Two-Stage Multimodal Approach for Surgical Video Question Answering

Team name: Medibot

Authors: Xiaoyao Liu, Xingyu Zhao, Xiaoliu Ding

Affiliations: Shanghai Microport MedBot(Group) Co., Ltd

Link to code repo: https://github.com/bravefox12138/surgvu2025vqa

Private dataset used: No

Introduction

Machine learning models that detect and track surgical context from endoscopic videos enable transformative applications. Automatically categorizing surgical progress (phase, step, task, or action) and identifying instruments allows precise assessment of performance, efficiency, and tool choreography. Video question answering (VQA) systems further support natural interfaces, automated feedback, and operating room planning. However, high-quality annotated datasets are challenging to obtain, requiring frame-by-frame labeling and instrument segmentation under diverse conditions, which is resource-intensive and demands expert training. Strategies like using robotic surgery logs and standardized ontologies provide partial solutions. In this challenge, we leverage large-scale multimodal models fine-tuned on surgical videos, adapting them to the domain's unique characteristics. Our approach reduces annotation overhead while achieving accurate and robust surgical visual understanding.

Methodology & Results

1. Dataset

We used only the official challenge dataset for model training and evaluation. No private datasets or external annotations were incorporated. However, the official dataset is inherently multi-dimensional. To enable the model to effectively learn diverse visual–language information, we applied the following preprocessing steps:

1) Video segmentation: We first extracted only the portions of the original videos corresponding to the annotated task intervals and further divided them into 30-second clips.
2) Text formatting: The accompanying textual information was standardized and organized into a structured JSON format, with different dimensions of information explicitly represented. In addition to the fixed fields of tool names, task name, task description, and matched description, we further enriched the dataset with two additional dimensions: tool functions and operated organs and tissues. These fields were automatically generated from the matched description using the Qwen3-8B[1] model.
3) Video–text pairing: Each video clip was paired with its corresponding JSON entry, along with a prompt.

The overall data format is illustrated below:

{ "video": path_to_the video,
"conversations": [
{"from": "human","value": "<video>\nAnalyze the surgical procedure in the video and provide a structured JSON output with keys: used_tools_and_function, operated_organ_and_tissue, task_name, task_description, matched_description"},

{"from": qwen", "value": "{\"used_tools_and_function\": {\"force bipolar\": \"Coagulation and tissue sealing\", \"permanent cautery hook\": \"Cutting and cauterizing tissue\", \"prograsp forceps\": \"Grasping and manipulating tissue during dissection\"}, \"operated_organ_and_tissue\": [\"gallbladder\", \"gallbladder tissue\", \"extraphepatic ducts\"], \"task_name\": \"skills application\", \"task_description\": \"These are large-scale steps focused on combining the taught skills in a clinically applicable scenario. These actions are relatively unstructured and variable, requiring adaptability and integration of multiple techniques.\", \"matched_description\": \"The surgeon will remove the gallbladder. If ICG was administered during the optional fluorescence imaging demo, then firefly can be used to visualize the extraphepatic ducts.\"}"}]}

In total, the preprocessing resulted in 162,000 video–question–answer pairs. of which 158,892 were used for training. The remaining 1,554 and 1,554 samples were allocated for validation and testing, respectively.

2. Method

Our objective was to design a multimodal model for surgical visual question answering. Given the hardware constraint of a 16 GB GPU memory, we selected Qwen2.5-VL-3B[2] as the base multimodal model, owing to its strong performance on a wide range of vision–language benchmarks. To adapt it to the surgical domain, we finetuned the model on our constructed dataset using two strategies: full-parameter fine-tuning and LoRA[3] (Low-Rank Adaptation).

To further improve both the quality and consistency of answers, we adopted a two-stage answering strategy:

Stage 1 (Scene Description): The model is first prompted to generate a structured description of the overall surgical scene, including all information in the JSON. Crucially, instrument information is not inferred solely by the vision–language model. Due to significant noise in the training dataset, particularly in tool annotations, the model is prone to hallucinations when identifying instruments. To mitigate this, we incorporated a dedicated instrument detection model, developed as part of our team's submission to Category 1: Surgical Tool Classification and Localization, which serves as the authoritative source for instrument presence. The resulting merged description thus combines the contextual understanding from the vision–language model with accurate and reliable tool identification from the detection model.

Stage 2 (Question-Specific Answering): The scene description generated in Stage 1, together with the user's question, is passed back into the model. At this stage, the model produces a concise and question-focused response. The style and format of the answer are further controlled through carefully designed prompt templates.

This two-stage pipeline enables the model to first establish a holistic representation of the surgical scene, and then refine its reasoning to deliver accurate, coherent, and stylistically consistent answers.

3. Result

We compared LoRA and full-parameter fine-tuning using three evaluation metrics: task recognition accuracy, as well as precision and recall for tools recognition, the results was listed in Table 1. Since the test set is large and contained considerable noise, we manually curated a subset of 25 samples to ensure data quality.

Table1.

Metric	Lora	Full para
task recognition accuracy	0.96	0.72
tools detect precision	0.69	0.63

tools detect recall	0.91	0.77

Table2 indicates whether the detection model was incorporated, allowing us to assess its impact on overall model performance.

Table2

Metric	Qwen2.5 VL	Qwen2.5 VL + detection
task recognition accuracy	0.96	0.96
tools detect precision	0.69	0.95
tools detect recall	0.91	0.87

Additionally, based on the 11 official test samples, we created 34 additional similar examples to evaluate Stage 2. By carefully adjusting the prompts, our local evaluation on a total of 45 test samples achieved a BLEU score of 0.73.

Conclusion &Discussion

Our submission explored fine-tuning strategies and multimodal model augmentation for surgical video question answering. Key takeaways include:

1. LoRA adaptation yielded better accuracy and stability. This suggests that smaller, efficient adaptation methods are preferable when annotated data is limited.

2. Adding a dedicated detection module improved tool recognition precision. While vision-language models are versatile, classical detection networks still excel in specialized recognition tasks.

3. Current training data contain considerable noise, such as mismatches between instrument names and video content. Future work could focus on more thorough data cleaning and leverage larger models to further improve performance.

4. The BLEU metric alone may not fully reflect model performance. Future evaluations should explore alternative assessment approaches, such as multiple-choice or open-ended question formats, to better capture the quality and correctness of model responses.

References

(1) Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; Qiu, Z. Qwen3 Technical Report. arXiv May 14, 2025. https://doi.org/10.48550/arXiv.2505.09388.

(2) Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; Lin, J. Qwen2.5-VL Technical Report. arXiv February 19, 2025. https://doi.org/10.48550/arXiv.2502.13923.
(3) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv October 16, 2021. https://doi.org/10.48550/arXiv.2106.09685.