Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/135/2025.21216

Advancements and Challenges of Multimodal Models in Medical Applications

Zimeng He[1,a,*]

[1]*College of Engineering, Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China*
a. 12112007@mail.sustech.edu.cn
**corresponding author*

Abstract: Multimodal models have demonstrated significant potential in the medical field integrating information from various modalities such as images and text to improve understanding and reasoning. This paper provides a comprehensive review of their applications, focusing on medical visual question answering (VQA), medical report generation, and surgical assistance systems. In VQA, multimodal models like MedFuseNet and XrayGPT enhance patient-doctor communication and assist in disease diagnosis. For medical report generation, models such as Medical-VLBERT and RadFM automate report writing, alleviating the workload of healthcare professionals while improving accuracy. In surgical assistance, models like Surgical-LVLM and PitVQA-Net support surgical localization, pathological analysis, and procedural annotations. Despite these advancements, challenges persist, including data scarcity, limited model interpretability, and difficulty adapting to dynamic medical scenarios. The lack of diverse and annotated datasets, particularly for rare diseases, hinders the models' generalization capabilities. Furthermore, ensuring patient privacy and compliance with regulatory frameworks is critical for broader adoption. This review synthesizes recent developments, highlights challenges, and provides insights into the future of multimodal AI in healthcare. By advancing intelligent healthcare systems, multimodal models have the potential to transform clinical practices, improve diagnostic accuracy, and enhance patient outcomes.

Keywords: Multimodal models, Vision-language models, Visual question answering, Surgical assistance.

1. Introduction

The advancement of deep learning has led to significant breakthroughs in areas like computer vision, natural language processing, and information retrieval. Among these, multimodal models have made incredible progress. Multimodal refers to the processing and fusion of information from different forms (modalities), such as images, text, language, video, and even sensory information like touch. Integrating information from various modalities enables a more comprehensive and accurate understanding and reasoning.

Taking vision-language models as an example, this model integrates information from both natural language processing and computer vision modalities. During training, the model can jointly train on images and descriptive texts, used to generate natural language descriptions of images or enhance

DOI: 10.54254/2755-2721/135/2025.21216

image classification tasks with text. Vision-language models typically consist of three key components: an image encoder, a text encoder, and a strategy for merging the information from both encoders. Some typical pretraining objectives and strategies include: contrastive learning – by contrastive learning, images and texts are mapped into a shared feature space, aiming to bring the embeddings of images and their corresponding textual descriptions closer, while embeddings of unrelated images and texts are pushed further apart; cross-attention-based multimodal fusion – by establishing interactions between text and image information, allowing the model to better focus on key information in the image when generating text descriptions; PrefixLM – treating image embeddings as prefixes to a text language model, thus combining image information with the text generation process[1-3]. Many mature pre-trained vision-language models have been developed, such as Contrastive Language-Image Pretraining (CLIP) [4], Fusion of Language and Vision (FLAVA) [5], Bootstrapping Language-Image Pretraining (BLIP) [6], and Vision-and-Language Transformer (ViLT) [7], among others.

In the past decade, deep learning has completely transformed medical data analysis, especially in the fields of medical image processing, disease diagnosis, classification, and prediction. The continuous evolution of deep learning architectures, particularly convolutional neural networks (CNNs) and transformer-based networks, has brought revolutionary progress to the medical field, providing new solutions, especially in image segmentation and complex task handling [8]. Vision-language models have made significant breakthroughs in performance, use cases, and applications. These models are increasingly being applied in the medical field. They effectively combine images, text, and other biomedical data to assist in medical diagnosis, report generation, and the intelligent development of surgical robotic systems, significantly improving the efficiency and accuracy of medical work. This paper reviews various applications of vision-language models in the medical field, including medical visual question answering (VQA), medical report generation, and surgical assistance systems [9]. This paper focuses on the current applications of vision-language models in disease diagnosis, medical image analysis, and report generation, analyzing the challenges and potential of multimodal technologies in medical scenarios. In addition, this paper also explores the application of these technologies in surgical assistance systems, discussing how visual and language models can improve the precision and safety of surgeries. Through this comprehensive review of the application of vision-language models in the medical field, this paper provides the latest technological advancements and challenges for researchers, doctors, and professionals in related fields. The discussion not only provides theoretical support for the optimization of medical image analysis, intelligent diagnosis, and surgical assistance systems but also offers guidance for the future innovative applications of multimodal medical AI. With the continuous advancement of technology, these models are expected to play a larger role in clinical medicine, driving the widespread adoption and popularization of medical AI technologies.

2. Medical visual question answering

In daily life, not everyone has sufficient medical knowledge to answer medical-related questions, and searching for answers online often does not yield timely or reliable results. At such times, a reliable VQA system can be of great help. A VQA system trained with multimodal models can process the image information provided by the questioner and descriptions related to diseases, providing accurate and reliable answers. It can serve as a medical assistant for the general public in daily life and also provide auxiliary advice and reference support for healthcare professionals during diagnostic analysis. In the medical process, a VQA system can help patients obtain medical explanations related to their images without directly asking the doctor. This can reduce communication pressure between patients and doctors, as well as alleviate anxiety and confusion caused by information asymmetry or incorrect information.

DOI: 10.54254/2755-2721/135/2025.21216

2.1. Application case

In recent studies, some powerful VQA systems have achieved promising application results. In the [10], MedFuseNet is an attention-based multimodal deep learning model specifically designed to address answer categorization and answer generation tasks in medical visual question answering (VQA). The model has demonstrated strong performance on both the MED-VQA and PathVQA datasets, covering both answer categorization and answer generation tasks. Through attention visualization, MedFuseNet can focus on key regions of the image relevant to the question, aiding in the interpretability of medical image analysis. In the [11], CGMVQA proposes a model that can perform both classification and answer generation, allowing complex medical VQA tasks to be broken down into multiple simpler tasks. Unlike diagnosing a single disease, the CGMVQA model can answer questions involving multiple types of medical images, transforming a strong AI problem into multiple weak AI problems, thereby addressing various types of complex issues. The model has achieved new state-of-the-art results on the ImageCLEF 2019 VQA-Med dataset. In the [12], LLaVA-Med is a visual-language conversational agent designed to respond to open-ended research inquiries related to biomedical images, trained efficiently and cost-effectively. The core idea is to leverage a large-scale, broad-coverage biomedical figure-caption dataset extracted from PubMed Central, using GPT-4 for self-instruction to generate open-ended instruction-following data from captions, and fine-tuning a large general-domain vision-language model. LLaVA-Med can assist in answering questions about biomedical images based on open-ended instructions. LLaVA-Med surpasses previous supervised models on specific metrics in the VQA-RAD, SLAKE, and PathVQA datasets. In the [13], XrayGPT is an innovative conversational medical vision-language model designed to analyze and answer open-ended questions related to chest X-rays. XrayGPT combines a medical vision encoder (MedCLIP) and a large language model (Vicuna). The model is trained on the MIMIC-CXR and OpenI X-ray datasets along with their associated reports, enabling efficient analysis of chest X-rays. On the MIMIC-CXR test set, XrayGPT outperforms the baseline model (Mini-GPT-4), and when compared to ChatGPT, the answers generated by XrayGPT are superior in terms of medical accuracy.

2.2. Potential problem

The application of VQA models in the medical field is rapidly developing, but it still faces some challenges and issues. In terms of language semantics, although large language models have matured and been widely applied, they still encounter challenges when applied to the medical domain. The medical field contains a vast amount of specialized terminology, which is often abstract and complex. The diagnosis of biomedical images requires an accurate understanding and interpretation of these terms. In medical applications, the interpretability of models is particularly important. Doctors need to understand the reasoning behind the diagnoses and recommendations provided by AI models, rather than just the model's output. Many current deep learning models, especially complex multimodal models, often struggle to understand the specific answer generation process, making it difficult to explain their decision-making process [9,14]. This creates challenges for the widespread adoption of these models in clinical settings, especially when decisions involving life and death are at stake, where the trust doctors place in the model is crucial. There are currently some excellent and reliable medical image datasets available for training and testing multimodal models. However, the scale and quality of these datasets are still far from sufficient to support the training of more complex and refined biomedical AI models. Additionally, medical image datasets often lack sufficiently diverse annotations, and many datasets have a limited number of pathological types, which restricts the model's generalization ability. This is especially problematic in the detection and diagnosis of rare diseases, where the lack of adequate annotated data can significantly impact the model's training performance. Current medical VQA tasks are mostly focused on a limited set of question categories,

DOI: 10.54254/2755-2721/135/2025.21216

such as disease classification and lesion localization within images. While these tasks are important for clinical diagnosis, they often fail to cover all the complex questions encountered in medicine. In real clinical scenarios, doctors may need to ask a wider range of questions about a patient's medical images, which could involve knowledge from multiple domains and require deeper understanding and reasoning abilities from the model. Therefore, the limited scope of question categories restricts the applicability of these models in real-world medical applications.

3. Medical report generation

Medical data itself has multimodal characteristics, including text (clinical notes, radiology reports), images (X-rays, CT scans), videos (surgical procedure recordings), and tabular data (vital signs, laboratory results). By combining visual and textual information, multimodal models can generate more accurate medical reports. The generation of medical reports is a tedious and repetitive task for doctors, with a significant amount of time spent writing reports during consultations and treatment. Multimodal models are capable of handling this task. By extracting information from both images and text, these models can generate contextually relevant medical reports. These reports not only provide simple descriptions of images but also integrate multiple types of information, offering a more comprehensive disease assessment and treatment recommendations. Under the supervision and guidance of doctors, the generated medical reports can provide patients with more detailed and accurate medical information, offering interpretability, and providing convenience to both doctors and patients.

3.1. Application case

Some researchers have begun exploring the field of report generation and have achieved promising results. In the [15], Medical-VLBERT is a model developed to detect abnormalities in COVID-19 scans and autonomously generate medical reports based on the identified lesion regions. During the COVID-19 pandemic, automated medical report generation can significantly alleviate the burden on doctors. The model adopts an alternating learning strategy that includes knowledge pretraining and knowledge transfer: it memorizes knowledge from medical texts and utilizes the acquired knowledge to generate professional medical sentences by observing medical images. The model is trained and fine-tuned on the COVID-19 CT dataset and the CX-CHR dataset, and the experimental results have been validated by three radiologists. In the [16], EMIXER is an end-to-end multimodal X-ray generation model that can simultaneously generate paired chest X-ray images and corresponding reports conditioned on diagnostic labels. The specific steps include generating images based on labels, producing corresponding text from image embeddings, and evaluating the quality of both the images and the text. Several radiologists rated the realism and quality of the synthetic data with an average score of 7.340/10, compared to 7.825/10 for real data. The synthetic datasets generated by EMIXER can be used in data-limited medical scenarios to provide augmented data for medical scenarios with scarce data, thereby improving the performance of downstream tasks. In the [17], RadFM is a general medical foundation model for radiology, designed to handle a wide range of clinical radiology tasks by learning from medical scans (such as X-rays, CT, MRI, PET, etc.) and their corresponding textual descriptions/reports. The model is trained at scale on the MedMD dataset to acquire extensive knowledge of medical terminology and imaging. MedMD covers a broad range of radiology modalities, involving 17 medical systems such as breast, cardiac, central nervous system, chest, etc., and includes over 5,000 diseases. RadFM supports multimodal report generation for both 2D and 3D medical images and outperforms existing models across various clinical tasks.

DOI: 10.54254/2755-2721/135/2025.21216

3.2. Potential problem

Similar to medical VQA, medical report generation also faces the issue of lacking multimodal datasets. Data scarcity limits the ability of vision-language models to understand complex and rare clinical scenarios. Additionally, a good multimodal medical report requires annotations and writing by professional medical personnel, which demands more human resources and time. Currently, the most common multimodal medical report generation models are mostly applied in radiology, such as X-rays and CT scan images. However, there are still many other types of medical images that require further study and the generation of corresponding medical reports, such as ultrasound images, angiography images, and fundus images. The diagnosis and analysis of related diseases still offer extensive research opportunities. Medical data involves sensitive personal information, so ensuring patient privacy and adhering to relevant laws and regulations when training multimodal medical report generation models is also a crucial issue [18].

4. Surgical assistance

In recent years, medical surgical robots have made significant progress and development, with an increasing number of surgical robots being applied in actual surgical procedures. Prominent surgical robots such as the da Vinci Surgical System, Sensei X robotic catheter System and RVIR-CI Vascular Interventional Robot have broad application prospects in fields like cardiac surgery, spinal surgery, and vascular interventions [19, 20]. With the assistance of surgical robots, doctors can perform surgeries without direct physical contact with the patient, operating the robot to carry out the procedure. Surgical robots offer advantages such as greater precision, stability, and support for remote operations, improving both the efficiency and safety of surgeries. During the operation of surgical robots, the doctor/operator can assess the current surgical area and pathological conditions through real-time imaging. An appropriate multimodal model can be integrated into the surgical robot to assist the operator in tasks such as surgical localization, pathological analysis, and image segmentation, providing valuable support during the procedure and helping to complete the surgery more effectively [14]. Additionally, while the doctor is performing the surgery, a multimodal model can combine video streams to annotate the surgical process, and analyze, and summarize the steps taken, which can be used for surgical review and experience summary.

4.1. Application case

The article will demonstrate the application of multimodal models in surgical robots. In the [21], Surgical-LVLM is a personalized large vision-language model tailored for complex surgical scenarios. By combining a pre-trained large vision-language model with a specially designed Visual Perception LoRA (VP-LoRA) module, this model effectively handles visual-language tasks in surgery, particularly addressing visual grounding issues. The model can answer questions related to surgical images (e.g., "What is the current surgical instrument?") and accurately locate the visual regions involved in the questions (e.g., "The tool is located in the upper left corner of the image"). The model was tested on the publicly available EndoVis 2017 and 2018 datasets, as well as the newly introduced EndoVis Conversations dataset. The results significantly demonstrate improvements in surgical visual-language understanding and localization accuracy, establishing state-of-the-art performance on these datasets. In the [22], the PitVQA dataset is a specialized dataset focusing on VQA in the context of endonasal pituitary surgery. It consists of 25 procedural videos and a large number of question-answer pairs covering key surgical concepts. PitVQA-Net is a model specifically designed for pituitary surgery scenarios, combining image and text in a generative model. The model incorporates an image-text embedding approach that enhances the contextual alignment between surgical questions and images. PitVQA-Net has been validated on the PitVQA and EndoVis18-VQA

DOI: 10.54254/2755-2721/135/2025.21216

datasets, demonstrating its superior performance in pituitary surgery VQA tasks. In the [23], Surgical-VQA focuses on visual question answering systems in surgical scenarios. During the research, the existing MICCAI EndoVis 2018 (Endoscopic Vision Challenge) dataset and the Cholec80 workflow recognition dataset were extended, resulting in the construction of two new Surgical-VQA datasets, targeting classification-based and sentence-based question answering tasks. The task design is based on surgical tools, interactions between tools and tissues, and surgical workflows. Experimental results demonstrate that the improved model outperforms the baseline model (VisualBERT) on multiple surgical datasets, achieving a visual question answering task tailored to surgical scenarios and providing a new intelligent system framework for surgical assistance.

4.2. Potential problem

Although multimodal models show promising potential in surgical assistance and could play a role in certain medical scenarios, they still face many challenges. The dynamic and constantly changing nature of the surgical environment presents significant challenges for multimodal models. For example, the rapid changes in surgical tools, tissues, and lesion areas, real-time decision-making by the surgeon, and variations in the patient's condition must be processed and understood by the model in real time. However, existing multimodal systems often struggle to adapt to the rapidly changing surgical scenes, failing to provide useful feedback or guidance in real-time, especially in complex and non-standard surgical procedures. Additionally, multimodal models are not yet suitable for emergency surgeries. This is because, at present, the accuracy of multimodal models is not high, and they may occasionally provide incorrect information that could adversely affect the surgeon's judgment. In emergency surgeries, even the smallest error can delay the best opportunity for surgical intervention, leading to potentially irreversible consequences.

5. Medical education

In the medical field, outstanding medical experts often face heavy workloads, balancing clinical diagnosis and surgeries with academic research. Under such circumstances, it is challenging for medical experts to dedicate themselves fully to student training or to address students' questions promptly. Existing large language models can help resolve simple theoretical questions, and open surgical videos provide resources for students to study and observe. However, specific issues arising during surgical procedures still require expert guidance and answers, as current materials may not address real-world teaching needs. A reliable VQA model could play a significant role in medical education. A well-trained model can provide answers tailored to different surgical scenarios, delivering accurate teaching for various surgical stages. By leveraging contextual information, a VQA model can better localize surgical scenes, offering students a more reliable question-and-answer experience. This helps medical students acquire knowledge more effectively while alleviating the teaching burden on medical experts.

5.1. Application case

Some studies have started exploring the use of VQA models in medical education. The article [24] introduces two medical video datasets, MedVidCL and MedVidQA, aimed at advancing research in medical video understanding and VQA systems. The MedVidCL dataset contains 6,117 finely annotated videos for medical video classification tasks, while the MedVidQA dataset includes 3,010 health-related questions and their visual answers from 899 medical videos. These datasets can be used for tasks such as medical video classification and medical visual answer localization. Extensive experiments and validations conducted on these datasets demonstrate that multimodal approaches (combining vision and language) significantly enhance localization and classification performance.

DOI: 10.54254/2755-2721/135/2025.21216

The MedVidCL and MedVidQA datasets provide powerful tools for medical video analysis and offer valuable resources for artificial intelligence applications in the medical field, particularly in medical education. In the [25], SurgicalGPT is an end-to-end trainable Language-Vision GPT(LV-GPT) model that extends the GPT-2 model to include visual input (images). The model places word tokens before visual tokens, mimicking the human thought process of understanding the question and inferring the answer from the image. During surgery, the model is capable of correctly identifying key information such as tool locations and surgical stages.In intricate surgical settings, the model's generated answers closely align with the ground truth. The LV-GPT model outperforms other leading VQA models on two publicly accessible surgical VQA datasets: the 2018 Endoscopic Vision Challenge robotic scene segmentation and CholecTriplet2021, as well as on a newly annotated dataset based on an extensive surgical scene collection. This paper explores the application of the model in the field of medical teaching. Combined with some surgical process image data, the model can well complete the task of answering students' real-time questions, providing more possibilities for students to learn. In the [26], the CAT-ViL model not only provides simple answers in surgical VQA, but also enables the localization of the answers. CAT-ViL proposes an end-to-end Transformer model that integrates a co-attention mechanism and a gated embedding module, using ResNet18 to extract global visual features from the entire image, avoiding errors that may arise from traditional object detection-based methods. Experimental results demonstrate that the CAT-ViL model outperforms existing models in localization tasks, exhibiting strong stability under various perturbation conditions. This research provides an effective solution for visual question localization tasks in surgical scenarios, contributing to enhanced medical education and improved understanding of surgical procedures.

6. Medical education

Multimodal models have demonstrated remarkable potential in the medical field, revolutionizing areas such as medical image analysis, disease diagnosis, report generation, and surgical assistance. By integrating information from different modalities, these models enhance understanding, improve decision-making, and provide robust support across various medical applications. Despite significant progress, challenges such as data scarcity, lack of model interpretability, and dynamic adaptation in real-time scenarios remain. Addressing these issues will be crucial to ensuring the reliable and safe deployment of multimodal AI systems in clinical environments. The significance of this review lies in its detailed exploration of cutting-edge multimodal technologies and their implications in medicine. By synthesizing recent developments and identifying key challenges, it provides valuable insights for researchers, clinicians, and AI developers. As technology advances, multimodal models are expected to play an increasingly prominent role in transforming medical practices, enhancing efficiency and accuracy, and ultimately improving patient outcomes. The continued development and optimization of these technologies will pave the way for a new era of intelligent healthcare solutions.

References

[1] *Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 14(3–4), 163-352.*

[2] *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.*

[3] *Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18030-18040).*

[4] *Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K. W., ... & Keutzer, K. (2021). How much can clip benefit vision-and-language tasks?. arXiv preprint arXiv:2107.06383.*

DOI: 10.54254/2755-2721/135/2025.21216

[5] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15638-15650).

[6] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning (pp. 12888-12900). PMLR.

[7] Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In International conference on machine learning (pp. 5583-5594). PMLR.

[8] Guo, Z., Li, X., Huang, H., Guo, N., & Li, Q. (2019). Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences, 3(2), 162-169.

[9] Lin, Z., Zhang, D., Tao, Q., Shi, D., et al. (2023). Medical visual question answering: A survey. Artificial Intelligence in Medicine, 143, 102611.

[10] Sharma, D., Purushotham, S., & Reddy, C. K. (2021). MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports, 11(1), 19826.

[11] Ren, F., & Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access, 8, 50626-50636.

[12] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... & Gao, J. (2024). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36.

[13] Thawakar, O. C., Shaker, A. M., Mullappilly, S. S., et al. (2024, August). XrayGPT: Chest radiographs summarization using large medical vision-language models. In Proceedings of the 23rd workshop on biomedical natural language processing (pp. 440-448).

[14] Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., & Huang, X. (2024). A comprehensive survey of large language models and multimodal large language models in medicine. arXiv preprint arXiv:2405.08603.

[15] Liu, G., Liao, Y., Wang, F., Zhang, B., Zhang, L., Liang, X., ... & Cui, S. (2021). 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.

[16] Biswal, S., Zhuang, P., Pyrros, A., Siddiqui, N., Koyejo, S., & Sun, J. (2022, December). EMIXER: End-to-end Multimodal X-ray Generation via Self-supervision. In Machine Learning for Healthcare Conference (pp. 297-324). PMLR.

[17] Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463.

[18] Hartsock, I., & Rasool, G. (2024). Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence, 7, 1430984.

[19] Peters, B. S., Armijo, P. R., Krause, C., Choudhury, S. A., & Oleynikov, D. (2018). Review of emerging surgical robotic technology. Surgical endoscopy, 32, 1636-1655.

[20] Bao, X., Guo, S., Xiao, N., Li, Y., Yang, C., & Jiang, Y. (2018). A cooperation of catheters and guidewires-based novel remote-controlled vascular interventional robot. Biomedical microdevices, 20, 1-19.

[21] Wang, G., Bai, L., Nah, W. J., Wang, J., Zhang, Z., Chen, Z., ... & Ren, H. (2024). Surgical-LVLM: Learning to Adapt Large Vision-Language Model for Grounded Visual Question Answering in Robotic Surgery. arXiv preprint arXiv:2405.10948.

[22] He, R., Xu, M., Das, A., Khan, D. Z., et al. (2024, October). PitVQA: Image-Grounded Text Embedding LLM for Visual Question Answering in Pituitary Surgery. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 488-498). Cham: Springer Nature Switzerland.

[23] Seenivasan, L., Islam, M., Krishna, A. K., & Ren, H. (2022, September). Surgical-vqa: Visual question answering in surgical scenes using transformer. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 33-43). Cham: Springer Nature Switzerland.

[24] Seenivasan, L., Islam, M., Kannan, G., & Ren, H. (2023, October). SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery. In International conference on medical image computing and computer-assisted intervention (pp. 281-290). Cham: Springer Nature Switzerland.

[25] Gupta, D., Attal, K., & Demner-Fushman, D. (2023). A dataset for medical instructional video classification and question answering. Scientific Data, 10(1), 158.

[26] Bai, L., Islam, M., & Ren, H. (2023, October). CAT-ViL: co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 397-407). Cham: Springer Nature Switzerland.