2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Surgical-VQLA: Transformer with Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery

Long Bai[1†], Mobarakol Islam[2†], Lalithkumar Seenivasan[3] and Hongliang Ren[1,3,4*],
Senior Member, IEEE

***Abstract*— Despite the availability of computer-aided simulators and recorded videos of surgical procedures, junior residents still heavily rely on experts to answer their queries. However, expert surgeons are often overloaded with clinical and academic workloads and limit their time in answering. For this purpose, we develop a surgical question-answering system to facilitate robot-assisted surgical scene and activity understanding from recorded videos. Most of the existing visual question answering (VQA) methods require an object detector and regions based feature extractor to extract visual features and fuse them with the embedded text of the question for answer generation. However, (i) surgical object detection model is scarce due to smaller datasets and lack of bounding box annotation; (ii) current fusion strategy of heterogeneous modalities like text and image is naive; (iii) the localized answering is missing, which is crucial in complex surgical scenarios. In this paper, we propose Visual Question Localized-Answering in Robotic Surgery (Surgical-VQLA) to localize the specific surgical area during the answer prediction. To deal with the fusion of the heterogeneous modalities, we design gated vision-language embedding (GVLE) to build input patches for the Language Vision Transformer (LViT) to predict the answer. To get localization, we add the detection head in parallel with the prediction head of the LViT. We also integrate generalized intersection over union (GIoU) loss to boost localization performance by preserving the accuracy of the question-answering model. We annotate two datasets of VQLA by utilizing publicly available surgical videos from EndoVis-17 and 18 of the MICCAI challenges. Our validation results suggest that Surgical-VQLA can better understand the surgical scene and localized the specific area related to the question-answering. GVLE presents an efficient language-vision embedding technique by showing superior performance over the existing benchmarks.**

†L. Bai and M. Islam are co-first authors.

*The work was supported by Hong Kong Research Grants Council (RGC) Collaborative Research Fund (CRF C4026-21GF and CRF C4063-18G), and General Research Fund (GRF #14211420 and GRF #14216022); Shun Hing Institute of Advanced Engineering (BME-p1-21/8115064) at the CUHK; and Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) Grant 202108233000303 awarded to Dr. H. Ren. M. Islam was funded by EPSRC grant [EP/W00805X/1]. We thank the CUHK Vice-Chancellor's Ph.D. Scholarship Scheme for conference travel support. (Corresponding author: Hongliang Ren)

[1] L. Bai and H. Ren are with the Dept. of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, China; (E-mail: b.long@ieee.org)

[2] M. Islam is with the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, UK. (E-mail: mobarakol.islam@ucl.ac.uk)

[3] L. Seenivasan and H. Ren are with Dept. of Biomedical Engineering, National University of Singapore, Singapore. (E-mail: lalithkumar_s@u.nus.edu)

[4] H. Ren is also with Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong 999077, China. (E-mail: hlren@ieee.org)

I. INTRODUCTION

In the absence of domain experts to answer pressing questions, the answer to "why?" could often be inferred by finding answers to "what?" and "where?". In an ideal situation, given the critical nature of the medical domain, every question on surgery and surgical procedures must be answered by expert surgeons. However, often overloaded with academic and clinical work, expert surgeons find it difficult to make time to clarify these questions [1], [2]. To address this to an extent, recorded surgical videos are shared with the student for them to learn by observation. To improve the student's learning experience, augmented/virtual reality-based training systems [3], automated eye tracking models [4] and automated surgical skill evaluation models [5] have also been introduced. However, these solutions still do not answer any particular questions a student might have. Their effectiveness in teaching a student relies heavily on the ability to infer from video observation and practice. Recently, MedFuseNet [1] was proposed that performs medical visual question answering (VQA) and unfolded the possibility of developing a reliable VQA model that could supplement medical experts in answering questions from patients and students.

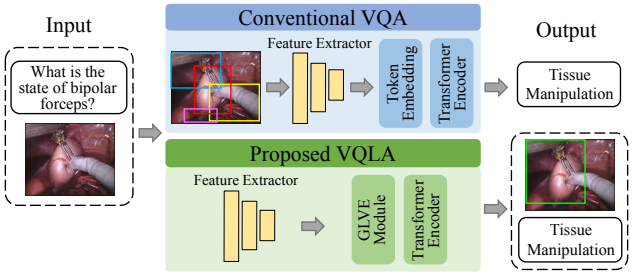

Fig. 1. An overview of our proposed VQLA pipeline, against the conventional VQA tasks. Object proposals are not required in our method, and bounding box prediction can be output together with the classification results.

Very recently, Surgical-VQA [2] has also been introduced that answers questionnaires on surgical tools, tool-tissue interactions and surgical phase based on the visual input. These two works have effectively unfolded the possibility of answering the "what?" of the questions. However, they still fail to address the "why?". For instance, while these models could potentially answer if a patient has COVID-19 based on the X-ray scan or answer the name of tissue of interest in a surgical procedure based on the input surgical scene, it is difficult to infer the answer for "why?" from those

answers. Although Surgical-VQA [2] offers the possibility to answer the "why?" using a sentence-based open-ended VQA model, the inherent lack of annotated dataset in the medical domain still makes it difficult and time-consuming to develop a robust open-ended Surgical-VQA model.

To outmaneuver the need for a massive annotated dataset and make it easier to infer the answer for "why?", we propose to answer the "what?" and the "where" using a Visual Question Localized-Answering (VQLA) model in the surgical domain. In addition to answering the questions, the VQLA model also highlights the specific areas in the image related to the question and answer. This allows a better understanding of complex medical diagnoses and surgical scenes. For instance, by answering the question "what is the tissue of interest?" in a surgical procedure and indicating its location in the surgical scene, the student could then easily compare it with the surrounding tissues or counterfactual surgical procedures (surgical scenes where the tissue of interest is different) and relate the tissue (even if partially occluded) to pre-operative scan for better inference on "why?". Localized-answer could also provide an additional advantage to students in inferring the reliability of predicted answers. For instance, if the localization is far-off from the surgical action or region of interest, it could mean that the predicted answer is less reliable. Fig. 1 presents the overall pipeline of our proposed VQLA.

Driven by readily available enormous datasets, tremendous progress has been made in developing VQA models [6]–[8] in the computer vision domain. Constructed using long-short term memory modules [9], [10] or attention modules [1], [11], most of these models rely heavily on object detection models, are time and resource expensive, are not end-to-end and perform a naive fusion of heterogenous features (visual and text features). Firstly, most of these models warrant employing object detection models to detect key objects in an image from which visual features are extracted. Thus, in addition to the question and answer annotations, bounding box annotations are needed to initially train the object detection model. The performance of the VQA model also relies heavily on the object detection model and a small detection error could exponentially influence the VQA model learning. Furthermore, extracting visual features only from the detected object regions and ignoring the key background features could limit the model's global scene understanding ability [12] that is crucial for VQA. Secondly, as these models are trained on outputs from pre-trained object detection and feature extraction models, they are not fully end-to-end, and warrant multiple stages of training, making the overall solution sub-optimal. Thirdly, as these models are often made of multiple sub-networks (object detection, feature extraction and VQA), they are resource and time heavy and limits usage in real-time application. Finally, these VQA models combine the heterogenous visual and text features using naive concatenation, addition, summation, averaging or attention techniques. While these naive techniques might perform effective feature fusion for homogenous features, their performance on heterogenous features is sub-optimal as each feature hold different significance. To this extent, attentional feature fusion (AFF) and iterative attentional feature fusion (iAFF) [13] have been recently proposed. To address the inherent limitations of using an object detection model and to perform effective heterogeneous feature fusion, we propose a detection-free Surgical VQLA model that can to trained in an end-to-end manner for localized answering based on input visual and question features. Furthermore, we propose a novel gated vision-language embedding for effective heterogeneous feature fusion and employ Generalized Intersection over Union (GIoU) [14] loss for robust localized-answering. Our key contributions and findings are:

- We design and propose a Surgical Visual Question Localized-Answering (Surgical-VQLA[1]) model that can predict localized-answer based on a given input question and surgical scene.
- Propose a detection-free GVLE-LViT model for VQLA tasks that effectively fuse heterogeneous features (visual and text) using our novel GVLE technique.
- Integrate GIoU loss with cross-entropy loss and $\mathcal{L}_1$ loss to improve both the prediction and localization performance of the VQLA model.
- With extensive validation, we find that (i) Surgical-VQLA can localize the context even when the answer is related to surgical interaction. (ii) Our detector-free VQLA demonstrates better feature learning by avoiding computationally expensive and prone to error detection modules and facilitates the end-to-end real-time application of the surgical question localized-answering system. (iii) Proposed GVLE effectively fuses the heterogeneous modalities of visual and word embedding and outperforms existing approaches.

II. METHODOLOGY

A. Preliminaries

1) VisualBERT ResMLP: VisualBERT ResMLP [2] is a Transformer encoder model that boosts the vision-and-language task performance of VisualBERT [6] with further enhancement of the input token interactions. BERT [15] is a Natural Language Processing model trained with subwords [16] as input. The input subwords e will be mapped to a set of embeddings $e \in E$, with each embedding computed by the sum of token embedding e_t, segment embedding e_s, and position embedding e_p. On top of BERT [15], VisualBERT [6] extracted visual features from object proposals to generate related visual embeddings F. Similarly, each embeddings $f \in F$ is the sum of visual features representation f_v, segment embedding f_s and position embedding f_p. Here, position embedding is unique for each token, but segment embedding is just used to distinguish sentence and visual features. The visual and word embeddings are then combined with concatenation operation, before being sent into the multilayer Transformer, and further establish the joint inference and representation of visual and text tokens. VisualBERT ResMLP [2] further emphasizes the token interactions based

[1]Official implementation at: github.com/longbai1006/Surgical-VQLA

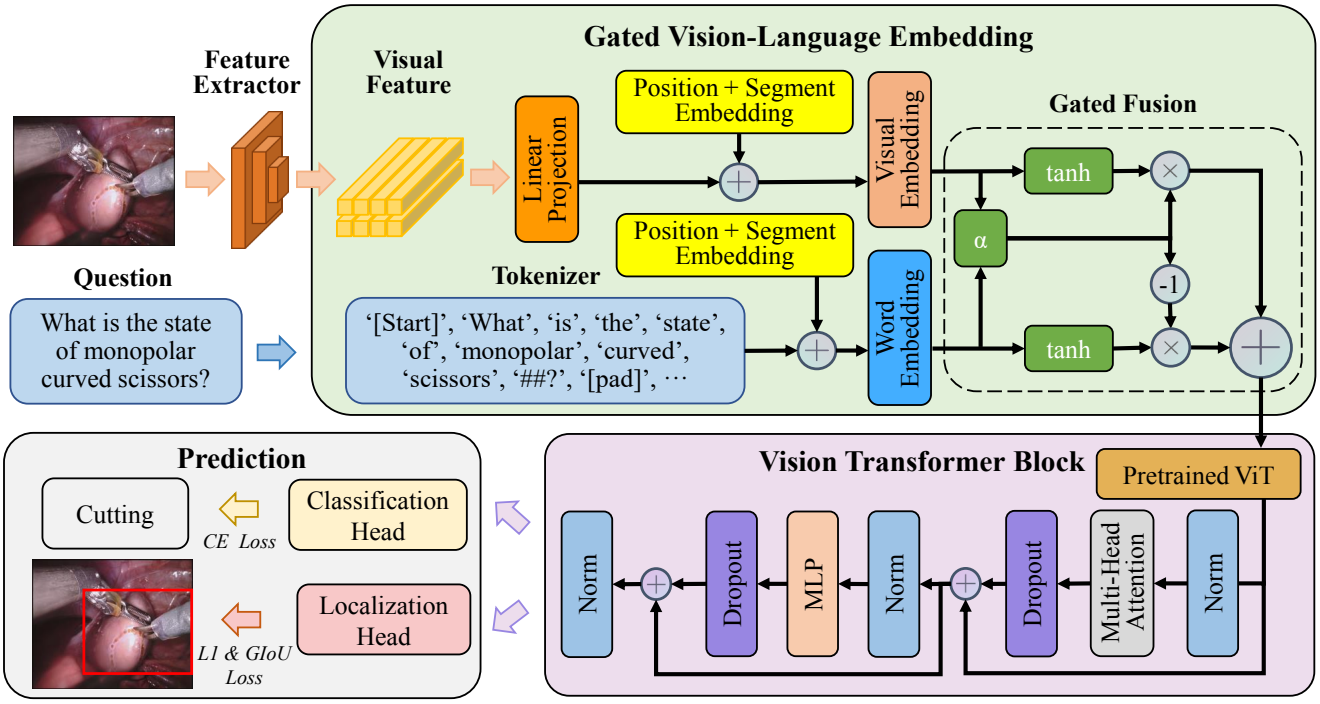

Fig. 2. The proposed network architecture. The robot surgery image feeds the pre-trained feature extractor and the question feeds the customized tokenizer. The GVLE module then embeds the input features and optimizes the combination of visual and word embeddings. Fused features are propagated through the pre-trained ViT module. Finally, the answer and bounding box prediction is given by a classification head with softmax and a localization head with FFN.

on the idea of residual MLP (ResMLP) [17] by adding cross-token and cross-channel modules in the Transformer block, which allows exchanging information between tokens.

2) Vision Transformer: Vision Transformer (ViT) [18] transfers the high performance of the Transformer [19] from language tasks to vision tasks by cutting images into flattened patches. ViT [18] is capable of capturing long-range dependency based on the self-attention mechanism, achieving notable success in vision-based tasks. After getting flattened image patches, ViT [18] conducts patch and position embedding to preserve positional encoding information before the data go into the Transformer encoder. Finally, a multilayer perceptron (MLP) head is used for classification prediction.

B. GVLE-LViT

We develop Language-Vision Transformer (GVLE-LViT) by proposing a Gated Vision-Language Embedding (GVLE) system for efficient embedding to perform Surgical-VQLA. GVLE-LViT forms of visual feature extractor with ResNet18 [20] pre-trained on ImageNet [21], a Tokenizer, GVLE for language-vision embedding, ViT [18] followed by a classification head and a localization head to localize spatial region while predicting the answer. Fig. 2 presents the detailed architecture of our model.

Instead of extracting visual features from object proposals like VisualBERT [6], we found that pre-trained ResNet18 [20] can achieve better performance in our task. The customized tokenizer has been trained on the surgical-specific dataset for the word embeddings. The extracted visual features and word embeddings then feed the GVLE module.

1) Gated Vision-Language Embedding (GVLE): Statistical representation usually does not span modalities [22]. Thus, the combination strategy between visual and word embeddings should be well explored. In VisualBERT [6] and VisualBERT ResMLP [2], after conducting the sum of embeddings, respectively, visual and word embeddings are combined by the naive concatenation. At the same time, they did not consider seeking better ways of fusing representation from different sources. Inspired by Gated Multimodal Unit [23], we borrow the idea from the flow control from recurrent neural networks. Here, the concatenation operation is replaced by a Gated Vision-Language Embedding (GVLE) module to find the best intermediate state from the visual and word embeddings. The right-top of Fig. 2 shows the GVLE module. The feature embeddings of each modality are propagated through a $tanh$ activation function, which encodes the internal representation of the modality features. The gate node α receives the information passed from the $tanh$ activation function and decides whether the corresponding embedding information is useful. The gate is therefore used to control the weights of the synthesized visual and word embeddings and constrain the model. Therefore, the equations to combine the visual and word embeddings are

as follows:

$$\begin{aligned} \omega &= \alpha\left(\theta_{\omega} \cdot [f \,\|\, e]\right) \\ \Upsilon &= \omega * \tanh\left(\theta_{f} \cdot f\right) + (1-\omega) * \tanh\left(\theta_{e} \cdot e\right) \end{aligned} \tag{1}$$

$(\theta_{\omega}, \theta_{f}, \theta_{e})$ are all learnable parameters. $[\cdot \,\|\, \cdot]$ denotes concatenation operation. f and e represents visual and word embeddings, respectively. Υ is the final output of the GVLE module. The model will be able to find the best intermediate representation during training with this architecture, coupling the visual and word embeddings. Subsequently, to fully exploit the power of pre-training, the output integrated embeddings will pass by the standard pre-trained ViT[2] [18] Transformer encoder and Layer-Normalization before the predication head.

2) Prediction Head: The prediction head can be divided into the classification head and localization head. In the classification head, the output of the ViT [18] block is propagated through a linear prediction layer with Softmax to achieve classification prediction. The feed-forward network (FFN) is employed as the localization head. The FFN possesses a 3-layer perceptron with ReLU activation before a linear projection layer. The localization head outputs the final prediction of the normalized coordinates of the bounding box: height, width, and centre coordinates. Therefore, the system is established as an end-to-end framework.

3) Loss Function: Firstly, a simple cross-entropy loss is employed as the classification loss. Then, in the detection task, we found that the sum of $\mathcal{L}_1$ loss and GIoU loss [14] lead to better performance. GIoU loss [14] focuses on both overlapping regions and other non-overlapping regions:

$$\mathcal{L}_{GIoU} = 1 - \left(\frac{|b_g \cap b_p|}{|b_g \cup b_p|} - \frac{|B\left(b_g, b_p\right) \backslash b_b \cup b_p|}{|B\left(b_b, b_p\right)|} \right) \tag{2}$$

b_g represents the ground truth bounding box, and b_p denotes the predicted bounding box. $|\cdot|$ represent the area, and the operation B indicate the largest box containing both b_g and b_p. We add the classification loss and detection loss together for the joint training. Therefore, the final loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_{CE} + \left(\mathcal{L}_{GIoU} + \mathcal{L}_1\right) \tag{3}$$

III. EXPERIMENT

A. Dataset

1) EndoVis-18-VQLA: MICCAI Endoscopic Vision Challenge 2018 [24] dataset is a public dataset with 14 video sequences on robotics surgery procedures. We combine the bounding box on tissue-instrument interaction detection tasks [25] and the question-answer pairs from surgical VQA classification tasks [12], generating **EndoVis-18-VQLA** with question-answer-bounding box annotations. Seenivasan et.al [12] annotated the question-answer pairs of EndoVis-18 and make it publicly accessible[3]. The answers are in single-word form with 18 distinct answer classes (1 organ, 13 tool interactions, and 4 tool locations). If the question-answer pair is only related to the organ or the tool locations, the corresponding detection bounding box will be directly given by the bounding box of the organ or the tool. Conversely, if the question-answer pair is related to the tissue-tool interaction, we adopt the operation B in Equation 2 to design a combined bounding box containing both the organ bounding box and the tool bounding box.

We split the training and validation set by following the setup in surgical VQA classification tasks [2]. Thus, we have 1560 frames with 9014 question-answer pairs in the training set, and 447 frames with 2769 question-answer pairs in the validation set. The EndoVis-18-VQLA dataset has been released publicly together with our official code implementation.

2) EndoVis-17-VQLA: EndoVis-2017 Dataset [26] is from the MICCAI Endoscopic Vision Challenge 2017. The original dataset contains 10 video sequences on robotics surgery scenes. To prove the generalization ability of our model, we manually select 97 frames with common tools and interactions from EndoVis-2017, and annotate the frames with question-answer-bounding box labels. Finally, we generate **EndoVis-17-VQLA** as an external validation dataset with 97 frames and 472 question-answer pairs. It is also publicly accessible with our official code implementation.

B. Implementation Details

For the fair comparison, we add our prediction head in Section II-B.2 with VisualBERT[4] [6] and VisualBERT ResMLP[2] [22] and train with loss function in Equation 3 to enable both classification and localization feature on these reference models.

All models are trained using the Adam optimizer [27] for 80 epochs with a batch size of 64 and a learning rate of 1×10^{-5}. The models are trained on EndoVis-18-VQLA training set, validated on both EndoVis-18-VQLA validation set and EndoVis-17-VQLA, an external validation dataset. All experiments are implemented by Python PyTorch framework, and conducted on a server with NVIDIA RTX 3090 GPU and Intel Core i9-10980XE CPU.

C. Results

The performance of our proposed GVLE-LViT model is both quantitatively (Table I) and qualitatively (Fig. 3 benchmarked against the state-of-the-art (SOTA) Transformers-based VisualBert [6] and VisualBert ResMLP [2] models for Visual Question localized-answering on EndoVis-18-VQLA and EndoVis-17-VQLA dataset. Table I shows that our proposed model outperforms other SOTA models on both datasets. Furthermore, comparing the performance of all three models using the features extracted from the object detection model output against the performance of 3 models using features from the entire image, we note that the performances of the models that use features from the entire image are consistently superior. This superior performance can be attributed to the model's ability to perform global

[2] github.com/rwightman/pytorch-image-models

[3] github.com/lalithjets/surgical_vqa

[4] github.com/uclanlp/visualbert

TABLE I

COMPARISON EXPERIMENTS OF OUR GVLE-LVIT MODEL, AGAINST VISUALBERT [6] AND VISUALBERT RESMLP [2] BASED MODEL. RN DENOTES RESNET.

Model	Visual Feature			EndoVis-18-VQLA			EndoVis-17-VQLA		
	Detection	Feature Extraction	FPS	Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT [6]	FRCNN [28]	RN [20]	18.09	0.5883	0.3012	0.7383	0.4428	**0.3844**	0.7094
VisualBERT ResMLP [2]				0.6049	0.3045	0.7287	0.4258	0.3702	0.6803
GVLE-LViT (Ours)				0.6079	**0.3677**	0.7122	0.4407	0.3273	0.6852
VisualBERT [6]	×	RN [20]	150.60	0.6215	0.3320	0.7356	0.3898	0.3169	0.7105
VisualBERT ResMLP [2]				0.6320	0.3311	0.7501	0.4195	0.3316	0.7035
GVLE-LViT (Ours)				**0.6659**	0.3614	**0.7625**	**0.4576**	0.2489	**0.7275**

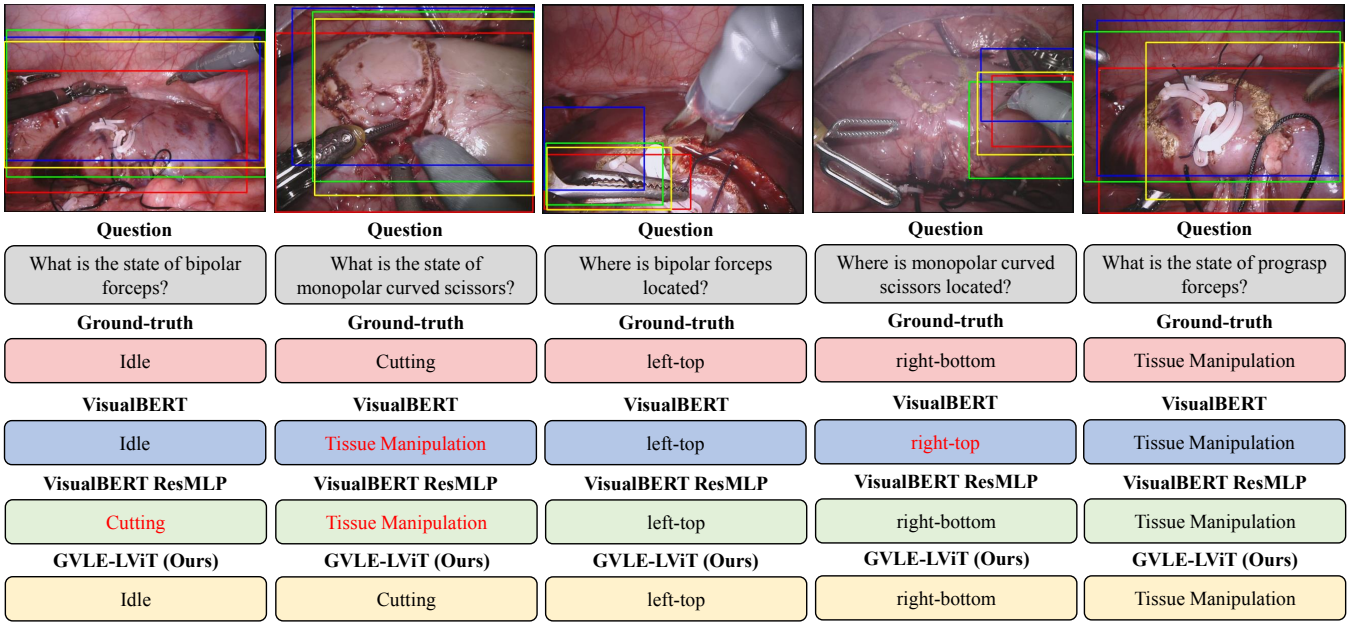

Fig. 3. Several examples of answer and bounding box generation by VisualBERT [6], VisualBERT ResMLP [2], and our GVLE-LViT model. Compared with the baseline models, the localization and classification prediction results of our model are more accurate. The denotation of bounding box color is as follows: red: Ground-truth, blue: VisualBERT [6], green: VisualBERT ResMLP [2], yellow: GVLE-LViT (Ours).

TABLE II

K-FOLD COMPARISON EXPERIMENTS OF OUR GVLE-LVIT MODEL ON VQLA TASKS, AGAINST VISUALBERT [6] AND VISUALBERT RESMLP [2] BASED MODEL.

Models	Fold	EndoVis-18-VQLA			EndoVis-17-VQLA		
		Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT [6]	1^{st} Fold	0.6215	0.3320	0.7356	0.3898	0.3169	0.7105
VisualBERT ResMLP [2]		0.6320	0.3311	0.7501	0.4195	**0.3316**	0.7035
GVLE-LViT (Ours)		**0.6659**	**0.3614**	**0.7625**	**0.4576**	0.2489	**0.7275**
VisualBERT [6]	2^{nd} Fold	0.6290	0.3458	0.7609	0.3898	0.3333	0.7141
VisualBERT ResMLP [2]		0.6174	0.3365	0.7667	0.4216	0.3787	0.7349
GVLE-LViT (Ours)		**0.6655**	**0.4122**	**0.7691**	**0.4831**	**0.3953**	**0.7433**
VisualBERT [6]	3^{rd} Fold	0.5771	0.3421	0.7440	**0.4470**	0.3488	0.7224
VisualBERT ResMLP [2]		0.5817	0.3794	0.7456	0.4025	0.3271	0.7159
GVLE-LViT (Ours)		**0.6247**	**0.4062**	**0.7636**	0.4449	**0.3546**	**0.7430**

scene understanding from the complete image (in-line with the observation made by Seenivasan et.al [12]) and optimal convergence of the model from end-to-end model training. Additionally, by removing the need for an object detection

TABLE III

ABLATION STUDIES WITH DIFFERENT LOCALIZATION LOSS FUNCTION COMBINATIONS ON OUR PROPOSED GVLE-LVIT MODEL, AGAINST VISUALBERT [6] AND VISUALBERT RESMLP [2] BASED MODEL

Models	Loss Function			EndoVis-18-VQLA			EndoVis-17-VQLA		
	VQA	Detection		Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT [6]	*CE*	$\mathcal{L}_1$	×	0.6244	**0.3681**	0.7234	0.4174	**0.3326**	0.7136
VisualBERT ResMLP [2]	*CE*	$\mathcal{L}_1$	×	0.6107	0.2977	0.7383	0.3877	0.3197	0.7089
GVLE-LViT (Ours)	*CE*	$\mathcal{L}_1$	×	0.6287	0.2965	0.7520	0.4407	0.2166	0.7120
VisualBERT [6]	*CE*	$\mathcal{L}_1$	*GIoU* [14]	0.6215	0.3320	0.7356	0.3898	0.3169	0.7105
VisualBERT ResMLP [2]	*CE*	$\mathcal{L}_1$	*GIoU* [14]	0.6320	0.3311	0.7501	0.4195	0.3316	0.7035
GVLE-LViT (Ours)	*CE*	$\mathcal{L}_1$	*GIoU* [14]	**0.6659**	0.3614	**0.7625**	**0.4576**	0.2489	**0.7275**

TABLE IV

COMPARISON EXPERIMENTS OF OUR GVLE LANGUAGE-VISION EMBEDDING FUSION AGAINST CONCAT [6], AFF [13] AND IAFF [13] BASED FUSION STRATEGIES.

Embedding Techniques	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-Score	mIoU	Acc	F-Score	mIoU
ConCAT [6]	0.6551	0.3591	0.7386	0.4258	0.3183	0.7035
AFF [13]	0.6295	0.3521	0.7459	0.3835	**0.3270**	0.7051
iAFF [13]	0.6356	0.3339	0.7498	0.4047	0.2948	0.7164
GVLE (Ours)	**0.6659**	**0.3614**	**0.7625**	**0.4576**	0.2489	**0.7275**

network, we increased the model's processing speed by more than 8 times, achieving 150.6 frames per second (FPS) and making it suitable for real-time applications. Qualitatively, as shown in Fig. 3, our model outperforms base models in both answering and localization (close to ground-truth bounding box annotation).

A K-fold study is also conducted to study and prove our model's superiority over the base model. We set up three different ways of splitting the training and test set. Table II proves that our proposed GVLE-LViT model generally outperforms the base Transformer-based models on all three folds on both datasets.

D. Ablation Studies

Firstly, an ablation study on the performance of the models trained using different loss function combinations is studied (Table III). As our GVLE-LViT and transformed-based baseline models (VisualBERT [6] and VisualBERT ResMLP [2]) aim to predict the localized answer, the loss function for both answer prediction and answer location is used during the training process. From Table III, it is observed that in addition to cross-entropy (CE) loss (for answer prediction) and $\mathcal{L}_1$ loss (for answer localization), incorporating GIoU [14] loss (for answer localization) significantly improves the model's performance in both answer prediction and answer localization.

Secondly, an ablation study on various techniques of heterogenous feature fusion is studied. Our proposed GVLE feature fusion technique is compared against ConCAT [6], AFF [13] and iAFF [13] techniques. Table IV proves that our proposed GVLE vision-language feature fusion technique outperforms other feature fusion techniques.

IV. CONCLUSIONS

We design and propose a Surgical Visual Question Localized-Answering (Surgical-VQLA) model that can answer "what?" and "where?" based on a given input question and surgical scene, making it easier for the student to infer "why?". Specifically, we propose a GVLE-LViT model that better fuses heterogeneous features (visual and text) using our proposed GVLE technique that outperforms existing SOTA models in Surgical-VQLA tasks on two surgical datasets. Additionally, we integrate GIoU loss with cross-entropy loss and $\mathcal{L}_1$ loss to improve both the prediction and localization performance of the model. Through extensive comparative, k-fold and ablation studies, we prove that our proposed GVLE-LViT trained using our proposed loss combination outperforms existing SOTA models. The Surgical-VQLA system may become an important auxiliary tool in surgical training.

While the proposed VQLA model aims to provide reliable answer prediction, to an extent, the localization of the answer could help quantify the reliability of prediction on new data, where if the localization is far-off than the target instrument or tissue, the user can infer that the prediction is probably wrong or the input data is out-of-distribution data. Therefore, using localization information to predict prediction reliability could be a possible future work. In an application stance, our proposed VQLA model opens novel possible applications for medical diagnosis. More complicated datasets and challenging QA pairs shall further boost the prospective of the Surgical-VQLA system.

REFERENCES

[1] D. Sharma, S. Purushotham, and C. K. Reddy, "Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, no. 1, pp. 1–18, 2021.

[2] L. Seenivasan, M. Islam, A. Krishna, and H. Ren, "Surgical-vqa: Visual question answering in surgical scenes using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 33–43.

[3] M.-C. Hsieh and Y.-H. Lin, "Vr and ar applications in medical practice and education," *Hu Li Za Zhi*, vol. 64, no. 6, pp. 12–18, 2017.

[4] H. Ashraf, M. H. Sodergren, N. Merali, G. Mylonas, H. Singh, and A. Darzi, "Eye-tracking technology in medical education: A systematic review," *Medical teacher*, vol. 40, no. 1, pp. 62–69, 2018.

[5] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.

[6] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[7] S. Sheng, A. Singh, V. Goswami, J. Magana, T. Thrush, W. Galuba, D. Parikh, and D. Kiela, "Human-adversarial visual question answering," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[9] H. Sharma and A. S. Jalal, "Image captioning improved visual question answering," *Multimedia Tools and Applications*, pp. 1–22, 2021.

[10] S. Barra, C. Bisogni, M. De Marsico, and S. Ricciardi, "Visual question answering: which investigated applications?" *Pattern Recognition Letters*, vol. 151, pp. 325–331, 2021.

[11] H. Sharma and A. S. Jalal, "Visual question answering model based on graph neural network and contextual attention," *Image and Vision Computing*, vol. 110, p. 104165, 2021.

[12] L. Seenivasan, S. Mitheran, M. Islam, and H. Ren, "Global-reasoned multi-task learning model for surgical scene understanding," *IEEE Robotics and Automation Letters*, 2022.

[13] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3560–3569.

[14] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[17] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek *et al.*, "Resmlp: Feedforward networks for image classification with data-efficient training," *arXiv preprint arXiv:2105.03404*, 2021.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Advances in neural information processing systems*, vol. 25, 2012.

[23] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[24] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[25] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 627–636.

[26] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.