SURGICAL-LVLM: LEARNING TO ADAPT LARGE VISION-LANGUAGE MODEL FOR GROUNDED VISUAL QUESTION ANSWERING IN ROBOTIC SURGERY

Guankun Wang[1,*] **Long Bai**[1,*] **Wan Jun Nah**[2] **Jie Wang**[1] **Zhaoxi Zhang**[1]
Zhen Chen[3] **Jinlin Wu**[3] **Mobarakol Islam**[4] **Hongbin Liu**[3] **Hongliang Ren**[1,†]

[1] The Chinese University of Hong Kong [2] Universiti Malaya
[3] Centre for Artificial Intelligence and Robotics, HKISI-CAS [4] University College London

`{gkwang, b.long}@link.cuhk.edu.hk, hlren@ee.cuhk.edu.hk`

ABSTRACT

Recent advancements in surgical Visual Question Answering (VQA) and related region grounding have shown great promise for robotic and medical applications, addressing the critical need for automated methods in personalized surgical mentorship. However, existing models primarily provide simple structured answers and struggle with complex scenarios due to their limited capability in recognizing long-range dependencies and aligning multimodal information. In this paper, we introduce Surgical-LVLM, a novel personalized large vision-language model tailored for complex surgical scenarios. Utilizing the pre-trained large vision-language model and specialized Visual Perception LoRA (VP-LoRA) blocks, our model excels in understanding complex visual-language tasks within surgical contexts. In addressing the visual grounding task, we propose the Token-Interaction (TIT) module, which strengthens the interaction between the grounding module and the language responses of the Large Visual Language Model (LVLM) after projecting them into the latent space. We demonstrate the effectiveness of Surgical-LVLM on several benchmarks, including EndoVis-17-VQLA, EndoVis-18-VQLA, and a newly introduced EndoVis Conversations dataset, which sets new performance standards. Our work contributes to advancing the field of automated surgical mentorship by providing a context-aware solution.

1 INTRODUCTION

The progress achieved in surgical vision question answering (VQA) (Seenivasan et al., 2022a) and related region grounding has attracted many researchers since these techniques demonstrate great potential for applications in robotics and healthcare. To integrate the above techniques and achieve them simultaneously, Bai et al. (2023d) pioneered the surgical visual question localized-answering (VQLA), which predicts answers with localized results through the input text and images. Current approaches in this field predominantly rely on the expertise of expert surgeons, often failing to acknowledge the challenges posed by real-world constraints, such as the limited availability of expert surgeons. Consequently, the demand for auto methods for personalized surgical mentorship stands as a critical, yet challenging issue. Despite their effectiveness in providing simple answers, these methods generally have difficulties when faced with challenging answer-location scenarios.

Building upon the groundwork of elementary Visual Question Answering (VQA) models, (Seenivasan et al., 2023; Yuan et al., 2023; Bai et al., 2023b;c) innovatively delineates a paradigm for the community in the study of Surgical Visual Question Localized-Answering utilizing deep learning methodologies. However, these methods are mainly subject to two common limitations. Firstly,

---

[*]Equal contribution.
[†]Corresponding author.

these methods have difficulty establishing accurate long-range dependencies when dealing with complex scenarios, due to occlusion and distortion, which may lead to the ignorance of relevant visual details that are significant to surgical environments. Additionally, they often overlook the alignment of multiple modalities, resulting in a mismatch between language and visual grounding. Considering these issues above, we ask: *Can we personalize a large vision-language model to automatically ground VQA in surgical scenarios in a simple and efficient manner?*
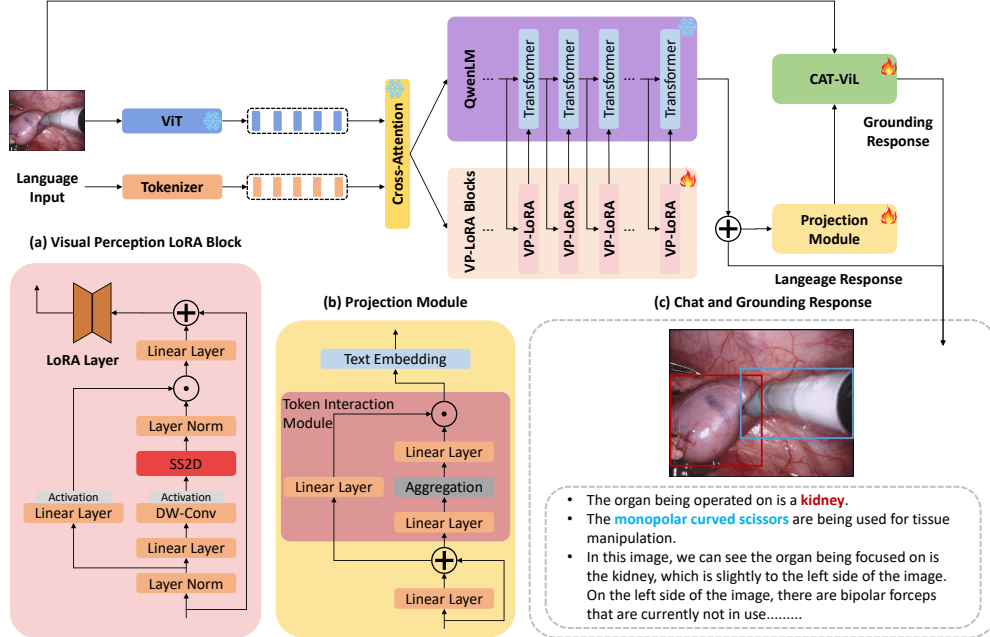

Figure 1: The flowchart of the proposed framework. It includes a pre-trained ViT, tokenizer, the Qwen-VL model, projection module, and the CAT-ViL embedding module.

With the rise of Large Language Models (LLMs) (Achiam et al., 2023), Large Vision-Language Models (LVLMs) have emerged as a new method for understanding the interplay between visual and textual information (Liu et al., 2024a; Zhang et al., 2023b; You et al., 2023). However, existing large models still struggle to grasp domain-specific semantic information in subdomains (Li et al., 2024b), making a surgical-focused LVLM a current necessity. To this end, we present Surgical-LVLM, a personalized surgical large vision-language model for highly specialized surgical scenarios. As shown in Fig. 1(c), our method efficiently customizes LVLM faced with occlusive medical samples. Specifically, we first obtain the describable visual-language tokens by a pre-trained vision transformer (ViT) and Qwen-VL (Bai et al., 2023a). Consequently, we utilize a cross-attention layer stack of specially designed Visual Perception LoRA (VP-LoRA) blocks to unleash attention's feature caption power with one core technique. Specifically, we integrate Visual State Space (VSS) blocks at each Low-Rank Adaptation layer, enabling the effective propagation of global contextual information and allowing Qwen-VL to better adapt to specialized surgical scenarios.

With the aforementioned designs, Surgical-LVLM exhibits favorable grounding performance for hard samples in a variety of complex scenes. Nevertheless, there might be occasional failure cases, specifically the misalignment problem that the language model may describe some correct, yet subimportant content. Such ambiguity challenges Surgical-LVLM's grounding capacity in determining the appropriate object to localize as output since both the important and sub-important parts can be regarded as correct answers. To alleviate this issue, we further propose an alignment variant of our approach. In detail, we project the tokens generated from Qwen-VL to latent space and take a token interaction (TIT) module to enable text tokens to emphasize the important vision information. By such efficient projection-based training, Surgical-LVLM selectively focuses on the relevant features or words that are most important for the VQLA task. Our contributions are as follows:

- We introduce Surgical-LVLM, a personalized foundational model tailored for VQA and grounding for intricate surgical contexts. Additionally, we further incorporate VP-LoRA to facilitate efficient long-range representation learning in complex surgical environments.

- We present the Token-Interaction Module, designed to effectively project and align language responses from LLMs with the visual grounding module, enhancing the synergy between textual and visual data.

- We have extensively validated the effectiveness of our proposed method on multiple benchmarks, including the public EndoVis 2017 and 2018 datasets, and our novel EndoVis Conversations dataset. The results significantly underscore improvements in accurate surgical visual-language understanding and grounding, and establish state-of-the-art performance on these datasets.

2 RELATED WORK

2.1 LARGE VISUAL-LANGUAGE MODEL

With the significant progress of LLMs such as GPT-4 (Achiam et al., 2023) and its open-source alternatives like LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), researchers have been increasingly motivated to exploit their integration of powerful language abilities with vision (Zhang et al., 2023a). Recently, the performance of the general large multimodal models has been continuously refreshing the expectations since the scaling up of the pretraining data (Bai et al., 2023a; Dai et al., 2024) and instruction-following data (Dai et al., 2024; Li et al., 2023a; Wang et al., 2024). However, most models can only excel in general domains and easily encounter challenges in specific scenarios, such as complex surgical scenarios. In addition, while much of the current work has a good performance of language responses based on images and user instructions, it is difficult to provide alignment to related image regions (visual grounding) simultaneously. It is noted that CogVLM-Grounding (Wang et al., 2023) and MiniGPT-v2 (Chen et al., 2023) have a limitation in the generation of concise captions during grounding tasks, attributable to its primary training on caption data from datasets such as Flickr30K (Plummer et al., 2015). Conversely, BuboGPT (Zhao et al., 2023) preserves conversational competency utilizing an external grounding model for grounding purposes, although the language encoder in the grounding model constrains the chat capability. These methods face challenges in achieving proficiency in both chat and grounding domains.

2.2 VQLA MODELS IN MEDICAL DOMAIN

Medical VQLA has emerged as a pivotal area in the medical domain, using the integration of textual questions and images to support clinical diagnostics and treatment. However, the unique characteristics inherent in medical data necessitate specialized expertise for accurate interpretation, thereby introducing additional complexities in developing effective medical VQLA models (Liu et al., 2023b; Bai et al., 2025). Recent advancements in medical VQLA research have been focused on addressing these inherent challenges. Surgical-VQLA (Bai et al., 2023d) incorporates a gated vision-language embedding approach to fuse multimodal features from visual and textual sources. CAT-ViL (Bai et al., 2023b) is subsequently proposed as a co-attention gated embedding module, allowing for instructive interaction between text and visual embeddings. These pioneering endeavors have contributed to several specialized improvements for medical scenarios and achieved superior representation learning. However, due to parameter limitations and the lack of specialized data, these models can just answer fixed-format questions and are unable to analyze comprehensive information about surgical scenarios to meet clinical needs. Therefore, the introduction of large language models (Liu et al., 2023a; Touvron et al., 2023; Zheng et al., 2024) into the field of surgical-VQLA is necessitated by their powerful capabilities in text generation and comprehension.

3 METHODOLOGY

3.1 PRELIMINARIES

3.1.1 QWEN-VL

Qwen-VL is a Large Vision-Language Mode designed to perceive and understand both texts and images (Bai et al., 2023a). The model architecture consists of Qwen-7B as the language model, a Vision Transformer (ViT) as the visual encoder, and a vision-language adapter to compress the image features. During training, the model is trained in a three-stage pipeline: pretraining with image-text pairs, multi-task pretraining with interleaved vision-language data, and supervised fine-tuning. This training process enables Qwen-VL to interact with users, perceive input images, and generate appropriate responses based on user intent.

3.1.2 SURGICAL-VQLA

Surgical-VQLA indicates models for Visual Question Localized-Answering in robotic surgery (Bai et al., 2023d), which uses a combination of VisualBERT (Li et al., 2019) and a Gated Vision-Language Embedding (GVLE) technique for effective fusion of visual and text features. The GVLE is designed to build input patches for the Language Vision Transformer and enables localized answering by incorporating a detection head parallel to the prediction head of the LViT. Besides, CAT-ViL (Bai et al., 2023b) further introduces the Co-Attention Gated Vision-Language (CAT-ViL) embedding module, which fuses multimodal features more effectively.

3.2 SURGICAL-LVLM

We propose Surgical-LVLM to perform complex scene comprehension as well as logical reasoning while performing VQLA tasks in surgical scenarios. As shown in Fig. 1, the framework consists of finetuned Qwen-VL (ViT (Dosovitskiy et al., 2020), tokenizer, QwenLM and VP-LoRA blocks), a projection module and the grounding module (CAT-ViL).

3.2.1 VISUAL PERCEPTION LORA

Due to problems such as occlusion and distortion, traditional Transformers have difficulties in establishing accurate long-range dependencies when dealing with images in complex surgical scenarios. Therefore, we design Visual Perception LoRA, which enhances the visual perception capabilities of Qwen-VL by integrating Visual State Space (VSS) blocks (Liu et al., 2024c) at each Low-Rank Adaptation (LoRA) layer (Hu et al., 2021) during finetuning.

As shown in Fig. 1 (a), the input first passes through an initial normalization layer. Then, the output is divided into two parallel information branches: the first branch contains linear transformation and activation, while the second branch incorporates linear transformation, depthwise separable convolution, and subsequent activation, followed by selective feature extraction through the 2D-Selective-Scan (SS2D) module. The SS2D is able to capture different features by expanding them and conducting feature extraction (Ruan & Xiang, 2024). Finally, the linear layer is utilized to integrate the merged features of the two branches, which are then combined via the residual connection to generate the VP-LoRA output following the LoRA layer. With X_{in} and X_{out} representing the input and output features, the equation of the VP-LoRA block is as follows:

$$X_{out} = \mathcal{F}_{\text{LoRA}}[\mathcal{F}_{\text{VSS}}(X_{in})] \tag{1}$$

The incorporation of VSS blocks into each LoRA layer of the Qwen-VL model facilitates the efficient learning of global contextual information. At each layer, the VSS blocks capture the dependencies between visual features and allow the model to refine its understanding of the input images. Besides, the VP-LoRA enables the model to focus on relevant visual details, understand complex visual scenes, and generate more accurate and contextually relevant responses.

3.2.2 MULTIMODAL ALIGNMENT FOR VISUAL GROUNDING

Currently, many LVLMs have separate chat and grounding functions when conducting the multi-model task. However, these models tend to perform poorly when both chatting and grounding are
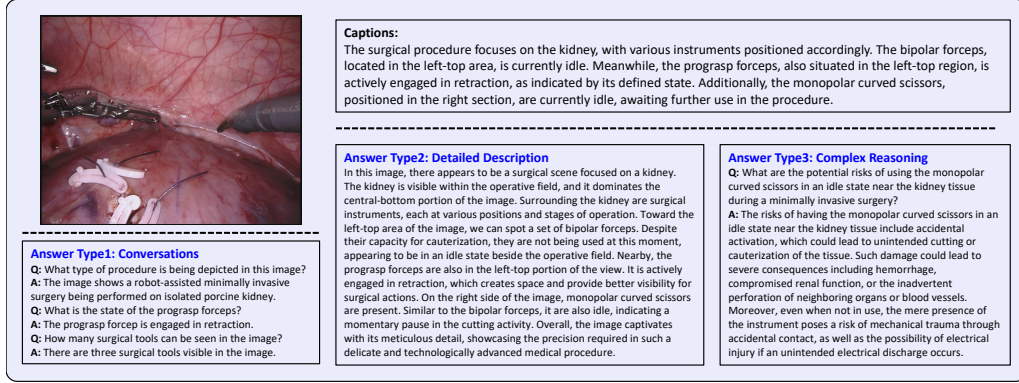
Figure 2: A sample of Instruction Tuning data generation. Contexts such as captions and images used to prompt the GPT are shown above the dashes, and three types of responses are shown below the dashes. We use images as prompts for GPT, which can help GPT comprehend the questions better and generate appropriate responses.

required due to the misalignment problem – the language model may describe some correct, yet subimportant content, which will mislead grounding model predictions. To tackle this problem, we propose a novel projection module to align the language response generated from Qwen-VL with the input of CAT-ViL. We design the projection module to enable the language response to support the prediction of related grounding responses.

In the alignment projection module, the language response is first processed through a linear layer and then combined with a residual connection. The combined features will be passed through the Token InTeraction (TIT) module (Yang et al., 2022) to further refine and emphasize the important information in the features. There are two parallel flows in the TIT module. The first flow contains two linear layers with an Aggregation block between them. With x_i and y_i representing the input and output features, the equations of the Aggregation block are as follows:

$$\boldsymbol{y}_i = \mathcal{F}_{inter}\left[\mathcal{A}(i, \mathbf{X}), \boldsymbol{x}_i\right] \tag{2}$$

where the contextual features for each location i are aggregated using $\mathcal{A}$ first, and then the query is computed via interaction $\mathcal{F}_{inter}$ to generate $\boldsymbol{y}_i$. $\mathcal{A}$ denotes the Aggregation operation. The second branch is just linear transformation. Finally, the merged features of these two flows will be output to the CAT-ViL model for text embedding. The TIT module selectively focuses on the relevant features or words that are most important for the VQLA task.

Through multimodal alignment with our projection module, we combine the capabilities of Qwen-VL and CAT-ViL to continuously generate language and grounding responses, thus achieving a fusion of chat and grounding abilities. This integration allows for a comprehensive understanding of the surgical scene and facilitates effective communication between the system and the user.

3.3 INSTRUCTION TUNING

In order to enhance the performance of LVLMs' capability for surgical scenarios understanding and logical reasoning, it is essential to create a dataset with high quality. For this purpose, we have established a dataset comprising question-answer pairs based on the images and annotations from the EndoVis 2017/2018 VQLA datasets. These pairs have been formatted according to the Qwen-VL structure. Guided by the methodology (Liu et al., 2024b), we have formulated questions that span a spectrum of complexity, ranging from easy conversations, detailed descriptions, and complex reasoning as illustrated in Figure 2. These instruction-tuning QA pairs have been synthesized utilizing the GPT-4, which is renowned for its exemplary linguistic proficiency. The detailed statistical information on this dataset can be found in Sec 4.1.

3.4 TRAINING STRATEGY

Our training strategy for Surgical-LVLM consists of two stages: instruction turning for LVLM and multimodel alignment for the grounding module.

Vision-Language Instruction Fine-tuning: In the first phase, we utilize the extensive dataset comprising surgical question-answer pairs alongside associated images to fine-tune the Qwen-VL model augmented with VP-LoRA. This stage exclusively modifies the parameters of adapter layers while keeping the original large language model layers frozen. Through instruction fine-tuning, the Qwen-VL model acquires comprehension of the semantic intricacies and linguistic structures specific to the surgical domain from both textual and visual encoding features. Thus, the model can generate meaningful representations, adeptly capturing the nuances and intricacies inherent in surgical terminology and context, thus establishing a robust foundation for effective vision-language grounding.

Visual Grounding Alignment: In the second stage of training, the fusion of chat and grounding capabilities is enacted by integrating the Qwen-VL and visual grounding module. Utilizing the pre-trained Qwen-VL model, language features extracted from its outputs are employed alongside corresponding surgical images to comprehensively optimize the parameters of CAT-ViL embedding. This optimization is geared towards enhancing the coherence between language responses and specific visual regions in surgical images. To achieve this alignment, we use the co-attention embedding fusion to effectively combine the multi-modal features. Through this process, CAT-ViL acquires the capacity to accurately ground textual descriptions to pertinent regions in the surgical images by learning the associations between Qwen-VL's output features and visual representations.

Overall, through the integration of the instruction fine-turned Qwen-VL and the trained CAT-ViL model, our proposed two-stage training strategy culminates in the synergistic enhancement of chat and grounding functionalities. This approach maximizes the utilization of Qwen-VL's logical reasoning VQA abilities and the language grounding capabilities of CAT-ViL, thereby achieving an integrated fusion of these components to fully harness their potential.

4 EXPERIMENTS

4.1 DATASET

EndoVis Conversations Dataset is a novel dataset generated from EndoVis-18-VQLA and EndoVis-17-VQLA Datasets. We designed questions encompassing three distinct types using GPT-4. Statistically, the training and test sets include 19,020 and 2,151 image-QA pairs, respectively. The image setup for the training and test sets is also referenced to the above two datasets.

EndoVis-18-VQLA Dataset comprises 14 robotic surgery videos obtained from the MICCAI Endoscopic Vision Challenge (Allan et al., 2020). Within this dataset, VQLA annotations are incorporated, integrating the question-answer pairs derived from (Seenivasan et al., 2022b) alongside bounding box annotations sourced from (Islam et al., 2020). These QA pairs encompass a wide spectrum of single-word responses that encapsulate information regarding organs, surgical instruments, and their interrelations. We adhere to the data partitioning strategy delineated in (Seenivasan et al., 2022a) for the training and test datasets.

EndoVis-17-VQLA Dataset is emanated from the MICCAI Endoscopic Vision Challenge 2017 (Allan et al., 2019). Utilizing this dataset as an external validation benchmark, we employ it to demonstrate the generalization efficacy of our proposed model across diverse surgical scenarios.

4.2 IMPLEMENTATION DETAILS

We utilize the mainstream GPT-4 Score in the comparison experiments against BLIP-2 (Li et al., 2023b), miniGPT-4 (Zhu et al., 2023), LLaVA-1.5 (Liu et al., 2023a) and LLaVA-Med (Li et al., 2024a) on the EndoVis Conversations Dataset. Besides, for the EndoVis-18-VQLA and EndoVis-17-VQLA Datasets, the baseline models for our quantitative experimental comparison refer to (Bai et al., 2023b). The evaluation metrics are accuracy, f-score, and mean intersection over union (mIoU) (Rezatofighi et al., 2019). All experiments are implemented by Python PyTorch frame-

work, and conducted on a server with NVIDIA A100 GPUs. The epoch, batch size, and learning rate are set to 20, 16, and 1×10^{-5}, respectively.

Table 1: Ablation study on EndoVis Conversation dataset. 'VP-LoRA' denotes 'Visual Perception LoRA' and 'IF' denotes 'Instruction Fine-tuning'.

Model	VP-LoRA	IF	EndoVis-18-C	EndoVis-17-C
			GPT-4 Score	
Surgical-LVLM	×	×	72.48	69.62
	✓	×	74.61	70.20
	×	✓	88.53	81.23
	✓	✓	**90.68**	**83.24**

Table 2: Comparison experiments of our Surgical-LVLM against LVLMs without fine-tuning on EndoVis Conversations Dataset.

Model	EndoVis-18-C	EndoVis-17-C
	GPT-4 Score	
BLIP-2 (Li et al., 2023b)	65.21	60.43
miniGPT-4 (Zhu et al., 2023)	70.79	66.51
LLaVA-1.5-7B (Liu et al., 2023a)	80.20	76.11
LLaVA-1.5-13B (Liu et al., 2023a)	81.29	77.59
LLaVA-Med (Li et al., 2024a)	83.54	80.63
Surgical-LVLM	**90.68**	**83.24**

Table 3: Quantitative evaluation experiments on EndoVis-18 and EndoVis-17 datasets.

Model	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT (Li et al., 2019)	0.6268	0.3329	0.7391	0.4005	0.3381	0.7073
VisualBERT R (Seenivasan et al., 2022a)	0.6301	0.339	0.7352	0.419	0.337	0.7137
MCAN (Ben-Younes et al., 2017)	0.6285	0.3338	0.7526	0.4137	0.2932	0.7029
VQA-DeiT (Touvron et al., 2021)	0.6104	0.3156	0.7341	0.3797	0.2858	0.6909
MUTAN (Ben-Younes et al., 2017)	0.6283	0.3395	0.7639	0.4242	0.3482	0.7218
MFH (Yu et al., 2018)	0.6283	0.3254	0.7592	0.4103	0.35	0.7216
BlockTucker (Ben-Younes et al., 2019)	0.6201	0.3286	0.7653	0.4221	0.3515	0.7288
CAT-ViL DeiT (Bai et al., 2023b)	0.6452	0.3321	0.7705	0.4491	**0.3622**	0.7322
GVLE-LViT (Bai et al., 2023d)	0.6659	**0.3614**	0.7625	**0.4576**	0.2489	0.7275
Surgical-LVLM (Ours)	**0.6947**	0.3325	**0.8416**	0.4068	0.3412	**0.7825**

4.3 RESULTS

We first evaluate the effectiveness of our proposed methods on the EndoVis Conversation dataset in Table 1. The Surgical-LVLM framework is evaluated based on whether VP-LoRA is introduced and whether it undergoes Instruction Fine-tuning. We can find that Instruction Fine-tuning is crucial for logical reasoning and comprehensive responses in surgical scenarios for the large language model. This is demonstrated by the huge gap in the GPT-4 Score. Therefore, the generalized large language models are not yet fully applicable in specialized scenarios such as surgery. Moreover, the introduction of VP-LoRA brings an effective enhancement to the model's language response, proving its effectiveness in accurate response generation.

Next, we compared the performance of our Surgical-LVLM framework against excellent LVLMs without Instruction Fine-tuning in surgical scenarios, which is shown in Table 2. Significantly, Surgical-LVLM has demonstrated remarkable performance with the highest GPT-4 Score recorded at 90.68 on the EndoVis-18-VQLA dataset and 83.24 on the EndoVis-17-VQLA dataset, surpassing

Table 4: Ablation study on different modified modules. 'VP-LoRA' denotes 'Visual Perception LoRA' and 'MA' denotes 'Multimodal Alignment'.

Model	VP-LoRA	MA	EndoVis-18-VQLA			EndoVis-17-VQLA		
			Acc	F-Score	mIoU	Acc	F-Score	mIoU
Surgical-LVLM	×	×	0.6878	0.3211	0.8303	0.4054	0.3386	0.7689
	✓	×	0.6947	0.3325	0.8322	0.4068	0.3412	0.7719
	×	✓	0.6878	0.3211	0.8336	0.4054	0.3386	0.7781
	✓	✓	**0.6947**	**0.3325**	**0.8416**	**0.4068**	**0.3412**	**0.7825**

all competing models evaluated on the EndoVis Conversations Dataset. This notable margin of improvement further underscores the imperative for instruction tuning within LVLMs frameworks when applied to specific domains.

Quantitative evaluation in Table 3 presents that our Surgical-LVLM outperforms other state-of-the-art models in traditional Surgical-VQLA tasks. On the EndoVis-18 dataset, our Surgical-LVLM model outperforms the others in terms of Accuracy and mIoU but has a lower F-Score. This indicates it is better at identifying the correct answers and detecting the images but might struggle slightly with balancing precision and recall. Its high mIoU suggests strong grounding capabilities. On the EndoVis-17 dataset, the performance of all models decreased due to suffering from domain migration. Despite the overall lower scores compared to EndoVis-18, the Surgical-LVLM still has the highest mIoU, confirming its robustness. However, other models, such as CAT-ViL DeiT and BlockTucker, have slightly better Accuracy and F-scores, hinting at potential areas of improvement for the Surgical-LVLM in terms of Accuracy and F-scores. Overall, the Surgical-LVLM model with Qwen large language model as well as multi-model alignment of CAT-ViL, which has excellent reasoning ability as well as strong inference-based grounding capability, as evidenced by mIoU metrics that are far superior to those of other state-of-the-art models.

Furthermore, an ablation study is conducted to observe the effects of the designed modules in EndoVis-18-VQLA and EndoVis-17-VQLA datasets. Surgical-LVLM is categorized by the inclusion of VP-LoRA and Multimodal Alignment and the results are shown in Table 4. Incorporating VP-LoRA consistently enhances performance across both datasets and metrics. While MA shows improvement, its impact is focused on detection performance. Models combining both VP-LoRA and MA demonstrate the highest performance, indicating synergistic effects. Overall, these findings emphasize the importance of VP-LoRA in the propagation of global contextual information for accurate response generation, with complementary benefits when combined with MA for grounding.

4.4 DISCUSSION AND LIMITATIONS

The quantitative evaluation demonstrates the superiority of our proposed Surgical-LVLM framework over other state-of-the-art models. The ablation analysis on the EndoVis Conversation dataset elucidates the pivotal role of VP-LoRA and Instruction Fine-tuning in augmenting language response performance across datasets and metrics. Notably, comparing Surgical-LVLM with other LVLMs further reinforces the necessity of Instruction Fine-tuning for domain-specific applications. Besides, Surgical-LVLM achieves higher accuracy and mIoU on the EndoVis-18 dataset. This achievement is primarily attributed to the VP-LoRA with the supplementary integration of Multimodal Alignment yielding further enhancements in grounding capabilities.

However, performance degradation across all models on the EndoVis-17 dataset underscores the challenge of maintaining robustness across different domains. While the Surgical-LVLM outperformed on various metrics, its slightly lower F-Score on the EndoVis-18 dataset suggests room for improvement in balancing precision and recall. In addition, further research could explore VP-LoRA's optimal integration and potential limitations, although its introduction enhances response accuracy. Finally, the deployment of LVLMs in real surgical scenarios is still limited by safety concerns necessitating a rigorous evaluation of their reliability in complex surgical environments to mitigate potential risks.

5 CONCLUSIONS

In this paper, we present Surgical-LVLM, a personalized vision-language model specifically designed for efficient grounded visual question answering in robotic surgery. The Surgical-LVLM uses the specifically designed VP-LoRA to enable the model to understand complex visual-language surgical tasks. Additionally, we introduce a novel token interaction module to facilitate the critical language information emphasis. Our experimental results, conducted on multiple benchmarks, manifest the superiority of our proposed approach. However, deploying LVLMs in real-world surgical scenarios remains a long way off due to the need to thoroughly review the reliability and decision-making process of LVLMs in dynamic and unpredictable surgical environments. Therefore, future works shall highlight the need for ethical considerations, and robust safety mechanisms to realize the potential of foundation models in surgical applications.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023a.

Long Bai, Mobarakol Islam, and Hongliang Ren. 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, 2023b.

Long Bai, Mobarakol Islam, and Hongliang Ren. Revisiting distillation for continual learning on visual question localized-answering in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 68–78. Springer, 2023c.

Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6859–6865. IEEE, 2023d.

Long Bai, Guankun Wang, Mobarakol Islam, Lalithkumar Seenivasan, An Wang, and Hongliang Ren. Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. *Information Fusion*, 113:102602, 2025.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2612–2620, 2017.

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8102–8109, 2019.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Mobarakol Islam, Lalithkumar Seenivasan, Lim Chwee Ming, and Hongliang Ren. Learning and reasoning with the graph structure representation in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 627–636. Springer, 2020.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024c.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pp. 445–456. Springer, 2023b.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. pp. 2641–2649, 2015.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio 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*, pp. 658–666, 2019.

Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya Krishna, and Hongliang Ren. 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, 2022a.

Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters*, 2022b.

Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery. *arXiv preprint arXiv:2304.09974*, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Guankun Wang, Han Xiao, Huxin Gao, Renrui Zhang, Long Bai, Xiaoxiao Yang, Zhen Li, Hongsheng Li, and Hongliang Ren. Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. *arXiv preprint arXiv:2410.07540*, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

Kun Yuan, Manasi Kattel, Joel L Lavanchy, Nassir Navab, Vinkle Srivastava, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *arXiv preprint arXiv:2312.10251*, 2023.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023a.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023b.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.