EndoChat: Grounded Multimodal Large Language Model for Endoscopic Surgery

Guankun Wang[1,2,†], **Long Bai**[1,3,†], **Junyi Wang**[1,†], **Kun Yuan**[3,4,†], **Zhen Li**[5], **Tianxu Jiang**[1], **Xiting He**[1], **Jinlin Wu**[6], **Zhen Chen**[6], **Zhen Lei**[6], **Hongbin Liu**[6], **Jiazheng Wang**[2], **Fan Zhang**[2], **Nicolas Padoy**[4], **Nassir Navab**[3], **and Hongliang Ren**[1,*]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Theory Lab, Central Research Institute, 2012 Labs, Huawei Technologies Co. Ltd., Hong Kong SAR, China.
[3]Chair of Computer Aided Procedures (CAMP), Technical University of Munich, Munich, Germany
[4]University of Strasbourg, CNRS, INSERM, ICube & IHU Strasbourg, Strasbourg, France
[5]Department of Gastroenterology, Qilu Hospital of Shandong University, Jinan, China
[6]Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong SAR, China
[*]Corresponding author: H. Ren, hlren@ee.cuhk.edu.hk.
[†]These authors contributed equally to this work.

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 lack of MLLMs specialized for surgical scene understanding in clinical applications. In this work, we introduce EndoChat to address various dialogue paradigms and subtasks in surgical scene understanding that surgeons encounter. 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 collected 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 eight surgical scene understanding tasks. Additionally, we conduct evaluations with professional surgeons, most of whom provide positive feedback on collaborating with EndoChat. Overall, these results demonstrate that our EndoChat has great potential to significantly advance training and automation in robotic-assisted surgery.

生成了新的数据集surg-396k
在五个任务中sota

Introduction

Robot-assisted surgery (RAS) offers unprecedented opportunities to enhance surgical precision, minimize patient trauma, and shorten postoperative recovery times[1]. 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. To ensure surgical safety and efficacy, surgeons must undergo rigorous training to acquire and refine the core skills required for robotic operation[2,3]. To improve the efficiency of this training process, various simulator-based surgical platforms[4] 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[5,6]. 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[6–17]. These systems analyze visual data from surgical scenarios to address trainees' questions. As the first work to propose a question-answering (QA) model specifically designed for surgical scenarios, Surgical-VQA represents key elements in surgical scenes—such as instruments, tissues, tools, and spatial positions—through textual descriptions[6]. Using the VisualBERT[18] framework, it integrates multimodal representations of text and images and generates corresponding answers through a decoder. Subsequent works, such as Surgical-VQLA, enhance this approach by incorporating bounding box outputs at the decoder level, providing explicit visual localization to assist surgeons[7,9,12,17]. Later developments explored different network designs or inference paradigms. However, these models often rely on structured VQA datasets,
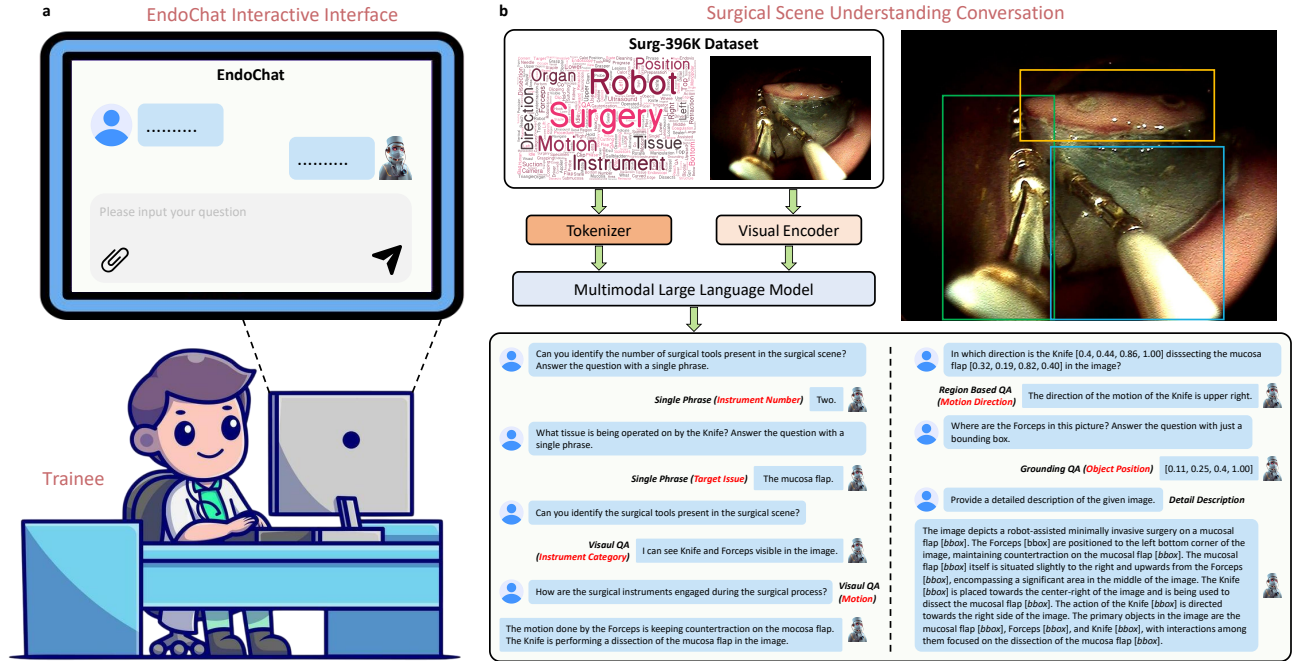
Figure 1. Overview of the EndoChat. **a** EndoChat is an interactive multimodal chatbot designed for surgical education and training. Users can interact with EndoChat by uploading images and formulating questions, enabling a comprehensive surgical scene understanding. **b** EndoChat is trained on Surg-396K, 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.

which are typically designed and trained for specific tasks. This reliance limits their ability to dynamically adapt to the wide range of questions posed by trainees[19,20]. 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 trained scope, these models often show significant performance drops, making them poorly suited for the complexity of diverse surgical scenarios[21,22]. 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.

Recently, 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[23–39]. 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 leveraging 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 real-time guidance from professional surgeons by simulating their knowledge and decision-making abilities. This could address the limitations of current training solutions and alleviate the burden on surgeons who are constrained by their demanding clinical schedules, ultimately improving the efficiency and quality of surgical training[26,40,41].

Some prior works have explored the capabilities of MLLMs in surgical scenarios[42–46]. For instance, LLaVA-Surg collects and annotates open-source videos and datasets, then enhances the model's conversational abilities through instruction tuning[42]. GP-VLS, on the other hand, unifies various tasks in surgical scenarios into a question-answering framework, modeling and expressing the results of different downstream surgical tasks in textual form[45]. However, we aim to further develop MLLMs to handle diverse queries from trainees in real-world scenarios. 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. These five distinct conversational paradigms encompass the majority of scenarios encountered in natural language dialogues[47]. By constructing them, we aim to design a surgical MLLM that can respond appropriately in different cases, creating a system that is better suited for human interaction. This would provide more practical and effective support for surgical training and education.

In this paper, we propose EndoChat (shown in Figure 1) to support diverse conversational paradigms in endoscopic surgical scenarios, including Single Phrase QA, Detailed Description, Visual QA, Region Based QA, and Grounding QA. This flexible framework addresses various interaction needs and accommodates a wide range of surgical tasks such as instrument recognition, motion recognition, target localization, tissue identification, instrument counting, and motion direction detection, which is summarized based on our designed surgical attributes and highly adaptable to various questions that the trainee might raise in different contexts. Our model is constructed based on SPHINX[48]. To better align the visual feature representation of surgical scenes, we introduce a novel strategy, Mixed Visual Token Engine (MVTE), to extract visual information at multiple scales. Unlike traditional frameworks that use pre-trained Vision Transformers (ViTs) to extract visual tokens, MVTE uses multiple visual towers to extract, interact, and fuse visual tokens. This approach improves the extraction of visual information before aligning it with text. Furthermore, to reduce hallucinations and enhance scene understanding in surgical tasks, we introduce a visual contrast-based method to suppress target hallucinations. To train our proposed EndoChat, we develop Surg-396K, a surgical multimodal instruction dataset specifically tailored for surgical contexts. This dataset is built upon publicly available Cholec80-VQA[6], EndoVis-VQLA[9], and CoPESD[49] datasets. Based on the original annotations provided by these datasets, we systematically extract surgical information through designed surgical attributes and generate instruction-tuning data with diverse conversational templates. Additionally, we expand the conversational data using ChatGPT-4. This process enabled us to simulate diverse questions that different surgeons might ask in various scenarios, covering five distinct conversational paradigms. As a result, Surg-396K provides comprehensive coverage of downstream tasks and conversational paradigms across a wide range of surgical scenarios. To validate the effectiveness of our proposed surgical understanding dialogue MLLM, we first conducted 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. Furthermore, we show that our model achieves state-of-the-art performance across various attribute-related sub-tasks of surgical scene understanding (e.g., instrument recognition and motion recognition, etc). 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.

贡献
1. 提出了MVTE混合视觉tokenizer
2. 通过一个新的策略加强了数据

Results

The system design of EndoChat

EndoChat is designed using an encoder-decoder architecture and optimized for surgical scene understanding and multi-modal interaction, as shown in Figure 4. The multi-modal encoder processes visual inputs, such as surgical images, and textual queries, mapping them into a shared latent space. Visual inputs are handled by the pre-trained mixed visual encoder that extracts semantic visual embeddings, while textual inputs are processed using large language models to understand human questions and queries. A multi-modal alignment layer integrates these visual embeddings with textual queries, synchronizing visual information and natural language understanding. The decoder, leveraging the fine-tuned LLaMA2-13B model[50] in SPHINX architecture[48], generates detailed, context-aware textual responses tailored to the query type. EndoChat is trained on Surg-396K, a diverse set of interactive instruction-tuning QA pairs, designed for surgical scene understanding and multi-modal interaction.

EndoChat tailors interactions to varying complexity levels of surgical conversation.

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 1, demonstrate that EndoChat outperforms the stat-of-the-art MLLMs such as BiomedGPT and LLaVA-Med. Specifically, on the EndoVis-17 part, where other models fail to answer questions effectively (0% 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). EndoChat also has superior performance on other parts of Surg-396K, and other public surgery datasets shown in Table 2 and 3, which demonstrates the robustness and effectiveness of the instruction-tuning process on our Surg-396K dataset.

Detailed Description caters to more complex scenarios where 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 4, 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 comprehensive advantage in generating detailed, context-aware, and high-quality descriptions.

Table 1. Comparison experiments with zero-shot MLLMs in Single Phrase QA and Grounding QA on three parts of 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[51]	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[52]	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[53]	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[54]	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[55]	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[48]	1.37	0.14	32.59	27.23	0.00	0.00	25.76	15.44	7.19	4.30	59.93	55.52
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**

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 hands-on training and surgical 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 Table 1 and 2, EndoChat outperforms other state-of-the-art 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 specialized models 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 4, 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.

EndoChat is a surgeon-like interaction tool

EndoChat leverages 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 2. Comparison experiments with zero-shot MLLMs (top) and specialized models (middle) in Single Phrase QA and Grounding QA on EndoVis-VQLA[9] dataset.

Model	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-score	mIoU	Acc	F-score	mIoU
BiomedGPT[51]	5.61	3.42	56.93	0.00	0.00	50.59
LLAVA-Med[52]	3.55	1.96	53.28	0.00	0.00	49.49
Qwen2-VL[53]	1.99	0.22	53.50	0.00	0.00	47.80
MiniGPTv2[54]	0.00	0.06	26.18	0.00	0.00	22.97
LLAVA-1.5[55]	2.31	1.09	45.04	0.00	0.00	48.09
SPHINX[48]	1.37	0.14	47.23	0.00	0.00	44.40
VisualBert[18]	62.68	33.29	73.91	40.05	33.81	70.73
VisualBert ResMLP[6]	63.01	33.90	73.52	41.90	33.70	71.37
MCAN[56]	62.85	33.38	75.26	41.37	29.32	70.29
VQA-DeiT[57]	61.04	31.56	73.41	37.97	28.58	69.09
MUTAN[56]	62.83	33.95	76.39	42.42	34.82	72.18
MFH[58]	62.83	32.54	75.92	41.03	35.00	72.16
BlockTucker[59]	62.01	32.86	76.53	42.21	35.15	72.88
CAT-ViL DeiT[7]	64.52	33.21	77.05	44.91	36.22	73.22
GVLE-LViT[9]	66.59	36.14	76.25	45.76	24.89	72.75
Surgical-LVLM[44]	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 3 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 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 4. Its results highlight a significant enhancement in metrics such as BLEU-4 and METEOR. Compared to SPHINX and other models, 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 these key evaluation metrics solidifies its position as a top choice for clinical and academic use in surgical domains.

EndoChat advances comprehensive understanding of surgical scenerios

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 eight 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 Figure 3 a, with each attribute defining one or two corresponding tasks[52]. 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 leveraging 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 5, 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

Table 3. Comparison experiments with zero-shot MLLMs (top) and specialized models (middle) in Single Phrase QA and Visual QA on Cholec80-VQA[6] dataset.

Model	Single Phrase QA		Visual QA			
	Acc	F-score	BLEU-3	BLEU-4	CIDEr	METEOR
BiomedGPT[51]	8.23	3.37	5.80	2.58	0.0159	19.62
LLAVA-Med[52]	10.05	4.09	13.30	10.54	0.1115	30.32
Qwen2-VL[53]	12.32	5.73	3.67	1.99	0.0005	18.62
MiniGPTv2[54]	0.00	0.00	1.57	1.03	0.0107	7.56
LLAVA-1.5[55]	9.99	5.52	8.44	5.90	0.0656	18.10
SPHINX[48]	11.67	4.12	5.18	1.13	0.0741	19.30
MedFuse[5]	86.10	30.90	37.80	33.30	1.2501	22.20
VisualBert[18]	89.70	63.30	96.30	95.60	8.8020	71.90
VisualBert ResMLP[6]	89.80	**63.40**	96.00	95.20	8.7592	71.10
Surgical-LVLM[44]	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 4. 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[51]	6.59	0.7301	13.07	37.17	26.39	2.20	0.1073	13.35	28.41	27.66	38.14
	LLAVA-Med[52]	13.54	1.1236	20.44	54.92	36.16	4.70	0.1595	17.35	37.23	35.84	46.40
	Qwen2-VL[53]	2.39	0.5803	11.71	50.72	43.73	4.09	0.2132	17.27	43.64	39.87	46.03
	MiniGPTv2[54]	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[55]	4.91	0.3627	15.93	41.21	32.94	3.19	0.2008	16.88	42.09	39.18	25.49
	SPHINX[48]	15.11	0.7862	15.53	32.18	30.14	2.57	0.1024	5.38	6.41	5.83	43.99
	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[51]	8.81	0.6362	15.16	40.61	32.34	3.57	0.1874	5.99	25.74	24.02	56.06
	LLAVA-Med[52]	12.92	0.8814	17.13	43.71	36.36	9.03	0.3583	17.27	42.07	37.18	64.71
	Qwen2-VL[53]	10.97	1.1381	20.26	43.69	42.71	7.26	0.4505	19.42	45.22	37.93	67.57
	MiniGPTv2[54]	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[55]	13.73	0.7942	17.78	47.23	37.38	8.72	0.5145	18.09	46.36	41.06	41.75
	SPHINX[48]	14.12	1.2109	16.74	42.15	37.32	3.34	0.0938	6.96	16.12	13.31	54.58
	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[51]	1.62	0.0064	5.88	19.23	16.25	1.69	0.0173	7.17	25.38	22.46	34.31
	LLAVA-Med[52]	4.56	0.2133	14.08	42.78	35.37	6.68	0.1489	17.42	50.70	44.04	71.10
	Qwen2-VL[53]	3.03	0.3411	14.51	48.81	38.20	4.69	0.0545	16.88	54.14	45.13	51.54
	MiniGPTv2[54]	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[55]	4.51	0.1774	14.99	45.98	35.74	6.14	0.1594	18.01	52.23	45.31	49.00
	SPHINX[48]	7.03	0.2601	14.98	42.30	34.75	6.19	0.0194	2.53	5.58	5.01	38.85
	EndoChat	**46.94**	**3.2134**	**39.61**	**73.56**	**66.79**	**49.79**	**3.4410**	**38.04**	**71.98**	**65.44**	**82.48**

distinguishing both instruments and their 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.

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 6. 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 (from 78.26 to 79.35) suggests that MVTE also 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, such

Table 5. Comparison experiments with zero-shot medical-specialized MLLMs in eight surgical scene understanding tasks on the CoPESD part of Surg-396K dataset. Eight 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[51]	78.84	24.03	12.92	9.86	13.41	5.61	7.68	3.06
LLaVA-Med[52]	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[51]	20.86	15.84	27.60	22.63	56.01	25.16	25.54	20.41
LLaVA-Med[52]	64.12	42.47	62.46	58.21	86.95	46.51	38.63	36.71
Yolov11[60]	\	\	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**

Table 6. Ablation study with zero-shot medical-specialized MLLMs on EndoVis-18 part of Surg-396K dataset.

Mixed Visual	Single Phrase		Grounding QA		Detail Description
Token Engine	Acc	F-score	acc@0.5	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	Visual QA		Region Based QA		Detail Description
Mitigation	CIDEr	ROUGE-L	CIDEr	ROUGE-L	GPT-4 Score
×	6.0068	79.47%	5.4069	80.83%	77.3853
✓	5.9904	79.62%	5.5735	81.21%	79.3463

as 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.

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 correct answers to these conversations are provided to assist the endoscopist in assessing EndoChat's descriptions, analyses, 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:

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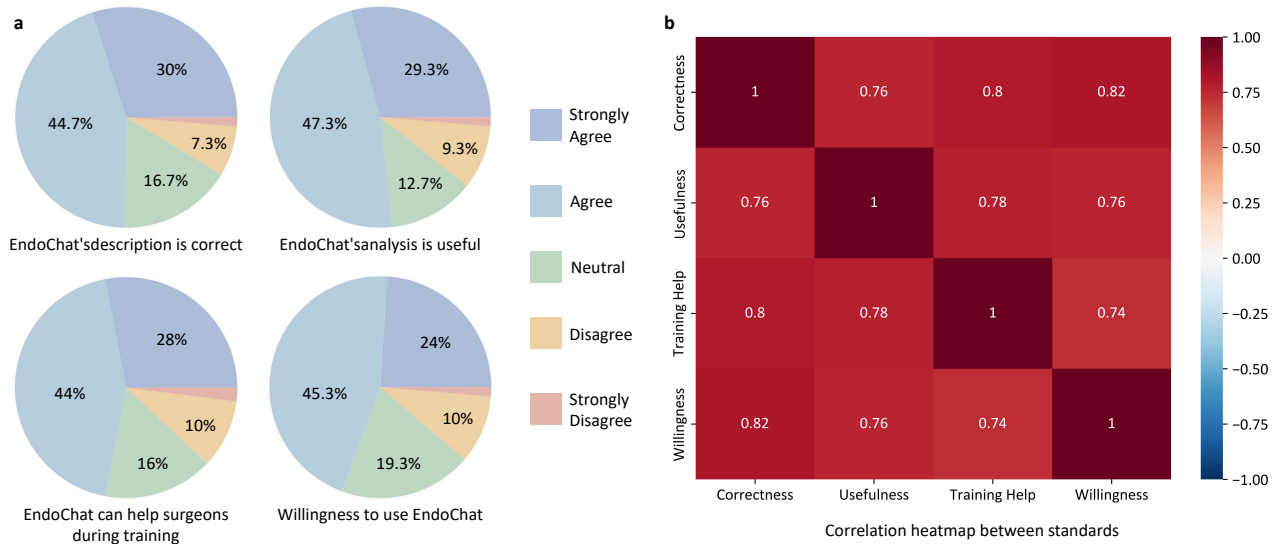- EndoChat can help surgeons during training.

Figure 2. Endoscopist evaluation of EndoChat in 150 cases. a Questionnaire-based evaluation of EndoChat conducted by endoscopists. The pie charts illustrate the distribution of cases that endoscopists express varying levels of agreement. **b** Correlation analysis of four evaluation standards.

- Willingness to use EndoChat.

As shown in Figure 2 a, the scores ranged from strongly agree to strongly disagree, and 74.7% cases are evaluated as having correct descriptions provided by EndoChat, while 76.6% 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, Figure 2 (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 help exhibits a robust correlation with both the usefulness of analysis and correctness, indicating that practical and 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 previous experiments, EndoChat shows clear advantages in bridging the gap between AI innovation and endoscopic surgical practice through such correlations, further solidifying its value as a versatile and impactful tool.

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 six MLLMs and over ten specialized models, demonstrating that our model achieved outstanding performance across various dialogue paradigms and surgical-specific 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 surgeons to evaluate their willingness to use EndoChat as an assistant during training or surgeries. In most evaluation cases, the surgeons provide positive feedback, further showcasing the clinical reliability, usability, and acceptability of our proposed EndoChat.

Generally, there are two 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. To enhance EndoChat's performance in these downstream tasks, we consider single-phrase QA to be the most

critical. This is 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 primary 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. In the future, we target to integrate EndoChat directly into surgical training or endoscopic surgery systems. By using monitors and voice-based dialogue systems, EndoChat could provide direct assistance to surgeons or trainees.

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. Such a large database facilitates tasks like action and instrument recognition; however, the limited number of cases may hinder the generalization ability of our model when applied to different surgical techniques. Expanding the database to include more surgical procedures and cases in the future could significantly enhance the generalizability and applicability of EndoChat across diverse surgical scenarios. 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 personal computers, or hosting the model in the cloud and enabling communication with mobile/computer terminals. Finding ways to deploy MLLMs in clinical settings with limited computational resources will remain a significant challenge. 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.

In conclusion, 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. Extensive experiments demonstrated the effectiveness of our approach, providing a more generalizable solution for understanding the surgical scene. Furthermore, 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.

Methods

Surg-396K: Surgical Multimodal Instruction Dataset

The AI-assisted surgery field has experienced a notable expansion in the availability of public multimodal datasets, particularly visual-question-answer (VQA) pairs, as evidenced by works ranging from[6] to[10]. 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 annotation for endoscopic surgery. Following the data generation and aggregation process shown in Figure 3, we compile a total of 341k image-instruction pairs for training and 55k for testing. The subsequent subsections will detail the generation process of Surg-396K.
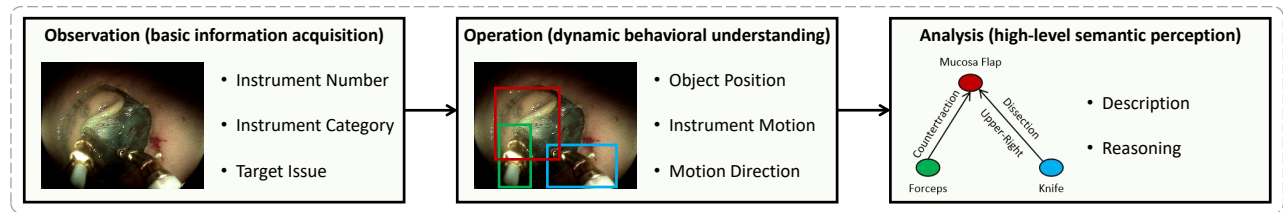
Preliminary for Constituent Datasets

In the construction of our Surg-396K dataset, we integrate three distinct datasets. Inspired by the achievements of recent MLLMs in text-annotation tasks[61], we utilize ChatGPT-4 to generate multimodal instruction-following data, resulting in five conversation types derived from EndoVis-VQLA[9] and CoPESD[49] datasets. The third dataset Cholec80-VQA[6], which lacks grounding information, is Surg-396K dataset.
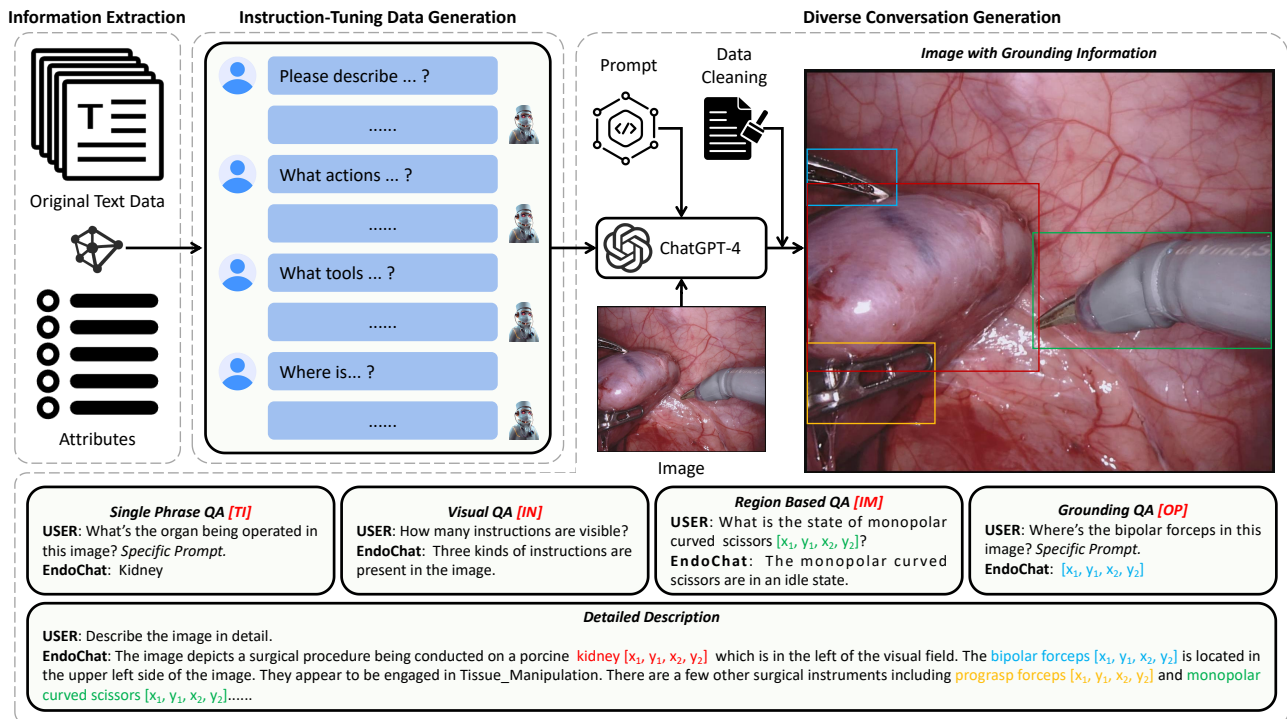
EndoVis-VQLA[9] is a publicly available dataset for endoscopic surgery, derived from the MICCAI Challenges of 2017[62] and 2018[63]. This dataset integrates VQA annotations with bounding box labels to create Visual Question Localized-Answering (VQLA) pairs, which encompass surgical actions, target tissues, 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[49] 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[6] is an innovative dataset generated from 40 video sequences of the Cholec80 dataset[64], 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 ChatGPT-4, but directly use Classification and Sentence as the conversation of the Single Phrase and Visual QA.

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

Single Phrase QA [TI]
USER: What's the organ being operated in this image? *Specific Prompt.*
EndoChat: Kidney

Visual QA [IN]
USER: How many instructions are visible?
EndoChat: Three kinds of instructions are present in the image.

Region Based QA [IM]
USER: What is the state of monopolar curved scissors $[x_1, y_1, x_2, y_2]$?
EndoChat: The monopolar curved scissors are in an idle state.

Grounding QA [OP]
USER: Where's the bipolar forceps in this image? *Specific Prompt.*
EndoChat: $[x_1, y_1, x_2, y_2]$

Detailed Description
USER: Describe the image in detail.
EndoChat: The image depicts a surgical procedure being conducted on a porcine kidney $[x_1, y_1, x_2, y_2]$ which is in the left of the visual field. The bipolar forceps $[x_1, y_1, x_2, y_2]$ is located in the upper left side of the image. They appear to be engaged in Tissue_Manipulation. There are a few other surgical instruments including prograsp forceps $[x_1, y_1, x_2, y_2]$ and monopolar curved scissors $[x_1, y_1, x_2, y_2]$......

(b) Construction pipeline of Surg-396K dataset with samples of various conversation types.
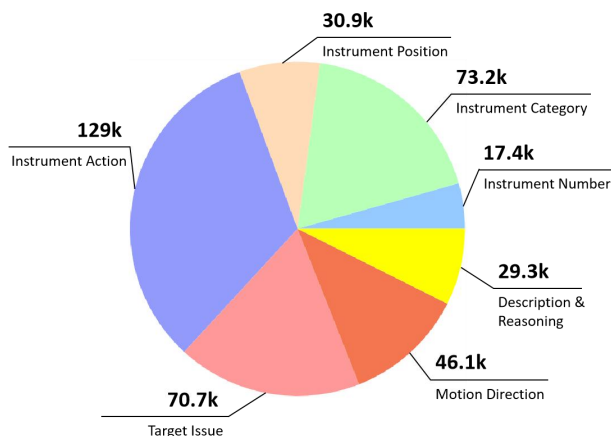
(c) Distribution of attributes in Surg-396K.
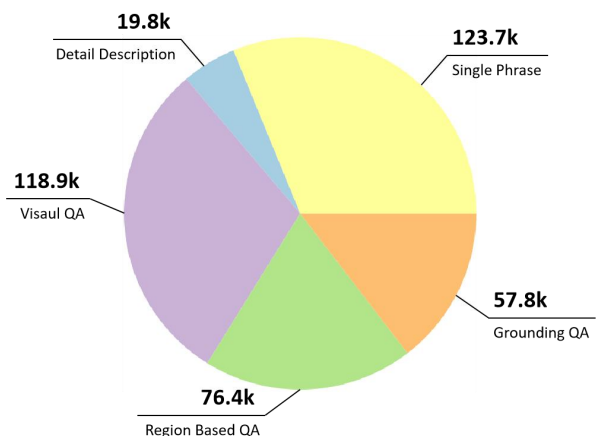
(d) Distribution of conversation types in Surg-396K.

Figure 3. 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 cleaning.

Table 7. List of attributes (with abbreviations) and conversation types designed for Surg-396K.

	Category	Sample
	Instrument Number [IN]	2, 3, etc.
	Instrument Category [IC]	Prograsp forceps, ultrasound probe, etc.
	Object Position [OP]	Right-bottom, left-top, etc.
Attribute	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.
	Category	Response Style
	Single Phrase	Answer the question with a single word or phrase.
Conversation Type	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].

Attribute Extraction

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 Figure 3 (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 *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 support semantic reasoning, encompassing tasks like *Description* and broader contextual reasoning. 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 the generation of instruction-tuning data. The distribution statistics for these attributes are shown in Figure 3 (c).

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. In order to emulate the natural language expression, we further interpret the instruction-tuning data through ChatGPT-4 and generate five conversation types. The distribution statistics and style of these conversation types are shown in Figure 3 (d) and Table 7

Single Phrase is designed to equip EndoChat with the capability to deliver concise, definitive answers to each query, leveraging 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 ChatGPT-4. 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:

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

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 ChatGPT-4 with guided prompts that take full advantage of instruction-tuning data. Therefore, we ensure the response covers all attributes in the corresponding images. The list of questions has also been diversified using ChatGPT-4. This type can augment 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, ChatGPT-4 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 instruments or the condition of specific tissues 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 value 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 also 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 Equation 1.

Data Cleaning

Following the generation of diverse conversations, a data cleaning process is implemented to ensure the integrity and reliability of the training data. Given the large scale of the Surg-396K, we conduct the sampling-based inspection with a ratio of 1/5. Specifically, we manually review the sampled text to evaluate its information completeness, relevance, and clarity. Information completeness refers to whether the text includes all essential content, such as operations and tools. We assign conversation type labels to each QA pair and assess relevance by verifying whether the content aligns with its assigned labels. Since the QA pairs are semantically enriched using ChatGPT-4, we also inspect whether the enriched text is semantically clear and accurate. During the data cleaning process, we document frequently occurring issues and apply modifications to non-sampled content accordingly. This procedure ensures that retained and revised annotations can provide meaningful information for MLLM training.

EndoChat

Architecture of 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 within the endoscopic surgery scenes. Therefore, we propose EndoChat, a novel large vision-language model capable of engaging in visual grounding conversations concerning endoscopic surgery. Given an input image, the mixed visual encoder first extract 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, which augments the representation by generating additional visual tokens. 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, we introduce a visual contrast mechanism to mitigate object hallucinations, further enhancing the model's visual understanding and ensuring consistency between visual inputs and language outputs in complex endoscopic surgery scenarios.

Mixed Visual Token Engine

In our EndoChat, we follow SPHINX that mixes 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[65] and OpenCLIP[66], 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 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 Figure 4, 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[67]. 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':

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

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:
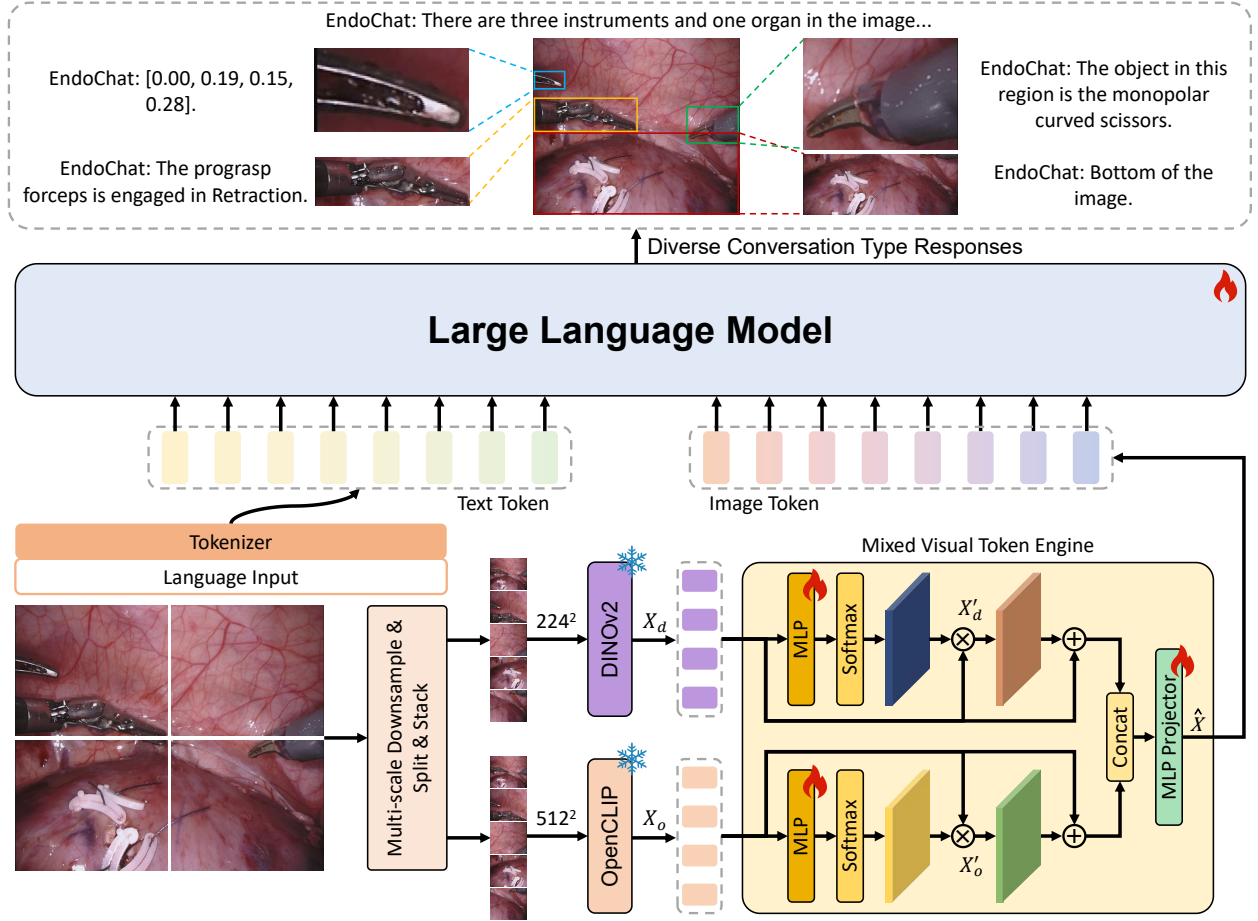
Figure 4. The overview of the proposed EndoChat. For each input image, we use a multi-scale downsampling strategy to generate different scales and sub-images. 224^2 and 512^2 indicate concatenated features with the shapes 5×224×224×3 and 5×512×512×3, respectively. These features are subsequently encoded 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.

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

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

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:

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

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[68]. 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:

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

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 against 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[69] is introduced:

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

where $\beta \in [0, 1]$ controls the truncation of the next token distribution. Larger β indicates more aggressive truncation, keeping only high-probability tokens. Incorporating adaptive plausibility constraints refines the contrastive distribution, enhancing confidence in decisions. 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.

Implementation Details

We adopt the open-source LLaMA2-13B[50] large language model in the SPHINX architecture 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)[70] 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 pre-trained LLaMA-2. 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. To train the LLM and the mixed visual token engine, we utilize an input resolution of 1024×1024. The training process is conducted on the Surg-396K dataset for a single epoch, employing four NVIDIA A800 GPUs. The initial learning rate is set to 2e-5 with a batch size of 16, requiring about 20 hours to complete.

Data availability

The Surg-396K dataset is publicly available at https://github.com/gkw0010/EndoChat.

Code availability

The codes of EndoChat can be downloaded at https://github.com/gkw0010/EndoChat.

Acknowledgements

Author contributions statement

Guankun Wang: Conceptualization, Methodology, Validation, Investigation, Original draft preparation. Long Bai: Conceptualization, Methodology, Investigation, Original draft preparation. Junyi Wang: Validation, Investigation, Dataset Generation. Kun Yuan: Methodology, Original draft preparation. Tianxu Jiang: Dataset Generation. Xiting He: Manuscript Review. Jinlin Wu: Computing Resource Support, Manuscript Review. Zhen Chen: Manuscript Review. Zhen Li: Endoscopist Evaluation, Manuscript Review. Hongbin Liu: Manuscript Review. Nicolas Padoy: Manuscript Review. Nassir Navab: Manuscript Review. Hongliang Ren: Conceptualization, Supervision, Review and editing, Resources.

References

1. Fiorini, P., Goldberg, K. Y., Liu, Y. & Taylor, R. H. Concepts and trends in autonomy for robot-assisted surgery. *Proc. IEEE* **110**, 993–1011 (2022).

2. Chen, R. *et al.* A comprehensive review of robotic surgery curriculum and training for residents, fellows, and postgraduate surgical education. *Surg. endoscopy* **34**, 361–367 (2020).

3. Aziz, H. *et al.* Effect of COVID-19 on surgical training across the United States: a national survey of general surgery residents. *J. surgical education* **78**, 431–439 (2021).

4. Mariani, A., Pellegrini, E. & De Momi, E. Skill-oriented and performance-driven adaptive curricula for training in robot-assisted surgery using simulators: A feasibility study. *IEEE transactions on biomedical engineering* **68**, 685–694 (2020).

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

6. Seenivasan, L., Islam, M., Krishna, A. & Ren, H. Surgical-VQA: Visual Question Answering in Surgical Scenes using Transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 33–43 (Springer, 2022).

7. Bai, L., Islam, M. & Ren, H. 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*, 397–407 (Springer, 2023).

8. Bai, L., Islam, M. & Ren, H. Revisiting distillation for continual learning on visual question localized-answering in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 68–78 (Springer, 2023).

9. Bai, L., Islam, M., Seenivasan, L. & Ren, H. Surgical-VQLA: Transformer with Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. *arXiv preprint arXiv:2305.11692* (2023).

10. Bai, L. *et al.* Surgical-VQLA++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. *Inf. Fusion* **113**, 102602 (2025).

11. Chen, K. *et al.* LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery. *arXiv preprint arXiv:2402.16664* (2024).

12. He, R. *et al.* PitVQA: Image-Grounded Text Embedding LLM for Visual Question Answering in Pituitary Surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 488–498 (Springer, 2024).

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

14. Seenivasan, L., Islam, M., Kannan, G. & Ren, H. SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, 281–290 (Springer, 2023).

15. Yuan, K. *et al.* Advancing surgical VQA with scene graph knowledge. *Int. J. Comput. Assist. Radiol. Surg.* 1–9 (2024).

16. Zhang, Y. *et al.* Dual modality prompt learning for visual question-grounded answering in robotic surgery. *Vis. Comput. for Ind. Biomed. Art* **7**, 9 (2024).

17. Zhu, Z. *et al.* 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)*, 711–721 (2024).

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

19. Pellegrini, C., Keicher, M., Özsoy, E. & Navab, N. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 409–419 (Springer, 2023).

20. Lin, Z. *et al.* Medical visual question answering: A survey. *Artif. Intell. Medicine* **143**, 102611 (2023).

21. Christopher, O. *et al.* Machine Translation with Large Language Models: Decoder Only vs. Encoder-Decoder. *arXiv preprint arXiv:2409.13747* (2024).

22. Gozalo-Brizuela, R. & Garrido-Merchan, E. C. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655* (2023).

23. Wang, P. *et al.* Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

24. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

25. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36** (2024).

26. Chen, J. *et al.* Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280* (2024).

27. Zhou, J. *et al.* Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).

28. Huang, W. *et al.* Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nat. Commun.* **15**, 7620 (2024).

29. Li, T. *et al.* GMAI-VL & GMAI-VL-5.5 M: A Large Vision-Language Model and A Comprehensive Multimodal Dataset Towards General Medical AI. *arXiv preprint arXiv:2411.14522* (2024).

30. Chen, P. *et al.* Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361* (2024).

31. Hu, Y. *et al.* Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183 (2024).

32. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. medicine* **29**, 1930–1940 (2023).

33. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

34. Ferber, D. *et al.* In-context learning enables multimodal large language models to classify cancer pathology images. *Nat. Commun.* **15**, 10104 (2024).

35. Jin, Q. *et al.* Matching patients to clinical trials with large language models. *Nat. Commun.* **15**, 9074 (2024).

36. Qiu, P. *et al.* Towards building multilingual language model for medicine. *Nat. Commun.* **15**, 8384 (2024).

37. Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D. & Mesgarani, N. Contextual feature extraction hierarchies converge in large language models and the brain. *Nat. Mach. Intell.* 1–11 (2024).

38. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat. medicine* **30**, 1134–1142 (2024).

39. Wan, P. *et al.* Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat. Medicine* 1–8 (2024).

40. Dou, Y. *et al.* ShennongGPT: A Tuning Chinese LLM for Medication Guidance. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 67–72 (IEEE, 2023).

41. He, S. *et al.* MedDr: Diagnosis-Guided Bootstrapping for Large-Scale Medical Vision-Language Learning. *arXiv preprint arXiv:2404.15127* (2024).

42. Li, J. *et al.* LLaVA-Surg: Towards Multimodal Surgical Assistant via Structured Surgical Video Learning. *arXiv preprint arXiv:2408.07981* (2024).

43. Jin, J. & Jeong, C. W. Surgical-LLaVA: Toward Surgical Scenario Understanding via Large Language and Vision Models. *arXiv preprint arXiv:2410.09750* (2024).

44. Wang, G. *et al.* Surgical-LVLM: Learning to Adapt Large Vision-Language Model for Grounded Visual Question Answering in Robotic Surgery. *arXiv preprint arXiv:2405.10948* (2024).

45. Schmidgall, S., Cho, J., Zakka, C. & Hiesinger, W. GP-VLS: A general-purpose vision language model for surgery. *arXiv preprint arXiv:2407.19305* (2024).

46. Hou, W. *et al.* Memory-Augmented Multimodal LLMs for Surgical VQA via Self-Contained Inquiry. *arXiv preprint arXiv:2411.10937* (2024).

47. Kuckreja, K. *et al.* Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840 (2024).

48. Lin, Z. *et al.* Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575* (2023).

49. Wang, G. *et al.* CoPESD: A Multi-Level Surgical Motion Dataset for Training Large Vision-Language Models to Co-Pilot Endoscopic Submucosal Dissection. *arXiv preprint arXiv:2410.07540* (2024).

50. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

51. Zhang, K. *et al.* Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints* arXiv–2305 (2023).

52. Li, J. *et al.* LLaVA-Surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981* (2024).

53. Wang, P. *et al.* Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

54. Chen, J. *et al.* Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* (2023).

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

56. Ben-Younes, H., Cadene, R., Cord, M. & Thome, N. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2612–2620 (2017).

57. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).

58. Yu, Z., Yu, J., Xiang, C., Fan, J. & Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks learning systems* **29**, 5947–5959 (2018).

59. Ben-Younes, H., Cadene, R., Thome, N. & Cord, M. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 8102–8109 (2019).

60. Khanam, R. & Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024).

61. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. neural information processing systems* **36** (2024).

62. Allan, M. *et al.* 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426* (2019).

63. Allan, M. *et al.* 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190* (2020).

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

65. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

66. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).

67. Song, M. *et al.* PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis. *Med. Image Analysis* 103248 (2024).

68. Leng, S. *et al.* 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*, 13872–13882 (2024).

69. Li, X. L. *et al.* Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097* (2022).

70. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).