SurgicalGPT: End-to-End Language-Vision GPT for Visual Question Answering in Surgery

Lalithkumar Seenivasan[1,*][0000−0002−0103−1234], Mobarakol Islam[2,*][0000−0002−7162−2822], Gokul Kannan[3][0000−0002−2284−703X], and Hongliang Ren[1,4,5,**][0000−0002−6488−1551]

[1] Dept. of Biomedical Engineering, National University of Singapore, Singapore.
[2] WEISS, University College London, United Kingdom.
[3] Dept. of Production Engineering, National Institute of Technology, Tiruchirappalli, India.
[4] Dept. of Electronic Engineering, Chinese University of Hong Kong.
[5] Shun Hing Institute of Advanced Engineering, Chinese University of Hong Kong.

Abstract. Advances in GPT-based large language models (LLMs) are revolutionizing natural language processing, exponentially increasing its use across various domains. Incorporating uni-directional attention, these autoregressive LLMs can generate long and coherent paragraphs. However, for visual question answering (VQA) tasks that require both vision and language processing, models with bi-directional attention or models employing fusion techniques are often employed to capture the context of multiple modalities all at once. As GPT does not natively process vision tokens, to exploit the advancements in GPT models for VQA in robotic surgery, we design an end-to-end trainable Language-Vision GPT (LV-GPT) model that expands the GPT2 model to include vision input (image). The proposed LV-GPT incorporates a feature extractor (vision tokenizer) and vision token embedding (token type and pose). Given the limitations of unidirectional attention in GPT models and their ability to generate coherent long paragraphs, we carefully sequence the word tokens before vision tokens, mimicking the human thought process of understanding the question to infer an answer from an image. Quantitatively, we prove that the LV-GPT model outperforms other state-of-the-art VQA models on two publically available surgical-VQA datasets (based on endoscopic vision challenge robotic scene segmentation 2018 and CholecTriplet2021) and on our newly annotated dataset (based on the holistic surgical scene dataset). We further annotate all three datasets to include question-type annotations to allow sub-type analysis. Furthermore, we extensively study and present the effects of token sequencing, token type and pose embedding for vision tokens in the LV-GPT model.

1 Introduction

The recent evolution of large language models (LLMs) is revolutionizing natural language processing and their use across various sectors (e.g., academia, health-

---

* Lalithkumar Seenivasan and Mobarakol Islam are co-first authors.
** Corresponding author.

care, business, and IT) and daily applications are being widely explored. In medical diagnosis, recent works [23] have also proposed employing the LLM models to generate condensed reports, interactive explanations, and recommendations based on input text descriptions (predicted disease and report). While the current single-modality (language) LLMs can robustly understand the questions, they still require prior text descriptions to generate responses and are unable to directly infer responses based on the medical image. Although language-only models can greatly benefit the medical domain in language processing, there is a need for robust multi-modality models to process both medical vision and language. In the surgical domain, in addition to the scarcity of surgical experts, their daily schedules are often overloaded with clinical and academic work, making it difficult for them to dedicate time to answer inquiries from students and patients on surgical procedures [3]. Although various computer-assisted solutions [1,10,11,16,17] have been proposed and recorded surgical videos have been made available for students to sharpen their skills and learn from observation, they still heavily rely on surgical experts to answer their surgery-specific questions. In such cases, a robust and reliable surgical visual question answering (VQA) model that can respond to questions by inferring from context-enriched surgical scenes could greatly assist medical students, and significantly reduce the medical expert's workload [19].

In the medical domain, MedfuseNet [19], an attention-based model, was proposed for VQA in medical diagnosis. Utilizing the advancements in the transformer models, VisualBert RM [18], a modified version of the VisualBert [12] model was also proposed for VQA in robotic surgery. Compared to most VQA models that require a region proposal network to propose vision patches, the VisualBert RM [18] performed VQA based on features extracted from the whole image, eliminating the need for a region proposal network. However, they were extracted using a non-trainable fixed feature extractor. While VisualBert [12] models and LLMs are transformer models, there are fundamentally different. VisualBert [12] transformers are bidirectional encoder models and are often employed for multi-modality tasks. In contrast, ChatGPT [6] (GPT3.5) and BARD (LaMDA [20]) are language-only uni-directional transformer decoder models employed for language generation. As they are proving to be robust in language generation, exploiting them to process the questions and enabling them to process vision could greatly improve performance in VQA tasks.

In this work, we develop an end-to-end trainable SurgicalGPT model by exploiting a pre-trained LLM and employing a learnable feature extractor to generate vision tokens. In addition to word tokens, vision tokens (embedded with token type and pose embedding) are introduced into the GPT model, resulting in a Language-Vision GPT (LV-GPT) model. Furthermore, we carefully sequence the word and vision tokens to leverage the GPT model's robust language processing ability to process the question and better infer an answer based on the vision tokens. Through extensive experiments, we show that the SurgicalGPT(LV-GPT) outperforms other state-of-the-art (SOTA) models by $\sim 3 - 5$ % on publically

---

[6] chat.openai.com

available EndoVis18-VQA [18] and Cholec80-VQA surgical-VQA [18] datasets. Additionally, we introduce a novel PSI-AVA-VQA dataset by adding VQA annotations to the publically available holistic surgical scene dataset(PSI-AVA) and observe similar performance improvement. Furthermore, we study and present the effects of token sequencing, where model performance improved by $\sim 2-4$ % when word tokens are sequenced earlier. Finally, we also study the effects of token type and pose embedding for vision tokens in the LV-GPT model.

2 Proposed Method

2.1 Preliminaries

GPT2 [6], a predecessor to GPT3.5 (ChatGPT), is a transformer decoder model that performs next-word prediction. Auto-regressive in nature, its self-attention blocks attend to earlier word tokens to predict the next word token iteratively, allowing the model to generate complex paragraphs [15]. Although robust in language generation, due to its unidirectional attention [13], in a given iteration, the generated token knows all earlier tokens but does not know any subsequent token (Fig. 1 (a)), restricting the model's ability to capture the entire context between all tokens. VisualBert [12], fundamentally different from GPT models, is a non-auto-regressive transformer encoder model. Its bidirectional self-attention blocks attend in both directions (earlier and subsequent tokens) [13], allowing the model to capture the entire context all at once (Fig. 1 (b)). Due to this, bi-directional attention models are often preferred for multi-modality tasks.

Vision-Language Processing: Employed mostly for language-only tasks, GPT models do not natively process vision tokens [8]. While it supports robust word embedding, it lacks vision tokenizer and vision embedding layers. This limits exploiting its language processing ability for multi-modality tasks. Alternate to GPT, as the VisualBert model is often preferred for multi-modality tasks, it encompasses dedicated embedding layers for both vision and word tokens.

2.2 LV-GPT: Language-Vision GPT

Overall Network: We design an end-to-end trainable multi-modality (language and vision) LV-GPT model (Fig. 2) for surgical VQA. We integrate a vision tokenizer (feature extractor) module and vision embedding with the GPT model to exploit its language processing ability in performing VQA tasks.

Fig. 1. Uni-directional attention in GPT language model vs bi-direction attention in VisualBert multi-modality model.

Fig. 2. End-to-End LV-GPT for Surgical VQA: The input question and surgical scene are tokenized, embedded, and sequenced to predict the answer.

Language-Vision Processing: The questions are tokenized using the inherent GPT2 tokenizer. The word tokens are further embedded based on token-id, token type (0) and token position by the inherent GPT2 word embedding layers. To tokenize the input surgical scene (image) into vision tokens, the LV-GPT includes a vision tokenizer (feature extractor): ResNet18 (RN18) [9] / Swin [14] / ViT [7]. Given an image, the tokenizer outputs vision tokens, each holding visual features from an image patch. Additionally, the vision tokens are further embedded based on token type (1) and token position (pos = 0) embeddings. The final embedded word and vision tokens (w_e and v_e) can be formulated as:

$$w_e = T_{t=0}(w_x) + P_{pos}(w_x) + w_x; \quad pos = 0, 1, 2, 3, ..., n.$$

$$v_e = T_{t=1}(v_x) + P_{pos=0}(v_x) + v_x; \quad v_x = \begin{cases} v_t, & dim(v_t^i) = dim(w_x^i) \\ f(v_t), & else \end{cases} \quad (1)$$

where, $T_t()$ is type embedding, $P_{pos}()$ is pose embedding, w_x and v_x are initial word and vision embedding, and v_t are vision tokens. Initial word embeds (w_x) are obtained using word embedding based on word token id. Depending on the size (dim) of each vision token, they undergo additional linear layer embedding ($f()$) to match the size of the word token.

Token Sequencing: LLMs are observed to process long sentences robustly and hold long-term sentence knowledge while generating coherent paragraphs/reports. Considering GPT's superiority in sequentially processing large sentences and its uni-directional attention, the word tokens are sequenced before the vision tokens. This is also aimed at mimicking human behaviour, where the model understands the question before attending to the image to infer an answer.

Classification: Finally, the propagated multi-modality features are then passed through a series of linear layers for answer classification.

3   Experiment

3.1   Dataset

EndoVis18-VQA: We employ publically available EndoVis18-VQA [18] dataset to benchmark the model performance. We use the classification subset that

includes classification-based question-and-answer (Q&A) pairs for 14 robotic nephrectomy procedure video sequences of the MICCAI Endoscopic Vision Challenge 2018 [2] dataset. The Q&A pairs are based on the tissue, actions, and locations of 8 surgical tools. The dataset includes 11783 Q&A pairs based on 2007 surgical scenes. The answers consist of 18 classes (1 kidney, 13 tool-tissue interactions, and 4 tool locations). Additionally, we further annotated the validation set (video sequences 1, 5, and 16) on question types to assist in additional analysis. We followed the EndoVis18-VQA [18] dataset's original train/test split.
Cholec80-VQA: The classification subset of the Cholec80-VQA [18] is also employed for model evaluation. It contains Q&A pairs for 40 video sequences of the Cholec80 dataset [21]. The subset consists of 43182 Q&A pairs on the surgical phase and instrument presence for 21591 frames. The answers include 13 classes (2 instrument states, 4 on tool count, and 7 on surgical phase). We additionally annotated the validation set (video sequences: 5, 11, 12, 17, 19, 26, 27 and 31) on the Q&A pairs types for further model analysis. The VQA [18] dataset's original train/test split is followed in this work.
PSI-AVA-VQA: We introduce a novel PSI-AVA-VQA dataset that consists of Q&A pairs for key surgical frames of 8 cases of the holistic surgical scene dataset (PSI-AVA dataset) [22]. The questions and answers are generated in sentence form and single-word (class) response form, respectively. They are generated based on the surgical phase, step, and location annotation provided in the PSI-AVA dataset [22]. The PSI-AVA-VQA consists of 10291 Q&A pairs and with 35 answer classes (4 locations, 11 surgical phases, and 21^1 surgical steps). The Q&A pairs are further annotated into 3 types (location, phase, and step). The fold-1 train/test split of parent PSI-AVA [22] dataset is followed in this work.

3.2   Implementation Details

All variants of our models [2] are trained based on cross-entropy loss and optimized using the Adam optimizer. The models were trained for 80 epoch, with a batch size of 64, except for LV-GPT (ViT) (batch size = 32 due to GPU limitation). learning rates lr = $1x10^{-5}$, $1x10^{-5}$ and $5x10^{-6}$ are used for EndoVis18-VQA, PSI-AVA-VQA and Cholec80-VQA dataset, respectively. The SOTA VisualBert [12] and VisualBert RM [18] models were implemented using their official code repositories. The Block [5], MUTAN [4], MFB [24] and MFH [25] were implemented using the official codes of Block [5].

4   Results

All our proposed LV-GPT model variants are quantitatively benchmarked (Table 1) against other attention-based/bi-directional encoder-based SOTA models on EndoVis18-VQA, Cholec80-VQA and PSI-AVA-VQA datasets based on the

---

[1] One class shares a common name with a surgical phase class.
[2] Code available: github.com/lalithjets/SurgicalGPT

Table 1. Quantitaive comparison of our LV-GPT (Swin), LV-GPT (RN18), and (LV-GPT (ViT) against state-of-the-art models.

MODELS	EndoVis18-VQA [18]			Cholec80-VQA [18]			PSI-AVA-VQA		
	Acc	Recall	FScore	Acc	Recall	FScore	Acc	Recall	FScore
VisualBert [12]	0.6143	0.4282	0.3745	0.9007	0.6294	0.6300	0.5853	0.3307	0.3161
VisualBert RM [18]	0.6190	0.4079	0.3583	0.9001	0.6573	0.6585	0.6016	0.3242	0.3165
Block [5]	0.6088	0.4884	0.4470	0.8948	0.6600	0.6413	0.5990	**0.5136**	**0.4933**
Mutan [4]	0.6303	**0.4969**	0.4565	0.8699	0.6332	0.6106	0.4971	0.3912	0.3322
MFB [24]	0.5238	0.4205	0.3622	0.8410	0.5303	0.4588	0.5712	0.4379	0.4066
MFH [25]	0.5876	0.4835	0.4224	0.8751	0.5903	0.5567	0.4777	0.2995	0.2213
LV-GPT (Swin)	0.6613	0.4460	0.4537	**0.9429**	**0.7339**	**0.7439**	0.6033	0.4137	0.3767
LV-GPT (RN18)	**0.6811**	0.4649	**0.4649**	0.8746	0.5747	0.5794	0.5933	0.3183	0.3168
LV-GPT (ViT)	0.6659	0.4920	0.4336	0.9232	0.6833	0.6963	**0.6549**	0.4132	0.3971

Fig. 3. Qualitative analysis: Comparison of answers predicted by VisualBERT [12], VisualBert RM [18], Block [5], and our LV-GPT (Swin) models against the ground truth based on input surgical scene and question.

accuracy (Acc), recall, and Fscore. In most cases, all our variants, LV-GPT (Swin), LV-GPT (RN18) and LV-GPT (ViT), are observed to significantly outperform SOTA models on all three datasets in terms of Acc. Specifically, the LV-GPT (Swin) variant (balanced performance across all datasets) is observed to outperform all SOTA models on all datasets and significantly improve the performance ($\sim 3 - 5\%$ improvement) on EndoVis18-VQA and Cholec80-VQA dataset. Additionally, it should be noted our model variants can be trained end-to-end, whereas, most of the SOTA models requires a region proposal network to process input image into vision tokens. Fig. 3 shows the qualitative performance of LV-GPT (Swin) against SOTA models on three datasets. A Comparison of our LV-GPT model performance on the EndoVis18-VQA dataset with default test queries vs rephrased test queries is presented in supplementary materials that highlight the model's robustness in language reasoning.

Early Vision vs Early Word: The performance of LV-GPT based on word and vision token sequencing (Table 2) is also studied. While all three variants of the LV-GPT models processing vision tokens earlier are observed to perform on par with SOTA models reported in Table 1, in most cases, their performances

Table 2. Comparison of LV-GPT model performance when vision tokens are sequenced earlier vs when word tokens are sequenced earlier.

Token sequencing	Model	EndoVis18-VQA			PSI-AVA-VQA		
		Acc	Recall	FScore	Acc	Recall	FScore
Early vision	LV-GPT (RN18)	0.6338	0.3600	0.3510	0.5542	0.2879	0.2886
	LV-GPT (Swin)	0.6208	0.4059	0.3441	**0.6068**	**0.4195**	**0.3813**
	LV-GPT (ViT)	0.6493	0.4362	0.3701	0.6023	0.2802	0.2628
Early word	LV-GPT (RN18)	**0.6811**	**0.4649**	**0.4649**	0.5933	0.3183	0.3168
	LV-GPT (Swin)	**0.6613**	**0.4460**	**0.4537**	0.6033	0.4137	0.3767
	LV-GPT (ViT)	**0.6659**	**0.4920**	**0.4336**	**0.6549**	**0.4132**	**0.3971**

on both datasets further improved by $\sim 2-4\%$ when word tokens are processed earlier. This improvement could be attributed to LLM's ability to hold sentence (question) context before processing the vision tokens to infer an answer. This behaviour, in our view, mimics the human thought process, where we first understand the question before searching for an answer from an image.

Pose Embedding for Vision tokens: The influence of positional embedding of the vision tokens (representing a patch region) in all the LV-GPT variants is studied by either embedded with position information (pos = 1, 2, 3, .., n.) or zero-position (pos = 0). Table 3 shows the difference in the performance of the best-performing LV-GPT variant in each dataset, with its vision tokens embedded with actual-position or zero-position. While we expected the positional embedding to improve the performance (dataset Q&A pairs related to

Table 3. Comparison of model performances on EndoVis18-VQA, Cholec80-VQA and PSI-AVA-VQA datasets when vision tokens are embedded with zero-positional embedding vs actual pose embedding.

Dataset	Model	Zero Pose Embedding			Actual Pose Embedding		
	Best LV-GPT	Acc	Recall	FScore	Acc	Recall	FScore
EndoVis18-VQA	LV-GPT (RN18)	**0.6811**	0.4649	0.4649	**0.6811**	**0.4720**	**0.4681**
Cholec80-VQA	LV-GPT (Swin)	**0.9429**	**0.7339**	**0.7439**	0.9414	0.7251	0.7360
PSI-AVA-VQA	LV-GPT (ViT)	**0.6549**	**0.4132**	**0.3971**	0.5905	0.3742	0.3463

Fig. 4. Sub-type performance analysis of LV-GPT model variants on Cholec80-VQA and PSI-AVA-VQA embedded with zero-pose embedding vs actual-pose embedding.

Table 4. Ablation study on vision token (VT) embedding.

VB-VE	C-VE	VT-TY + VT-PE	VT-TY + VT-ZPE	LV-GPT (RN18)			LV-GPT (ViT)		
				Acc	Recall	FScore	Acc	Recall	FScore
✓				0.6287	0.4061	0.4063	0.6147	0.4199	0.3679
	✓			0.6728	0.4366	0.4455	0.6504	0.4792	0.4323
	✓	✓		**0.6811**	**0.4720**	**0.4681**	0.6259	0.4306	0.3805
	✓		✓	**0.6811**	0.4649	0.4649	**0.6659**	**0.4920**	**0.4336**

tool location), from the results, we observe that embedding vision tokens with zero-position embedding results in better performance. In-depth analysis shows that our CNN-based LV-GPT (RN18) model improved with positional embedding (Table 3 and Table 4). In the transformer-based LV-GPT (Swin) / LV-GPT (ViT) models, positional embedding is already incorporated at the vision tokenizer (VIT/Swin) layer, and adding positional embedding at the GPT level results in double Position embedding. Thus, "zero-position" can be interpreted as "LV-GPT only requires one layer of positional embedding". A sub-type analysis (Fig. 4) is also performed on the model performance to analyze the effect of positional embedding of the vision tokens. The model in which the vision tokens were embedded with zero-position (at the GPT level), performed marginally better/similar on all sub-types in the Cholec80-VQA dataset. However, its performance improvement was significant in the PSI-AVA-VQA dataset sub-types, including the 'tool location' sub-types that contain questions on tool location.

Ablation Study on Vision Token Embedding: An ablation study on the vision token embedding in the LV-GPT model on the EndoVis18-VQA dataset is also shown in Table 4. VB-VE refers to vision token embedding using VisualBert vision embedding. The C-VE refers to custom embedding, where, in LV-GPT (RN18), the vision token undergoes additional linear layer embedding to match the word-token dimension, and in other variants, vision tokens from the Swin/VIT are directly used. The subsequent VT-TY + VT-PE and VT-TY + VT-ZPE refers to the additional vision token type (TY) and actual-position (PE)/zero-position (ZPE) embedding. We observe that employing C-VE with VT-TY + VT-ZPE results in better performance.

5   Conclusion

We design an end-to-end trainable SurgicalGPT, a multi-modality Language-Vision GPT model, for VQA tasks in robotic surgery. In addition to GPT's inherent word embeddings, it incorporates a vision tokenizer (trainable feature extractor) and vision token embedding (type and pose) to perform multi-modality tasks. Furthermore, by carefully sequencing the word tokens earlier to vision tokens, we exploit GPT's robust language processing ability, allowing the LV-GPT to significantly perform better VQA. Through extensive quantitative analysis,

we show that the LV-GPT outperforms other SOTA models on three surgical-VQA datasets and sequencing word tokens early to vision tokens significantly improves the model performance. Furthermore, we introduce a novel surgical-VQA dataset by adding VQA annotations to the publically available holistic surgical scene dataset. While multi-modality models that process vision and language are often referred to as "vision-language" models, we specifically name our model "language-vision GPT" to highlight the importance of the token sequencing order in GPT models. Integrating vision tokens into GPT also opens up future possibilities of generating reports directly from medical images/videos.

6   Acknowledgement

References

1. Adams, L., Krybus, W., Meyer-Ebrecht, D., Rueger, R., Gilsbach, J.M., Moesges, R., Schloendorff, G.: Computer-assisted surgery. IEEE Computer graphics and applications **10**(3), 43–51 (1990)
2. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
3. Bates, D.W., Gawande, A.A.: Error in medicine: what have we learned? (2000)
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. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: Block Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8102–8109 (2019)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.C.: From images to textual prompts: Zero-shot vqa with frozen large language models. arXiv preprint arXiv:2212.10846 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hong, M., Rozenblit, J.W., Hamilton, A.J.: Simulation-based surgical training systems in laparoscopic surgery: a current review. Virtual Reality **25**, 491–510 (2021)

11. Kneebone, R.: Simulation in surgical training: educational issues and practical implications. Medical education **37**(3), 267–277 (2003)
12. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
13. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. arXiv preprint arXiv:2103.10385 (2021)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., Gao, J.: Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. arXiv preprint arXiv:2005.05298 **3** (2020)
16. Rogers, D.A., Yeh, K.A., Howdieshell, T.R.: Computer-assisted learning versus a lecture and feedback seminar for teaching a basic surgical technical skill. The American Journal of Surgery **175**(6), 508–510 (1998)
17. Sarker, S., Patel, B.: Simulation and surgical training. International journal of clinical practice **61**(12), 2120–2125 (2007)
18. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII. pp. 33–43. Springer (2022)
19. Sharma, D., Purushotham, S., Reddy, C.K.: Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports **11**(1), 1–18 (2021)
20. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
21. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
22. Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., Santander, J., Caicedo, J., Fernández, N., Arbeláez, P.: Towards holistic surgical scene understanding. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII. pp. 442–452. Springer (2022)
23. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)
24. 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)
25. 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)