flamingo framework [25], for instance, can be prompted to generate both answers and reasoning rationales, allowing clinicians to verify the AI's thought process against textual or visual evidence. This dual-output approach strengthens user trust, facilitates error identification, and supports iterative model refinement, making AI systems more accountable and debuggable.

Beyond rationale generation, the Uni-Med framework [36] introduces a user-driven approach to interpretability by explicitly mapping user instructions to specific answer components. This targeted mapping ensures that responses align with the clinician's intent, providing a more personalized and comprehensible reasoning process. By integrating user instructions into model inference, Uni-Med enhances AI customization and adaptability, making it a valuable tool for diverse clinical assessments.

Moving forward, the development of reasoning and interpretability in MedVQA is expected to take a more hybrid and clinically informed approach. Standardizing reasoning protocols to align intermediate inferences with established guidelines can strengthen consistency and bolster clinical trust. Integrating expert knowledge can help models distinguish between subtle diagnostic categories and reduce misinterpretations. Designing interfaces that let clinicians probe or challenge a model's reasoning fosters transparency, especially when paired with confidence estimation to highlight areas of uncertainty. Building iterative feedback loops into the model-training process allows users to correct errors and guide updates, improving relevance to ever-evolving medical practices. As MedVQA systems take on more significant roles in clinical decision support, advancements in reasoning and interpretability will be crucial in ensuring safety, reliability, and real-world usability.

5.4. Efficiency and Scalability

As LLMs and MLLMs continue to advance, their enhanced capabilities have made them a common choice for MedVQA and other medical AI applications. However, these improvements often come at the cost of increasing model size, leading to higher computational demands and significant resource constraints, particularly in training and deployment. To address these challenges, many recent studies have adopted PEFT techniques, such as LoRA, QLoRA, and prefix tuning, as well as the MoE technique to reduce the computational burden while maintaining model performance [13,29,37]. These methods allow models to adapt to new tasks with minimal parameter updates, making fine-tuning more feasible for resource-limited clinical settings.

Despite these optimizations, there remains a pressing need for lightweight and efficient models that can deliver strong reasoning capabilities without excessive computational overhead. The adaptation of general-domain generative models for specialized clinical tasks remains an open challenge, requiring more targeted fine-tuning strategies that align model outputs with domain-specific medical knowledge. Future research must continue refining adaptive fine-tuning approaches, improving knowledge integration, and developing efficient model architectures to ensure that MedVQA systems are both scalable and clinically applicable in real-world medical workflows. Incorporating additional strategies like quantization, model pruning, or knowledge distillation can further cut memory usage and expedite inference, facilitating broader adoption even in smaller healthcare facilities where computational resources are at a premium.

5.5. Hallucination of LLMs/MLLMs

Hallucination in LLMs and MLLMs refers to the generation of incorrect, misleading, or non-existent information that appears plausible but lacks factual grounding [62]. In medical settings, hallucinations can manifest as fabricated diagnoses, misinterpretation of medical images, or incorrect treatment recommendations, posing significant risks to patient safety. Unlike traditional rule-based or discriminative AI models that select from predefined outputs, generative models produce free-form responses, increasing the likelihood of hallucinations when encountering ambiguous, incomplete, or out-of-distribution medical queries. This issue is particularly concerning in Med-VQA, where models must interpret complex visual and textual inputs, and any inaccurate responses can lead to misdiagnosis or improper medical decisions.

To mitigate hallucinations, various strategies have been proposed [63]. Retrieval-augmented generation (RAG) integrates external, verified medical knowledge sources, ensuring responses are grounded in evidence-based medical literature. Fine-tuning with high-quality, domain-specific datasets improves the model's understanding of medical concepts while reducing reliance on statistical guessing. Confidence scoring mechanisms allow models to flag uncertain or low-confidence responses, prompting human oversight. Human-in-the-loop verification, where medical professionals validate AI-generated responses, is crucial before deployment in high-stakes clinical applications. Despite these advancements, hallucination remains a critical barrier to real-world adoption, necessitating ongoing evaluation and rigorous testing to ensure reliability, accountability, and trustworthiness in clinical environments.

5.6. Regulation and Ethical Considerations

A critical avenue for future MedVQA research lies in addressing the challenges of regulatory certifications, legal liability, and ethical considerations. From a regulatory standpoint, MedVQA systems must demonstrate consistent accuracy and safety to gain certification from bodies such as the U.S. Food and Drug Administration (FDA) or analogous international agencies, which necessitates rigorous validation studies and standardized evaluation metrics. Legal liability also poses a substantial barrier, as misdiagnoses or oversights by AI-driven systems may expose healthcare providers and developers to malpractice suits. Consequently, transparent error reporting and robust clinical oversight mechanisms are essential for mitigating risk [63]. Ethical concerns further complicate deployment, including potential biases stemming from non-representative training datasets and the need to safeguard patient privacy, particularly with sensitive imaging data and protected health information. Moving forward, interdisciplinary collaboration among technologists, clinicians, legal experts, and policymakers will be pivotal for integrating MedVQA systems into real-world medical settings in a manner that is both equitable and compliant with evolving regulatory frameworks.

Author Contributions: W.D.: Investigation, writing—original draft, writing—review and editing. Y.H.: Writing—review and editing, supervision. S.S.: Writing—review and editing. T.T.: Writing—review and editing, supervision. J.W.: Supervision. H.X.: Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by National Natural Science Foundation of China under grants No. 82202984, Zhejiang Key R&D Program of China under grant No. 2023C03053, and Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hasan, S.A.; Ling, Y.; Farri, O.; Liu, J.; Müller, H.; Lungren, M. Overview of imageclef 2018 medical domain visual question answering task. In Proceedings of the CLEF 2018 Working Notes, Avignon, France, 10–14 September 2018.
2. Lau, J.J.; Gayen, S.; Ben Abacha, A.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **2018**, *5*, 1–10. [CrossRef] [PubMed]
3. Ossowski, T.; Hu, J. Multimodal Prompt Retrieval for Generative Visual Question Answering. *arXiv* **2023**, arXiv:2306.17675.
4. Lin, Z.; Zhang, D.; Tao, Q.; Shi, D.; Haffari, G.; Wu, Q.; He, M.; Ge, Z. Medical visual question answering: A survey. *Artif. Intell. Med.* **2023**, *143*, 102611. [CrossRef]
5. Khare, Y.; Bagal, V.; Mathew, M.; Devi, A.; Priyakumar, U.D.; Jawahar, C. Mmbert: Multimodal bert pretraining for improved medical vqa. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 1033–1036.
6. Ren, F.; Zhou, Y. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access* **2020**, *8*, 50626–50636. [CrossRef]
7. Nguyen, B.D.; Do, T.T.; Nguyen, B.X.; Do, T.; Tjiputra, E.; Tran, Q.D. Overcoming data limitation in medical visual question answering. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part IV 22; Springer: Berlin/Heidelberg, Germany, 2019; pp. 522–530.
8. 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, Virtual, 12–16 October 2020; pp. 2345–2354.
9. Eslami, S.; Meinel, C.; De Melo, G. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 1181–1193.
10. Gong, H.; Chen, G.; Liu, S.; Yu, Y.; Li, G. Cross-modal self-attention with multi-task pre-training for medical visual question answering. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 456–460.
11. Moon, J.H.; Lee, H.; Shin, W.; Kim, Y.H.; Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 6070–6080. [CrossRef] [PubMed]
12. Ha, C.; Asaadi, S.; Karn, S.K.; Farri, O.; Heimann, T.; Runkler, T. Fusion of Domain-Adapted Vision and Language Models for Medical Visual Question Answering. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*; Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., Bitterman, D., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 246–257. [CrossRef]
13. Van Sonsbeek, T.; Derakhshani, M.M.; Najdenkoska, I.; Snoek, C.G.; Worring, M. Open-ended medical visual question answering through prefix tuning of language models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 726–736.
14. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. Development of a large-scale medical visual question-answering dataset. *Commun. Med.* **2024**, *4*, 277. [CrossRef]
15. Chen, J.; Yang, D.; Jiang, Y.; Lei, Y.; Zhang, L. MISS: A Generative Pre-training and Fine-Tuning Approach for Med-VQA. In *Proceedings of the International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 299–313.
16. Sharma, D.; Purushotham, S.; Reddy, C.K. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Sci. Rep.* **2021**, *11*, 19826. [CrossRef]
17. Haridas, H.T.; Fouda, M.M.; Fadlullah, Z.M.; Mahmoud, M.; ElHalawany, B.M.; Guizani, M. MED-GPVS: A deep learning-based joint biomedical image classification and visual question answering system for precision e-health. In Proceedings of the ICC 2022-IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022; pp. 3838–3843.
18. Al-Hadhrami, S.; Menai, M.E.B.; Al-Ahmadi, S.; Alnafessah, A. A critical analysis of benchmarks, techniques, and models in medical visual question answering. *IEEE Access* **2023**, *11*, 136507–136540. [CrossRef]
19. Hong, X.; Song, Z.; Li, L.; Wang, X.; Liu, F. BESTMVQA: A Benchmark Evaluation System for Medical Visual Question Answering. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 435–451.
20. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; pp. 7871–7880. [CrossRef]
21. Ni, J.; Abrego, G.H.; Constant, N.; Ma, J.; Hall, K.; Cer, D.; Yang, Y. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 1864–1874.

22. Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B.D.; Ren, H.; et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **2024**, *30*, 3129–3141. [CrossRef]
23. Ossowski, T.; Hu, J. Retrieving multimodal prompts for generative visual question answering. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 2518–2535.
24. 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. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 28541–28564.
25. Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E.P.; Rajpurkar, P. Med-flamingo: A multimodal medical few-shot learner. In Proceedings of the Machine Learning for Health (ML4H), PMLR, New Orleans, LA, USA, 10 December 2023; pp. 353–367.
26. Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA 12–16 November 2024; pp. 7346–7370.
27. Xu, D.; Chen, Y.; Wang, J.; Huang, Y.; Wang, H.; Jin, Z.; Wang, H.; Yue, W.; He, J.; Li, H.; et al. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Virtual Meeting, Bangkok, Thailand, 11–16 August 2024; pp. 4977–4997.
28. Yan, Q.; Duan, J.; Wang, J. Multi-modal Concept Alignment Pre-training for Generative Medical Visual Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Virtual Meeting, Bangkok, Thailand, 11–16 August 2024; pp. 5378–5389.
29. He, J.; Li, P.; Liu, G.; Zhao, Z.; Zhong, S. Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv* **2024**, arXiv:2401.02797.
30. Talafha, B.; Al-Ayyoub, M. JUST at VQA-Med: A VGG-Seq2Seq Model. In Proceedings of the CLEF (Working Notes), Avignon, France, 10–14 September 2018.
31. Zhou, Y.; Kang, X.; Ren, F. TUA1 at ImageCLEF 2019 VQA-Med: A Classification and Generation Model based on Transfer Learning. In Proceedings of the CLEF (Working Notes), Lugano, Switzerland, 9–12 September 2019.
32. Verma, H.; Ramachandran, S. HARENDRAKV at VQA-Med 2020: Sequential VQA with Attention for Medical Visual Question Answering. In Proceedings of the CLEF (Working Notes), Thessaloniki, Greece, 22–25 September 2020.
33. Li, P.; Liu, G.; He, J.; Zhao, Z.; Zhong, S. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 374–383.
34. Li, P.; Liu, G.; Tan, L.; Liao, J.; Zhong, S. Self-supervised vision-language pretraining for medial visual question answering. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; pp. 1–5.
35. Wei, L.; Wang, W.; Shen, X.; Xie, Y.; Fan, Z.; Zhang, X.; Wei, Z.; Chen, W. MC-CoT: A Modular Collaborative CoT Framework for Zero-shot Medical-VQA with LLM and MLLM Integration. *arXiv* **2024**, arXiv:2410.04521.
36. Wu, Z.; Xu, H.; Long, Y.; You, S.; Su, X.; Long, J.; Luo, Y.; Xu, C. Detecting any instruction-to-answer interaction relationship: universal instruction-to-answer navigator for med-VQA. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; JMLR.org, ICML'24.
37. Jiang, S.; Zheng, T.; Zhang, Y.; Jin, Y.; Yuan, L.; Liu, Z. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 3843–3860.
38. Said, E.T.; Soufiane, A.E.A.; Jamal, E.T. MEDIFICS: Model Calling Enhanced VLM for Medical VQA. In Proceedings of the 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 23–24 October 2024; pp. 1–6.
39. Ben Abacha, A.; Hasan, S.A.; Datla, V.V.; Demner-Fushman, D.; Müller, H. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In Proceedings of the CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes, Lugano, Switzerland, 9–12 September 2019.
40. Ionescu, B.; Müller, H.; Péteri, R.; Rückert, J.; Abacha, A.B.; de Herrera, A.G.S.; Friedrich, C.M.; Bloch, L.; Brüngel, R.; Idrissi-Yaghir, A.; et al. Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 541–564.

41. Pelka, O.; Koitka, S.; Rückert, J.; Nensa, F.; Friedrich, C.M. Radiology objects in context (roco): A multimodal image dataset. In Proceedings of the Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 16 September 2018; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2018; pp. 180–189.
42. Subramanian, S.; Wang, L.L.; Bogin, B.; Mehta, S.; van Zuylen, M.; Parasa, S.; Singh, S.; Gardner, M.; Hajishirzi, H. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 2112–2120.
43. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 590–597.
44. Hua, S.B.Z.; Lu, A.X.; Moses, A.M. CytoImageNet: A large-scale pretraining dataset for bioimage transfer learning. *arXiv* **2021**, arXiv:2111.11646.
45. Jing, B.; Xie, P.; Xing, E. On the Automatic Generation of Medical Imaging Reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, VIC, Australia, 15–20 July 2018; pp. 2577–2586.
46. Shamshad, F.; Khan, S.; Zamir, S.W.; Khan, M.H.; Hayat, M.; Khan, F.S.; Fu, H. Transformers in medical imaging: A survey. *Med. Image Anal.* **2023**, *88*, 102802. [CrossRef]
47. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 58–65.
48. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef]
49. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 1–9. [CrossRef]
50. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
51. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Vancouver, BC, Canada, 8–12 October 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 525–536.
52. Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Wang, G.; Li, H.; Zhu, J.; Chen, J.; et al. Yi: Open foundation models by 01. ai. *arXiv* **2024**, arXiv:2403.04652.
53. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
54. He, X.; Zhang, Y.; Mou, L.; Xing, E.; Xie, P. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv* **2020**, arXiv:2003.10286.
55. 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 Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13 April 2021; pp. 1650–1654.
56. Bae, S.; Kyung, D.; Ryu, J.; Cho, E.; Lee, G.; Kweon, S.; Oh, J.; Ji, L.; Chang, E.; Kim, T.; et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 3867–3880.
57. Huang, J.; Chen, Y.; Li, Y.; Yang, Z.; Gong, X.; Wang, F.L.; Xu, X.; Liu, W. Medical knowledge-based network for patient-oriented visual question answering. *Inf. Process. Manag.* **2023**, *60*, 103241. [CrossRef]
58. Huang, Y.; Wang, X.; Liu, F.; Huang, G. OVQA: A clinically generated visual question answering dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 2924–2938.
59. 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, Bucharest, Romania, 21–24 September 2021.
60. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image—Text Pairs. *NEJM AI* **2024**, *2*, AIoa2400640. [CrossRef]
61. 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* **2023**, arXiv:2305.10415.

62. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
63. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings Bioinform.* **2024**, *25*, bbad493. [CrossRef]