



PitVQA: Image-Grounded Text Embedding LLM for Visual Question Answering in Pituitary Surgery

Runlong He^{1,2(✉)}, Mengya Xu³, Adrito Das¹, Danyal Z. Khan^{1,4},
Sophia Bano^{1,5}, Hani J. Marcus^{1,4}, Danail Stoyanov^{1,5},
Matthew J. Clarkson^{1,2}, and Mobarakol Islam^{1,2(✉)}

¹ UCL Hawkes Institute, University College London, London, UK

² Department of Medical Physics and Biomedical Engineering, University College London, London, UK

{runlong.he.23,mobarakol.islam}@ucl.ac.uk

³ Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

⁴ Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK

⁵ Department of Computer Science, University College London, London, UK

Abstract. Visual Question Answering (VQA) within the surgical domain, utilizing Large Language Models (LLMs), offers a distinct opportunity to improve intra-operative decision-making and facilitate intuitive surgeon-AI interaction. However, the development of LLMs for surgical VQA is hindered by the scarcity of diverse and extensive datasets with complex reasoning tasks. Moreover, contextual fusion of the image and text modalities remains an open research challenge due to the inherent differences between these two types of information and the complexity involved in aligning them. This paper introduces PitVQA, a novel dataset specifically designed for VQA in endonasal pituitary surgery and PitVQA-Net, an adaptation of the GPT2 with a novel image-grounded text embedding for surgical VQA. PitVQA comprises 25 procedural videos and a rich collection of question-answer pairs spanning crucial surgical aspects such as phase and step recognition, context understanding, tool detection and localization, and tool-tissue interactions. PitVQA-Net consists of a novel image-grounded text embedding that projects image and text features into a shared embedding space and GPT2 Backbone with an excitation block classification head to generate contextually relevant answers within the complex domain of endonasal pituitary surgery. Our image-grounded text embedding leverages joint embedding, cross-attention and contextual representation to understand the contextual relationship between questions and surgical images. We demonstrate the effectiveness of PitVQA-Net on both the PitVQA and the publicly available EndoVis18-VQA dataset, achieving improvements in balanced accuracy of 8% and 9% over the most recent baselines, respectively. Our code and dataset is available at <https://github.com/mobarakol/PitVQA>.

Keywords: Surgical VQA · Pituitary Tumor · Surgical Data Science · Vision Language Model

1 Introduction

Pituitary surgery, particularly through the endonasal approach, is a complex and delicate procedure that demands high precision and situational awareness [10]. Surgeons must navigate through critical anatomical structures, requiring not only an in-depth understanding of the surgical field but also the ability to adapt to dynamic intra-operative conditions [5]. In this context, the application of Visual Question Answering (VQA) technologies can offer substantial benefits, such as providing instant information on surgical phases and steps, tool usage, and tissue interactions, as well as offering predictive guidance on forthcoming phases, steps, and instrument requirements, thus enhancing the surgical workflow. The integration of Artificial Intelligence (AI), specifically large language model (LLM) or vision language models (VLM), into the operating room, promises to revolutionize surgical practices by providing real-time decision support, enhancing surgical precision, and fostering a more intuitive interaction between surgeons and technology [6, 11, 20]. A pivotal aspect of this integration is the development of systems capable of understanding and responding to complex visual and procedural contexts akin to human experts. VQA emerges as a promising field in this regard, especially within the surgical domain, where it can significantly augment intra-operative decision-making processes and post-operative surgical education [1, 2, 9, 20, 21].

The question-answer pairs in the VQA dataset include questions related to surgical images and correct answers. The surgical VQA dataset is crucial for building large language models (LLM) or vision language models (VLM) that can understand and reason about surgical images, answer questions related to surgical procedures and assist surgeons in performing complex tasks in the operating room. Several existing VQA datasets, such as EndoVis18-VQA [21], Cholec80-VQA [21], SSG-VQA [25], focus on nephrectomy and laparoscopic cholecystectomy surgeries. These datasets contain questions such as identification tasks (e.g., “What is the name of the white object that is retracted by the top-mid instrument?” [25]), and phase recognition (e.g., “What is the surgical phase of the image?” [20]). However, most of these datasets are limited in their representation of complex surgical tasks, size, and diversity. There are also a couple of VQA models associated with these datasets, including SurgicalVQA [21], SurgicalGPT [20], Surgical-VQLA [3], SSG-VQA-Net [25]. Nonetheless, these models often utilize poor image-text fusion and naive classification heads to convert the language generation model into VQA classification.

In this paper, we introduce a large and diverse surgical VQA dataset and design a VQA LLM network for pituitary surgery. Our contribution can be summarized as below:

- Builds a PitVQA dataset, a specialized dataset focused on VQA in the context of endonasal pituitary surgery, featuring 25 procedural videos and extensive Q&A pairs addressing key surgical concepts.

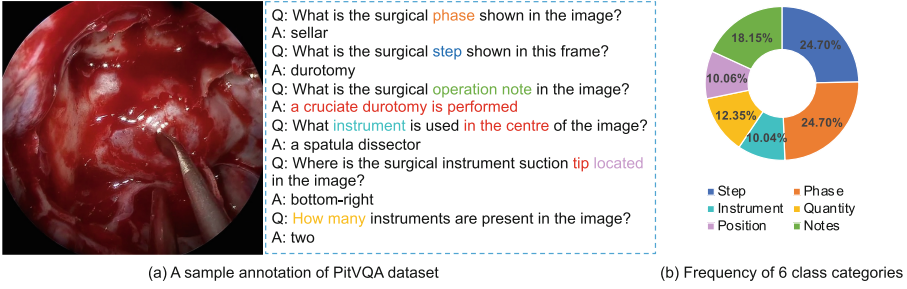


Fig. 1. PitVQA dataset of visual questions answering for pituitary surgery. There are overall 59 classes in the 6 class categories of phases, steps, instruments, quantity, positions and operation notes.

- Develops a PitVQA-Net, an adaptation of GPT2 [18] incorporating a novel image-text embedding and gated-attention excitation block (EB) classification head for surgical VQA.
- Design a image-grounded text embedding approach within PitVQA-Net, enhancing contextual alignment between surgical questions and imagery.
- Validates PitVQA-Net’s superior surgical VQA performance on both PitVQA and EndoVis18-VQA datasets, proving its efficacy in supporting surgeon-AI interactions and decision-making.

2 Method

2.1 Preliminaries

Vision-language processing seeks to establish a deep understanding of the connection between visual content and natural language. Contrastive Language-Image Pre-training (CLIP) [17] and Bootstrapping Language-Image Pre-training (BLIP) [12] are two most significant vision-language processing models that have revolutionized this field. CLIP excels in learning robust image and text representations from large, uncurated datasets, enabling tasks like zero-shot image classification and image-text similarity search. BLIP takes this further by introducing a bootstrapping approach where captions are automatically generated and then filtered. In this work, we utilize BLIP to design our Image-grounded Text Embedding for the surgical VQA task. On the other hand, there is evidence that the squeeze & excitation (SE) [8] block, a form of lightweight gated attention mechanism, enhances the representation power of the network and boosts the prediction accuracy [19]. We have designed an excitation block (EB) using gated attention and integrated it into the classification head to amplify the significant features in the model learning.

2.2 Proposed Method: PitVQA

PitVQA Dataset: Our PitVQA dataset comprises 25 videos of endoscopic pituitary surgeries from the National Hospital of Neurology and Neurosurgery in London, United Kingdom, similar to the dataset used in the MICCAI PitVis challenge¹. All patients provided informed consent, and the study was registered with the local governance committee. The surgeries were recorded using a high-definition endoscope (Karl Storz Endoscopy) with a resolution of 720p and stored as MP4 files. All videos were annotated for the surgical phases, steps, instruments present and operation notes guided by a standardised annotation framework, which was derived from a preceding international consensus study on pituitary surgery workflow [16]. Annotation was performed collaboratively by 2 neurosurgical residents with operative pituitary experience and checked by an attending neurosurgeon. We extracted image frames from each video at 1 fps and removed any frames that were blurred or occluded. Ultimately, we obtained a total of 109,173 frames, with the videos of minimum and maximum length yielding 2,443 and 7,179 frames, respectively. We acquired frame-wise question-answer pairs for all the categories of the annotation. Overall, there are 884,242 question-answer pairs from 109,173 frames, which is around 8 pairs for each frame. There are 59 classes overall, including 4 phases, 15 steps, 18 instruments, 3 variations of instruments present in a frame, 5 positions of the instruments, and 14 operation notes in the annotation classes. The length of the questions ranges from a minimum of 7 words to a maximum of 12 words. A comparison of the unique classes between our PitVQA and a publicly available dataset of EndoVis18-VQA is presented in the Table 1. A sample frame and corresponding Q&A pairs are presented in Fig. 1(a). The class frequency distribution is illustrated in Fig. 1(b), where the lowest and highest classes are instruments and phases with 10.04% and 24.70%.

Table 1. The comparison between our PitVQA and a publicly available dataset of EndoVis18-VQA [20].

Dataset	EndoVis18-VQA	PitVQA
Average length (words)	5.8	10.3
Average #Questions	5.0	8.1
#Steps	0	15
#Phases	0	4
#Instruments/Objects	1	18
#Quantity	0	3
#Positions	4	5
#Operation notes	13	14

¹ <https://www.synapse.org/Synapse:syn51232283>.

PitVQA-Net: Our PitVQA-Net comprises an Image-grounded Text Embedding, a GPT2 Backbone, and an Excitation Block(EB) Classification Head as shown in the Fig. 2. Detailed descriptions of these modules can be found below:

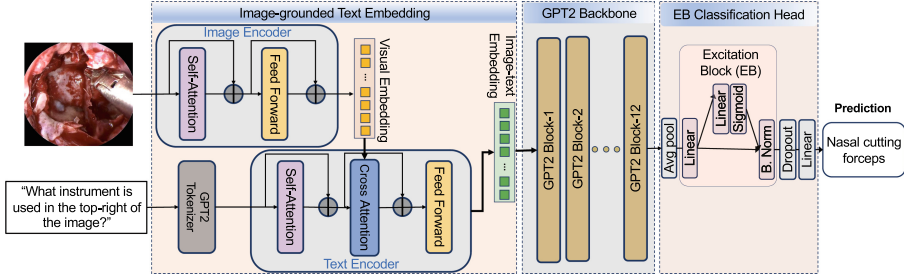


Fig. 2. PitVQA-Net: The network forms of Image-grounded Text Embedding, GPT2 Backbone and Classification Head. The image-grounded text embedding leverages joint embedding, cross-attention and contextual representation.

Image-Grounded Text Embedding: Image-grounded Text Embedding forms of image encoder and cross-attention text encoder. The image encoder is a vision transformer which analyzes visual content and produces a detailed representation of the image features. The cross-attention text encoder (e.g. Bert Model [7]) processes the text input with the interaction of the image features through cross-attention. During the process, the image features attend to relevant parts of the text sequence, and textual features attend to specific image regions. This helps the model “ground” the language in corresponding visual elements. Through multiple layers of cross-attention, the image-grounded text encoder produces updated representations where both the text and image features are deeply contextualized with respect to one another. The output of the Image-grounded Text Encoder is a set of text embeddings that have been enriched with visual context. In our PitVQA-Net, these embeddings then pass to the GPT2, followed by the EB classification head for the VQA prediction.

GPT2 Backbone: The GPT2 backbone consists of 12 GPT2 blocks, which are the decoder-only transformer blocks. It is specifically designed for language generation tasks. Each GPT2 block contains multi-head attention, followed by a layer normalization layer and a feed-forward neural network. Multi-head attention computes self-attention multiple times in parallel with different learned projections of the query (Q), key (K), and value (V) vectors, then concatenates and linearly projects the results to produce a final representation. The self-attention [22] mechanism is formulated as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. It effectively captures and processes different aspects of the input data, leveraging the self-attention mechanism to model the dependencies between tokens in the sequence. The stacking of multiple such blocks allows GPT2 to model complex

language patterns and generate coherent and contextually relevant text based on the input provided. In our PitVQA-Net, the GPT2 backbone receives the embedding from our Image-grounded Text Embedding module, then processes them and produces hidden state features. These features are then passed to the SE classification head to predict the answer class in our VQA task.

EB Classification Head: We design a lightweight Excitation Block (EB) layer with a linear layer, gated attention, and a batch normalization layer. The gated-attention forms of a sigmoid activation succeed a linear layer, which is capable of amplifying the significant features and suppressing the weak features by multiplying feature maps with gated weights obtained from the sigmoid function. We integrate EB into our classification head, including an average pooling, dropout, and a linear layer, as illustrated in Fig. 2. In our PitVQA-Net, the feature maps from GPT2 Backbone are passed through the EB classification head to transform the feature maps into logits. Then, we use a softmax function to obtain the final probability distribution.

3 Experiments and Results

3.1 Dataset

In addition to our PitVQA dataset (details in the Sect. 2.2), we also validate our model on a publicly available dataset of EndoVis18-VQA [20]. The dataset consists of 11,783 Q&A pairs derived from 2,007 surgical scenes from 14 video sequences of nephrectomy surgery procedures. The answers are in the form of single words with 18 distinct labels (1 kidney, 13 tool-tissue interactions, and 4 instrument locations). We split the training and validation set by following the original setup [20]. Thus, the training set contains 1560 frames and 9014 question-answer pairs, while the validation set consists of 447 frames and 2769 question-answer pairs. Both PitVQA and EndoVis18-VQA exhibit significant class imbalance, which limits the reliability of traditional accuracy metrics for robustness evaluation across the classes.

3.2 Implementation Details

The implementation and pre-trained weights of our backbone networks are adopted from the official repositories of the BLIP [12] and Huggingface GPT2². Our model is trained on cross-entropy loss and Adam optimizer with the learning rate of 1×10^{-5} . For the performance comparison, we selected closely related state-of-the-art (SOTA) surgical VQA models such as SurgicalGPT [20], VisualBert RM [21], and VisualBert [13] to retrain using official repositories. Additionally, we adopted the results of other recent baselines, including MFH [24], MFB [23], and Mutan [4], as reported in [20]. All experiments are conducted with the PyTorch framework on an NVIDIA RTX A6000 GPU.

² <https://huggingface.co/openai-community/gpt2>.

3.3 Evaluation Metrics

To demonstrate the model’s generalizability and robustness across datasets with imbalanced class frequency, balanced accuracy has proven to be an effective evaluation metric [15]. It is a prevalence-independent measure that computes the prediction accuracy by equally weighting the contribution of each class. Given the significant class imbalance in the datasets, we have employed metrics such as balanced accuracy (B. Acc), and Fscore, including Accuracy and Recall.

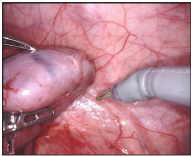
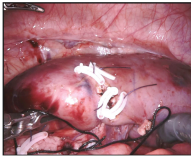
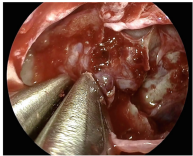
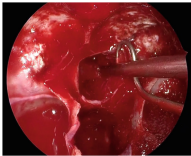
	EndoVis18-VQA		PitVQA	
Surgical scene				
Question	What is the state of monopolar curved scissors?	Where is grasper forceps located?	What is the surgical step shown in this frame?	What instrument is used in the top-right of the image?
Ground truth	Idle	left-top	sellotomy	ring curette
VisualBert	Cutting	left-bottom	sellotomy	suction
VisualBert RM	Idle	left-bottom	sellotomy	suction
SurgicalGPT	Idle	left-top	sellotomy	suction
PitVQA-Net	Idle	left-top	sellotomy	ring curette

Fig. 3. Qualitative visualization of our model prediction comparing with closely related works with datasets of our PitVQA and EndoVis18-VQA.

3.4 Results

Table 2 presents the results of the proposed method compared with other SOTA surgical VQA models. We emphasize on the metric of balanced accuracy to highlight the robustness of the model prediction across the classes considering highly imbalanced dataset. There are significant performance improvements in our method with a balanced accuracy of 58.82% and 45.06% for the datasets of PitVQA and EndoVis18-VQA, respectively. The performance improvements are around 8% on PitVQA and 9% on EndoVis18-VQA over the closely related work of SurgicalGPT. Similar trends are observed in other metrics of Fscore and recall.

The qualitative performance comparison for both PitVQA and EndoVis18-VQA is illustrated in Fig. 3. It appears that most of the models are capable of accurately recognizing surgical steps, whereas the identification of instruments with localization reasoning mostly fails. However, PitVQA-Net demonstrates robust prediction across various types of question answering in both datasets.

Table 2. The performance comparison of our method with other SOTA surgical VQA models in our PitVQA dataset. The balanced accuracy [15] is denoted as B. Acc.

MODELS	EndoVis18-VQA				PitVQA			
	FScore	B. Acc	Acc	Recall	FScore	B. Acc	Acc	Recall
Mutan [4]	0.4565	—	0.6303	0.4969	—	—	—	—
MFB [23]	0.3622	—	0.5238	0.4205	—	—	—	—
MFH [24]	0.4224	—	0.5876	0.4835	—	—	—	—
VisualBert [13]	0.3745	0.3474	0.6143	0.4282	0.4286	0.4358	0.6338	0.4549
VisualBert RM [21]	0.3583	0.3422	0.6190	0.4079	0.4281	0.3892	0.6318	0.4103
SurgicalGPT [20]	0.4649	0.3543	0.6811	0.4649	0.5261	0.5090	0.7232	0.5397
PitVQA-Net (Ours)	0.6165	0.4506	0.6851	0.4849	0.5952	0.5882	0.7601	0.5917

3.5 Ablation Study

To assess the effectiveness of our proposed method, we conducted an ablation study comparing different vision-language embeddings, including CLIP [17], as well as other large language model variants such as BioGPT [14] and BERT [7], as detailed in Table 3. Additionally, we explored the benefits of using LLM pre-trained weights by conducting an experiment with PitVQA-Net without leveraging any pre-trained weights. We also investigate the effect of EB in our PitVQA-Net. Our experiments show that our EB block significantly enhances performance on the EndoVis18-VQA dataset. Overall, the results demonstrate the effectiveness of each innovation within our proposed method.

Table 3. Ablation study with other vision-language embedding of CLIP [17] and LLMs like BioGPT [14], and BERTZ [7]. We also observe the performance of the pitVQA-Net without pretrained weights (w/o pretrain.), and without excitation block (w/o EB).

MODELS	EndoVis18-VQA			PitVQA		
	B. Acc	Recall	FScore	B. Acc	Recall	FScore
CLIP-GPT	0.4342	0.5503	0.4513	0.4138	0.4405	0.4176
CLIP-BioGPT	0.4145	0.4876	0.4728	0.4072	0.4295	0.4168
BLIP-BERT	0.3752	0.4525	0.5317	0.5574	0.5957	0.5663
PitVQA-Net (w/o pretrain.)	0.3650	0.4265	0.5345	0.5419	0.5710	0.5647
PitVQA-Net (w/o EB)	0.3978	0.4577	0.5523	0.5800	0.5982	0.5939
PitVQA-Net	0.4506	0.4849	0.6165	0.5882	0.5917	0.5952

4 Discussion and Conclusion

In this study, we introduced PitVQA, a targeted dataset for VQA in the domain of endonasal pituitary adenoma surgery and PitVQA-Net, an innovative adapta-

tion of the GPT2 that incorporates a novel image-grounded text embedding and gated-attention excitation block. Our experimental results demonstrate a clear performance advantage for PitVQA-Net, achieving superior results on both the PitVQA dataset and the publicly available EndoVis18-VQA dataset when compared to existing surgical VQA models. The ablation study further reinforces the significance of our proposed method, highlighting the effectiveness of our image-grounded text embedding, excitation block, selection of LLM, and importance of pretrained weights. This work paves the way for the development of intuitive and collaborative surgical AI assistants. By enabling accurate and contextually aware responses to complex surgical questions, PitVQA-Net demonstrates significant promise for enhancing intra-operative decision-making and ultimately improving patient outcomes. Future directions for this research include expanding the PitVQA dataset to generate sentences by covering a wider range of surgical scenarios, exploring more advanced vision-language fusion techniques, and investigating the potential for real-time deployment of such systems within the operating room.

Acknowledgments. This work was supported in whole, or in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [20-3145/Z/16/Z] and the Engineering and Physical Sciences Research Council (EPSRC) [EP/W00805X/1, EP/Y01958X/1]; Horizon 2020 FET Open [863146]; the UCLH/UCL NIHR Biomedical Research Centre; the Department of Science, Innovation and Technology (DSIT); and the Royal Academy of Engineering Chair in Emerging Technologies Scheme. AD is supported by EPSRC [EP/S021612/1]. DZK is supported by an NIHR Academic Clinical Fellowship. With thanks to Digital Surgery Ltd, a Medtronic company, for access to Touch SurgeryTM Enterprise for both video recording and storage.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Bai, L., Islam, M., Ren, H.: 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. Springer (2023)
3. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery (2023)
4. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2612–2620 (2017)
5. Das, A., Khan, D.Z., Williams, S.C., Hanrahan, J.G., Borg, A., Dorward, N.L., Bano, S., Marcus, H.J., Stoyanov, D.: A multi-task network for anatomy identification in endoscopic pituitary surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 472–482. Springer (2023)

6. Decker, H., Trang, K., Ramirez, J., Colley, A., Pierce, L., Coleman, M., Bongiovanni, T., Melton, G.B., Wick, E.: Large language model- based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Network Open* **6**(10), e2336997–e2336997 (2023)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)* (2018)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018)
9. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6700–6709 (2019)
10. Khan, D.Z., Hanrahan, J.G., Baldeweg, S.E., Dorward, N.L., Stoyanov, D., Marcus, H.J.: Current and future advances in surgical therapy for pituitary adenoma. *Endocrine Reviews* (2023)
11. Lawson McLean, A.: Artificial intelligence in surgical documentation: A critical review of the role of large language models. *Annals of Biomedical Engineering* pp. 1–2 (2023)
12. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International Conference on Machine Learning*. pp. 12888–12900. PMLR (2022)
13. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. *arXiv preprint [arXiv:1908.03557](https://arxiv.org/abs/1908.03557)* (2019)
14. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* **23**(6), bbac409 (2022)
15. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. *Nature methods* pp. 1–18 (2024)
16. Marcus, H.J., Khan, D.Z., Borg, A., Buchfelder, M., Cetas, J.S., Collins, J.W., Dorward, N.L., Fleseriu, M., Gurnell, M., Javadpour, M., et al.: Pituitary society expert delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. *Pituitary* **24**(6), 839–853 (2021)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019)
19. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*. pp. 421–429. Springer (2018)
20. Seenivasan, L., Islam, M., Kannan, G., Ren, H.: Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery. *arXiv preprint [arXiv:2304.09974](https://arxiv.org/abs/2304.09974)* (2023)

21. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–43. Springer (2022)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
23. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 1821–1830 (2017)
24. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* **29**(12), 5947–5959 (2018)
25. Yuan, K., Kattel, M., Lavanchy, J.L., Navab, N., Srivastav, V., Padoy, N.: Advancing surgical vqa with scene graph knowledge (2024)