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 naïve; and (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.

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.

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?”.

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.
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 counterfeit 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 the tremendous development in deep learning and readily available enormous datasets [2], [6]–[8], significant progress has been made in developing VQA models [9]–[11] in the computer vision domain. Constructed using long-short term memory modules [12], [13] or attention modules [1], [14], 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 heterogeneous 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 [15] which 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 limit usage in real-time applications. 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) [16] 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 (GloU) [17] loss for robust localized-answering.

Overall, 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 GloU loss with cross-entropy loss and $L_1$ 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 [9] with further enhancement of the input token interactions. BERT [18] is a Natural Language Processing model trained with sub-words [19] 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 [18], VisualBERT [9] 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.

\[ F = \text{concat}(f_v, f_s, f_p) \]

Official implementation at: github.com/longbai1006/Surgical-VQLA
What is the state of monopolar curved scissors?

**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 FPN.


2) **Vision Transformer:** Vision Transformer (ViT) [21] transfers the high performance of the Transformer [22] from language tasks to vision tasks by cutting images into flattened patches. ViT [21] 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 [21] 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 [23] pre-trained on ImageNet [24], a Tokenizer, GVLE for language-vision embedding, ViT [21] 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 [9], we found that pre-trained ResNet18 [23] 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 [25]. Thus, the combination strategy between visual and word embeddings should be well explored. In VisualBERT [9] 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 [26], 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 $\alpha$ 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 obtained visual and word embeddings and constrain the model. Therefore, the equations to combine the visual and word embeddings
are as follows:
\[
\omega = \alpha (\theta_w \cdot [f \parallel e])  \\
Υ = \omega \ast \tanh (\theta_f \cdot f) + (1 - \omega) \ast \tanh (\theta_e \cdot e)  \\
(\theta_w, \theta_f, \theta_e) \text{ are all learnable parameters. } [\cdot \parallel \cdot] \text{ denotes concatenation operation. } f \text{ and } e \text{ represents visual and word embeddings, respectively. } \text{T is the final output of the GVLE module.} \]

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 [21] block is propagated through a linear projection 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 \(L_1\) loss and GIoU loss [17] lead to better performance. GIoU loss [17] focuses on both overlapping regions and other non-overlapping regions:

\[
\mathcal{L}_{GIoU} = 1 - \left( \frac{|b_g \cap b_p|}{|b_g \cup b_p|} - \frac{|B(b_g, b_p) \setminus b_g \cup b_p|}{|B(b_g, b_p)|} \right)  \\
b_g \text{ represents the ground truth bounding box, and } b_p \text{ denotes the predicted bounding box. } |\cdot| \text{ represent the area, and the operation } B \text{ indicate the largest box containing both } b_g \text{ and } b_p. \text{ We add the classification loss and detection loss together for the joint training. Therefore, the final loss function can be defined as:}
\]

\[
\mathcal{L} = \mathcal{L}_{CE} + (\mathcal{L}_{GIoU} + L_1)  \\
\text{III. EXPERIMENT}  \\
\text{A. Dataset}  \\
1) \text{EndoVis-18-VQLA: MICCAI Endoscopic Vision Challenge 2018 [27] dataset is a public dataset with 14 video sequences on robotics surgery procedures. We combine the bounding box on tissue-instrument interaction detection tasks [28] and the question-answer pairs from surgical VQA classification tasks [15], generating EndoVis-18-VQLA with question-answer-bounding box annotations. Seenivasan et.al [15] annotated the question-answer pairs of EndoVis-18 and make it publicly accessible. 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 [29] 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\[9\] and VisualBERT ResMLP\[25\] 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 [30] for 80 epochs with a batch size of 64 and a learning rate of \(1 \times 10^{-5}\). 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 i9-10980X CPU.

C. Results

The performance of our proposed GVLE-LViT model is both quantitatively (Table 1) and qualitatively (Fig. 3) benchmarked against the state-of-the-art (SOTA) Transformers-based VisualBert [9] and VisualBert ResMLP [2] models for Visual Question localized-answering on EndoVis-18-VQLA and EndoVis-17-VQLA dataset. Table 1 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

\[\text{github.com/rwrightman/pytorch-image-models}\]
\[\text{github.com/uclanlp/visualbert}\]
\[\text{github.com/lalithjets/surgical_vqa}\]
### TABLE I

<table>
<thead>
<tr>
<th>Model</th>
<th>Visual Feature</th>
<th>EndoVis-18-VQLA</th>
<th>EndoVis-17-VQLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Detection</td>
<td>F-Score</td>
<td>mIoU</td>
</tr>
<tr>
<td></td>
<td>Extraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td></td>
<td>0.5883</td>
<td>0.3012</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td></td>
<td>0.6049</td>
<td>0.3045</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td>FRCNN [31] RN [23]</td>
<td>18.09</td>
<td><strong>0.3677</strong></td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td></td>
<td>0.6215</td>
<td>0.3320</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>× RN [23]</td>
<td>150.60</td>
<td><strong>0.6659</strong></td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Several examples of answer and bounding box generation by VisualBERT [9], 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 [9], green: VisualBERT ResMLP [2], yellow: GVLE-LViT (Ours).

### TABLE II

<table>
<thead>
<tr>
<th>Models</th>
<th>Fold</th>
<th>EndoVis-18-VQLA</th>
<th>EndoVis-17-VQLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc</td>
<td>F-Score</td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td></td>
<td>0.6215</td>
<td>0.3320</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>1st Fold</td>
<td>0.6320</td>
<td>0.3311</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td><strong>0.6659</strong></td>
<td><strong>0.3614</strong></td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td></td>
<td>0.6290</td>
<td>0.3458</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>2nd Fold</td>
<td>0.6174</td>
<td>0.3365</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td><strong>0.6655</strong></td>
<td><strong>0.4122</strong></td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td></td>
<td>0.5771</td>
<td>0.3421</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>3rd Fold</td>
<td>0.3817</td>
<td>0.3794</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td><strong>0.6247</strong></td>
<td><strong>0.4062</strong></td>
</tr>
</tbody>
</table>

scene understanding from the complete image (in-line with the observation made by Seenivasan et.al [15]) and optimal convergence of the model from end-to-end model training. Additionally, by removing the need for an object detection
TABLE III

<table>
<thead>
<tr>
<th>Models</th>
<th>Loss Function</th>
<th>EndoVis-18-VQLA</th>
<th>EndoVis-17-VQLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Detection</td>
<td>Acc</td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td>CE</td>
<td>0.6244</td>
<td>0.3681</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>$L_1$</td>
<td>0.6107</td>
<td>0.2977</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td>0.6287</td>
<td>0.2965</td>
</tr>
<tr>
<td>VisualBERT [9]</td>
<td>$L_1$</td>
<td>0.6215</td>
<td>0.3320</td>
</tr>
<tr>
<td>VisualBERT ResMLP [2]</td>
<td>$L_1$</td>
<td>0.6320</td>
<td>0.3311</td>
</tr>
<tr>
<td>GVLE-LViT (Ours)</td>
<td></td>
<td>0.6659</td>
<td>0.3614</td>
</tr>
</tbody>
</table>

TABLE IV
COMPARISON EXPERIMENTS OF OUR GVLE LANGUAGE-VISION EMBEDDING FUSION AGAINST ConCAT [9], AFF [16] and iAFF [16] BASED FUSION STRATEGIES

<table>
<thead>
<tr>
<th>Embedding Techniques</th>
<th>EndoVis-18-VQLA</th>
<th>EndoVis-17-VQLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>F-Score</td>
</tr>
<tr>
<td>ConCAT [9]</td>
<td>0.6551</td>
<td>0.3591</td>
</tr>
<tr>
<td>AFF [16]</td>
<td>0.6295</td>
<td>0.3521</td>
</tr>
<tr>
<td>iAFF [16]</td>
<td>0.6356</td>
<td>0.3339</td>
</tr>
<tr>
<td>GVLE (Ours)</td>
<td>0.6659</td>
<td>0.3614</td>
</tr>
</tbody>
</table>

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 $L_1$ 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.