Medical Visual Question Answering in Healthcare: A Survey

Abstract

Medical Visual Question Answering (Med-VQA) represents a significant advancement in healthcare AI, integrating visual and textual data to answer clinical questions based on medical images. This survey examines the current landscape of Med-VQA, focusing on the challenges and innovations within this domain. The scarcity of labeled datasets remains a critical hurdle, limiting the training of robust models. However, advancements such as Medical Large Vision Language Models (Med-LVLMs) and multimodal learning approaches are enhancing diagnostic accuracy and report generation. Techniques like self-supervised learning and medical visual language pre-training are pivotal in overcoming data limitations, enabling models to learn from unlabeled data and improve feature representations. The integration of large language models (LLMs) into Med-VQA systems enhances answer generation and reporting, although challenges related to model interpretability, reliability, and ethical concerns persist. Future research should focus on developing more efficient training methods, improving data quality, and addressing ethical and privacy issues to ensure the responsible deployment of Med-VQA systems in clinical settings. By leveraging these advancements, Med-VQA systems have the potential to significantly improve clinical decision-making and patient care, highlighting the importance of interdisciplinary collaboration in advancing healthcare AI.

1 Introduction

1.1 Concept and Importance of Med-VQA

Medical Visual Question Answering (Med-VQA) represents a significant advancement in healthcare, merging visual and textual data to address clinically relevant inquiries based on medical images. This interdisciplinary domain utilizes Medical Large Vision Language Models (Med-LVLMs), enhancing diagnostic accuracy and the quality of medical assessments [1]. The complexity of Med-VQA arises from the need to accurately interpret intricate medical images, such as computed tomography (CT) scans and chest X-rays (CXR), which are essential for diagnosing conditions like COVID-19 [2].

Med-VQA finds application across diverse clinical settings, offering critical support to healthcare professionals and augmenting their capabilities [3]. A primary challenge in advancing Med-VQA lies in the limited availability of high-quality, publicly labeled datasets, which are crucial for training and evaluating these models [4]. Despite these obstacles, Med-VQA systems play a vital role in generating accurate and coherent medical reports, essential for documenting findings from medical images [5].

Moreover, incorporating expert knowledge into vision-language models (VLMs) mitigates the limitations of existing models that lack specialized training for healthcare tasks, thereby enhancing their relevance in medical contexts [6]. The capability of Med-VQA to interpret medical images and produce contextually aware responses not only aids clinical decision-making but also supports

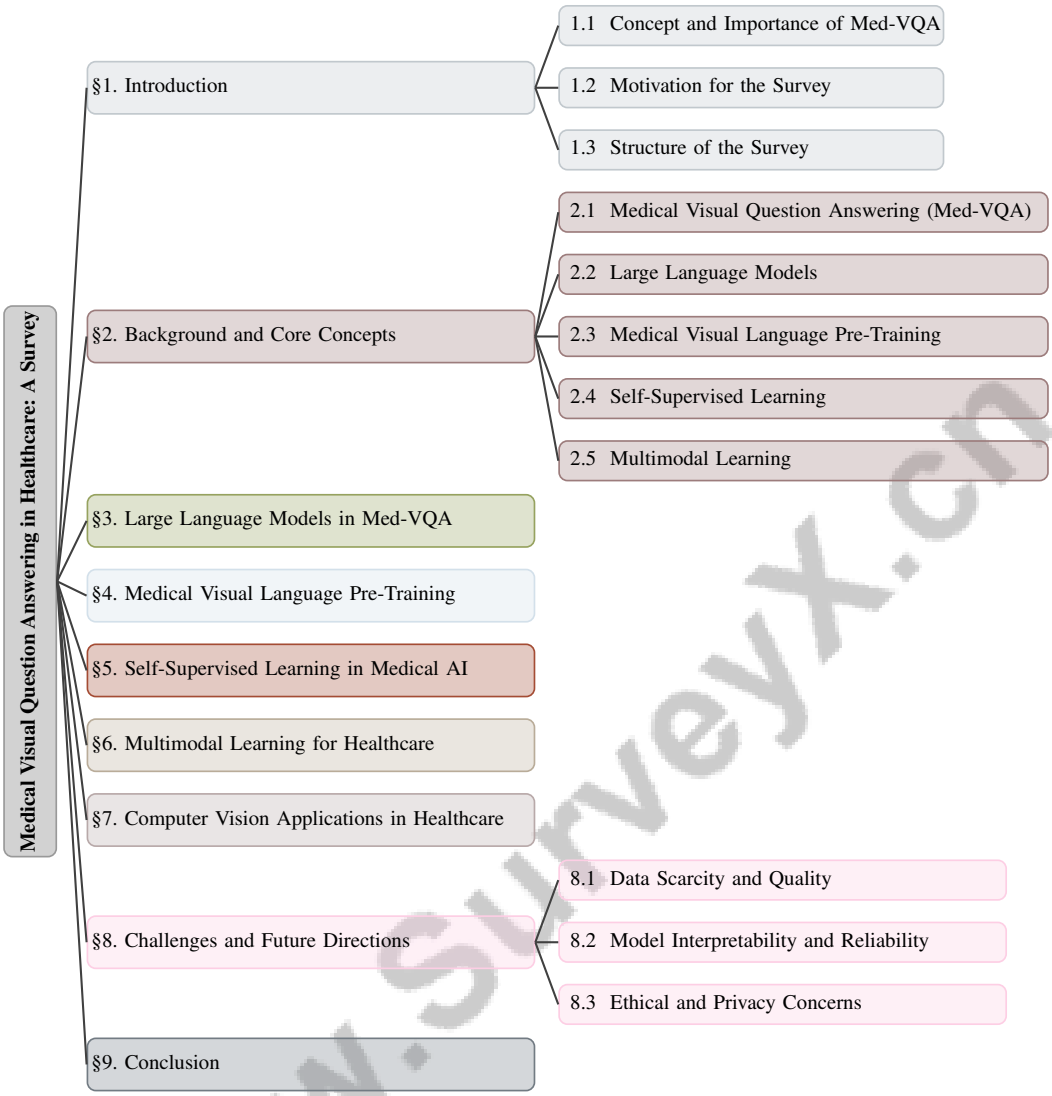


Figure 1: chapter structure

healthcare professionals in resource-limited environments, underscoring its significance in the global healthcare landscape.

1.2 Motivation for the Survey

The motivation behind this survey on Medical Visual Question Answering (Med-VQA) stems from the pressing need to address challenges and gaps within healthcare AI, particularly concerning radiology and pathology images. Current VQA systems often struggle to effectively utilize both visual and textual information, a limitation this survey aims to investigate and rectify [7]. Establishing benchmarks such as VQA-RAD and PathVQA is essential for progressing in answering clinical and pathology-related questions, thereby improving patient care and clinical decision-making.

This survey also addresses inefficiencies in medical report generation, exacerbated by the diversity of image modalities and the scarcity of labeled medical image-report pairs. Such scarcity restricts the training of deep neural networks and hinders the development of robust Med-VQA systems capable of managing complex clinical inquiries [5]. The manual nature of report writing, especially for time-sensitive diagnoses like COVID-19, highlights the urgent need for automated solutions [2].

Furthermore, the survey underscores the necessity of creating standardized benchmarks to assess the trustworthiness and safety of Medical Large Vision Language Models (Med-LVLMs) in healthcare

AI-generated, for reference only.

applications [1]. The limited functionality of current text-based VQA systems in scenarios requiring hands-free interaction emphasizes the need for advancements in Med-VQA to enhance accessibility and interaction in clinical environments [8].

This survey aims to tackle the multifaceted challenges associated with Med-VQA by providing a comprehensive analysis of the current research landscape, including a review of publicly available datasets and methodologies, as well as an exploration of specific medical challenges and future research directions. Through this, it seeks to advance the development of AI-driven healthcare solutions, fostering a deeper understanding of the effective application of these technologies in the medical field [9, 10].

1.3 Structure of the Survey

This survey is systematically organized to offer an extensive examination of Medical Visual Question Answering (Med-VQA), detailing its unique challenges, current methodologies, and future research directions, thereby underscoring its implications for enhancing decision-making processes in healthcare through improved interpretability and contextual understanding of medical images [9, 11]. Following the introduction, which establishes the foundational concept and significance of Med-VQA, Section 2 explores the background and core concepts necessary for understanding Med-VQA, including large language models, medical visual language pre-training, self-supervised learning, and multimodal learning, all of which enhance the capabilities of Med-VQA systems.

Section 3 delves into the application of large language models within Med-VQA, emphasizing their integration in reporting and answer generation while discussing challenges and potential future directions. Section 4 provides a thorough examination of techniques employed for pre-training models on medical visual language data, highlighting the use of paired and unpaired vision and text datasets through self-supervised learning. It also reviews various frameworks and benchmarks utilized to evaluate these pre-trained models, discussing their effectiveness in enhancing downstream medical tasks and addressing challenges in the medical domain [12, 13, 14].

In Section 5, the survey investigates the role of self-supervised learning in Med-VQA, detailing its advantages, such as improved model generalization with limited labeled data, and the challenges it encounters, including effective integration of visual and textual information. Section 6 further explores the integration of multimodal learning approaches within healthcare, assessing their impact on medical outcomes, reviewing relevant frameworks and models, and discussing the complexities associated with data integration, which are critical for advancing medical AI [14, 15, 16, 17, 9].

The survey further examines current applications of computer vision in healthcare in Section 7, focusing on advancements in medical image analysis and diverse applications in medical imaging. Section 8 identifies key challenges in Med-VQA development, such as data scarcity, model interpretability, and ethical concerns, proposing future research directions. In conclusion, Section 9 encapsulates the pivotal themes discussed throughout the document, emphasizing the critical role of Med-VQA in transforming healthcare practices by enhancing diagnostic accuracy and facilitating improved patient outcomes through advanced AI technologies [18, 19, 20, 10, 9].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Medical Visual Question Answering (Med-VQA)

Medical Visual Question Answering (Med-VQA) is a pivotal element in medical AI, facilitating the integration of visual and textual data to resolve intricate clinical queries [21]. Its core objective is the precise interpretation of medical images and the generation of responses ranging from simple binary to complex descriptive questions, thereby enhancing diagnostic accuracy and medical assessments [22, 23]. The scarcity of high-quality annotated datasets poses a significant challenge in developing robust Med-VQA systems capable of interpreting radiological images [23]. Techniques such as transfer learning and cross-modal fusion are employed to improve feature representation and integration of visual and linguistic data [22]. Vision-language pre-training (VLP) models further automate anomaly detection and report generation from medical scans, streamlining diagnostic processes [22].

Efforts to enhance Med-VQA systems' generalizability across clinical contexts include incorporating domain-specific knowledge into the training of Medical Vision Language Models (VLMs), which significantly boosts performance by integrating specialized medical insights [21]. Additionally, methodologies that learn and fuse information from 3D medical images and corresponding textual descriptions expand Med-VQA applications [22]. Beyond diagnostics, Med-VQA also facilitates medical report generation from images, necessitating high-quality cross-modal annotations like medical image-report pairs [23]. This capability enhances clinical decision-making and provides reliable, context-aware responses, improving medical practice efficiency.

2.2 Large Language Models

Large Language Models (LLMs) significantly enhance Med-VQA capabilities by leveraging advanced linguistic and contextual understanding to interpret complex medical images and generate precise responses. These models, based on the Transformer architecture, are categorized into Encoder-only, Decoder-only, and Encoder-Decoder structures, each contributing uniquely to medical LLMs and Multimodal Large Language Models (MLLMs) [24]. Their integration facilitates processing both textual and visual data, improving Med-VQA systems' accuracy and efficiency.

Despite their potential, integrating LLMs and vision-language models (VLMs) into clinical practice faces obstacles such as high costs, lack of transparency, and privacy concerns [25]. The reliability of Large Multimodal Models (LMMs) in Med-VQA is under scrutiny, as evaluations reveal they can perform worse than random guessing despite high benchmark accuracy [26]. Rigorous benchmarks comparing medical LLMs and VLMs to general-domain counterparts are essential for assessing domain-adaptive pretraining (DAPT) effectiveness in medical question-answering tasks [25].

Benchmarks focusing on clinical questions related to radiology images are crucial for evaluating model performance across categories like Modality, Plane, Organ System, and Abnormality [27]. These evaluations are vital for assessing LLMs' effectiveness in interpreting medical images and generating clinically relevant answers. New pretraining paradigms for Large Multimodal Models (LMMs) enhance visual comprehension, essential for advancing Med-VQA systems [28].

In ophthalmology, frameworks categorizing Visual Question Answering (VQA) research highlight LLMs' potential to enhance Med-VQA tasks [29]. However, the absence of specialized LLMs for diagnosing ophthalmic diseases using multimodal data remains a challenge [30]. Furthermore, current VLMs' inadequacy in accurately interpreting medical images, due to generalist training and static datasets, underscores the need for domain-specific adaptations [6].

Integrating LLMs into Med-VQA systems is a transformative advancement in medical AI, enabling the synthesis of multimodal data to enhance diagnostic accuracy and robustly support healthcare professionals in clinical decision-making. By capitalizing on multimodal large language models (MLLMs), these technologies facilitate improved patient health insights and optimize applications like diagnosis classification and medical imaging. However, challenges such as data limitations and ethical considerations must be addressed to ensure effective and responsible implementation in medical practice [31, 32]. Ongoing LLM development and adaptation are crucial for tackling complex Med-VQA challenges, advancing AI-driven healthcare solutions.

2.3 Medical Visual Language Pre-Training

Medical visual language pre-training is crucial for enhancing Med-VQA systems' performance by utilizing extensive datasets encompassing various medical modalities and conditions. This approach addresses the challenge of limited, high-quality annotated datasets, which are resource-intensive to produce [13]. By integrating visual and textual data, pre-training improves diagnostic accuracy and efficiency in clinical workflows, especially when navigating complex and heterogeneous medical data [4].

Advancements in medical contrastive vision-language pre-training have significantly guided representation learning without human annotations. These methods leverage diagnostic report information to enhance Med-VQA systems' interpretability and accuracy, aligning visual data with domain-specific insights [33]. Frameworks like Multi-Aspect Vision-Language Pre-training (MAVL) exemplify this approach by aligning disease descriptions with corresponding medical images, improving recognition accuracy [13].

The MuVAM model illustrates multi-view attention mechanisms focusing on image-to-question and word-to-text relationships, enhancing the model's capacity to process complex medical queries [33]. Benchmarks like Med-HallMark, designed for evaluating hallucinations in medical multimodal contexts, emphasize the necessity of rigorous evaluation frameworks [21].

Frameworks such as EHRXQA, comprising extensive question-answer pairs across various modalities, are crucial for facilitating diverse multimodal reasoning tasks within medical visual language contexts [23]. These comprehensive datasets and pre-training strategies ensure Med-VQA systems can manage medical data intricacies, ultimately supporting healthcare professionals in clinical decision-making.

2.4 Self-Supervised Learning

Self-supervised learning (SSL) is a transformative strategy in developing Med-VQA systems, utilizing large amounts of unlabeled data to learn meaningful feature representations without extensive human annotations [34]. This approach is particularly beneficial in the medical field, where acquiring labeled datasets is challenging and resource-intensive. SSL techniques enhance Med-VQA models' robustness and accuracy, especially with limited labeled data [35].

A notable SSL application in Med-VQA involves innovative proxy tasks for feature learning from raw medical data. For instance, the Rubik's cube recovery task improves feature extraction from 3D volumetric data, showcasing SSL frameworks' potential to enhance performance in complex medical imaging scenarios [36]. Such techniques are crucial for improving Med-VQA systems' interpretability and diagnostic accuracy.

In dental panoramic radiograph analysis, SSL methods address insufficient training data challenges. Enhanced masked image modeling techniques improve radiograph analysis, demonstrating SSL approaches' adaptability to various medical imaging modalities [34]. These advancements highlight SSL's significance in addressing data scarcity issues and boosting Med-VQA systems' capabilities to generate accurate clinical insights.

Integrating self-supervised learning techniques into Med-VQA represents a substantial advancement in medical AI, enhancing model performance by leveraging limited training data through innovative pretraining methods like masked image modeling and contrastive learning, achieving state-of-the-art results across multiple medical VQA datasets [14, 9, 37, 16]. By continually exploring and refining SSL methodologies, researchers can further enhance Med-VQA systems' diagnostic capabilities and reliability, contributing to more effective and efficient healthcare delivery.

2.5 Multimodal Learning

Multimodal learning is foundational in developing Med-VQA systems, integrating diverse data modalities such as text, images, and genomic data to enhance interpretability and diagnostic accuracy in healthcare applications. This approach addresses the inherent complexity and richness of medical data, which often includes modalities like MRI, CT, X-Ray, and histopathology, each providing unique insights into patient health [13]. Multimodal learning overcomes unimodal systems' limitations, expanding their utility in varied clinical scenarios.

Recent advancements in multimodal learning include models like Med-2E3, enhancing 3D medical image analysis by integrating 3D and 2D features [38]. This integration allows comprehensive analysis of complex medical images, improving Med-VQA systems' capacity to provide accurate and context-aware diagnostic support. The MedBLIP model exemplifies using a query mechanism to align 3D medical image features with textual descriptions, facilitating effective multimodal representation learning and medical data interpretation [22].

Innovative approaches, such as fine-tuning the LLaMA2 model on a specialized ophthalmic dataset to create 'Ophtha-LLaMA2,' highlight multimodal learning's potential to deliver tailored diagnostic support for specific medical domains like ophthalmology [30]. This specialization enhances the model's ability to address domain-specific challenges and improve diagnostic accuracy.

Exploring spatial and semantic relationships in medical images is another critical focus area, as deeper understanding significantly enhances VQA systems' performance. Integrating advanced image enhancement techniques with multimodal architectures, particularly those employing Convolutional Neural Networks (CNN) and Transformer-based models for feature extraction, improves Med-VQA

systems' performance and diagnostic accuracy. Studies presented at the MEDVQA-GI 2023 challenge demonstrate that multimodal approaches combining image enhancement with BERT and various vision models significantly increase accuracy and F1-Score, enhancing systems' ability to provide precise clinical question answers alongside medical images [39, 32, 11, 40, 41]. Additionally, frameworks like the Multimodal Co-Attention Transformer (MCAT) integrate gigapixel Whole Slide Images (WSIs) and genomic features, utilizing a co-attention mechanism to learn interactions between histology and genomic data, enriching Med-VQA models' interpretative power.

In recent years, the application of large language models in the domain of Medical Visual Question Answering (Med-VQA) systems has garnered significant attention. This integration is crucial for advancing the capabilities of AI in healthcare, particularly in enhancing diagnostic accuracy and patient care. As illustrated in Figure 2, the figure delineates the multifaceted aspects of model integration, benchmarking, and applications within Med-VQA systems. Furthermore, it underscores current challenges faced by researchers, such as the necessity for improved data quality, the development of efficient training methods, and the enhancement of robustness in AI-driven healthcare solutions. Such insights not only inform ongoing research but also pave the way for future investigations aimed at optimizing these technologies for practical healthcare applications.

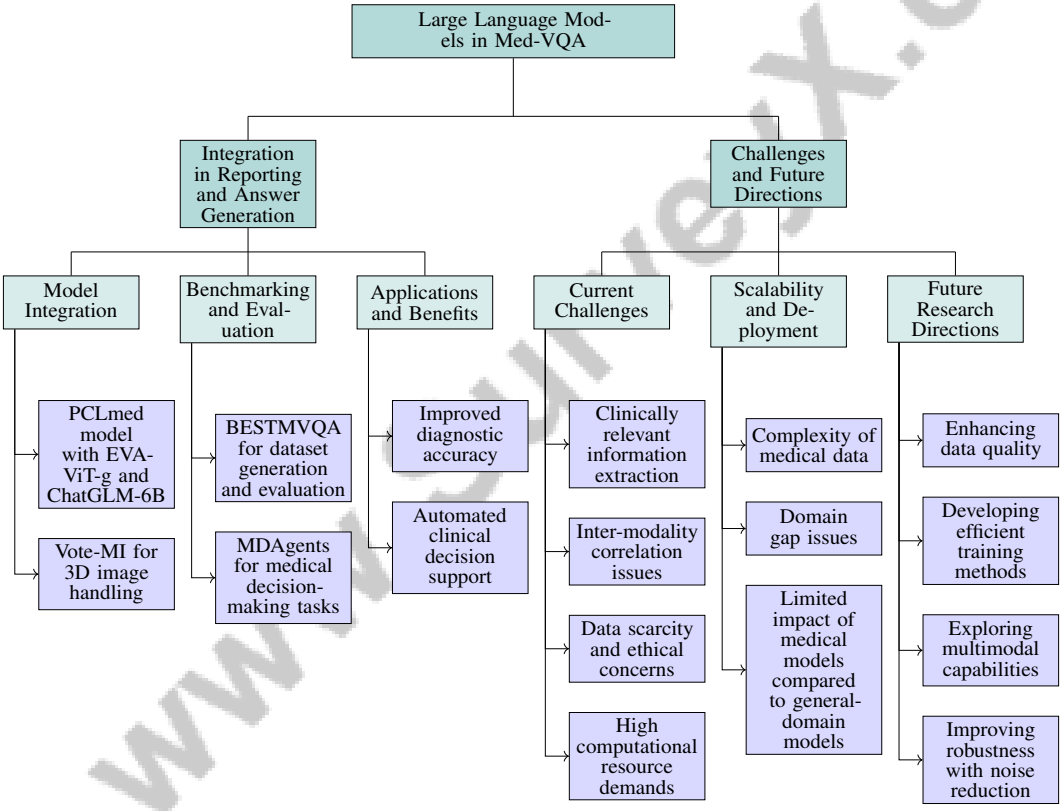


Figure 2: This figure illustrates the integration of large language models in Medical Visual Question Answering (Med-VQA) systems, highlighting key areas of model integration, benchmarking, and applications. It also outlines current challenges and future research directions, emphasizing the need for enhanced data quality, efficient training methods, and improved robustness for AI-driven healthcare solutions.

3 Large Language Models in Med-VQA

3.1 Integration in Reporting and Answer Generation

The integration of large language models (LLMs) into Medical Visual Question Answering (Med-VQA) systems enhances the precision and efficiency of reporting and answer generation by leveraging advanced linguistic capabilities to interpret complex medical images and produce contextually

relevant responses. The PCLmed model exemplifies this integration, utilizing a vision encoder (EVA-ViT-g) paired with a bilingual language model (ChatGLM-6B) through a lightweight query Transformer, thereby improving medical report generation [5]. Another approach, exemplified by Vote-MI, enhances the handling of 3D medical images by selecting representative 2D slices that retain critical information, demonstrating the adaptability of vision-language models (VLMs) in specialized medical applications [42]. Benchmarks like BESTMVQA systematically generate and evaluate datasets, addressing challenges of data insufficiency and reproducibility in Med-VQA, thus facilitating model performance assessment [21].

Evaluations of MDAgents across various medical decision-making tasks highlight the effectiveness of different models in addressing complex medical queries, underscoring the versatility of LLMs in medical applications [43]. The integration of LLMs into Med-VQA systems marks a significant advancement in medical AI. Utilizing advanced multimodal data synthesis techniques, such as those in the Medical Vision Language Learner (MedViLL), these models improve diagnostic accuracy by analyzing radiology images alongside unstructured reports. This integration enhances the understanding of complex medical data and provides healthcare professionals with automated support for critical clinical decision-making tasks, including diagnosis classification, medical image-report retrieval, and report generation [32, 44, 31, 23, 45]. Continuous innovation and refinement of these models promise to significantly enhance the efficacy and reliability of AI-driven healthcare solutions.

3.2 Challenges and Future Directions

Integrating large language models (LLMs) into Medical Visual Question Answering (Med-VQA) systems presents challenges necessitating innovative solutions and future research. A primary challenge is the effective extraction of clinically relevant information from medical images, where current methods exhibit limitations [46]. Existing Med-VLP approaches often prioritize intra-modality features over inter-modality correlations, restricting performance in tasks requiring a comprehensive understanding of both images and text [47]. This issue is compounded by the scarcity of high-quality training data, ethical concerns, and the substantial computational resources needed for model training and deployment [24].

The scalability and practical deployment of LLMs in clinical environments pose another significant hurdle. The complexity and high dimensionality of medical data, along with the high costs of performance computing devices, impede widespread adoption [48]. Furthermore, the domain gap between training data and real-world clinical data limits knowledge transfer, reducing model effectiveness in practical applications [49]. Benchmarks comparing medical LLMs and VLMs with general-domain counterparts reveal that most medical models do not significantly outperform their general-domain equivalents across various tasks, indicating a need for more specialized models [25].

The phenomenon of hallucination in Med-VQA models, especially in the NOTA (None of the Above) scenario, presents an additional challenge. Current models often struggle to accurately distinguish between confusing image pairs, sometimes performing below random guessing. Existing benchmarks tend to focus on general LVLM hallucinations, neglecting the unique complexities and critical nature of medical hallucinations [50]. Additionally, the reliance on 2D representations of 3D medical data in existing transfer learning methods results in significant loss of crucial anatomical information necessary for effective analysis and diagnosis [51].

Future research should aim to enhance data quality and develop more efficient training methods, alongside exploring multimodal capabilities in clinical settings [24]. Improving the robustness of Med-VQA frameworks could involve sophisticated noise reduction techniques and expanding datasets to encompass a broader variety of medical conditions and imaging modalities. Moreover, future work could focus on developing more effective data augmentation techniques and exploring different types of multi-modal transformers to enhance representation learning for both images and textual questions [41]. Addressing these multifaceted challenges will significantly improve the accuracy and reliability of AI-driven healthcare solutions, ultimately supporting clinicians in delivering more effective and efficient patient care.

Category	Feature	Method
Techniques and Strategies	Multimodal Integration	LQ-VQA[11], MB[22]
Evaluation and Benchmarking	Multimodal Assessment	VILA-M3[6]

Table 1: This table presents a summary of methods utilized in the development and evaluation of Medical Visual Question Answering (Med-VQA) systems. It categorizes the techniques into multimodal integration strategies and assessment methodologies, highlighting key contributions from recent studies in the field.

4 Medical Visual Language Pre-Training

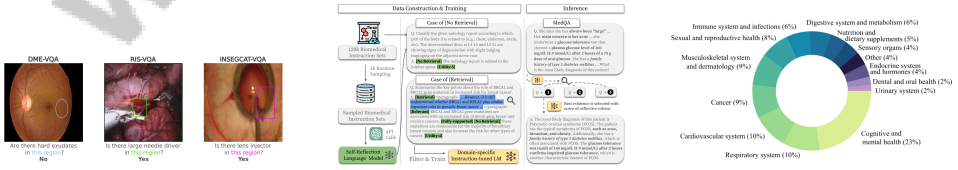
Pre-training in Medical Visual Question Answering (Med-VQA) systems is crucial for enhancing their performance in complex medical inquiries. Table 3 offers a comparative overview of different pre-training methods utilized in Med-VQA systems, emphasizing their data types, pre-training techniques, and evaluation metrics. Additionally, Table 1 provides a concise overview of the methods employed in Med-VQA systems, illustrating both the integration of multimodal data and the approaches used for their evaluation and benchmarking. This section explores the techniques and strategies that form the foundation of effective Med-VQA systems, focusing on how these methodologies optimize model capabilities.

4.1 Techniques and Strategies

The advancement of Med-VQA systems is significantly enhanced by sophisticated pre-training techniques that improve model performance across clinical scenarios. Utilizing multimodal datasets, such as SLAKE, which includes 642 radiology images with detailed annotations, is pivotal for developing models that require comprehensive visual and textual representations [4]. MedBLIP, a bootstrapping language-image pre-training model, exemplifies the integration of 3D medical images with textual information, enhancing diagnostic accuracy by aligning visual data with textual context [22].

The benchmark introduced by [23] employs an automatic pipeline for dataset construction, resulting in a dataset significantly larger and more diverse than previous efforts, emphasizing the importance of dataset diversity for model robustness. Cross-modal self-attention (CMSA) modules in multi-task pre-training frameworks improve the fusion of visual and linguistic features, enhancing answer prediction by aligning medical images with textual descriptions [52, 31, 32, 53].

The integration of pre-trained transformer-based models for joint embedding generation significantly improves performance in biomedical vision-and-language tasks compared to traditional CNN-RNN approaches. Models like LXMERT, VisualBERT, UNIER, and PixelBERT outperform CNN-RNN models in classifying thoracic findings, highlighting the advantages of multimodal representation [54, 55, 33]. Various pre-training objectives, such as masked prediction and contrastive learning, effectively capture complex interrelationships between visual and textual data, enhancing Med-VQA system performance [11, 40, 56, 57, 58].



(a) Three VQA questions about medical images[11]

(b) Data Construction Training vs Inference: A Comparative Study of Biomedical Instruction Sets[52]

(c) Healthcare and Medical Conditions[59]

Figure 3: Examples of Techniques and Strategies

As illustrated in Figure 3, various strategies are employed in medical visual language pre-training to enhance VQA models and domain-specific language models in healthcare. The examples highlight the capabilities of different VQA models, the methodologies of data construction and training, and the

distribution of healthcare conditions, emphasizing the multifaceted strategies employed to advance medical visual language pre-training [11, 52, 59].

4.2 Evaluation and Benchmarking

Benchmark	Size	Domain	Task Format	Metric
BESTMVQA[21]	42,000	Medical Imaging	Question Answering	Accuracy
SSL-OCT[60]	108,312	Retinal Disease Classification	Image Classification	Accuracy, Weighted F1
PMC-VQA[61]	227,000	Radiology	Question Answering	ACC, BLEU-1
GMAI-MMBench[62]	26,000	Medical Imaging	Visual Question Answering	Accuracy, F1-score
MedThink[63]	3,515	Medical Imaging	Visual Question Answering	Accuracy, BLEU-4
PathVQA[20]	32,799	Pathology	Visual Question Answering	Accuracy, F1
GEMeX[64]	1,605,575	Chest X-ray	Visual Question Answering	AR-score, A-score
MultiMedBench[65]	1,000,000	Radiology	Question Answering	Micro-F1-14, BLEU-1

Table 2: Table illustrating a selection of benchmarks utilized in the evaluation of Med-VQA systems, detailing their respective sizes, domains, task formats, and performance metrics. These benchmarks provide a comprehensive overview of the datasets used to assess model performance across various medical imaging and question-answering tasks.

Evaluating and benchmarking Med-VQA systems are essential for assessing model effectiveness across clinical tasks. Table 2 presents a detailed summary of key benchmarks used for evaluating Med-VQA systems, highlighting their significance in advancing model performance across diverse clinical tasks. Metrics such as accuracy, F1-score, BLEU-4, and ROUGE provide a comprehensive evaluation of model performance [6]. Datasets like BESTMVQA, with diverse medical images and questions, are crucial for rigorous model testing [21].

Cross-modal evaluation techniques enhance understanding of the alignment between visual and textual data, supporting improved diagnostic accuracy and various medical tasks. Advancements in vision-language pre-training frameworks, such as MedViLL, demonstrate significant performance improvements across medical imaging tasks, underscoring the importance of effective cross-modal interactions [53, 66, 32, 47]. These techniques are vital for assessing approaches that enhance Med-VQA systems, ultimately aiding healthcare professionals in clinical decision-making.

Feature	MedBLIP	CMSA	LXMERT
Data Type	3D Medical Images	Multimodal Datasets	Visual And Text
Pre-training Technique	Bootstrapping	Cross-modal Self-attention	Joint Embedding
Evaluation Metric	Not Specified	Not Specified	Not Specified

Table 3: This table provides a comparative analysis of three pre-training models used in Medical Visual Question Answering (Med-VQA) systems: MedBLIP, CMSA, and LXMERT. It highlights key features such as data type, pre-training technique, and evaluation metrics, offering insights into the methodologies employed for integrating multimodal data and optimizing model performance.

5 Self-Supervised Learning in Medical AI

5.1 Self-Supervised and Contrastive Learning Approaches

Self-supervised learning (SSL) and contrastive learning are pivotal in advancing Medical Visual Question Answering (Med-VQA) systems by leveraging extensive unlabeled data, thus enhancing performance across clinical tasks. These methodologies are particularly advantageous in the medical domain, where acquiring labeled datasets is resource-intensive. SSL significantly boosts performance while minimizing dependence on human annotations, effectively managing task-relevant and irrelevant information to enhance the generalization capacity of Med-VQA systems for more accurate interpretations of complex medical images [67].

The integration of SSL with vision-language models (VLMs) has notably improved diagnostic accuracy, particularly in disease identification from medical images. Techniques such as Rubik's Cube Recovery, which involves partitioning 3D medical volumes into cubes, permuting them, and training networks to restore the original configuration, enhance feature extraction and interpretability

in 3D medical imaging [36]. This method is crucial for advancing the diagnostic capabilities of Med-VQA systems in complex imaging contexts.

Contrastive learning, a subset of SSL, effectively aligns medical images with corresponding textual descriptions, facilitating comprehensive medical report creation [68]. Detailed prompts and auxiliary predictions improve diagnostic accuracy for minority pathologies and mitigate hallucinations, thereby enhancing Med-VQA reliability [34]. The SD-SimMIM framework exemplifies SSL's adaptability, combining self-distillation with masked image modeling to improve feature representation in dental panoramic X-ray analysis [34].

Furthermore, benchmarks such as the one established by [8] promote hands-free interaction in medical diagnostics, enhancing accessibility and usability. This benchmark underscores the critical role of integrating SSL and contrastive learning approaches to improve Med-VQA systems' functionality and usability in clinical settings.

5.2 Role in Med-VQA

SSL is instrumental in advancing Med-VQA systems by extracting meaningful semantic information from vast amounts of unlabeled medical data, significantly enhancing label efficiency and generalization across various tasks without extensive labeled datasets. However, its effectiveness is contingent on the availability of quality training data, particularly for specific cases like COVID-19, which can impact model performance [2]. Despite this limitation, SSL techniques empower Med-VQA systems to interpret complex medical images and generate accurate clinical responses, thereby assisting healthcare professionals in decision-making.

Frameworks such as SD-SimMIM illustrate the application of SSL, enhancing feature representation with limited training data and leading to improved performance in dental imaging tasks [34]. Additionally, SSL frameworks proposed by [36] significantly boost the performance of 3D deep learning networks on volumetric medical datasets without requiring additional annotations, highlighting SSL's potential in enhancing diagnostic capabilities in 3D medical imaging.

The integration of SSL with VLMs is exemplified by the MedBLIP model, where the MedQFormer module effectively connects 3D medical images with 2D pre-trained models, facilitating enhanced multi-modal learning [22]. This integration is crucial for developing robust feature representations, enabling Med-VQA systems to generalize effectively across diverse clinical scenarios.

Moreover, the benchmark introduced by [33] emphasizes the advantages of joint embeddings in enhancing model performance on multimodal tasks, particularly within the biomedical domain. This underscores the efficacy of SSL in improving the interpretative capabilities of Med-VQA systems.

6 Multimodal Learning for Healthcare

6.1 Frameworks and Models

Multimodal learning frameworks are integral to enhancing Medical Visual Question Answering (Med-VQA) systems by integrating diverse data modalities such as images, text, and clinical records. This integration improves the interpretative and diagnostic capabilities of AI in healthcare, providing comprehensive analyses that support clinical decision-making and improve patient outcomes [31, 19, 43, 32]. For example, the Med-2E3 framework combines 3D and 2D features for refined medical image analysis, thereby boosting diagnostic accuracy [38]. Similarly, MedBLIP aligns 3D medical images with textual data through query mechanisms, facilitating effective multimodal representation learning [22].

The fine-tuning of large language models on specialized datasets, such as 'Ophtha-LLaMA2' for ophthalmology, illustrates the potential of multimodal learning to deliver tailored diagnostic support in specific medical domains, addressing domain-specific challenges and enhancing diagnostic precision [30]. A key focus of these frameworks is exploring spatial and semantic relationships within medical images, significantly enhancing VQA system performance. The integration of image enhancement techniques with multimodal architectures, including CNN and Transformer-based models, further augments Med-VQA capabilities [13]. Frameworks like the Multimodal Co-Attention Transformer (MCAT) employ a co-attention mechanism to integrate gigapixel Whole Slide Images (WSIs) and genomic features, enriching the interpretative power of Med-VQA models.

6.2 Challenges in Data Integration

Integrating multimodal data in Med-VQA systems presents challenges that impede AI-driven healthcare solutions. A significant issue is the noise in unpaired data, complicating the alignment and fusion of diverse modalities such as images, text, and clinical records [69]. This noise, arising from discrepancies in data quality, format, and source, can lead to inconsistencies affecting Med-VQA performance and reliability. Additionally, reliance on external models for data alignment limits Med-VQA systems' generalizability across clinical scenarios [69]. These models often lack specificity to medical data, resulting in suboptimal multimodal input integration and interpretation, potentially introducing biases and errors that compromise diagnostic accuracy.

The complexity of medical data, characterized by high-dimensional and heterogeneous information from diverse imaging modalities and textual descriptions, poses significant integration challenges. Developing a unified framework that accurately reflects the clinical context requires advanced algorithms and substantial computational resources. The intricate semantics of biomedical texts can cause misalignments between medical images and their textual descriptions. Recent advancements in medical vision-language pre-training frameworks address these issues by decomposing disease descriptions into fundamental aspects, enhancing alignment between visual data and textual representations. These approaches utilize large-scale datasets and sophisticated models to improve disease recognition accuracy, facilitating more effective clinical decision-making and patient care [31, 23, 70, 53].

To address these challenges, efforts should focus on developing advanced techniques for noise reduction and data harmonization, as well as designing specialized models capable of seamlessly integrating and processing multimodal medical data. Tackling inherent challenges in Med-VQA, including diverse clinical queries and varying visual reasoning skills, can significantly enhance accuracy and reliability, ultimately improving clinical decision-making processes and patient care outcomes. Recent developments, such as conditional reasoning frameworks and multimodal learning methods, have shown promising results, with notable accuracy improvements for complex open-ended questions, highlighting the potential of Med-VQA systems to assist healthcare professionals in extracting relevant information from medical images and providing precise diagnostic answers [19, 39, 9, 37, 71].

7 Computer Vision Applications in Healthcare

7.1 Advancements in Medical Image Analysis

Recent advancements in medical image analysis, driven by the integration of computer vision and natural language processing, have substantially advanced Medical Visual Question Answering (Med-VQA) systems. These systems interpret medical images and respond to clinically relevant inquiries, leveraging transformer-based architectures to enhance diagnostic accuracy and reduce clinical errors [9, 41]. Despite their promise, Med-VQA systems are still in early stages, requiring further exploration to address domain-specific challenges. The integration of advanced computer vision techniques and the availability of high-quality medical imaging datasets have facilitated these advancements, with deep learning algorithms enabling precise interpretation of complex medical images.

Convolutional neural networks (CNNs) and transformer architectures have become pivotal in feature extraction and image classification, enhancing applications such as image captioning, object detection, and visual question answering through multimodal intelligence [12, 72]. These models demonstrate superior performance in identifying and classifying medical conditions across imaging modalities like MRI, CT, and X-rays, where accurate segmentation and analysis are crucial for generating contextually relevant answers in Med-VQA systems.

The introduction of multi-view learning and attention mechanisms has refined image analysis, enabling models to focus on relevant regions within medical images. This targeted approach enhances interpretation accuracy through region-based questioning, fostering a nuanced understanding of specific image areas while reducing computational complexity [18, 26, 53, 11].

Comprehensive benchmarks and evaluation frameworks are crucial in assessing image analysis models' performance, providing standardized datasets and metrics for comparative validation [37, 19,

57]. Insights from these evaluations enhance Med-VQA systems by addressing challenges related to reasoning capabilities and the integration of visual regions of interest, leading to more reliable and clinically relevant applications.

The integration of multimodal data, such as genomic information and electronic health records with medical imaging data, has opened new avenues for informed decision-making in healthcare. By amalgamating diverse data sources and employing advanced reasoning frameworks like Triangular Reasoning VQA (Tri-VQA), Med-VQA systems deliver precise diagnostic insights, improving the accuracy of responses to clinical questions associated with medical images [19, 37].

7.2 Diverse Applications in Medical Imaging

The application of computer vision in medical imaging has significantly diversified due to advancements in machine learning and deep learning techniques. These innovations have enhanced the processing and interpretation of complex medical data, leading to specialized models such as Medical Vision Language Learner (MedViLL) and various large visual language models (VLMs). These models improve diagnostic classification and automate report generation, integrating visual and textual information from medical images and reports to address challenges like image noise and resolution issues that can result in misdiagnosis [45, 32, 73]. Applications span various modalities, including MRI, CT, X-ray, and ultrasound, each providing unique insights into patient health.

In radiology, computer vision models automate the detection and classification of abnormalities in chest X-rays and CT scans, significantly aiding rapid diagnoses of conditions such as pneumonia and COVID-19 [2]. The integration of CNNs and transformer architectures refines model accuracy, facilitating detailed feature extraction and precise diagnostic interpretations [13].

In pathology, computer vision techniques analyze histopathological images, assisting pathologists in identifying cancerous cells and other conditions. These systems leverage high-resolution imaging and advanced algorithms to detect subtle patterns indicative of disease, supporting early diagnosis and treatment planning [30].

In ophthalmology, computer vision models assess retinal images for signs of diabetic retinopathy and other ocular diseases. Frameworks like 'Ophtha-LLaMA2' exemplify tailored diagnostic support in specific medical domains, enhancing the accuracy and efficiency of ophthalmic evaluations [30].

Beyond diagnostics, computer vision plays a role in surgical planning and navigation, where 3D reconstructions from imaging data assist surgeons in visualizing and planning complex procedures. This capability is particularly beneficial in minimally invasive surgeries, where precise navigation is critical for successful outcomes [38].

The integration of multimodal data—including genomic, clinical, and medical imaging information—has advanced personalized medicine by enabling comprehensive analyses of patient health. This integration supports innovative applications such as automated diagnosis classification, radiology report generation, and enhanced patient engagement, ultimately leading to better-tailored treatments and outcomes [31, 32, 65, 44]. By synthesizing diverse data sources, computer vision models offer comprehensive insights into patient health, facilitating accurate and individualized treatment plans.

8 Challenges and Future Directions

Addressing the challenges in Medical Visual Question Answering (Med-VQA) is essential for the development of effective systems. Key barriers include data scarcity and quality, which critically affect model performance and reliability in clinical environments. Successful Med-VQA systems must deliver accurate medical inquiry responses, with interpretability and transparency being vital for healthcare professionals' trust and patient safety [21, 19, 37, 63]. Analyzing these challenges is crucial for identifying pathways for innovation.

8.1 Data Scarcity and Quality

Data scarcity and quality present significant obstacles to advancing Med-VQA systems. The lack of large-scale labeled datasets hampers the training of models for accurate interpretation of complex medical images [23]. This issue is particularly pronounced in specialized domains like 3D brain

scans, where limited textual descriptions impede the generation of contextually accurate answers [22]. Additionally, datasets that do not match the unique characteristics of medical images can detrimentally impact performance.

Current benchmarks often lack sufficient data for training state-of-the-art models and fail to provide a comprehensive evaluation framework [21]. The complexity of medical data annotations, which require domain expertise, is exacerbated by privacy concerns [23]. Furthermore, the need for high-quality, specialized data to fine-tune general large language models (LLMs) poses a challenge, as poor data quality can lead to suboptimal performance.

Innovative strategies that leverage both labeled and unlabeled data, alongside advancements in personalized training methodologies and self-supervised learning, are essential to overcome these challenges. Utilizing large-scale datasets, such as the newly developed PMC-OA with 1.6 million image-caption pairs, can facilitate efficient training processes and improve outcomes in tasks like image-text retrieval and medical classification [23, 35]. Surmounting these obstacles will enhance Med-VQA systems' accuracy, ultimately aiding clinicians in delivering effective patient care.

8.2 Model Interpretability and Reliability

For Med-VQA systems to be integrated into clinical practice, they must provide transparent and reliable outputs for medical decision-making. A challenge is the superficial integration of image features into the language model's embedding space, which may result in a lack of deep understanding of visual tokens due to discrepancies between visual and textual modalities [28]. This highlights the need for sophisticated methods to harmonize these modalities.

Efforts to enhance interpretability include techniques like Vote-MI, which improve diagnostic accuracy and stability in zero-shot and few-shot learning tasks by selecting representative slices, thereby bolstering model reliability [42]. However, the complexity of open-ended questions in Med-VQA necessitates further research to effectively address these challenges.

The potential for overfitting, particularly when models rely heavily on expert feedback, remains a notable limitation, compounded by inadequate hyper-parameter tuning and the lack of domain-specific pre-training due to computational constraints. Additionally, reliance on curated gene sets may restrict the distinctiveness of learned features, impacting interpretability [38]. Moreover, the complexity of these models can complicate transparency and trustworthiness, critical for clinical applications, while issues like catastrophic forgetting, hallucination, and code-switching generation highlight the need for robust solutions to enhance reliability [33].

8.3 Ethical and Privacy Concerns

Deploying Med-VQA systems in clinical settings requires a thorough examination of ethical and privacy concerns to ensure responsible AI integration in healthcare. A significant ethical issue is the potential for hallucinations in AI-generated outputs, which can lead to misleading diagnostic information and compromise patient safety [22]. Addressing hallucinations is essential for improving model reliability, and future research should focus on expanding datasets and refining categorization methods to enhance detection in diverse medical scenarios [22]. Incorporating manual verification processes is also crucial for ensuring the accuracy of LLMs in real-world applications [6].

Privacy concerns are paramount in Med-VQA systems, which often require access to sensitive patient data, including medical images and electronic health records. Ensuring the confidentiality and security of this data is vital for maintaining patient trust and complying with regulatory standards. Research highlights the retrospective use of open-access human subject data, emphasizing the need for careful data governance [34]. Future studies could explore frameworks like Retrieval-Augmented Generation (RAG) to dynamically incorporate relevant information during inference, enhancing representation learning while maintaining data privacy [6].

Moreover, the ethical implications of deploying survival outcome prediction models in clinical settings, particularly concerning genomic data privacy, require careful consideration [22]. Integrating additional modalities or larger datasets to improve model performance necessitates a thoughtful approach to data governance and ethical use [22]. Additionally, the computational demands of Med-VQA systems raise ethical considerations related to resource allocation and environmental

impact. Future research should investigate methods to dynamically compute sparse attention to mitigate computational costs, especially in scenarios with numerous answer classes [6].

Future research should prioritize pre-training vision-and-language models with broader biomedical datasets and explore applications across a wider range of tasks [22]. Investigating additional data augmentation strategies and extending approaches to domains beyond medical VQA are promising directions [34]. Enhancing federated learning approaches and developing unified models for diverse medical modalities can also contribute to addressing ethical and privacy concerns [6].

9 Conclusion

The exploration of Medical Visual Question Answering (Med-VQA) underscores its pivotal role in revolutionizing healthcare by integrating visual and textual data to address intricate clinical questions. These systems significantly enhance diagnostic accuracy by generating precise and contextually aware medical reports, thereby aiding radiologists and optimizing diagnostic processes. The development of comprehensive datasets, such as SLAKE, has been instrumental in advancing the field, providing semantically rich data crucial for the training and evaluation of Med-VQA models, with the ultimate goal of refining clinical decision-making.

The potential of Multimodal Vision Language Models (MVLMs) to transform healthcare services is evident, highlighting the necessity for interdisciplinary efforts to overcome existing challenges and further the domain. Innovations in 3D medical image analysis exemplify the capabilities of Med-VQA systems to enhance clinical applications, often outperforming traditional models in report generation and medical VQA tasks. Moreover, efforts to improve generalizability and reduce bias in clinical radiology are essential for strengthening the robustness of Med-VQA systems.

The survey also emphasizes the importance of incorporating expert knowledge into methodologies for medical image analysis, thereby enhancing diagnostic support for clinicians. The improvements in VQA performance underscore the value of integrating vision and language models to improve healthcare outcomes. This survey provides a comprehensive resource for researchers in computer vision and natural language processing, highlighting the critical contribution of Med-VQA to the advancement of healthcare.

Despite significant progress, challenges such as data scarcity and the need for diverse question categories remain, emphasizing the necessity for continuous innovation and the development of high-quality datasets to improve patient care. With ongoing advancements and the integration of diverse data sources, Med-VQA systems hold considerable promise for enhancing clinical decision-making and patient outcomes.

References

[1] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

[2] Guangyi Liu, Yinghong Liao, Fuyu Wang, Bin Zhang, Lu Zhang, Xiaodan Liang, Xiang Wan, Shaolin Li, Zhen Li, Shuixing Zhang, et al. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):3786–3797, 2021.

[3] Deepak Gupta, Swati Suman, and Asif Ekbal. Hierarchical deep multi-modal network for medical visual question answering, 2020.

[4] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[5] Bang Yang, Asif Raza, Yuexian Zou, and Tong Zhang. Pclmed at imageclefmedical 2023: Customizing general-purpose foundation models for medical report generation. In *CLEF (Working Notes)*, pages 1754–1766, 2023.

[6] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915*, 2024.

[7] Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi. Muvam: A multi-view attention-based model for medical visual question answering, 2021.

[8] Tonmoy Rajkhowa, Amartya Roy Chowdhury, Sankalp Nagaonkar, and Achyut Mani Tripathi. Tm-pathvqa:90000+ textless multilingual questions for medical visual question answering, 2024.

[9] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey, 2023.

[10] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[11] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Localized questions in medical visual question answering, 2023.

[12] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

[13] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A. Linte, and Binod Bhattarai. Medical vision language pretraining: A survey, 2023.

[14] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A Linte, and Binod Bhattarai. Medical vision language pretraining: A survey. *arXiv preprint arXiv:2312.06224*, 2023.

[15] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering, 2023.

[16] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medical visual question answering, 2022.

[17] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023.

[18] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Targeted visual prompting for medical visual question answering, 2024.

[19] Lin Fan, Xun Gong, Cenyang Zheng, and Yafei Ou. Tri-vqa: Triangular reasoning medical visual question answering for multi-attribute analysis, 2024.

[20] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

[21] Xiaojie Hong, Zixin Song, Liangzhi Li, Xiaoli Wang, and Feiyan Liu. Bestmvqa: A benchmark evaluation system for medical visual question answering, 2023.

[22] Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts, 2023.

[23] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[24] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[25] Daniel P. Jeong, Pranav Mani, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. The limited impact of medical adaptation of large language and vision-language models, 2025.

[26] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa, 2024.

[27] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

[28] Yin Xie, Kaicheng Yang, Ninghua Yang, Weimo Deng, Xiangzi Dai, Tiancheng Gu, Yumeng Wang, Xiang An, Yongle Zhao, Ziyong Feng, Roy Miles, Ismail Elezi, and Jiankang Deng. Croc: Pretraining large multimodal models with cross-modal comprehension, 2024.

[29] Xiaolan Chen, Ruoyu Chen, Pusheng Xu, Weiyi Zhang, Xianwen Shang, Mingguang He, and Danli Shi. Visual question answering in ophthalmology: A progressive and practical perspective, 2024.

[30] Huan Zhao, Qian Ling, Yi Pan, Tianyang Zhong, Jin-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, Shi-Nan Wu, Min Kang, Zihao Wu, Zhengliang Liu, Xi Jiang, Tianming Liu, and Yi Shao. Ophtha-llama2: A large language model for ophthalmology, 2023.

[31] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[32] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multimodal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

[33] Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports, 2020.

[34] Amani Almalki and Longin Jan Latecki. Enhanced masked image modeling for analysis of dental panoramic radiographs, 2023.

[35] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube, 2019.

[37] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.

[38] Yiming Shi, Xun Zhu, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. Med-2e3: A 2d-enhanced 3d medical multimodal large language model, 2024.

[39] Triet M. Thai, Anh T. Vo, Hao K. Tieu, Linh N. P. Bui, and Thien T. B. Nguyen. Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering, 2023.

[40] Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. Miss: A generative pre-training and fine-tuning approach for med-vqa. In *International Conference on Artificial Neural Networks*, pages 299–313. Springer, 2024.

[41] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.

[42] Yuli Wang, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Wen-Chi Hsu, Maliha Imami, Zhicheng Jiao, Paul Zhang, et al. Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. *Advances in Neural Information Processing Systems*, 37:99947–99964, 2024.

[43] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, et al. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

[44] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multimodal understanding and generation for medical images and text via vision-language pre-training, 2022.

[45] Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, Yutong Xie, and Qi Wu. A survey of medical vision-and-language applications and their techniques, 2024.

[46] Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. Lapa: Latent prompt assist model for medical visual question answering, 2024.

[47] Ke Zhang, Yan Yang, Jun Yu, Hanliang Jiang, Jianping Fan, Qingming Huang, and Weidong Han. Multi-task paired masking with alignment modeling for medical vision-language pre-training. *IEEE Transactions on Multimedia*, 26:4706–4721, 2023.

[48] Bingqian Lin, Zicong Chen, Mingjie Li, Haokun Lin, Hang Xu, Yi Zhu, Jianzhuang Liu, Wenjia Cai, Lei Yang, Shen Zhao, Chenfei Wu, Ling Chen, Xiaojun Chang, Yi Yang, Lei Xing, and Xiaodan Liang. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining, 2023.

[49] Xinpeng Ding, Ziwei Liu, and Xiaomeng Li. Free lunch for surgical video understanding by distilling self-supervisions, 2022.

[50] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models, 2024.

[51] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis, 2019.

[52] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, 2024.

[53] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2024.

[54] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984, 2024.

[55] Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1999–2004. IEEE, 2020.

[56] Xinyue Hu, Lin Gu, Kazuma Kobayashi, Qiyuan An, Qingyu Chen, Zhiyong Lu, Chang Su, Tatsuya Harada, and Yingying Zhu. Interpretable medical image visual question answering via multi-modal relationship graph learning, 2023.

[57] Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. R-llava: Improving med-vqa understanding through visual region of interest, 2025.

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

[59] Juraj Vladika, Phillip Schneider, and Florian Matthes. Medreqal: Examining medical knowledge recall of large language models via question answering, 2024.

[60] Luffina C. Huang, Darren J. Chiu, and Manish Mehta. Self-supervised learning featuring small-scale image dataset for treatable retinal diseases classification, 2024.

[61] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[62] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[63] Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: Explaining medical visual question answering via multimodal decision-making rationale, 2024.

[64] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis, 2024.

[65] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[66] Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lei Wang, Lingqiao Liu, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. A systematic evaluation of gpt-4v's multimodal capability for medical image analysis, 2024.

[67] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

[68] Tim J. M. Jaspers, Ronald L. P. D. de Jong, Yasmina Al Khalil, Tijn Zeelenberg, Carolus H. J. Kusters, Yiping Li, Romy C. van Jaarsveld, Franciscus H. A. Bakker, Jelle P. Ruurda, Willem M. Brinkman, Peter H. N. De With, and Fons van der Sommen. Exploring the effect of dataset diversity in self-supervised learning for surgical computer vision, 2024.

[69] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey, 2024.

[70] Tianyuan Yao, Chang Qu, Jun Long, Quan Liu, Ruining Deng, Yuanhan Tian, Jiachen Xu, Aadarsh Jha, Zuhayr Asad, Shunxing Bao, Mengyang Zhao, Agnes B. Fogo, Bennett A. Landman, Haichun Yang, Catie Chang, and Yuankai Huo. Compound figure separation of biomedical images: Mining large datasets for self-supervised learning, 2022.

[71] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder, 2023.

[72] Moseli Mots'oehli. Assistive image annotation systems with deep learning and natural language capabilities: A review, 2024.

[73] Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 172–176. IEEE, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.