EndoChat: Grounded multimodal large language model for endoscopic surgery

Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al.

Medical Image Analysis 107 (2026) 103789

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

EndoChat: Grounded multimodal large language model for endoscopic surgery

Guankun Wang [a,b,1], Long Bai [a,c,1], Junyi Wang [a,1], Kun Yuan [c,d,1], Zhen Li [e], Tianxu Jiang [a], Xiting He [a], Jinlin Wu [f], Zhen Chen [f], Zhen Lei [f], Hongbin Liu [f], Jiazheng Wang [b], Fan Zhang [b], Nicolas Padoy [d], Nassir Navab [c], Hongliang Ren [a,*]

[a] The Chinese University of Hong Kong, 999077, Hong Kong Special Administrative Region of China
[b] Theory Lab, Central Research Institute, 2012 Labs, Huawei Technologies Co. Ltd., 999077, Hong Kong Special Administrative Region of China
[c] Chair of Computer Aided Procedures (CAMP), Technical University of Munich, Munich, 81927, Germany
[d] University of Strasbourg, CNRS, INSERM, ICube & IHU Strasbourg, Strasbourg, 67200, France
[e] Department of Gastroenterology, Qilu Hospital of Shandong University, Jinan, 250000, China
[f] Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, 999077, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Multimodal large language model
Endoscopic surgery
Surgical scene understanding
Dialogue paradigm

ABSTRACT

Recently, Multimodal Large Language Models (MLLMs) have demonstrated their immense potential in computer-aided diagnosis and decision-making. In the context of robotic-assisted surgery, MLLMs can serve as effective tools for surgical training and guidance. However, there is still a deficiency of MLLMs specialized for surgical scene understanding in endoscopic procedures. To this end, we present EndoChat, an MLLM tailored to address various dialogue paradigms and subtasks in understanding endoscopic procedures. To train our EndoChat, we construct the Surg-396K dataset through a novel pipeline that systematically extracts surgical information and generates structured annotations based on large-scale endoscopic surgery datasets. Furthermore, we introduce a multi-scale visual token interaction mechanism and a visual contrast-based reasoning mechanism to enhance the model's representation learning and reasoning capabilities. Our model achieves state-of-the-art performance across five dialogue paradigms and seven surgical scene understanding tasks. Additionally, we conduct evaluations with professional surgeons, who provide positive feedback on the majority of conversation cases generated by EndoChat. Overall, these results demonstrate that EndoChat has the potential to advance training and automation in robotic-assisted surgery. Our dataset and model are publicly available at https://github.com/gkw0010/EndoChat.

1. Introduction

Robot-assisted surgery (RAS) offers unprecedented opportunities to enhance surgical precision, minimize patient trauma, and shorten post-operative recovery times (Nwoye et al., 2023). However, the effective application of this technology places significant demands on the skills of surgeons, particularly in mastering the operation of robotic systems during procedures (Alabi et al., 2025; Wagner et al., 2023). To ensure surgical safety and efficacy, surgeons must undergo rigorous training to acquire the core skills required for robotic operation (Chen et al., 2020; Aziz et al., 2021). To improve the efficiency of this training process, various simulator-based surgical platforms (Mariani et al., 2020) have been developed. However, when trainees encounter challenges during training, they often require immediate feedback and guidance from professional surgeons to resolve doubts or correct mistakes. Unfortunately, professional surgeons typically face significant time constraints due to their heavy workload in clinical, teaching, and research responsibilities, making it difficult for them to provide consistent, real-time support during training (Sharma et al., 2021; Seenivasan et al., 2022). As a result, there is a pressing need for technological solutions that can deliver flexible, real-time, and efficient support in surgical training.

Recently, artificial intelligence (AI)-based dialogue systems with structured Visual Question Answering (VQA) have been introduced into surgical training (Seenivasan et al., 2022; Bai et al., 2023b). These systems analyze visual data from surgical scenarios to address trainees' questions. However, their reliance on simple structured, and object

* Corresponding author.
E-mail address: hlren@ee.cuhk.edu.hk (H. Ren).
[1] Co-first authors.

https://doi.org/10.1016/j.media.2025.103789
Received 10 March 2025; Received in revised form 20 August 2025; Accepted 25 August 2025
Available online 31 August 2025

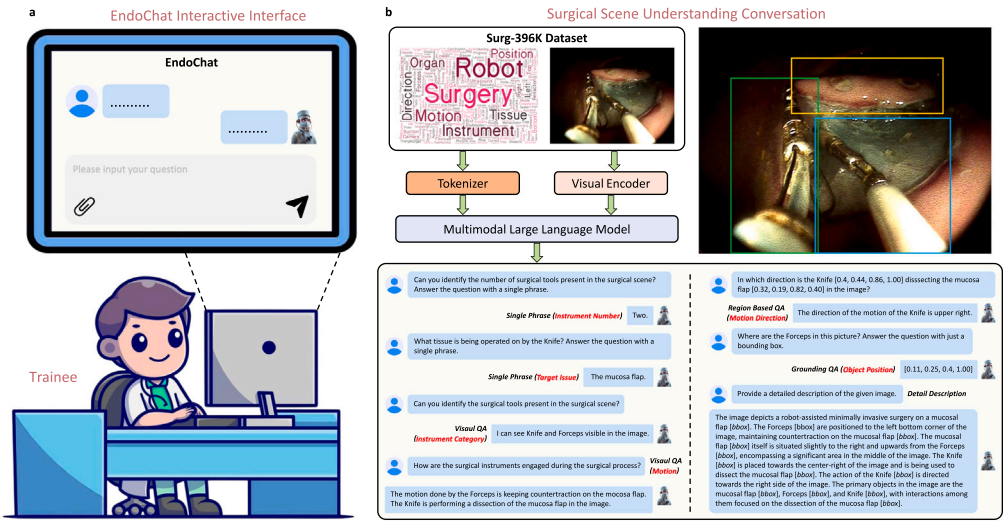


Fig. 1. Overview of the EndoChat. **a** EndoChat is an interactive multimodal large language model designed for surgical education and training. Users can interact with EndoChat by uploading images and formulating questions, enabling a comprehensive surgical scene understanding. **b** EndoChat is trained on Surg-396K, a large-scale multimodal instruction dataset. Surg-396K includes five conversation paradigms, enabling EndoChat to effectively perform natural language and visual grounding conversations with trainees. On the bottom is an example of the multi-turn conversation.

presence-based VQA datasets, typically trained for specific tasks, limits their ability to dynamically adapt to the wide range of questions posed by trainees (Lin et al., 2023b). Moreover, most of these approaches are based on encoder–decoder architectures, which require explicitly defined input/output formats. This rigidity makes them less flexible for handling highly open-ended generative tasks and limits their scalability. When applied beyond their designed scope, these models often show significant performance drops, making them poorly suited for the complexity of diverse surgical scenarios (Gozalo-Brizuela and Garrido-Merchan, 2023). When trainees ask open-ended questions, existing VQA systems lack the flexibility and contextual understanding needed for such interactions. As a result, they struggle to handle open-ended questions or complex multi-turn dialogues, significantly limiting their usefulness in surgical training.

Currently, medical Multimodal Large Language Models (MLLMs) are emerging as a promising solution, offering significant potential through large-scale pretraining to perform complex reasoning and understanding across tasks (Li et al., 2024b; Chen et al., 2024b; Ye et al., 2024; Liu et al., 2023). Specifically, MLLMs can extract information from multimodal data in surgical scenarios and perform advanced reasoning. Unlike structured question-answering systems, MLLMs are capable of processing unstructured and complex contextual information. For example, trainees can ask open-ended questions in natural language, and MLLMs can generate targeted responses by using their pre-trained knowledge and multimodal reasoning capabilities. These powerful natural language processing abilities enable MLLMs to handle multi-turn conversations and dynamically adjust responses based on context. This interaction is akin to receiving guidance from professional surgeons, significantly enhancing the training experience. Overall, MLLMs have the potential to partially replace guidance from professional surgeons by simulating their knowledge and decision-making abilities. This can address the limitations of current training solutions and alleviate the burden on surgeons who are constrained by demanding clinical schedules, ultimately improving the efficiency and quality of surgical training (Chen et al., 2024b; Dou et al., 2023).

Surgical MLLMs have attracted significant attention from researchers (Li et al., 2024a; Jin and Jeong, 2024; Wang et al., 2025). However, existing surgical MLLMs primarily focus on one of three approaches: utilizing the powerful capabilities of MLLMs to enhance question-answering performance (Hou et al., 2024), extending models for specific predefined tasks through instruction tuning (Wang et al., 2025), or using web-sourced data to construct general-purpose descriptive and conversational datasets (Li et al., 2024a). These methods face two major challenges: (i) In real-world scenarios, the queries from trainees are highly diverse. Solely relying on predefined query formats or generic captioning is insufficient to handle the variety of queries from trainees in practical settings. (ii) Many current MLLM applications rely on pre-trained visual encoders to extract visual features (Li et al., 2024a; Liu et al., 2024b). However, due to the domain gap, semantic information from general-purpose visual scenes cannot be directly applied to surgical scenarios. This leads to an inadequate understanding of visual information and the occurrence of reasoning hallucinations.

To address these challenges, we approach them from two perspectives: constructing open-ended, knowledge-intensive surgical vision-language datasets and developing vision-enhanced MLLM models tailored for surgical scenarios. Firstly, we develop Surg-396K, a surgical multimodal instruction dataset specifically for surgical contexts. We systematically extract surgical attribute information from three public datasets and generate instruction-tuning data using diverse conversational templates. To better simulate real-world scenarios, as presented in Table 1, we propose five distinct conversational paradigms to capture the majority of natural language dialogue scenarios and define seven attribute-specific, surgery-related downstream tasks to ensure comprehensive coverage of surgical scene understanding. As a result, the MLLM trained on Surg-396K is better equipped for surgeon–system interaction, enabling it to respond effectively in surgical contexts and provide practical, reliable support for surgical training and education. Secondly, in the model architecture of MLLMs, we design the Mixed Visual Token Engine (MVTE) to extract visual information at multiple scales for better vision-language alignment in surgical contexts. Unlike traditional frameworks that directly use pre-trained Vision Transformers (ViTs) to extract visual tokens, MVTE uses multiple visual towers to extract, interact, and fuse visual tokens, improving the visual information extraction before aligning with text. Furthermore, to minimize model hallucinations, we propose a visual contrast-based approach that compares outputs generated from both original and distorted visual inputs. This method refines the token selection process by applying adaptive plausibility constraints, ensuring better consistency between visual inputs and language outputs in endoscopic surgery scenarios.

As a result, we propose EndoChat (shown in Fig. 1) to support diverse conversational paradigms in endoscopic surgical scenarios. This flexible framework effectively addresses diverse interaction needs and

supports a wide range of surgical tasks, making it highly adaptable to the varied questions that trainees may pose in different contexts. To validate the effectiveness of EndoChat, we first conduct rigorous comparisons with commercial and open-source MLLMs across different dialogue paradigms. The results demonstrate that our approach surpasses existing general-purpose and medical MLLMs in terms of both surgical understanding accuracy and dialogue capability. In addition, we show that our model achieves state-of-the-art performance across various attribute-related sub-tasks of surgical scene understanding. Ablation studies further confirm the effectiveness of our innovative architectural design within the MLLM framework. Moreover, we invite experienced practicing surgeons to independently evaluate whether the assistant is beneficial for advancing surgical training procedures and whether they would be willing to adopt it. The evaluation results indicate that surgeons hold a positive attitude toward our proposed EndoChat, further demonstrating that EndoChat is a qualified assistant for various surgical training and education scenarios. In summary, EndoChat marks a notable advancement in applying MLLMs to surgical training, delivering intelligent, context-aware assistance to trainees. The key contributions of this work are summarized as follows:

- We propose EndoChat, a novel grounded MLLM that supports five conversational paradigms and seven attribute-related surgical sub-tasks in endoscopic surgical scenarios, addressing the requirement for effective dialogue systems in surgical training and guidance.
- We develop Surg-396K, a comprehensive multimodal surgical dataset containing 396K image-instruction pairs through a multi-conversation construction pipeline that systematically extracts surgical information and generates structured annotations.
- EndoChat incorporates the Mixed Visual Token Engine that enhances multi-scale visual information extraction and fusion. Additionally, a visual contrast-based method is integrated to address object hallucinations within MLLMs.
- Extensive experiments on our proposed dataset demonstrate that EndoChat outperforms existing general-purpose and medical MLLMs across various dialogue paradigms and surgical scene understanding sub-tasks. We also validate the practical efficacy of EndoChat through expert evaluations by experienced endoscopists, confirming its potential as an effective tool for enhancing surgical training and education.

2. Related work

2.1. Multimodal large language models

Recent advancements in MLLMs demonstrate diverse strategies for integrating vision and language. Firstly, LLaVA-1.5 (Liu et al., 2024a) refines visual instruction tuning within the LLaVA framework, using a fully-connected MLP connector and lightweight data strategies to achieve efficient multimodal performance. In parallel, Qwen2.5-VL (Bai et al., 2025a) and its successor Qwen3 (Yang et al., 2025a) offer strong document parsing, dynamic resolution handling, and robust tool-use capabilities; Qwen3 (Yang et al., 2025a) also introduces a unified reasoning-control mechanism and improved multilingual visual alignment. Beyond instruction tuning, several models explore improved pretraining techniques and scalable multimodal architectures. InternVL3 (Zhu et al., 2025) adopts a native multimodal pretraining approach with variable visual position encoding, leading to state-of-the-art results across multiple benchmarks. DeepSeek-VL2 (Wu et al., 2024) enhances its MoE-based architecture with dynamic vision tiling and high-resolution understanding, supporting tasks such as document parsing and visual grounding. To further improve training efficiency and deployment scalability, SPHINX-X (Liu et al., 2024c) simplifies previous SPHINX models (Lin et al., 2023a) by merging training stages and optimizing vision encoders, producing scalable MLLMs across different LLM backbones. Gemma 3 (Team et al., 2025) introduces multimodality to lightweight models through SigLIP-based visual embeddings (Zhai et al., 2023) and achieves strong results with low memory cost via compressed token representations. Finally, LLaMA 3 (Grattafiori et al., 2024) introduces a herd of large-scale language models with early-stage multimodal capabilities integrated via image, video, and speech adapters. LLaMA 4 (Meta, 2025) builds on this foundation by more deeply integrating vision-language modeling, enabling stronger long context understanding and reasoning performance across diverse tasks.

These advancements have sparked significant interest in MLLM research. The architectures of current MLLMs are relatively standardized, typically comprising four main components: a visual encoder, a text tokenizer, an alignment module, and an LLM (Li et al., 2023b; Xue et al., 2024; Zhu et al., 2024a; Chen et al., 2023). The visual encoder uses pre-trained vision models to transform images into tokens that are interpretable by LLMs. Commonly employed vision models include EVA-CLIP (Sun et al., 2024), DINOv2 (Oquab et al., 2024), and InternViT (Chen et al., 2024c). Similarly, the text tokenizer utilizes methods like Byte Pair Encoding (BPE) (Shibata et al., 1999) or WordPiece (Song et al., 2021) to convert textual inputs into tokenized representations. Before integrating visual and textual tokens into the LLM, a vision-language alignment process is necessary to enable effective multimodal semantic understanding. Several studies have explored alignment strategies, including the Perceiver Resampler from Flamingo (Alayrac et al., 2022), the Q-Former from BLIP-2 (Li et al., 2023b), and the simple linear layer of LLaVA (Liu et al., 2024a,b), to enhance the model's capabilities of attending to visual information conditioned on text prompts. Additionally, SPHINX-X (Liu et al., 2024c) has introduced architectural refinements in vision-text alignment modules, contributing to more robust image-conditioned reasoning. After alignment, the vision-language embeddings are passed into the LLM, which uses self-attention mechanisms and autoregressive language modeling to generate the final textual outputs (Vaswani et al., 2017; Brown et al., 2020; Yang et al., 2025b). Furthermore, training MLLMs requires large-scale vision-language paired datasets (Zhou et al., 2025a; Kuang et al., 2024). To address the challenge of data scarcity, the LLaVA models employed visual instruction tuning and used the advanced visual understanding capabilities of commercial MLLMs to extract diverse vision-language paired datasets (Liu et al., 2024b,a). This approach significantly enriched the open-source community by contributing foundational data resources. However, these standardized architectures face notable limitations in surgical scenarios, as their visual encoders are pretrained on general-domain datasets, resulting in a domain gap that hinders the capture of fine-grained semantic details in surgical images. Besides, the alignment modules often struggle to model interactions among instruments and tissues. To address these challenges, it is necessary to design multi-scale, domain-specific visual encoding strategies tailored for the unique surgical scenes.

Based on the strong visual understanding capabilities of MLLMs in general scenarios, researchers have developed powerful multimodal medical assistants (Lu and Wang, 2025; Zhou et al., 2025b; Song et al., 2024; McDuff et al., 2025; Liu et al., 2025). Early attempts focused on dataset-level efforts, such as collecting and organizing new medical datasets and performing text augmentation through instruction tuning (Tu et al., 2024; Li et al., 2024b). In addition, common computer-assisted diagnosis tasks have been widely explored, including image diagnosis for X-rays (Wang et al., 2024d), cancer prediction (Skourti, 2025), detection of eye diseases (Shi et al., 2025), and dermatological diagnosis tasks (Zhou et al., 2024). For medical report generation tasks, XrayGPT utilized advanced vision-language models to produce interactive summaries from radiology reports, offering concise findings and supporting follow-up questions (Thawakar et al., 2024). Huang et al. aimed to refine medical reports by focusing on key semantic information, which enhanced the accuracy and interpretability of the generated content (Huang et al., 2024). Other studies have worked on improving classification tasks by integrating privacy-preserving LLMs

and multi-type annotations into datasets, helping address the challenge of noisy labels (Lanfredi et al., 2025). However, due to the limited availability of high-quality annotations in endoscopic surgery, the significant domain gap between endoscopic surgery and general scenarios, and the presence of hallucinations (Zhao et al., 2024), the MLLM applications in the surgical field, while having seen some initial exploration, still face substantial challenges.

2.2. Surgical vision-language models

Early surgical vision-language models primarily focused on developing VQA systems to address question-answering and dialogue requirements during surgical procedures (Chen et al., 2024a; Peng et al., 2024; Zhu et al., 2024b). As the first approach to introduce a question-answering model specifically tailored for surgical scenarios, Surgical-VQA represented key elements in surgical environments — such as instruments, tissues, tools, and spatial positions — using textual descriptions (Seenivasan et al., 2022). Built on the VisualBERT (Li et al., 2019) framework, it integrated multimodal representations of text and images to generate corresponding answers through a decoder. Subsequent works, such as Surgical-VQLA, improved upon this method by incorporating bounding box outputs at the decoder stage, enabling explicit visual localization to better assist surgeons (Bai et al., 2023b; Zhu et al., 2024b). SSG-VQA addressed the shortage of vision-language datasets in surgical scenarios by generating synthetic surgical dialogue data through predefined attributes and templates. Later advancements explored alternative network architectures (Zhang et al., 2024a; Seenivasan et al., 2023) and optimal vision-language alignment paradigms (Yuan et al., 2025). However, as discussed in Section 1, structured dialogue datasets combined with encoder–decoder architectures limit the model's capabilities, confining it to queries within predefined content. This restricts its ability to handle complex interactions, thereby reducing its clinical applicability.

With the rise of MLLMs (Wang et al., 2024a; Achiam et al., 2023; Liu et al., 2024a,b), the medical community is beginning to explore their potential applications in the field of healthcare (Li et al., 2024b; Chen et al., 2024b; Ye et al., 2024; Hu et al., 2024; Ferber et al., 2024). Several prior studies have demonstrated the potential of MLLMs in surgical scenarios (Li et al., 2024a; Jin and Jeong, 2024; Wang et al., 2025; Schmidgall et al., 2024; Hou et al., 2024). For example, Surgical-LLaVA (Jin and Jeong, 2024) extended the instruction-tuning framework of LLaVA (Liu et al., 2024b) by incorporating existing classification annotations or structured texts into the training of MLLMs. Moreover, Surgical-LVLM integrated an additional localization module, enabling the generation of bounding boxes for specific targets. SCAN (Hou et al., 2024) introduced a memory-augmented query mechanism to enhance the VQA performance by employing self-contained queries within the MLLM. LLaVA-Surg collected and annotated open-source videos and datasets to improve the model's conversational capabilities through instruction tuning (Li et al., 2024a). In contrast, GP-VLS unified various surgical tasks within a question-answering framework, representing the outputs of different downstream surgical tasks in textual form (Schmidgall et al., 2024). CoPESD introduced a novel surgical scene understanding dataset. After annotating the instrument-tissue-action information, it further combined instruction tuning and professional surgeons' descriptions to build a multi-granularity surgical motion analysis dataset (Wang et al., 2024c). However, existing methods are primarily based on general captioning or predefined surgical scene understanding tasks (e.g., action analysis, instrument recognition), aiming to enhance dialogue diversity through techniques such as instruction tuning. These approaches overlook the dialogue paradigm required for real-world interactions with surgeons, which is the key issue we aim to address in this paper.

3. Methods

3.1. Surgical multimodal instruction dataset: Surg-396K

The AI-assisted surgery field has experienced a notable expansion in the availability of public multimodal datasets, particularly VQA pairs, as evidenced by works ranging from Seenivasan et al. (2022) to Yuan et al. (2024). However, the availability of multimodal instruction data remains limited, primarily due to the time-intensive and less standardized processes involved in human crowd-sourcing. To promote the development of MLLMs tailored for surgical understanding, we propose Surg-396K, a surgical multimodal instruction dataset incorporating 41K images and 396K instruction-following annotations for endoscopic surgery. Following the data generation process shown in Fig. 2, we define five conversational paradigms and seven attribute-related sub-tasks to capture the majority of natural language dialogues while ensuring comprehensive coverage of surgical scene understanding. The style of these conversation types and sub-tasks is shown in Table 1.

3.1.1. Preliminary for constituent datasets

In the construction of our Surg-396K dataset, we integrate three distinct datasets. Inspired by the achievements of recent MLLMs in text-annotation tasks (Liu et al., 2024b), we utilize GPT-4V to expand multimodal instruction-following data, resulting in five conversation types derived from EndoVis-VQLA (Bai et al., 2023b) and CoPESD (Wang et al., 2024c) datasets. The third dataset Cholec80-VQA (Seenivasan et al., 2022), which lacks grounding information, is directly employed in two of the conversation types in Surg-396K dataset.

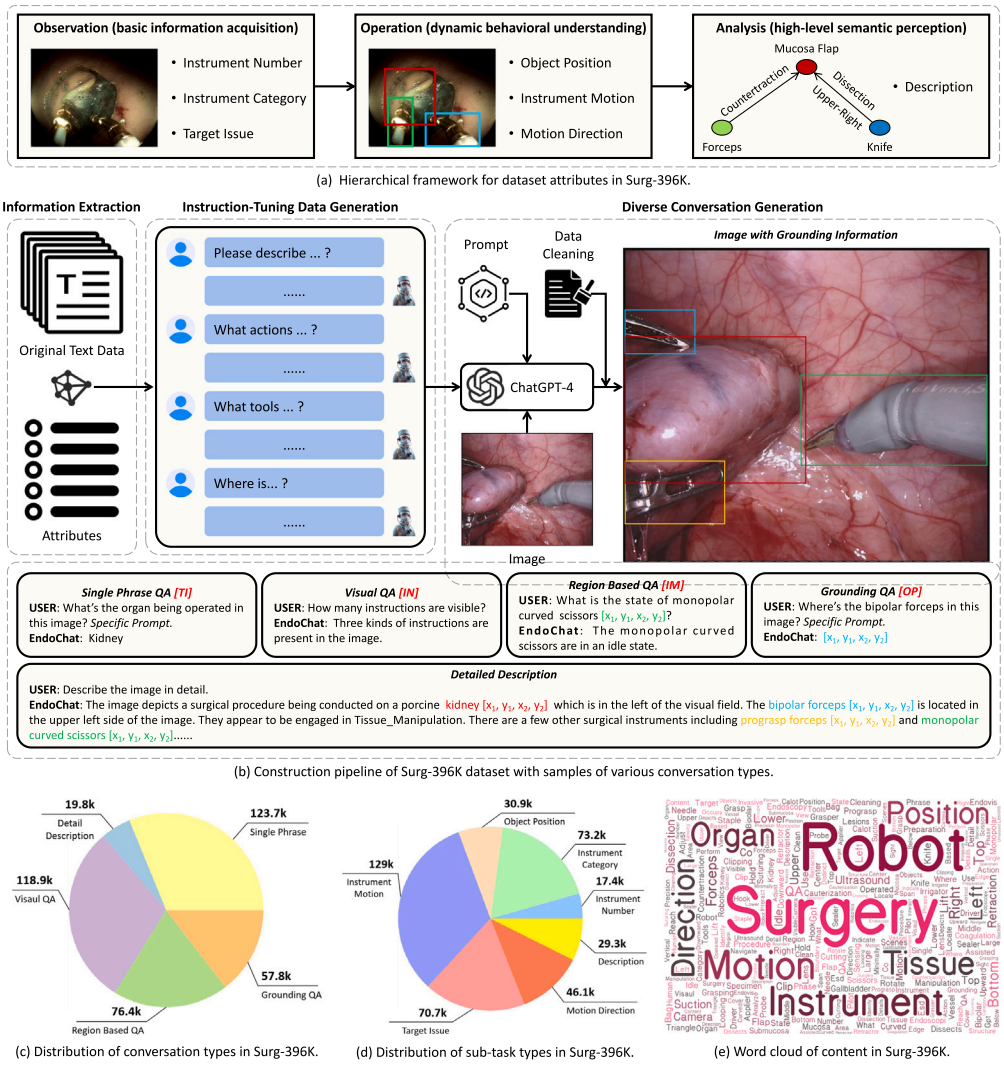
EndoVis-VQLA (Bai et al., 2023b) is a publicly available dataset for endoscopic surgery, derived from the MICCAI Challenges of 2017(Allan et al., 2019) and 2018 (Allan et al., 2020). This dataset integrates VQA annotations with bounding box labels to create Visual Question Localized-Answering (VQLA) pairs, which encompass surgical actions, target tissues, instruments, and their respective bounding boxes. The images in EndoVis-VQLA have a resolution of 1280×1024 pixels. The dataset is comprised of two parts: EndoVis-18-VQLA, containing 2007 frames, and EndoVis-17-VQLA, including 97 frames.

CoPESD (Wang et al., 2024c) is a comprehensive multi-level surgical motion dataset specifically designed for the training of MLLMs in the context of Endoscopic Submucosal Dissection (ESD). It comprises 17,679 images accompanied by detailed motion annotations derived from over 35 hours of ESD videos. The resolution of these images is 1306×1009 pixels. The motion annotations include information on target tissues, instruments, surgical motions, motion directions and the corresponding bounding boxes.

Cholec80-VQA (Seenivasan et al., 2022) is an innovative dataset generated from 40 video sequences of the Cholec80 dataset (Twinanda et al., 2016), encompassing a total of 21,591 frames. The images in Cholec80-VQA have a resolution of 854×480 pixels. Utilizing original tool-operation and phase annotations from the Cholec80 dataset, Cholec80-VQA proposes two types of question–answer pairs for each frame: Classification, which features 14 unique single-word answers; Sentence, which is presented in full sentence form. Due to the absence of grounding information and less content in the annotation of each image, we do not expand it with GPT-4V, but directly use Classification and Sentence as the conversation of the Single Phrase and Visual QA.

3.1.2. Attribute retrieval

To ensure that the annotations encompass comprehensive surgical information, we adopt a hierarchical framework for attribute analysis, from basic observation to dynamic operation and high-level perception, as shown in Fig. 2(a). At the Observation level, foundational attributes are defined, including *Instrument Number* (IN, count of visible instruments), *Instrument Category* (IC, classification of instrument types), and *Target Issue* (TI, anatomical targets of surgical focus). The Operation level focuses on dynamic behaviors and spatial characteristics, such as



(a) Hierarchical framework for dataset attributes in Surg-396K.

(b) Construction pipeline of Surg-396K dataset with samples of various conversation types.

(c) Distribution of conversation types in Surg-396K.

(d) Distribution of sub-task types in Surg-396K.

(e) Word cloud of content in Surg-396K.

Fig. 2. Overview of the construction pipeline and distribution statistics for our Surg-396K dataset. The pipeline involves five key steps: annotation attribute analysis, information extraction, instruction-tuning data generation, diverse conversation generation, and data validation.

Object Position (OP, spatial mapping within a 3 × 3 grid), *Instrument Motion* (IM, functional roles inferred from motion), and *Motion Direction* (MD, trajectories across eight cardinal and diagonal directions). Finally, the Analysis level integrates these attributes to comprehensively analyze the surgical scenes, summarized as *Description*. This structured design achieves a seamless integration of the dataset content that maximizes the extraction of surgical information from the original annotations. Furthermore, we develop attribute-specific QA templates tailored for generating instruction-tuning data.

3.1.3. Diverse conversation generation

The expression format of the instruction-tuning data obtained by information extraction through attributes is limited to human-designed templates, resulting in a homogeneous structure. To emulate natural language expression, we diversify the instruction-tuning data by incorporating interaction requirements specific to surgical scenarios, ensuring coverage of different levels of inquiry and information needs. For example, a trainee might ask a brief question about a specific target, which does not require any redundant information. Therefore, *Single Phrase QA* is designed to provide concise, accurate, and direct responses relevant to surgical contexts. When a trainee requests a detailed explanation of an entire surgical image, *Detailed Description*, on the other hand, offers surgeons comprehensive explanations of all sub-tasks present within the current surgical scene. For routine queries about image content, *Visual QA* delivers context-aware answers by combining user queries with image information. When a trainee needs a focused conversation about a specific region, *Region Based QA* focuses on targeted responses for specific areas of interest, while *Grounding QA* generates bounding box descriptions based on user-provided prompts. Based on this diversification, we interpret the instruction-tuning data through GPT-4V and generate five conversation types:

Single Phrase is designed to equip EndoChat with the capability to deliver concise, definitive answers to each query, using a rapid analysis of the surgical image. This type of conversation can be directly sourced from the instruction-tuning data. Additionally, we have enriched the diversity of the questions utilizing GPT-4V. To guide the model to provide the answer in a single word or phrase, we introduce the task-specific prompt *"Answer the question with a single phrase."* at the end of the questions, which can be represented by:

$$\texttt{Human}: \mathbf{T}_q\ [Prompt] \backslash \texttt{n}\ \texttt{EndoChat}: \mathbf{T}_a \backslash \texttt{n} \tag{1}$$

Detailed Description provides comprehensive, grounded responses to queries that delve into the intricate details of the visual scene. Answers in this type are entirely generated by GPT-4V with prompts that take full advantage of instruction-tuning data. Therefore, we ensure the response covers all attributes in the surgical images. The list of questions has also been diversified using GPT-4V. This type can augment

Table 1
List of attributes (with abbreviations) and conversation types designed for Surg-396K.

	Category	Response Style
Conversation Type	Single Phrase	Answer the question with a single word or phrase.
	Detail Description	Describe the image in detail with [grounding].
	Visual QA	Question and answer without [grounding].
	Region Based QA	Question with [grounding], answer without [grounding].
	Grounding QA	Answer the question with a [grounding].
	Category	**Sample**
Sub-task Type	Instrument Number [IN]	2, 3, etc.
	Instrument Category [IC]	Prograsp forceps, ultrasound probe, etc.
	Object Position [OP]	Right-bottom, left-top, etc.
	Instrument Motion [IM]	Idle, lift, etc.
	Target Issue [TI]	Kidney, the mucosal flap, etc.
	Motion Direction [MD]	Upward, lower left, etc.
	Description	Illustrate the image through a descriptive explanation., etc.

EndoChat's proficiency in articulating complex visual information as if it is observing the scene in real-time.

Visual QA emphasizes straightforward question–answer pairs that provide general insights into the surgical scene without specific grounding. This mode differs from Single Phrase by allowing more contextual information in the responses while still maintaining conciseness. Therefore, the generation process for Visual QA entails an additional step compared to Single Phrase. Specifically, GPT-4V is utilized to elaborate the single-word or single-phrase response into a complete sentence while incorporating the content related to description and reasoning attributes from the instruction-tuning data.

Region Based QA incorporates grounding information within the question to guide the model's attention to a specific region of the image. This conversation type facilitates targeted analysis of visual content, pinpointing the precise location of surgical objects compared to Visual QA. For text expression, we insert the bounding box of the target after its text, e.g., *"kidney* $[x_1, y_1, x_2, y_2]$*"*. x_1 and y_1 denote the coordinates of the top-left corner of the bounding box, while x_2 and y_2 specify the bottom-right corner. Each coordinate is normalized to the interval $[0, 1]$.

Grounding QA delivers responses solely through bounding boxes, training EndoChat to provide accurate spatial answers based on the interplay between the visual content and the posed questions. We introduce the task-specific prompt *"Answer the question with just a bounding box."* at the end of the questions to guide the model to provide the answer in a bounding box. This conversation type has the same format as Single Phrase QA in Eq. (1).

More details for these conversation templates are provided in the supplementary material.

3.1.4. Surgical sub-task formulation

While we have established diverse conversation paradigms to enable MLLMs to handle a wide range of natural language dialogue scenarios, it is critical to ensure a comprehensive understanding of surgical scenes. Given that various elements relevant to surgical scenes have already been extracted through prior attribute retrieval, we formulate seven attribute-related sub-tasks to systematically evaluate MLLMs from different aspects of surgical understanding. The QA pairs for most sub-tasks are derived from Single Phrase QA and Grounding QA since Single Phrase QA can provide concise responses capturing fundamental visual attributes such as instrument types, quantities, and basic spatial relationships, whereas those from Grounding QA offer precise spatial annotations that delineate object positions and regions of interest. Additionally, description-related sub-tasks are directly derived from the Detailed Description paradigm. These attribute-driven sub-tasks encompass a broad spectrum of surgical scene understanding, ranging from basic observation to high-level analytical reasoning, facilitating a comprehensive assessment of MLLMs in understanding complex surgical environments.

3.1.5. Data validation

Following the generation of diverse conversations, a rigorous data validation process is implemented to ensure the integrity and reliability of the Surg-396K dataset. Given the large scale of the dataset, we adopt a 1/5 random sampling strategy for each conversation type across each source dataset. The validation is carried out by four trained medical reviewers, organized into two independent groups. Each group is responsible for reviewing half of the sampled annotations. Within each group, cross-validation is conducted to ensure internal consistency, and any uncertainties or disagreements are resolved through collaborative discussion with an experienced endoscopic surgeon. During validation, we assess information completeness (whether the annotations contained all essential and accurate content, such as tools and procedures), relevance (alignment between the QA content and the assigned conversation type), and semantic clarity (ensuring the GPT-4V-enriched content is linguistically and contextually coherent). Notably, since the annotated content is generated based on structured attributes extracted from source datasets, our validation can perform systematic attribute-level comparisons to identify and eliminate any GPT-4V-induced hallucinations or factual inconsistencies. Frequent annotation issues observed during sampling are recorded, and corresponding corrections are applied to the non-sampled data accordingly. This hybrid validation approach ensures that both sampled and non-sampled annotations meet high standards of quality and consistency, supporting reliable downstream training of the MLLM.

3.1.6. Comparison with existing surgical scene understanding datasets

We present a comprehensive comparison between Surg-396K and existing surgical scene understanding datasets, evaluating key factors such as surgery type, dataset scale, and annotation diversity. Table 2 provides a detailed summary of Surg-396K in relation to both earlier and more recent benchmarks developed for surgical scene understanding tasks. Surg-396K exhibits a substantial advantage in both scale and annotation diversity compared to existing datasets. It comprises 41.4K images and 396K annotations spanning multiple surgical procedures, including Laparoscopic Cholecystectomy, Nephrectomy, and Submucosal Dissection, significantly surpassing datasets such as Cholec80-VQA and EndoVis-18-VQA in both volume and variety. Furthermore, in contrast to datasets like CoPESD and PSI-AVA-VQA, Surg-396K's extensive annotation set facilitates more comprehensive training for MLLMs. While SSG-VQA contains a larger number of QA pairs, its annotation format is unimodal and the QA structure is relatively monotonous. In contrast, Surg-396K integrates diverse conversational formats alongside grounding annotations, enabling a broader range of multimodal tasks. This combination of scale, diversity, and multimodal richness establishes Surg-396K as a more comprehensive and versatile benchmark, advancing research in fine-grained surgical scene understanding.

Table 2
The comparison of Surg-396K with existing surgical scene understanding benchmarks. In the "Surgery Type" column, "LC" indicates Laparoscopic Cholecystectomy, "Ne" indicates Nephrectomy. "Pr" indicates Prostatectomy. "SD" indicates Submucosal Dissection.

Dataset	Years	Surgery type	Image size	Annotations	Annotation size
VQA-Med-2018 (Hasan et al., 2018)	2018	\	2.9K	QA Pairs	6.4K
VQA-Med-2019 (Ben Abacha et al., 2019)	2019	\	4.2K	QA Pairs	15.3K
Cholec80-VQA (Seenivasan et al., 2022)	2022	LC	21.6K	QA Pairs	43K
EndoVis-18-VQA (Seenivasan et al., 2022)	2022	Ne	2K	QA Pairs	11.8K
EndoVis-VQLA (Bai et al., 2023b)	2023	Ne	2.2K	QA Pairs; Bbox	9.5K
PSI-AVA-VQA (Seenivasan et al., 2023)	2024	Pr	2.2K	QA Pairs	10.3K
CoPESD (Wang et al., 2024c)	2024	SD	17.7K	QA Pairs; Bbox	17.7K
SSG-VQA (Yuan et al., 2024)	2024	LC	25K	QA Pairs	960K
Surg-396K (Ours)	2025	LC; Ne; SD	41.4K	QA Pairs; Bbox	396K

3.2. Visual enhanced MLLM: EndoChat

Visual grounding conversations in endoscopic surgery involve the intricate interaction between visual and linguistic modalities, requiring a comprehensive understanding of knowledge about distinct objects or regions. Therefore, we propose EndoChat, a novel multimodal large language model designed for visual grounding conversations within the endoscopic surgery scenes, as shown in Fig. 3. Given an input image, the mixed visual encoder extracts source tokens, denoted as $X_d \in \mathbb{R}^{N\times D\times L_1}$ and $X_o \in \mathbb{R}^{N\times D\times L_2}$, where N represents the number of frames, D denotes the hidden dimension, and L_1 and L_2 correspond to the sequence lengths of the respective token sets. Subsequently, the extracted source tokens are processed by our proposed Mixed Visual Token Engine. The resulting enhanced image tokens are represented as $X' \in \mathbb{R}^{N\times(D+m)\times(L_1+L_2)}$, where m denotes the number of newly generated tokens. These enriched visual tokens are then aligned with language tokens and input to the language model to produce the final response. Moreover, a visual contrast mechanism is introduced to mitigate object hallucinations.

3.2.1. Preliminary for EndoChat

EndoChat is built upon the SPHINX architecture (Lin et al., 2023a), a versatile multi-modal large language model designed to support a range of visual instruction-following tasks. The architecture of SPHINX builds upon the large language model LLaMA-2 (Touvron et al., 2023), incorporating multiple vision encoders and employing a joint mixing strategy for weights, tasks, and visual embeddings. To enhance its visual understanding, SPHINX mixes visual embeddings from different vision backbones and processes high-resolution images through a novel strategy of dividing the image into multi-scale sub-images. The integration of multi-task training and the joint mixing strategies empower SPHINX with robust multi-modal capabilities, encompassing tasks such as object detection, diagram interpretation and region-level captioning.

3.2.2. Mixed visual token engine

In our EndoChat, we mix visual embedding to more scales from high-resolution sub-images, thereby enhancing the encoding of high-resolution images. For input images with high resolution, we implement two parallel pathways to generate five corresponding images at resolutions of 224 × 224 and 512 × 512, respectively. Then, these images are fed into a mixed visual encoder, which consists of DINOv2 (Oquab et al., 2024) and OpenCLIP (Radford et al., 2021), resulting in outputs X_d and X_o. For MLLMs, visual encoders typically summarize the visual embeddings after encoding image tokens by extracting an aggregated representation through operations like a multi-layer perceptron (MLP). Although this direct representation is computationally efficient, it struggles to capture multi-scale information and often overlooks crucial spatial relationships between different positions or regions. Thus, it may confuse the LLM and underutilize its capabilities. To address these limitations, we introduce the Mixed Visual Token Engine (MVTE). MVTE dynamically generates global visual tokens based on the source token produced by the mixed visual encoder, which seamlessly integrates and maximizes the informational utility of multi-scale visual tokens.

Specifically, as shown in the right bottom of Fig. 3, there are two parallel pathways to process source tokens X_d and X_o from the mixed encoder. In each path, a contextual MLP network (Linear-ReLU-Linear) followed by Softmax normalization is employed to generate the contextual attention map (Song et al., 2024). Subsequently, we utilize matrix multiplication to compute the output visual tokens which are spatial-wisely concatenated with their source token to obtain the combined tokens X':

$$X' = softmax(MLP(X)) \cdot X \oplus X \tag{2}$$

Finally, we channel-wisely concatenate two pathways' combined tokens: X'_o and X'_d, followed by an MLP Projector for dimension alignment to obtain the final image tokens $\hat{X}$. The process can be described in the following equation:

$$\hat{X} = Proj(X'_o \oplus X'_d) \tag{3}$$

The inclusion of MVTE enables the LLM to generate more complementary features, enhancing its comprehension of complex endoscopic surgical scenes and improving effectiveness in complex reasoning tasks.

3.2.3. Hallucination mitigation through visual contrast

The MLLMs, parameterized by θ, are adept at capturing intricate visual patterns x and textual query q, translating them into coherent linguistic representations y. Specifically, MLLMs sample the response y auto-regressively from the probability distribution, predicting the next word step by step based on x and q, expressed as:

$$y_t \sim p_\theta\left(y_t \mid x, q, y_{<t}\right), \propto \exp \operatorname{logit}_\theta\left(y_t \mid x, q, y_{<t}\right) \tag{4}$$

where y_t represents the token at time step t, and $y_{<t}$ denotes the sequence of generated tokens up to time step $t-1$. In challenging visual scenarios like endoscopic surgery, MLLMs suffer from Object Hallucination, a phenomenon that arises from their reliance on statistical biases and unimodal priors. This dependency leads to generated text that, while semantically coherent, can be inconsistent with the objects in a given image. Due to the complexity of the endoscopic surgery scenario, ambiguous visual features can lead the MLLMs to ignore critical visual cues, instead favoring linguistic priors in natural pretraining datasets when generating outputs.

To address object hallucinations within MLLMs, we introduce the contrast of the model's output generated based on the original and distorted visual input to counteract the statistical biases and language priors (Leng et al., 2024). Visual contrast is a training-free approach that is grounded in generating two parallel output distributions: one based on the original visual input x and another based on a distorted version x'. The distorted input x' is produced by applying controlled Gaussian noise to x, which amplifies language priors and statistical biases that contribute to hallucinations. The contrastive probability distribution p is computed through the logit differences between the original and distorted visual inputs as follows:

$$p(y|x, x', q) = \text{softmax}\Big[(1+\alpha)\cdot \text{logit}_\theta(y|x, q) - \alpha \cdot \text{logit}_\theta(y|x', q)\Big] \tag{5}$$

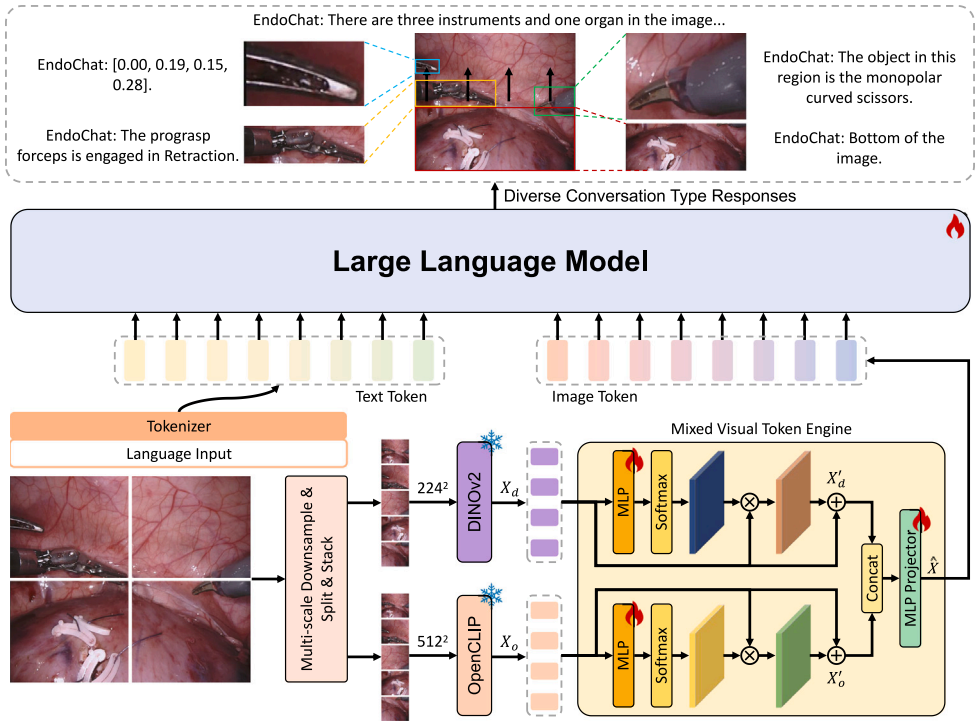


Fig. 3. The overview of the proposed EndoChat. For each input image, we use a multi-scale downsampling strategy to generate different scales and sub-images. 224^2 and 512^2 indicate concatenated features with the shapes $5 \times 224 \times 224 \times 3$ and $5 \times 512 \times 512 \times 3$, respectively. These features are subsequently encoded using a mixed visual backbone, followed by the Mixed Visual Token Engine. The resulting vision tokens are then transformed into language space, suitable for input to the Large Language Model. In addition to visual inputs, region coordinates can be auxiliary inputs, along with specific prompts to guide user-defined tasks. This enables the LLM to generate language responses for related object regions.

where α is a hyperparameter that adjusts the weighting between the two distributions, with higher α values enhancing the distinction between the two distributions. Such visual contrast serves as a corrective mechanism, reducing hallucinations by contrasting with a distribution predisposed to favoring them. Furthermore, to prevent p from punishing valid outputs and facilitate the generation of a correct token, an adaptive constraint (Li et al., 2023a) is introduced:

$$\mathcal{L}\left(y_{<t}\right) = \left\{ y_t \in \mathcal{L} : p_\theta\left(y_t \mid x, q, y_{<t}\right) \geq \beta \max_w p_\theta\left(w \mid x, q, y_{<t}\right) \right\}, \\ p\left(y_t \mid x, x', q\right) = 0, \text{ if } y_t \notin \mathcal{L}\left(y_{<t}\right) \tag{6}$$

where $\beta \in [0, 1]$ controls the truncation of the next token distribution. Larger β indicates more aggressive truncation, keeping only high-probability tokens. Adaptive plausibility constraints refine the contrastive distribution, enhancing decision confidence. This streamlines the candidate pool, often retaining a single high-probability token, and neutralizes potential adverse effects of visual contrast, preventing the generation of implausible tokens and preserving content integrity.

3.2.4. Training strategy

We adopt the open-source LLaMA2-13B (Touvron et al., 2023) large language model as our EndoChat's foundational component. LLaMA2-13B serves as a unified interface for diverse vision-language tasks. To ensure the model's responses are aligned and contextually effective, task-specific prompts are appended to input data, which helps guide the LLM's responses. For LLM fine-tuning, we employ the Low-Rank Adaptation (LoRA) (Hu et al., 2022) technique that introduces two smaller matrices as a low-rank approximation of the large, original matrix. We optimize parameters of low-rank matrices instead of all parameters in the LLM. This adaptation method reduces training time and computational overhead. At the same time, it preserves the model's broader knowledge of generic object categories and spatial landmarks, thereby enhancing its vision-language reasoning capabilities in the endoscopic surgery domain.

4. Results

4.1. Implementation details

We conduct comparative experiments against multiple MLLMs, including BiomedGPT (Zhang et al., 2024b), LLAVA-Med (Li et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025a), LLaMA-3.2-11B (Grattafiori et al., 2024), Gemma-3-12B (Team et al., 2025), Qwen2-VL-7B (Wang et al., 2024a), GLM-4V-9B (GLM et al., 2024), MiniGPTv2 (Chen et al., 2023), LLAVA-1.5-13B (Liu et al., 2024a), and SPHINX (Lin et al., 2023a). These benchmarks cover the current state-of-the-art within open-source, general-purpose models of comparable parameter scale to our EndoChat. Additionally, we benchmark our method against a range of specialized models, namely VisualBert (Li et al., 2019), VisualBert ResMLP (Seenivasan et al., 2022), MCAN (Ben-Younes et al., 2017), VQA-DeiT (Touvron et al., 2021), MUTAN (Ben-Younes et al., 2017), MFH (Yu et al., 2018), BlockTucker (Ben-Younes et al., 2019), CAT-ViL DeiT (Bai et al., 2023a), GVLE-LViT (Bai et al., 2023b), and Surgical-LVLM (Wang et al., 2025). All evaluation metrics in the results are presented in percentage form, except for CIDEr. To train the LLM and the Mixed Visual Token Engine, we utilize an input resolution of 1024×1024. We conduct the training on the Surg-396K dataset for a single epoch, employing four NVIDIA A800 GPUs. We utilize AdamW (Loshchilov and Hutter, 2019) optimizer with an initial learning rate of 2e-5, following a cosine decay schedule and a linear warm-up phase. The training process employs a batch size of 16 and is completed in approximately 20 hours.

We conduct four types of MLLM comparison experiments: zero-shot comparison (Tables 3–7), fine-tuning comparison (Table 8), generalization comparison (Table 9) and ablation study (Table 10). Apart from the MLLM comparison, in Tables 4 and 5, we also introduce specialized models (e.g., VisualBERT, MCAN) that are trained on the training set of the comparison dataset before testing. For generalization comparison,

Table 3
Comparison experiments with zero-shot MLLMs in Single Phrase QA and Grounding QA on three parts of the Surg-396K dataset.

Model	EndoVis-18				EndoVis-17				CoPESD			
	Acc	F-score	AP@50	mIoU	Acc	F-score	AP@50	mIoU	Acc	F-score	AP@50	mIoU
BiomedGPT (Zhang et al., 2024b)	5.61	3.42	39.96	32.35	0.00	0.00	38.65	36.55	5.44	1.39	36.70	28.81
LLAVA-Med (Li et al., 2024b)	3.55	1.96	32.22	28.28	0.00	0.00	28.39	17.49	8.39	3.78	54.01	51.49
Qwen2-VL-7B (Wang et al., 2024a)	1.99	0.22	42.13	35.35	0.00	0.00	44.49	37.82	16.72	6.29	63.59	57.59
MiniGPTv2 (Chen et al., 2023)	0.00	0.06	12.06	10.05	0.00	0.00	15.02	9.77	3.37	0.30	37.89	33.36
LLAVA-1.5-13B (Liu et al., 2024a)	2.31	1.09	33.98	30.04	0.00	0.00	27.63	18.09	8.19	4.53	54.98	50.28
SPHINX (Lin et al., 2023a)	1.37	0.14	32.59	27.23	0.00	0.00	25.76	15.44	7.19	4.30	59.93	55.52
Qwen2.5-VL-7B (Bai et al., 2025a)	27.48	20.44	60.75	63.95	39.19	17.30	61.02	64.15	32.47	10.50	82.47	70.09
LLaMA-3.2-11B (Grattafiori et al., 2024)	24.05	2.52	62.56	56.28	22.46	4.34	60.59	53.89	6.66	0.08	79.18	60.19
Gemma-3-12B (Team et al., 2025)	36.37	23.17	58.58	53.19	23.09	16.58	46.18	48.91	20.22	7.33	83.25	64.29
EndoChat	**71.47**	**43.74**	**93.22**	**86.89**	**55.51**	**29.78**	**90.25**	**86.62**	**75.34**	**31.18**	**99.43**	**93.64**

all MLLMs are trained on the Surg-396K dataset before evaluating on the out-of-distribution datasets PSI-AVA-VQA (Seenivasan et al., 2023) and EndoVis-17-VQLA-Extend (Bai et al., 2025b).

For training and evaluation on the Surg-396K dataset, three constituent datasets all follow the data splitting scheme of their source dataset (Bai et al., 2023b; Wang et al., 2024c; Seenivasan et al., 2022). Consequently, Surg-396K comprises 34,277 training images with 339,098 QA pairs, and 7097 test images with 57,494 QA pairs.

4.2. EndoChat tailors interactions to varying complexity levels of surgical conversations

EndoChat adapts its interactions to meet the diverse complexities of surgical scenarios through two significant paradigms: Single Phrase QA and Detailed Description. These complementary approaches enable EndoChat to provide precise, actionable insights for both real-time surgical guidance and in-depth educational purposes.

Single Phrase QA focuses on delivering concise and definitive answers, which is ideal for straightforward queries during surgical procedures. By employing the task-specific prompt: "*Answer the question with a single phrase.*", EndoChat provides succinct responses, including key aspects such as instrument types, counts, actions, or relative positions within the surgical scene. This capability relies on rapid analysis of visual content, ensuring efficiency without redundant elaboration. For instance, queries like "How many instruments are visible?" are answered directly, such as "three". Empirical evaluations, shown in Table 3, demonstrate that EndoChat outperforms the SOTA open-source MLLMs such as LLaMA-3.2 and Gemma-3 under comparable parameter counts. Specifically, on the EndoVis-17 part, where other models struggle to answer questions effectively (0% or very low accuracy and F-score), EndoChat achieves a remarkable 55.51% accuracy, along with F-score (29.78), AP@50 (90.25), and mIoU (86.62). It also has superior performance on other parts of Surg-396K, and public datasets shown in Tables 4 and 5, which demonstrates the robustness and effectiveness of the instruction-tuning process on our Surg-396K dataset.

Detailed Description caters to more complex scenarios where a comprehensive understanding is necessary. This interaction type provides in-depth explanations grounded in the visual content of the surgical scene, making it essential for training scenarios and complex procedures. Answers generated by EndoChat offer detailed insights into tissues, instruments, motions, and other surgical elements, aiding in decision-making and contextual understanding, and also providing the reasoning to justify the answers. As shown in Table 6, EndoChat's performance in generating detailed descriptions is rigorously assessed using GPT-4 Score, where it significantly outperforms all MLLMs across all parts of Surg-396K. This demonstrates EndoChat's advantage in generating detailed, context-aware, and high-quality descriptions.

4.3. EndoChat enhances interactions through multiple modalities

EndoChat also enhances surgical educational interactions with Grounding QA and Region Based QA. Grounding QA outputs bounding boxes to accurately localize surgical elements, providing context-aware guidance for trainees. Region Based QA, on the other hand, uses input bounding boxes to focus on specific areas, such as tools or tissues, aiding in precise localization for surgical training and navigation.

EndoChat demonstrates superior performance in Grounding QA, a crucial capability for real-time surgical guidance. By delivering responses through bounding boxes, Grounding QA enables EndoChat to ensure accurate spatial localization based on both the visual content and the posed questions. The task-specific prompt, "*Answer the question with just a bounding box.*", directs the model to focus on precise spatial information. As shown in Tables 3 and 4, EndoChat outperforms other SOTA models by a significant margin. Notably, EndoChat achieves the highest mIoU on both the EndoVis-18-VQLA and EndoVis-17-VQLA datasets, with scores of 86.89 and 86.62, respectively. The consistently superior performance of EndoChat over medical-specialized models and general-purpose SOTA MLLMs highlights its more effective integration of both visual and language-based reasoning, suggesting EndoChat is better suited for surgical navigation tasks.

Region Based QA, another key interaction, enables EndoChat to provide more targeted analysis by guiding its attention to specific areas within a surgical image. By incorporating the bounding box from the user into the question, Region Based QA helps focus the model on regions of particular interest, such as a surgical tool or tissue. This approach is essential for tasks requiring precise localization or assessment of anatomical structures. As shown in Table 6, EndoChat outperforms the other models in Region Based QA, achieving the highest on different downstream surgical datasets. This highlights its ability to accurately focus on and analyze localized areas, providing valuable insights for surgical navigation, such as tracking instruments or assessing tissue conditions during procedures.

4.4. EndoChat is a surgeon-like interaction tool

EndoChat uses the multi-modal conversational instruction-tuning dataset to become a surgeon-like interaction model through Visual Question Answering, where the system answers general questions about the surgical scene while maintaining a balance between conciseness and contextual clarity. Unlike Single Phrase QA, which provides brief responses, Visual QA allows for more detailed insights, elaborating on key aspects of the image. This approach mimics a more natural conversational flow, enabling EndoChat to deliver informative yet succinct answers, resembling how a human expert would provide both simple and insightful feedback during surgery.

Table 5 demonstrates that EndoChat significantly outperforms both zero-shot MLLMs and specialized models in Visual QA on the Cholec80-VQA dataset. EndoChat achieves the highest score in all evaluation metrics, surpassing SOTA specialized models VisualBert and VisualBert

Table 4
Comparison experiments with zero-shot MLLMs (top) and specialized models (middle) in Single Phrase QA and Grounding QA on EndoVis-VQLA (Bai et al., 2023b) dataset.

Model	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-score	mIoU	Acc	F-score	mIoU
BiomedGPT (Zhang et al., 2024b)	5.61	3.42	56.93	0.00	0.00	50.59
LLAVA-Med (Li et al., 2024b)	3.55	1.96	53.28	0.00	0.00	49.49
Qwen2-VL-7B (Wang et al., 2024a)	1.99	0.22	53.50	0.00	0.00	47.80
MiniGPTv2 (Chen et al., 2023)	0.00	0.06	26.18	0.00	0.00	22.97
LLAVA-1.5-13B (Liu et al., 2024a)	2.31	1.09	45.04	0.00	0.00	48.09
SPHINX (Lin et al., 2023a)	1.37	0.14	47.23	0.00	0.00	44.40
Qwen2.5-VL-7B (Bai et al., 2025a)	27.48	20.44	63.95	39.19	17.30	64.15
LLaMA-3.2-11B (Grattafiori et al., 2024)	24.05	16.52	56.28	22.46	14.34	53.89
Gemma-3-12B (Team et al., 2025)	36.37	23.17	53.19	23.09	16.58	48.91
VisualBert (Li et al., 2019)	62.68	33.29	73.91	40.05	33.81	70.73
VisualBert ResMLP (Seenivasan et al., 2022)	63.01	33.90	73.52	41.90	33.70	71.37
MCAN (Ben-Younes et al., 2017)	62.85	33.38	75.26	41.37	29.32	70.29
VQA-DeiT (Touvron et al., 2021)	61.04	31.56	73.41	37.97	28.58	69.09
MUTAN (Ben-Younes et al., 2017)	62.83	33.95	76.39	42.42	34.82	72.18
MFH (Yu et al., 2018)	62.83	32.54	75.92	41.03	35.00	72.16
BlockTucker (Ben-Younes et al., 2019)	62.01	32.86	76.53	42.21	35.15	72.88
CAT-ViL DeiT (Bai et al., 2023a)	64.52	33.21	77.05	44.91	**36.22**	73.22
GVLE-LViT (Bai et al., 2023b)	66.59	36.14	76.25	45.76	24.89	72.75
Surgical-LVLM (Wang et al., 2025)	69.47	33.25	84.16	40.68	34.12	78.25
EndoChat	**71.47**	**43.74**	**86.89**	**55.51**	29.78	**86.62**

Table 5
Comparison experiments with zero-shot MLLMs (top) and specialized models (middle) in Single Phrase QA and Visual QA on Cholec80-VQA (Seenivasan et al., 2022) dataset.

Model	Single Phrase QA		Visual QA			
	Acc	F-score	BLEU-3	BLEU-4	CIDEr	METEOR
BiomedGPT (Zhang et al., 2024b)	8.23	3.37	5.80	2.58	0.0159	19.62
LLAVA-Med (Li et al., 2024b)	10.05	4.09	13.30	10.54	0.1115	30.32
Qwen2-VL-7B (Wang et al., 2024a)	12.32	5.73	3.67	1.99	0.0005	18.62
MiniGPTv2 (Chen et al., 2023)	0.00	0.00	1.57	1.03	0.0107	7.56
LLAVA-1.5-13B (Liu et al., 2024a)	9.99	5.52	8.44	5.90	0.0656	18.10
SPHINX (Lin et al., 2023a)	11.67	4.12	5.18	1.13	0.0741	19.30
Qwen2.5-VL-7B (Bai et al., 2025a)	42.36	26.78	1.30	0.64	0.0169	10.34
LLaMA-3.2-11B (Grattafiori et al., 2024)	45.28	10.63	4.48	3.36	0.0501	33.91
Gemma-3-12B (Team et al., 2025)	40.30	19.85	3.36	1.84	0.0142	14.36
MedFuse (Sharma et al., 2021)	86.10	30.90	37.80	33.30	1.2501	22.20
VisualBert (Li et al., 2019)	89.70	63.30	96.30	95.60	8.8020	71.90
VisualBert ResMLP (Seenivasan et al., 2022)	89.80	**63.40**	96.00	95.20	8.7592	71.10
Surgical-LVLM (Wang et al., 2025)	87.53	60.10	96.00	95.13	8.7755	70.88
EndoChat	**92.05**	61.64	**97.28**	**96.81**	**9.6702**	**72.16**

Table 6
Comparison experiments with zero-shot MLLMs in Visual QA, Region based QA, and detailed description on Surg-396K dataset.

Dataset	Model	Visual QA					Region based QA					Detailed description
		BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-L	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-L	GPT-4 Score
EndoVis-18	BiomedGPT (Zhang et al., 2024b)	6.59	0.7301	13.07	37.17	26.39	2.20	0.1073	13.35	28.41	27.66	38.14
	LLAVA-Med (Li et al., 2024b)	13.54	1.1236	20.44	54.92	36.16	4.70	0.1595	17.35	37.23	35.84	46.40
	Qwen2-VL-7B (Wang et al., 2024a)	2.39	0.5803	11.71	50.72	43.73	4.09	0.2132	17.27	43.64	39.87	46.03
	MiniGPTv2 (Chen et al., 2023)	1.05	0.0235	12.89	52.33	25.25	0.88	0.0157	8.09	6.56	2.21	18.03
	LLAVA-1.5-13B (Liu et al., 2024a)	4.91	0.3627	15.93	41.21	32.94	3.19	0.2008	16.88	42.09	39.18	25.49
	SPHINX (Lin et al., 2023a)	15.11	0.7862	15.53	32.18	30.14	2.57	0.1024	5.38	6.41	5.83	43.99
	Qwen2.5-VL-7B (Bai et al., 2025a)	1.50	0.0112	12.17	52.16	42.66	0.83	0.0009	10.88	51.11	43.96	23.33
	LLaMA-3.2-11B (Grattafiori et al., 2024)	3.91	0.0842	36.82	65.12	62.20	2.35	0.2226	27.50	48.03	44.95	22.27
	Gemma-3-12B (Team et al., 2025)	1.67	0.0069	10.56	60.61	53.89	0.78	0.0013	6.56	50.42	45.81	12.77
	EndoChat	**52.20**	**5.9904**	**40.11**	**81.20**	**79.62**	**59.65**	**5.5735**	**41.05**	**82.01**	**81.21**	**79.35**
EndoVis-17	BiomedGPT (Zhang et al., 2024b)	8.81	0.6362	15.16	40.61	32.34	3.57	0.1874	5.99	25.74	24.02	56.06
	LLAVA-Med (Li et al., 2024b)	12.92	0.8814	17.13	43.71	36.36	9.03	0.3583	17.27	42.07	37.18	64.71
	Qwen2-VL-7B (Wang et al., 2024a)	10.97	1.1381	20.26	43.69	42.71	7.26	0.4505	19.42	45.22	37.93	67.57
	MiniGPTv2 (Chen et al., 2023)	1.80	0.0114	12.89	39.08	18.90	1.08	0.0256	9.80	10.16	9.36	29.35
	LLAVA-1.5-13B (Liu et al., 2024a)	13.73	0.7942	17.78	47.23	37.38	8.72	0.5145	18.09	46.36	41.06	41.75
	SPHINX (Lin et al., 2023a)	14.12	1.2109	16.74	42.15	37.32	3.34	0.0938	6.96	16.12	13.31	54.58
	Qwen2.5-VL-7B (Bai et al., 2025a)	4.33	0.0513	16.65	54.69	41.87	1.65	0.0012	15.38	55.90	44.09	28.73
	LLaMA-3.2-11B (Grattafiori et al., 2024)	5.76	0.2259	34.88	57.02	49.07	3.18	0.0219	29.43	53.41	45.41	21.29
	Gemma-3-12B (Team et al., 2025)	3.34	0.0319	14.22	51.94	41.67	2.69	0.0118	12.19	50.85	45.24	20.32
	EndoChat	**21.75**	**1.5083**	**23.41**	**52.05**	**46.65**	**18.12**	**1.4149**	**21.25**	**48.34**	**43.91**	**68.67**
CoPESD	BiomedGPT (Zhang et al., 2024b)	1.62	0.0064	5.88	19.23	16.25	1.69	0.0173	7.17	25.38	22.46	34.31
	LLAVA-Med (Li et al., 2024b)	4.56	0.2133	14.08	42.78	35.37	6.68	0.1489	17.42	50.70	44.04	71.10
	Qwen2-VL-7B (Wang et al., 2024a)	3.03	0.3411	14.51	48.81	38.20	4.69	0.0545	16.88	54.14	45.13	51.54
	MiniGPTv2 (Chen et al., 2023)	2.45	0.0936	15.03	37.92	35.45	3.75	0.0055	8.16	29.36	31.25	27.16
	LLAVA-1.5-13B (Liu et al., 2024a)	4.51	0.1774	14.99	45.98	35.74	6.14	0.1594	18.01	52.23	45.31	49.00
	SPHINX (Lin et al., 2023a)	7.03	0.2601	14.98	42.30	34.75	6.19	0.0194	2.53	5.58	5.01	38.85
	Qwen2.5-VL-7B (Bai et al., 2025a)	0.70	0.0282	9.96	51.92	37.47	0.61	0.0053	9.12	52.19	41.77	24.74
	LLaMA-3.2-11B (Grattafiori et al., 2024)	1.86	0.0425	26.39	52.50	48.62	2.07	0.0303	24.81	58.96	60.08	7.84
	Gemma-3-12B (Team et al., 2025)	1.73	0.0029	12.54	46.97	39.37	2.71	0.0038	12.26	51.20	49.69	29.81
	EndoChat	**46.94**	**3.2134**	**39.61**	**73.56**	**66.79**	**49.79**	**3.4410**	**38.04**	**71.98**	**65.44**	**82.48**

Table 7
Comparison experiments with zero-shot medical-specialized MLLMs in various surgical scene understanding tasks (excluding the Description task, which has been shown in Table 6) on the CoPESD part of Surg-396K dataset. Six types of tasks include: the number of instruments, the object location in the surgical scene (textual form), the current motion of the instrument, the direction of instruction motion, the identification of instruments, instrument detection, the recognition of issues and issue detection.

Model	CoPESD							
	Instrument number		Object position		Instrument motion		Motion direction	
	Acc	F-score	Acc	F-score	Acc	F-score	Acc	F-score
BiomedGPT (Zhang et al., 2024b)	78.84	24.03	12.92	9.86	13.41	5.61	7.68	3.06
LLAVA-Med (Li et al., 2024b)	49.88	14.05	26.66	17.62	17.13	7.78	14.52	4.90
EndoChat	**85.14**	**32.57**	**39.88**	**17.80**	**68.53**	**32.47**	**43.21**	**22.64**
	Instrument category				Target tissue			
	Acc	F-score	AP@50	mIoU	Acc	F-score	AP@50	mIoU
BiomedGPT (Zhang et al., 2024b)	20.86	15.84	27.60	22.63	56.01	25.16	25.54	20.41
LLAVA-Med (Li et al., 2024b)	64.12	42.47	62.46	58.21	86.95	46.51	38.63	36.71
Yolov11 (Khanam and Hussain, 2024)	\	\	90.29	82.63	\	\	96.52	85.28
EndoChat	**91.78**	**91.77**	**99.78**	**93.70**	**97.54**	**94.09**	**98.79**	**93.52**

ResMLP. This highlights EndoChat's effectiveness in tasks that require both fine-grained understanding and detailed responses. Other than that, EndoChat also exhibits a clear advantage in other parts of Surg-396K, which is presented in Table 6. Its results highlight a significant enhancement in metrics such as BLEU-4 and METEOR. Compared to Qwen2.5-VL and other SOTA MLLMs, which achieve relatively lower scores, EndoChat demonstrates a more comprehensive capability in extracting and reasoning over multimodal information. EndoChat sets a new benchmark for surgical assistance and educational applications. Its strong performance across key evaluation metrics solidifies its position as a top choice for clinical and academic use in surgical domains.

4.5. EndoChat advances comprehensive understanding of surgical scenarios

In order to comprehensively evaluate the capacity of EndoChat in addressing the challenges inherent in surgical environments, we conduct an in-context learning comparison between EndoChat and medical-specialized MLLMs in seven surgical scene understanding tasks on the CoPESD part of the Surg-396K dataset. The CoPESD part is chosen since it encompasses the full range of surgical understanding challenges. The tasks are formulated based on the dataset attributes illustrated in Fig. 2(a), with each attribute defining one or two corresponding tasks (Li et al., 2024a). These tasks reflect the essential components of surgical scenarios, such as instrument recognition, motion understanding, and issue detection. In-context learning is utilized for its advantage to dynamically adapt to new tasks and queries by utilizing task-specific prompts like "*The answer must be one of the following words or phrases: 'Reach', 'Rotate', 'Grasp', 'Lift', 'Hold', 'Stay idle', 'Dissect'.*". As summarized in Table 7, EndoChat consistently outperformed other medical-specialized MLLMs, showcasing its adaptability and precision in handling diverse understanding challenges.

Firstly, tasks such as instrument counting and object localization highlight fundamental scene comprehension abilities. While all models achieved reasonable performance in instrument counting, EndoChat leads with an accuracy of 85.14%, surpassing BiomedGPT (78.84%) and LLaVA-Med (49.88%). For object localization, EndoChat's accuracy (39.88%) exceeded BiomedGPT by over 26%, demonstrating its superior integration of spatial and semantic cues. Secondly, in more complex tasks such as motion recognition and direction prediction, EndoChat achieved the highest accuracy and F-scores, indicating its robustness in dynamic and context-sensitive scenarios. These results suggest that EndoChat effectively deciphers intricate spatial relationships within surgical scenes. Furthermore, in instrument category identification and target tissue recognition, EndoChat also achieves excellent performance, surpassing all other models. These findings demonstrate its proficiency in distinguishing both instruments and associated anatomical targets in complex surgical environments. Overall, these findings highlight EndoChat's capacity to generalize across diverse task types, making it a reliable tool for real-world surgical training and guidance.

4.6. Comparative analysis with fine-tuned MLLMs

To further verify the advantages of EndoChat under matched supervision, we fine-tune several SOTA open-source MLLMs with comparable parameter counts on the Surg-396K dataset and evaluate them across all five surgical conversation types. As shown in Table 8, these models exhibit substantial performance gains compared to their zero-shot baselines, demonstrating that Surg-396K offers effective supervision for generic architectures and can significantly enhance their ability to understand endoscopic surgical scenes. Despite consistent improvement, EndoChat still outperforms these fine-tuned competitors in most cases.

This performance gap is particularly evident in high-level reasoning tasks such as Detailed Description and Visual QA, where EndoChat exhibits superior GPT-4 evaluation scores and semantic consistency. While Qwen2.5-VL and GLM-4V improve substantially after fine-tuning, their generative outputs often suffer from incomplete alignment between visual content and language output, especially when interpreting spatial dynamics or subtle tissue-tool interactions. In contrast, EndoChat demonstrates strong alignment between visual grounding and language generation, as reflected by higher CIDEr and ROUGE-L scores in Visual QA tasks. EndoChat also demonstrates strong performance in Single Phrase QA and Grounding QA, where precise identification of scene attributes is critical. Benefiting from the Mixed Visual Token Engine, which aggregates multi-scale features from high-resolution inputs, EndoChat effectively captures both semantic and spatial cues across the full scene. However, in Region Based QA, the advantage of EndoChat narrows. Models like Gemma-3 and LLaMA-3.2, with simpler patch-based encoders and stronger reliance on explicit region prompts, may better preserve localized spatial focus in these tasks. Overall, these results highlight the effectiveness of EndoChat's tailored architecture, underscoring its advantage as a task-adaptive and semantically grounded MLLM for surgical applications.

To further assess the generalization capability of our model, we evaluate EndoChat and SOTA MLLMs on unrelated datasets, including PSI-AVA-VQA (Seenivasan et al., 2023) and EndoVis-17-VQLA-Extend (Bai et al., 2025b), after training on the Surg-396K dataset. As shown in Table 9, EndoChat achieves the best overall performance across both datasets, with the highest accuracy (39.92%) and recall (31.06%) on PSI-AVA-VQA, and the best accuracy (37.99%), F-score (26.39%), and mIoU (86.62%) on EndoVis-17-VQLA-Extend. These results confirm that EndoChat's surgical scene understanding capabilities are not confined to the training data but can generalize to novel, real-world surgical scenarios, making it a reliable tool for surgical education and guidance in diverse environments.

Table 8
Comparison experiments with fine-tuned MLLMs on Surg-396K dataset.

Dataset	Model	Single Phrase QA		Visual QA					Region based QA					Grounding QA		Detailed description
		Acc	F-score	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-L	BLEU-4	CIDEr	METEOR	ROUGE-1	ROUGE-L	AP@50	mIoU	GPT-4 Score
EndoVis-18	Qwen2.5-VL-7B (Bai et al., 2025a)	71.98	38.67	49.65	5.5663	39.26	81.09	78.64	58.87	5.0427	42.10	80.87	79.04	90.59	84.23	71.9
	LLaMA-3.2-11B (Grattafiori et al., 2024)	**72.59**	42.99	49.87	4.5486	38.69	80.97	78.36	59.56	5.3736	42.58	81.53	80.42	88.87	79.78	63.52
	Gemma-3-12B (Team et al., 2025)	68.58	37.56	52.15	5.9191	38.69	80.97	78.36	**60.64**	5.4801	**43.39**	81.05	80.64	88.99	78.46	65.32
	GLM-4V-9B (GLM et al., 2024)	45.76	15.15	37.42	3.1599	29.83	56.91	54.47	35.57	3.2589	28.89	59.41	57.62	52.99	51.44	57.86
	EndoChat	71.47	**43.74**	**52.20**	**5.9904**	**40.11**	**81.20**	**79.62**	59.65	**5.5735**	41.05	**82.01**	**81.21**	**93.22**	**86.89**	**79.35**
EndoVis-17	Qwen2.5-VL-7B (Bai et al., 2025a)	54.99	27.49	**22.58**	1.4695	23.28	**53.46**	**48.71**	16.20	1.0655	19.71	43.59	38.59	87.29	84.19	57.97
	LLaMA-3.2-11B (Grattafiori et al., 2024)	54.24	29.24	22.16	1.4606	23.17	51.75	45.12	**18.78**	1.4055	**23.90**	47.85	43.16	**90.68**	81.27	59.04
	Gemma-3-12B (Team et al., 2025)	51.06	28.22	21.55	1.3054	22.71	51.16	45.85	17.70	1.3948	21.81	48.04	42.70	86.44	76.60	63.97
	GLM-4V-9B (GLM et al., 2024)	39.19	25.37	18.16	1.0888	18.96	47.99	42.32	14.31	0.4845	17.32	42.37	37.24	38.14	48.01	32.04
	EndoChat	**55.51**	**29.78**	21.75	**1.5083**	**23.41**	52.05	46.65	18.12	**1.4149**	21.25	**48.34**	**43.91**	90.25	**86.62**	**68.67**
CoPESD	Qwen2.5-VL-7B (Bai et al., 2025a)	75.26	31.03	38.91	3.1714	34.84	67.93	59.04	46.19	3.3689	38.13	70.95	64.48	99.57	92.88	76.89
	LLaMA-3.2-11B (Grattafiori et al., 2024)	75.04	31.38	46.15	2.9494	36.51	69.09	63.01	49.52	3.4331	**39.95**	**73.83**	**68.73**	**99.58**	88.69	72.52
	Gemma-3-12B (Team et al., 2025)	75.29	**31.72**	46.70	3.1247	37.41	70.57	66.46	48.77	3.4177	39.26	72.60	68.33	99.19	88.84	76.09
	GLM-4V-9B (GLM et al., 2024)	74.44	28.58	43.83	2.9476	37.82	69.64	62.85	44.96	3.2398	37.61	71.34	63.98	94.88	78.54	46.13
	EndoChat	**75.34**	31.18	**46.94**	**3.2134**	**39.61**	**73.56**	**66.79**	**49.79**	**3.441**	38.04	71.98	65.44	99.43	**93.64**	**82.48**

Table 9
Generalization experiments with fine-tuned MLLMs on PSI-AVA-VQA and EndoVis-17-VQLA-Extend dataset.

Model	PSI-AVA-VQA (Seenivasan et al., 2023)			EndoVis-17-VQLA-Extend (Bai et al., 2025b)		
	Acc	Recall	F-score	Acc	F-score	mIoU
Qwen2.5-VL-7B	0.3964	0.1739	**0.0951**	0.3769	0.2422	84.19
LLaMA-3.2-11B	0.3374	0.145	0.0907	0.3616	0.2077	81.27
Gemma-3-12B	0.3366	0.1113	0.0811	0.3404	0.1183	76.6
GLM-4V-9B	0.1612	0.1055	0.0152	0.2613	0.1649	78.54
EndoChat	**0.3992**	**0.3106**	0.0655	**0.3799**	**0.2639**	**86.62**

Table 10
Ablation study with zero-shot medical-specialized MLLMs on EndoVis-18 part of Surg-396K dataset.

Mixed visual token engine	Single phrase		Grounding QA		Detail description
	Acc	F-score	AP@50	mIoU	GPT-4 Score
×	66.74%	33.75%	93.03%	86.86%	78.26
✓	71.47%	43.74%	93.22%	86.89%	79.35
Hallucination Mitigation	**Visual QA**		**Region Based QA**		**Detail Description**
	CIDEr	ROUGE-L	CIDEr	ROUGE-L	GPT-4 Score
×	6.01	79.47%	5.41	80.83%	77.39
✓	5.99	79.62%	5.57	81.21%	79.35

4.7. Ablation study for the effectiveness of core modules in EndoChat

In the ablation study, we evaluate the effectiveness of our proposed modules in EndoChat: Mixed Visual Token Engine and Visual Contrast Hallucination Mitigation, utilizing the EndoVis-18 subset of the Surg-396K dataset. The results are shown in Table 10. For the evaluation of MVTE, we focus on three conversation types: Single Phrase, Grounding QA, and Detail Description, since these are particularly sensitive to the quality of image feature extraction and perception. Specifically, for the Single Phrase, accuracy increases from 66.74% to 71.47% and F-score from 33.75% to 43.74%, demonstrating that MVTE significantly enhances the model's ability to generate more accurate phrase-level descriptions. While the improvement in Grounding QA is modest, the increase in GPT-4 score suggests that MVTE contributes to refining the model's reasoning capabilities, particularly for complex surgical scenarios. These results indicate that MVTE strengthens the model's capacity to capture high-quality image features, thus improving performance in tasks requiring fine-grained image perception.

The Hallucination Mitigation module is evaluated on conversation types that are particularly prone to hallucinations, including Visual QA, Region-Based QA, and Detail Description. These conversation types are susceptible to the model generating irrelevant or inaccurate responses due to the challenge of aligning visual content with textual queries. In Visual QA, this module leads to an increase in CIDEr (from 5.9068 to 5.9904) and ROUGE-L (from 79.47% to 79.62%), indicating improvements in the semantic accuracy and contextual relevance of the generated responses. The effect is more pronounced in Region Based QA, where CIDEr increases from 5.4069 to 5.5735 and ROUGE-L rises from 80.83 to 81.21, demonstrating that the hallucination mitigation module significantly improves the model's ability to produce reliable, region-specific answers. Additionally, the increase in the GPT-4 score (from 77.39 to 79.35) further underscores this module's contribution to enhancing the model's overall reasoning capabilities. These findings highlight the critical role of hallucination mitigation in improving performance across visual grounding tasks.

4.8. Expert evaluation of EndoChat by endoscopists

To validate EndoChat's potential in advancing surgical training and education, we conduct an expert evaluation involving 150 endoscopic surgery cases, evaluated by experienced endoscopists from Qilu Hospital. Each surgery case comprised a surgical image along with the corresponding five rounds of conversation. To ensure the evaluation's comprehensiveness, the five rounds of conversation included a detailed description of the scenario, supplemented by four rounds of randomly selected Visual QA and Region Based QA, which encompass all attributes of the surgical data in Surg-396K. Additionally, the ground-truth of these conversations is provided to assist the endoscopist in assessing EndoChat's descriptive accuracy, analytical depth, and applicability in training scenarios.

During the endoscopist evaluation, the conversations generated by EndoChat are presented with the indication that the results are produced by MLLMs. The subsequent process is to assess the usability of EndoChat by comparing its generated outputs with the correct answers. Endoscopists then evaluated the results and assigned scores to each case for the following standards:

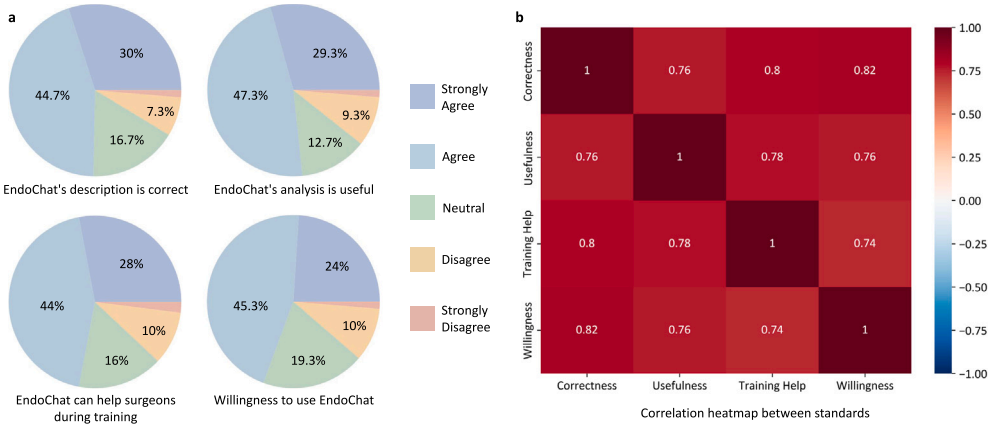


Fig. 4. Endoscopist evaluation of EndoChat in 150 cases. a Questionnaire-based evaluation of EndoChat conducted by endoscopists. The pie charts illustrate the distribution of cases in which endoscopists express varying levels of agreement. **b** Correlation analysis of four evaluation standards.

- EndoChat's description is correct.
- EndoChat's analysis is useful.
- EndoChat can help surgeons during training.
- Willingness to use EndoChat.

As shown in Fig. 4(a), the scores ranged from strongly agree to strongly disagree, and 74.7% of cases are evaluated as having correct descriptions provided by EndoChat, while 76.6% of cases feature useful analysis that enhances the understanding of surgical scenes. Additionally, for 72% of the cases, endoscopists agreed that EndoChat could effectively assist trainees in surgical training, helping to refine procedural skills and improve educational outcomes. Finally, 69.3% of the cases reflected a willingness to incorporate EndoChat into surgical training, indicating its potential for real-world adoption. These findings highlight EndoChat's role as a reliable tool and its effectiveness in advancing surgical training procedures and education in endoscopic surgery. Other than that, Fig. 4(b) illustrates the pairwise correlations among the evaluation standards. A strong positive correlation is observed between the correctness of answers and the willingness to use EndoChat, highlighting that its capacity to provide accurate and reliable information directly drives its potential for real-world implementation. Additionally, the training helps exhibit a robust correlation with both the usefulness of analysis and correctness, indicating that insightful analysis is pivotal for acceptance and education. These relationships emphasize that EndoChat's ability to generate precise, contextually relevant outputs aligns with the endoscopists' need for training support. Given the excellent performance demonstrated in experiments, EndoChat shows clear advantages in bridging the gap between AI innovation and surgical practice through such correlations, further solidifying its value as a versatile and impactful tool.

5. Discussion

This study aims to develop an intelligent surgical chatbot and copilot for surgical education and training. We first construct a high-quality, multi-paradigm dataset, Surg-396K, for surgical scene understanding and dialogue, along with a comprehensive framework for vision-language data collection and annotation in surgical scenarios. Furthermore, we develop the enhanced visual representation extraction and inference strategy, which serves as the foundation for EndoChat, our chatbot system designed to perform multimodal understanding and dialogue in surgical contexts.

Our analysis and comparisons include ten MLLMs and ten specialized models, showing that our model achieves outstanding performance across various dialogue paradigms and surgical scene understanding tasks. We further validate the effectiveness of our proposed visual feature learning approach and the visual contrast-based MLLM reasoning method through ablation experiments. Additionally, we invite experienced endoscopists to evaluate their willingness to use EndoChat as an assistant during training. In most evaluation cases, the surgeons provide positive feedback, further showcasing the clinical reliability, usability, and acceptability of our proposed EndoChat. There are demonstrations of qualitative comparison in the supplementary material.

Generally, there are three key factors in designing a precise and surgeon-friendly chatbot for dialogue and usage. The first is to ensure accuracy in downstream tasks for surgical scene understanding, such as instrument recognition and action identification (Saab et al., 2024). To enhance EndoChat's performance in these downstream tasks, we consider single-phrase QA to be the most critical. This is primarily because the answers in such dialogue data are often simple words or short phrases, making it easier for the model to link visual information with text annotations, thereby achieving higher accuracy in sub-tasks. The second is making the chatbot better suited to how surgeons use it, which is a main focus of this study. Establishing different dialogue paradigms helps respond to questions from surgeons and trainees in various contexts, and also helps constrain the divergent dialogue tendencies typical of MLLMs. This allows the model to focus more on the questions posed by surgeons and provide relevant answers. Finally, deployment is another crucial aspect when considering real-world usage in clinical environments. Therefore, we evaluate the practical feasibility of deploying EndoChat by measuring its inference resource consumption on a single NVIDIA RTX A6000 GPU (48 GB). Under typical usage, the model requires approximately 10,269 MB of GPU memory to process a single 1024×1024 resolution input image in real time, without memory overflow or batch accumulation. In addition, the average inference times for five conversation types are as follows: *Single Phrase* (0.45 s), *Detailed Description* (7.52 s), *Visual QA* (0.76 s), *Region-Based QA* (0.71 s), and *Grounding QA* (0.94 s). This demonstrates that EndoChat can be deployed on workstation-grade or high-end consumer GPUs to provide real-time guidance without the need for distributed inference infrastructure.

Despite the impressive performance of our EndoChat on various surgical dialogue tasks, it still faces several limitations. First, although we possess a large surgical image database, the number of unique surgical cases included is relatively small, which may affect its generalizability to a wider range of surgical techniques. While the dataset includes a large number of image-text pairs, which is beneficial for tasks such as instrument recognition, motion understanding, and spatial reasoning, the diversity of surgical cases is limited. This restricts the model's ability to generalize to surgeries involving different anatomical regions, pathological conditions, or procedural techniques. For example, endoscopic procedures that require specialized workflows or involve rare anatomical variations may not be adequately represented in the dataset, leading to reduced performance when applied to such

scenarios (Wang et al., 2023; Goetz et al., 2024). Moreover, the lack of case diversity could hinder the model's robustness in adapting to novel surgical environments or less frequently performed procedures, where the visual and contextual cues might differ significantly from the dataset. To overcome this, future work should prioritize expanding the dataset to include a broader range of surgical cases, encompassing various surgical specialties, techniques, and patient demographics. This would improve the model's adaptability and ensure its applicability across diverse clinical settings, finally enhancing its utility as a reliable tool in surgical training and guidance. Second, while EndoChat achieves SOTA performance and surpasses recent multimodal models in surgical scene understanding, we acknowledge that its architecture, built on SPHINX and LLaMA2-13B, may benefit from integration with more recent advances. Emerging models such as LLaMA-3 and Qwen3-VL offer enhanced multimodal reasoning, alignment, and efficiency. EndoChat's modular design — including the Mixed Visual Token Engine and visual contrast mechanism — provides a flexible foundation to incorporate these advances, potentially strengthening both reasoning capabilities and domain adaptability. Future work will explore such upgrades to sustain and amplify EndoChat's performance in line with the rapid progress of multimodal learning. In addition, MLLMs often rely on substantial computational power, which poses challenges for deployment in resource-constrained edge environments. Existing deployment approaches include developing lightweight MLLMs for deployment on mobile devices, or hosting the model in the cloud and enabling communication with mobile/computer terminals (Wang et al., 2024b; Yao et al., 2025). Lastly, as more and more diverse data are introduced, issues concerning the privacy and ethical use of clinical data need to be carefully studied and reviewed to ensure compliance during their application.

6. Conclusion

In this paper, we introduce Surg-396K, a comprehensive surgical multimodal dataset that includes 396K image-instruction pairs across multiple conversation paradigms. Based on Surg-396K, we present a flexible surgical understanding MLLM, EndoChat, designed to integrate various downstream tasks in surgical scene understanding and support different dialogue paradigms that may occur between surgeons and chatbots. Endochat integrates the Mixed Visual Token Engine (MVTE) and the visual contrast-based hallucination mitigation strategy, allowing it to effectively capture high-quality visual features and reduce inaccuracies in generated responses. Extensive experiments demonstrated the effectiveness of our approach, providing a more generalizable solution for understanding the surgical scene. Furthermore, the positive feedback from expert evaluations underscores the model's practical applicability in real-world surgical environments.

Moving forward, we will open-source our model weights, training code, and data to promote the development of multimodal AI systems in the surgical domain. In the future, we will collaborate with surgeons and clinical systems to conduct more rigorous and extensive validations to ensure the safety, reliability, and usability of the dialogue model. We aim to integrate EndoChat into surgical training or endoscopic surgery systems. By using monitors and voice-based dialogue systems, EndoChat could provide direct assistance to surgeons or trainees.

CRediT authorship contribution statement

Guankun Wang: Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Long Bai:** Writing – original draft, Methodology, Investigation, Conceptualization. **Junyi Wang:** Writing – original draft, Validation, Investigation, Data curation. **Kun Yuan:** Writing – original draft, Methodology. **Zhen Li:** Writing – review & editing, Validation. **Tianxu Jiang:** Writing – review & editing, Data curation. **Xiting He:** Writing – review & editing, Software. **Jinlin Wu:** Writing – review & editing, Resources. **Zhen Chen:** Writing – review & editing, Conceptualization. **Zhen Lei:** Writing – review & editing, Formal analysis. **Hongbin Liu:** Writing – review & editing, Formal analysis. **Jiazheng Wang:** Writing – review & editing, Conceptualization. **Fan Zhang:** Writing – review & editing, Conceptualization. **Nicolas Padoy:** Writing – review & editing, Data curation. **Nassir Navab:** Writing – review & editing, Data curation. **Hongliang Ren:** Writing – original draft, Supervision, Resources, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by HK RGC, Collaborative Research Fund (CRF C4026-21GF), General Research Fund (GRF 14203323, GRF 14216022, and GRF 14206125), NSFC/RGC Joint Research Scheme N_CUHK420/22, Research Grants Council (RGC) - Research Impact Fund (RIF) Grant R4020-22, InnoHK program and National Natural Science Foundation of China (No. 82261160396).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2025.103789.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Alabi, O., Vercauteren, T., Shi, M., 2025. Multitask learning in minimally invasive surgical vision: A review. Med. Image Anal. 103480.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. 35, 23716–23736.

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

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

Aziz, H., James, T., Remulla, D., Sher, L., Genyk, Y., Sullivan, M.E., Sheikh, M.R., 2021. Effect of COVID-19 on surgical training across the United States: a national survey of general surgery residents. J. Surg. Educ. 78 (2), 431–439.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al., 2025a. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Bai, L., Islam, M., Ren, H., 2023a. 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. Springer, pp. 397–407.

Bai, L., Islam, M., Seenivasan, L., Ren, H., 2023b. 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, IEEE, pp. 6859–6865.

Bai, L., Wang, G., Islam, M., Seenivasan, L., Wang, A., Ren, H., 2025b. Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. Inf. Fusion 113, 102602.

Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H., 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. 9-12 September 2019.

Ben-Younes, H., Cadene, R., Cord, M., Thome, N., 2017. Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2612–2620.

Ben-Younes, H., Cadene, R., Thome, N., Cord, M., 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (01), pp. 8102–8109.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Chen, K., Du, Y., You, T., Islam, M., Guo, Z., Jin, Y., Chen, G., Heng, P.-A., 2024a. LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery. In: 2024 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 10772–10778.

Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G., Wang, X., Cai, Z., Ji, K., Wan, X., et al., 2024b. Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 7346–7370.

Chen, R., Rodrigues Armijo, P., Krause, C., SAGES Robotic Task Force, Siu, K.-C., Oleynikov, D., 2020. A comprehensive review of robotic surgery curriculum and training for residents, fellows, and postgraduate surgical education. Surg. Endosc. 34, 361–367.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al., 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Sci. China Inf. Sci. 67 (12), 220101.

Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M., 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478.

Dou, Y., Zhao, X., Zou, H., Xiao, J., Xi, P., Peng, S., 2023. ShennongGPT: A tuning Chinese LLM for medication guidance. In: 2023 IEEE International Conference on Medical Artificial Intelligence. MedAI, IEEE, pp. 67–72.

Ferber, D., Wölflein, G., Wiest, I.C., Ligero, M., Sainath, S., Ghaffari Laleh, N., El Nahhas, O.S., Müller-Franzes, G., Jäger, D., Truhn, D., et al., 2024. In-context learning enables multimodal large language models to classify cancer pathology images. Nat. Commun. 15 (1), 10104.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al., 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793.

Goetz, L., Seedat, N., Vandersluis, R., van der Schaar, M., 2024. Generalization—a key challenge for responsible AI in patient-facing clinical applications. Npj Digit. Med. 7 (1), 126.

Gozalo-Brizuela, R., Garrido-Merchan, E.C., 2023. ChatGPT is not all you need. a state of the art review of large generative AI models. arXiv preprint arXiv:2301.04655.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M., 2018. Overview of imageclef 2018 medical domain visual question answering task. In: Proceedings of CLEF 2018 Working Notes.

Hou, W., Cheng, Y., Xu, K., Hu, Y., Li, W., Liu, J., 2024. Memory-augmented multimodal LLMs for surgical VQA via self-contained inquiry. arXiv preprint arXiv:2411.10937.

Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P., 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22170–22183.

Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al., 2022. LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations.

Huang, W., Li, C., Yang, H., Liu, J., Liang, Y., Zheng, H., Wang, S., 2024. Enhancing the vision–language foundation model with key semantic knowledge-emphasized report refinement. Med. Image Anal. 97, 103299.

Jin, J., Jeong, C.W., 2024. Surgical-LLaVA: Toward surgical scenario understanding via large language and vision models. In: Advancements in Medical Foundation Models: Explainability, Robustness, Security, and beyond.

Khanam, R., Hussain, M., 2024. Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.

Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y., 2024. Natural language understanding and inference with MLLM in visual question answering: A survey. ACM Comput. Surv..

Lanfredi, R.B., Mukherjee, P., Summers, R.M., 2025. Enhancing chest x-ray datasets with privacy-preserving large language models and multi-type annotations: a data-driven approach for improved classification. Med. Image Anal. 99, 103383.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L., 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882.

Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T.B., Zettlemoyer, L., Lewis, M., 2023a. Contrastive decoding: Open-ended text generation as optimization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 12286–12312.

Li, J., Li, D., Savarese, S., Hoi, S., 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. PMLR, pp. 19730–19742.

Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J., 2024a. LLaVA-Surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. Neural Inf. Process. Syst. 36.

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

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al., 2023a. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575.

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z., 2023b. Medical visual question answering: A survey. Artif. Intell. Med. 143, 102611.

Liu, H., Li, C., Li, Y., Lee, Y.J., 2024a. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2024b. Visual instruction tuning. Adv. Neural Inf. Process. Syst. 36.

Liu, X., Liu, H., Yang, G., Jiang, Z., Cui, S., Zhang, Z., Wang, H., Tao, L., Sun, Y., Song, Z., et al., 2025. A generalist medical language model for disease diagnosis assistance. Nature Med. 1–11.

Liu, D., Zhang, R., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., et al., 2024c. SPHINX-X: scaling data and parameters for a family of multi-modal large language models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 32400–32420.

Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., Lu, L., Liu, Z., Zheng, Y., Sun, X., et al., 2023. A medical multimodal large language model for future pandemics. Npj Digit. Med. 6 (1), 226.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations.

Lu, Y., Wang, A., 2025. Integrating language into medical visual recognition and reasoning: A survey. Med. Image Anal. 103514.

Mariani, A., Pellegrini, E., De Momi, E., 2020. Skill-oriented and performance-driven adaptive curricula for training in robot-assisted surgery using simulators: A feasibility study. IEEE Trans. Biomed. Eng. 68 (2), 685–694.

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., et al., 2025. Towards accurate differential diagnosis with large language models. Nature 1–7.

Meta, A., 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. p. 2025, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, Checked on, 4, (7).

Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al., 2023. CholecTriplet2021: A benchmark challenge for surgical action triplet recognition. Med. Image Anal. 86, 102803.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2024. DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. J..

Peng, P., Fan, W., Liu, W., Yang, X., Zhou, D., 2024. Prior-Posterior Knowledge Prompting-and-Reasoning for Surgical Visual Question Localized-Answering. In: 2024 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–9.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al., 2024. Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416.

Schmidgall, S., Kim, J.W., Kuntz, A., Ghazi, A.E., Krieger, A., 2024. General-purpose foundation models for increased autonomy in robot-assisted surgery. Nat. Mach. Intell. 6 (11), 1275–1283.

Seenivasan, L., Islam, M., Kannan, G., Ren, H., 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. Springer, pp. 281–290.

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

Sharma, D., Purushotham, S., Reddy, C.K., 2021. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. 11 (1), 19826.

Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., et al., 2025. A multimodal visual–language foundation model for computational ophthalmology. Npj Digit. Med. 8 (1), 1–13.

Shibata, Y., Kida, T., Fukamachi, S., Takeda, M., Shinohara, A., Shinohara, T., Arikawa, S., 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical Report DOI-TR-161, Department of Informatics, Kyushu University.

Skourti, E., 2025. A vision–language foundation model for clinical oncology. Nat. Cancer 1–1.

Song, X., Salcianu, A., Song, Y., Dopson, D., Zhou, D., 2021. Fast WordPiece tokenization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 2089–2103.

Song, M., Wang, J., Yu, Z., Wang, J., Yang, L., Lu, Y., Li, B., Wang, X., Wang, X., Huang, Q., et al., 2024. PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis. Med. Image Anal. 103248.

Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X., 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al., 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F., 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. pp. 440–448.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al., 2024. Towards generalist biomedical AI. Nejm Ai 1 (3), AIoa2300138.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36 (1), 86–97.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wagner, M., Müller-Stich, B.-P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al., 2023. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Med. Image Anal. 86, 102770.

Wang, G., Bai, L., Nah, W.J., Wang, J., Zhang, Z., Chen, Z., Wu, J., Islam, M., Liu, H., Ren, H., 2025. Surgical-LVLM: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. In: ICLR 2025 Workshop on Foundation Models in the Wild.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al., 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

Wang, G., Liu, J., Li, C., Zhang, Y., Ma, J., Wei, X., Zhang, K., Chong, M., Zhang, R., Liu, Y., et al., 2024b. Cloud-device collaborative learning for multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12646–12655.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T., 2023. Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975.

Wang, G., Xiao, H., Gao, H., Zhang, R., Bai, L., Yang, X., Li, Z., Li, H., Ren, H., 2024c. CoPESD: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. arXiv preprint arXiv:2410.07540.

Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., Shen, D., 2024d. Interactive computer-aided diagnosis on medical image using large language models. Commun. Eng. 3 (1), 133.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al., 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302.

Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al., 2024. Xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al., 2025a. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., et al., 2025b. Qwen2. 5-1m technical report. arXiv preprint arXiv:2501.15383.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Chen, C., Li, H., Zhao, W., et al., 2025. Efficient GPT-4V level multimodal large language model for deployment on edge devices. Nat. Commun. 16 (1), 5509.

Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., Wang, B., et al., 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. Adv. Neural Inf. Process. Syst. 37, 94327–94427.

Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Netw. Learn. Syst. 29 (12), 5947–5959.

Yuan, K., Kattel, M., Lavanchy, J.L., Navab, N., Srivastav, V., Padoy, N., 2024. Advancing surgical VQA with scene graph knowledge. Int. J. Comput. Assist. Radiol. Surg. 1–9.

Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N., 2025. Learning multi-modal representations by watching hundreds of surgical video lectures. Med. Image Anal. 103644.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986.

Zhang, Y., Fan, W., Peng, P., Yang, X., Zhou, D., Wei, X., 2024a. Dual modality prompt learning for visual question-grounded answering in robotic surgery. Vis. Comput. Ind. Biomed. Art 7 (1), 9.

Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al., 2024b. A generalist vision–language foundation model for diverse biomedical tasks. Nature Med. 30 (11), 3129–3141.

Zhao, L., Deng, Y., Zhang, W., Gu, Q., 2024. Mitigating object hallucination in large vision-language models via image-grounded guidance. arXiv preprint arXiv:2402.08680.

Zhou, Y., Bai, L., Cai, S., Deng, B., Xu, X., Shen, H.T., 2025a. TAU-106k: A new dataset for comprehensive understanding of traffic accident. In: The Thirteenth International Conference on Learning Representations.

Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., et al., 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat. Commun. 15 (1), 5649.

Zhou, Y., Song, L., Shen, J., 2025b. Training medical large vision-language models with abnormal-aware feedback. arXiv preprint arXiv:2501.01377.

Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M., 2024a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al., 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479.

Zhu, Z., Zhang, Y., Cheng, X., Huang, Z., Xu, D., Wu, X., Zheng, Y., 2024b. Alignment before awareness: Towards visual question localized-answering in robotic surgery via optimal transport and answer semantics. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 711–721.