Surgical-LLaVA: Toward Surgical Scenario Understanding via Large Language and Vision Models

Anonymous ACL submission

Abstract

Conversation agents powered by large language models are revolutionizing the way we interact with visual data. Recently, large vision-language models (LVLMs) have been extensively studied for both images and videos. However, these studies typically focus on common scenarios. In this work, we introduce an LVLM specifically designed for surgical scenarios. We integrate visual representations of surgical images and videos into the language feature space. Consequently, we establish a LVLM model, Surgical-LLaVA, fine-tuned on instruction following data of surgical scenarios. Our experiments demonstrate that Surgical-LLaVA exhibits impressive multi-modal chat abilities in surgical contexts, occasionally displaying multi-modal behaviors on unseen instructions. We conduct a quantitative evaluation of visual question-answering datasets for surgical scenarios. The results show superior performance compared to previous works, indicating the potential of our model to tackle more complex surgery scenarios.

1 Introduction

The rapid advancements in AI have increasingly focused on developing versatile assistants that can effectively understand and interact with the world through multiple sensory modalities, such as vision (Li et al., 2022) and language (Brown et al., 2020). This multi-modal approach harnesses the unique strengths of each channel, enhancing the AI's ability to perform a wide range of real-world tasks more accurately and efficiently (Askell et al., 2021; Li et al., 2024a). Despite significant progress with large language models (LLMs) like GPT-3 (Liu et al., 2021), GPT-4 (Achiam et al., 2023), and open-source alternatives such as LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), these models typically handle language tasks in isolation, limiting their potential in applications that require a comprehensive understanding of multimodal data. Recent efforts have attempted to bridge this gap by integrating visual comprehension within a single model, aiming to create a unified representation that captures both visual and linguistic information. For example, models such as LLaVA (Liu et al., 2024) and Video-LLaMA ((Zhang et al., 2023)) utilize shared visual encoders to process images and videos.

In the surgical applications, the ability to understand and process both images and videos is of paramount importance (Saab et al., 2024; Li et al., 2024b). Surgical procedures generate a wealth of visual data, including static images and dynamic videos. While general-domain vision-language models have been successful, they are less effective in surgical contexts because surgical visual-text pairs differ significantly from typical web content. This discrepancy can cause general-domain visual assistants to act like laypersons, either avoiding surgical questions or providing incorrect or completely fabricated responses. Despite significant advances in surgery visual question answering (VQA), prior methods often treat the problem as a classification task (e.g., choosing among specific answers from the training set) (Kirtac et al., 2022; Valderrama et al., 2022). As a result, conversational generative AI for surgical applications is often restricted to specific tasks.

In this paper, we present Surgical-LLaVA, a first attempt to extent multimodal instruction-tuning to the surgical domain for multimodal conversational assistant. Inspired by recent work in instruction-tuning, Surgical LLaVA uses GPT-3.5 to generate diverse surgical multimodal instruction-following data using image/video-pairs, and fine-tune a surgical domain vision-langauge model using LoRA method. Specifically, our paper contributed follows as:

- We propose Surgical-LLaVA, a multimodal model capable of engaging in meaningful conversations about surgical scenarios. It combines the language understanding capabilities of LLMs with a pretrained visual encoder tailored for spatiotemporal representations of surgical procedures.

- We present datasets consisting of high-quality surgical visual instruction pairs, generated through a scalable and diverse annotation framework specifically designed for the surgical scenarios.

- We achieved superior performance compared to existing instruction-following agents in video reasoning for surgery scenario and visual question-answering.

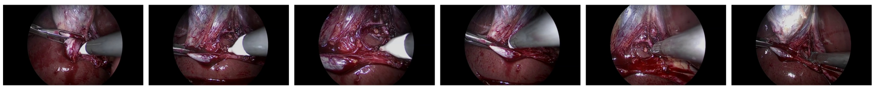
2 Related Work

Large Langauge Models The emergence of large-scale language models (LLMs) such as GPT , LLaMA and OPT (Zhang et al., 2022) has led to a paradigm shift in the field of natural language processing. These models excel in language generation and in-context learning, and demonstrate the ability to understand complex tasks. The high adaptability and generalisability of LLMs has led researchers to fine-tune these models for optimal performance.

One of the key strategies in such research is instructional tuning. This approach focuses on improving the model's alignment with user intent and optimising the quality of its output. For example, InstructGPT (Ouyang et al., 2022) and ChatGPT use this technique to improve their ability to interact with a variety of dialogues and answer complex questions. This effective approach has recently been applied to open source models such as Alpaca (Peng et al., 2023) and Vicuna, resulting in performance improvements.

Leveraging LLMs for Multimodal Understanding The recent advancements in multimodal understanding have been primarily driven by the integration of image-based vision models with large language models (LLMs). Pioneering contributions, such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023), have demonstrated the power of leveraging web-scale image-text data and cross-modal alignment techniques to exhibit impressive capabilities in conversational and few-shot learning settings. Equally noteworthy is the emergence of Large Language and Vision Assistant (LLaVA) (Liu et al., 2024), a model derived from the LLaMa architecture, which capitalizes on GPT-4's language proficiency to generate multimodal instruction-following data. Through instruction tuning on the derived data, LLaVA has showcased promising multimodal chat capabilities, hinting at the scalability potential of such an approach. Furthermore, the InstructBLIP (Dai et al., 2024) model has demonstrated strong image-based dialogue capabilities through vision-language instruction tuning and innovative instruction-aware visual feature extraction. Inspired by these success, several medical vision-language model have been studied (Shu et al., 2023; Yunxiang et al., 2023; Wu et al., 2023). LLaVA-Med (Liu et al., 2024) fine-tuned from biomedical data to instruction-following data and achieved superior performance on a variety of prompts.

Surgical Scenario Visual Question Answering Early surgery video datasets primarily consisted of images and their corresponding annotations, focusing on tasks such as instrument detection, segmentation, and procedural step recognition. The Cholec80 dataset (Twinanda et al., 2016) and the EndoVis18 dataset (Allan et al., 2020) were pioneering efforts in this domain, providing annotated laparoscopic videos and surgical scenes for instrument recognition and segmentation, respectively. However, the creation and annotation processes for these datasets were labor-intensive and time-consuming, limiting their scalability and diversity. To address these limitations, researchers shifted their focus towards leveraging the abundance of visual-text resources available in the medical domain. (Seenivasan et al., 2022) and (Seenivasan et al., 2023) pioneered the integration of visual and textual information by constructing datasets tailored for visual-question answering tasks in surgical settings. These datasets aim to capture the rich multimodal information present during surgical procedures, enabling the development of models capable of simultaneously understanding and reasoning about complex visual and textual cues, thereby opening new avenues for research and allowing the exploration of novel tasks and applications that leverage the synergy between visual and textual information. Surgical-LLaVA aimed to develop an effective vision-language assistant for various complex prompts by generating multimodal instruction-following data for surgical scenarios by utilizing the language capabilities of LLMs such as GPT.



Caption:

The current phase is CalotTriangleDissection, and the tools used are: grasper, clipper and irrigator.

Instruction-following Annotation:

User: How does the grasper assist in the Calot's Triangle dissection phase?

Assistant: The grasper is a surgical tool that is used to securely hold and manipulate tissues within the Calot's Triangle, allowing for precise dissection and identification of anatomical structures.

User: In what way does the irrigator contribute to the success of the Calot's Triangle dissection phase?

Assistant: The irrigator helps to maintain a clear surgical field by flushing away blood and debris, allowing for better visualization and accurate dissection of critical structures within the Calot's Triangle.

Figure 1: An example to illustrate the instruction-following data. We utilized the original caption to create an annotation that follows instructions with various prompts. The video and caption were acquired from Cholec80 dataset (Hong et al., 2020). The instruction-following data generated by GPT-3.5 using the text only (captions).

3 Surgical Visual Instruction Data Generation

This section describes a data-driven approach for multimodal directive follow-up data collection using LLMs using a novel framework specifically tailored to the surgical scenarios. Inspired by the recent success of visual language models in text annotation tasks, our approach is based on widely available image pair data, but with one important difference: medical data requires a specific and specialized context, so creating instructions using LLMs trained only on general data may result in the loss of important medical information. To address these issues, we adopted the LLaVA approach (Peng et al., 2023) for data generation and incorporated annotation information as input to facilitate the generation of instructional data tailored to the surgical scenario. Specifically, our framework is the basis for generating a variety of contextualized instructions using expert-annotated surgical image data.

Recognizing the lack of comprehensive information in the original annotations, we attempted to leverage LLM's medical and background knowledge, such as GPT-3.5. We leveraged the original annotations to create instruction-following annotations with various prompts and instructions, as shown in Figure 1. By leveraging LLM's powerful language understanding and generation capabilities, it plays a key role in expanding the original annotations and incorporating relevant medical knowledge, procedural details, and contextual cues to create comprehensive and informative guideline-following annotations. To achieve this, we create a test set based on the ActivityNet-200 dataset (Caba Heilbron et al., 2015), which contains videos accompanied by detailed descriptive captions and human-annotated question-answer pairs. Moreover, we construct an evaluation pipeline utilizing the GPT-3.5 model. This approach not only allows us to generate high-quality, multimodal guidance data specific to the surgical scenarios, but also effectively utilizes existing annotation resources.

4 Surgical-LLaVA

Surgical-LLaVA is a vision-language model that enhances surgical scenario analysis and conversation capabilities by aligning visual representations with a LLM. To achieve this, we leverage existing approaches used in the development of vision-language (VL) models for visual tasks. Given the scarcity of visual-caption pairs and the significant resources required for training from scratch, our strategy involves adapting pretrained image-based VL models for visual applications, as seen in previous works (Rasheed et al., 2023; Ni et al., 2022). We specifically build upon the LLaVA, an Large Multimodal Model (LMM) that combines the visual encoder of CLIP (Radford et al., 2021) with the Vicuna language decoder (Chiang et al., 2023), and is fine-tuned end-to-end on generated instructional vision-language data. We further fine-tune LLaVA with our visual-instruction data to tailor it for conversation tasks.

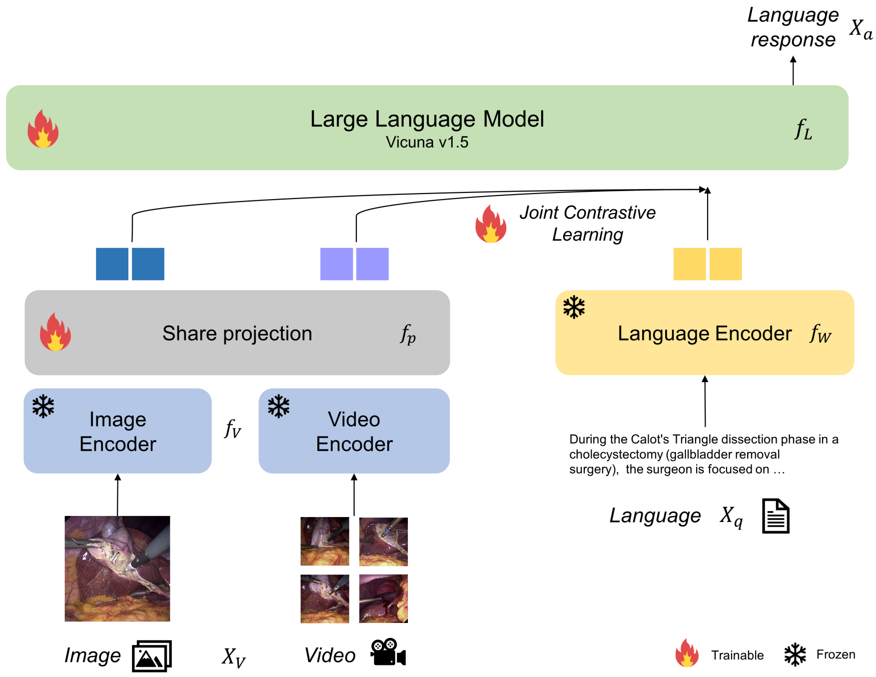


Figure 2: Architecture of Surgical-LLaVA. We adopted llava as the baseline, which vicuna as the LLM model and the pre-trained CLIP visual encoder ViT-L/14 as the visual model. The training involves encoding these inputs into token representations, followed by joint contrastive learning to align modalities within the semantic space. LoRA fine-tuning is applied to enhance the model's efficiency and performance.

4.1 Architecture

The primary goal is to effectively apply the capabilities of the pre-trained LLM and visual model to surgical scenarios. The architecture is illustrated in Figure 2. We adopted LLaVA as the baseline, which vicuna as the LLM model and the pre-trained CLIP visual encoder ViT-L/14 as the visual model. Our visual encoder, originally designed for image processing, is extended to handle video inputs. Given a video sample $V_i \in \mathbb{R}^{T\times H\times W\times C}$ with T frames, the encoder generates both temporal and spatial features. To derive video-level features, we perform average pooling on the frame-level embeddings along the temporal dimension, resulting in video-level temporal representations $t_i \in \mathbb{R}^{N\times D}$. Similarly, average pooling along the spatial dimension produces video-level spatial representations $z_i \in \mathbb{R}^{T\times D}$. By concatenating the temporal and spatial features, we obtain comprehensive video-level features.

4.2 Visual Understanding Training

The overall training process for Surgical-LLaVA follows a similar approach to LLM models like GPT. The model takes as input a text seqeunce X_T and visual data X_V (image or videos). These inputs are encoded into a token representation according to Eq 1. The training objective is to maximize the likelihood probability in Eq 2.

$$\mathbf{Z}_T = f_T(\mathbf{X}_T), \quad \mathbf{Z}_V = f_P(f_V(\mathbf{X}_V)) \tag{1}$$

$$p(\mathbf{X}_A \mid \mathbf{X}_V, \mathbf{X}_T) = \prod_{i=1}^{L} p_\theta\left(\mathbf{X}_A^{[i]} \mid \mathbf{Z}_V, \mathbf{Z}_T^{[1:i-1]}\right) \tag{2}$$

where L represents the length of the generated sequence, and θ denotes the trainable model parameters. This phase focuses on enabling the model to interpret visual representation from an extensive dataset comprising image/video-text pairs. Each visual sample corresponds to a single round of conversation data (X_q, X_a), where $X_T = X_q$ and X_a serves as the ground truth.

Joint Contrastive Learning In our approach, we employ a dynamic joint training that includes both image and video samples within each batch. We employ a transformer model for our language encoder. The language encoder transforms these tokens into a text logit $y \in \mathbb{R}^{L\times\mathcal{C}}$, where L is the length of the sequence. To align different modalities, we leverage contrastive learning techniques (Chen et al., 2020). This approach aims to increase the similarity between paired data, bringing them into closer proximity within the semantic space, while decreasing the similarity between unpaired data. By using contrastive learning, we can associate each modality with the language component.

$$L_{\mathrm{M2T}} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{K}\exp(x_i^\top y_j/\tau)} \quad (3)$$

In this context, x_i refers to the i-th modality data (image and video) and y_j to the j-th text, with both their features being normalized. K stands for the batch size, and τ is the temperature parameter. By aligning each modality M directly with language T, we achieve significant improvements. This ensures a stronger alignment than a one-way alignment.

4.3 Visual Instruction Tuning

We employ instruction-tuning of the LLM on the prediction tokens, utilizing its original auto-regressive training objective. The pretrained model is finetuned with curated, high-quality visual-text pairs. During the fine-tuning phase, we use pre-defined prompts based on the following template:

```
USER:               <Instruction>
<Visual-tokens> Assistant:
```

In this framework, the <Instruction> signifies a query related to the visual content, randomly selected from a dataset of visual-question-answer pairs. The predicted <Answer> corresponds specifically to the query posed. During training, the weights for both the visual encoder and the language model remain fixed, and the model aims to maximize the likelihood of predicting the tokens that form the answer by adjusting the linear layer.

LoRA fine-tuning We apply the LoRA (Hu et al., 2021) technique to expedite the fine-tuning process. For an encoder with a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, we keep the weight matrix W_0 fixed while learning an additional weight matrix BA. Specifically, for a modality-agnostic encoder $h(\cdot)$ and input x, the forward pass is defined as follows:

$$h(x) = W_0 x + BAx$$

Here, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, where r is the minimum of d and k. It is crucial to note that both W_0 and BA share the same input and output dimensions, allowing their outputs to be summed to produce the final result.

5 Experiments

Implementation Details We use LLaVA as our baseline model. We finetune the model for 3 epochs using a learning rate of 1e-5 and overall batch size of 16. The training of our 7B model took around 16 hours on 4 RTX3090 24GB GPUs. During inference, for memory efficiency, we load the model in FP16 mode. The data in each batch is random combination of images and videos.

Data Description We utilized three datasets as visual datasets for our surgical scenario.

- Cholec80-VQA (Twinanda et al., 2016) contains Q&A pairs for 80 video sequences of the Cholec80 dataset. The videos are configured at 25 frames per second (fps), while the annotations are provided at 1 fps. To align with the annotation frame rate, we extracted frames from the videos at 1 fps.

- EndoVis-18-VQA (Allan et al., 2020) consist of Q&A pairs for 18 robotic nephrectomy procedure video sequences from the MICCAI Endoscopic Vision Challenge 2018 dataset. For this dataset, we utilized 2,600 images and leveraged multiple annotations per single image.

- PSI-AVA-VQA (Valderrama et al., 2022) consists of 10291 Q&A pairs with 35 answer classes of holistic surgical scenario. They are constructed based on the surgical phase, step and location annotation provided in the PSI-AVA dataset.

5.1 Surgical Video Understanding

To evaluate the performance of Surgical-LLaVA on surgical scenario conversation, we present a benchmark designed to assess the text generation capabilities of visual models. The evaluation pipeline for video understanding follows Video-ChatGPT (Maaz et al., 2023). This pipeline evaluates the model's performance and assigns relative scores to the generated responses on a scale of 1-5, in the following three dimensions:

(i) *Conversation:* We assesses the accuracy and relevance of the model's responses during the visual dialogue, ensuring it accurately reflects the video content without any misinterpretations or false information.

Table 1: Comparison between different large visual language models on video reasoning benchmarks.

Methods	LLM size	Conversation		Detail description		Complex reasoning	
		Accuracy	Score	Accuracy	Score	Accuracy	Score
Video-ChatGPT (Maaz et al., 2023)	7B	42.7	3.1	38.0	2.6	39.8	2.5
Video-LLaVA (Lin et al., 2023)	7B	51.4	3.3	40.3	2.7	41.1	2.6
Surgical-LLAVA	7B	**58.3**	**3.9**	**47.1**	**3.2**	**46.5**	**3.1**

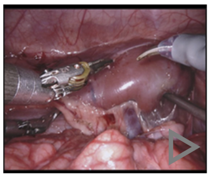
Visual input example, Surgical video	
	
User	This video is about kidney cancer surgery, describe about this video.
Surgical-LLaVA	In this image, the surgeon is using a bipolar forceps located at the left-bottom to grasp and manipulate tissues during the kidney cancer surgery. The prograsp forceps, located at the right-bottom, are idle, indicating that they are not currently being used. The monopolar curved scissors, located at the right top, are also idle, suggesting that no cutting or cauterization is currently taking place. The overall setup indicates a careful and precise approach to the surgical procedure, with the tools strategically positioned for efficient and safe manipulation of the kidney during the operation.
User	What is purpose of using a prograsp forceps?
Surgical-LLaVA	The prograsp forceps are used to securely hold and manipulate tissues or organs during the surgical procedure, allowing for precise movements and control.
User	This video is about kidney cancer surgery, describe about this video.
Video-LLaVA	This is an video of a surgical procedure for kidney cancer. The surgeon is using a robotic arm to remove the tumor from the patient's abdomen. The patient is under anesthesia and the surgeon is using a surgical instrument to remove the tumor. The surgery is being performed in a hospital operating room.

Figure 3: Example comparison of surgical visual chat and reasoning capabilities. Compared to Video-LLaVA (Lin et al., 2023), Surgical-LLaVA offers specific and accurate answers to surgical scenarios.

(ii) *Detail description:* We evaluate the thoroughness of the model's responses, checking for completeness by ensuring all major points from the video are covered, and for specificity by including precise details rather than generic statements.

(iii) *Complex reasoning:* We assess the model's ability to engage in complex reasoning, ensuring its responses demonstrate an understanding of the video's context and logical connections between the content points.

Among the models evaluated, Surgical-LLaVA stands out with the highest scores across all three dimensions as shown in Table 1. The Surgical-LLaVA model not only demonstrates superior conversation and detailed descriptions but also excels in complex reasoning, particularly in understanding and articulating intricate surgical scenarios . This ability to grasp and reason through complex medical content is critical, showcasing its potential for applications in surgical environments where accurate and nuanced interpretation of video content is paramount. In Figure 3, we illustrate example of surgical visual conversations using different representative chatbot on image. Surgical-LLaVA responds to questions accurately, leveraging medical knowledge, whereas Vidoe-LLaVA (Lin et al., 2023) responds more like a layperson, often producing commonsense-based hallucinations.

5.2 Evaluation on Visual Question-Answering Benchmarks

In this evaluation, we assess the performance of various models on visual question-answering (VQA) tasks, particularly focusing on the Cholec80-VQA, EndoVis18-VQA, and PSI-AVA-VQA datasets. Table 2 provides a comparative analysis of different models based on their performance metrics. Table 2 shows that Surgical-LLaVA significantly outperforms existing models, achieving the highest accuracy rates in all three datasets. The ability to maintain high accuracy across different datasets highlights the versatility and reliability of Surgical-LLaVA in handling various types of visual and contextual information in surgical videos. Its con-

Table 2: Comparison of various models on visual question-answering.

Methods	Cholec80-VQA	EndoVis18-VQA	PSI-AVA-VQA
VisualBert (Li et al., 2019)	89.7	61.4	58.5
Block (Peng et al., 2020)	89.5	60.1	59.9
MFH (Yu et al., 2018)	87.5	58.8	47.8
Surgical-VQA (Seenivasan et al., 2022)	89.8	63.2	65.6
Surgical-LLaVA	**92.2**	**68.7**	**67.1**

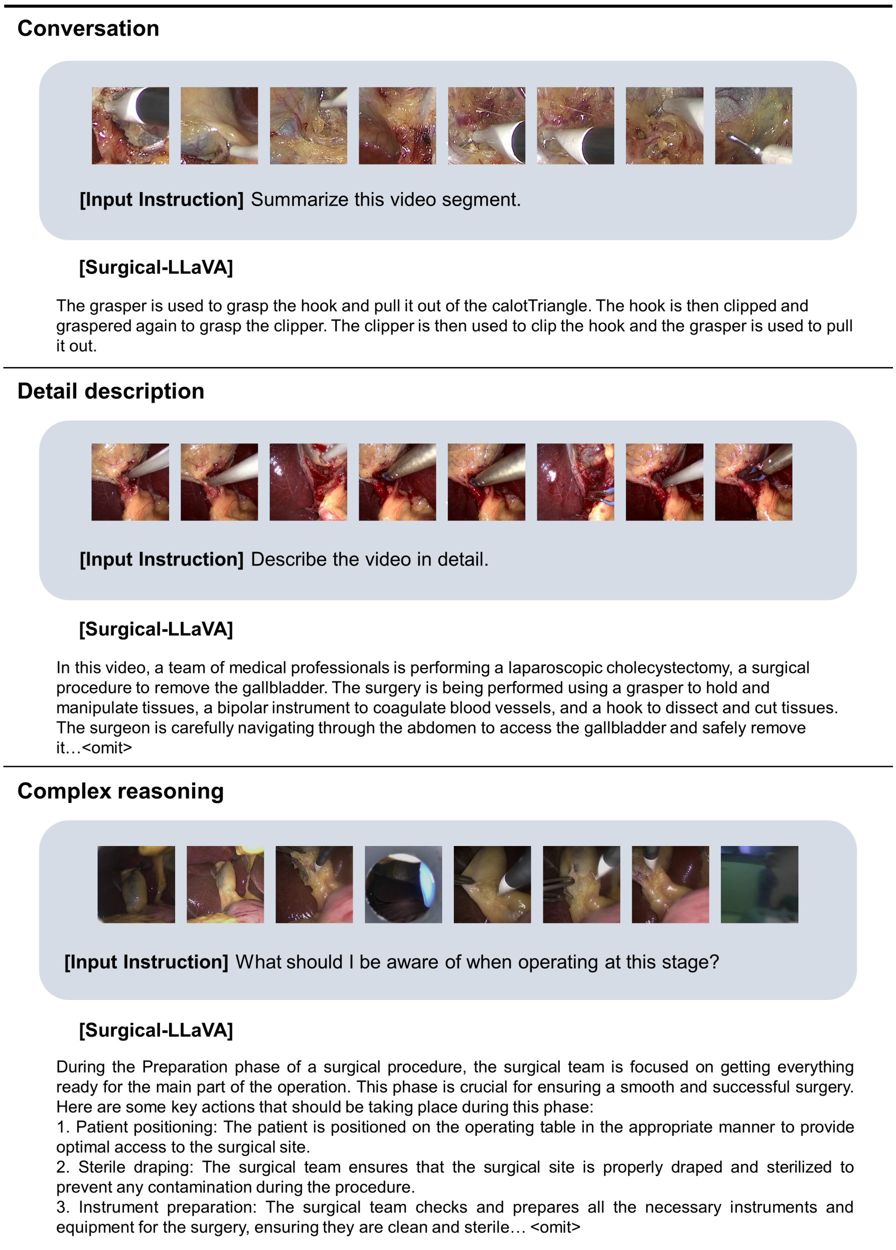


Figure 4: Examples from Surgical-LLaVA's demonstration of video reasoning. It shows conversation, detail description and complex reasoning cases.

sistent performance across multiple benchmarks signifies a major leap forward in interaction with visual surgery data.

5.3 Qualitative Evaluation

To comprehensively assess the capabilities of our proposed Surgical-LLaVA model, we conducted an extensive qualitative evaluation spanning a diverse array of open-ended video question-answering tasks.

Conversation We confirmed whether the model accurately reflects the content of the surgical videos without introducing any hallucinations or misinterpretations. This involves verifying that the generated text stays true to the visual information presented and is contextually appropriate as illustrated in top of Figure 4.

Detail Description We evaluated the model's capacity to generate detailed and descriptions of the surgical scenes. Surgical-LLaVAs describe the tools, steps, and even a description of the surrounding tissues in a surgery as illustrated in middle of Figure 4.

Complex Reasoning These tasks focused on the model's capability to perform complex reasoning based on the visual information and contextual knowledge, . Surgical-LLaVA identified the current phase from the visual data and effectively suggest things to watch out for at that stage, as exemplified in bottom of Figure 4.

Throughout the evaluation, our Surgical-LLaVA model demonstrated remarkable proficiency in comprehending the visual content of the surgical videos and generating accurate, informative, and contextually relevant responses across the various tasks. The model effectively leveraged the visual information present in the videos to provide precise answers, detailed descriptions, and reasoned insights, showcasing its capability in understanding and reasoning about complex surgical procedures.

5.4 Ablation Study

We conducted an ablation study on joint contrastive learning. As shown in Table 3, we compared the performance of Surgical-LLaVA* without image training. The model trained with both images and videos shows significant improvements across all metrics. These findings indicate that combining image and video training enhances the LLM's ability to comprehend visual representations in surgical scenarios.

Table 3: Effect of joint training. We evaluate on three visual question-answering datasets. * denotes that we utilized only video data in both the first and second stages.

Methods	Conversatoin	Detail description	Complex reasoning
Surgicla-LLaVA*	57.5	44.5	42.0
Joint with image	58.3	47.1	46.5
Δ Acc.	+0.8%	+2.6%	3.5%

6 Conclusion

In this work, we introduced Surgical-LLaVA, a multimodal model designed for engaging in meaningful conversations and reasoning about surgical scos. By integrating the language understanding capabilities of LLMs with pretrained visual encoders tailored for spatiotemporal representations of surgical procedures, Surgical-LLaVA exhibits impressive multi-modal chat abilities in surgical contexts. A contribution of our work is the introduction of a novel dataset consisting of high-quality surgical visual instruction pairs, generated through a scalable and diverse annotation framework specifically designed for the medical domain. Through quantitative and qualitative evaluations, we demonstrated Surgical-LLaVA's superior performance compared to existing state-of-the-art models in various tasks, including visual question-answering, video reasoning about surgical scenarios.

Limitations

The success of Surgical-LLaVA underscores the potential of combining large language models with specialized visual encoders for domain-specific applications. However, current public surgical datasets have limitations in providing limited information such as phase, tool, etc. The ability to include specific and diverse information in surgical datasets will greatly improve scalability. In addition, the study should actually be reviewed by clinicians for its utility. This work is anticipated to provide valuable insights into multi-modal approaches for surgical scenarios within the LLM framework, paving the way for advancements in AI-assisted surgical training, decision-making processes, and patient care.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol

Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2020. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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

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

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. 2020. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*.

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

Kadir Kirtac, Nizamettin Aydin, Joël L Lavanchy, Guido Beldi, Marco Smit, Michael S Woods, and Florian Aspart. 2022. Surgical phase recognition: From public datasets to real-world data. *Applied Sciences*, 12(17):8746.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301.

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

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

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

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*, 3.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. 2023. Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 281–290. Springer.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. 2022. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43. Springer.

Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. 2023. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities.

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

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97.

Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. 2022. Towards holistic surgical scene understanding. In *International conference on medical image computing and computer-assisted intervention*, pages 442–452. Springer.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

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

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.