RESEARCH-ARTICLE

CoPESD: A Multi-Level Surgical Motion Dataset for Training Large Vision-Language Models to Co-Pilot Endoscopic Submucosal Dissection

GUANKUN WANG, Chinese University of Hong Kong, Hong Kong, Hong Kong

HAN XIAO, Chinese University of Hong Kong, Hong Kong, Hong Kong

RENRUI ZHANG, Chinese University of Hong Kong, Hong Kong, Hong Kong

HUXIN GAO, Chinese University of Hong Kong, Hong Kong, Hong Kong

LONG BAI, Chinese University of Hong Kong, Hong Kong, Hong Kong

XIAOXIAO YANG, Qilu Hospital of Shandong University, Jinan, Shandong, China

View all

Published: 27 October 2025

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3746027.3758200
ISBN: 9798400720352

CoPESD: A Multi-Level Surgical Motion Dataset for Training Large Vision-Language Models to Co-Pilot Endoscopic Submucosal Dissection

Guankun Wang*
The Chinese University of Hong Kong
Hong Kong, Hong Kong
gkwang@link.cuhk.edu.hk

Han Xiao*
Renrui Zhang
The Chinese University of Hong Kong
Hong Kong, Hong Kong
Shanghai AI Laboratory
Shanghai, China
xiaohan@pjlab.org.cn
renruizhang@link.cuhk.edu.hk

Huxin Gao
Long Bai
The Chinese University of Hong Kong
Hong Kong, Hong Kong
huxingao@cuhk.edu.hk
b.long@link.cuhk.edu.hk

Xiaoxiao Yang
Zhen Li
The Qilu Hospital of Shandong University
Jinan, China
yangxiaoxiao10286@qiluhospital.com
qilulizhen@sdu.edu.cn

Hongsheng Li†
The Chinese University of Hong Kong
Hong Kong, Hong Kong
Shanghai AI Laboratory
Shanghai, China
hsli@ee.cuhk.edu.hk

Hongliang Ren†
The Chinese University of Hong Kong
Hong Kong, Hong Kong
hlren@ee.cuhk.edu.hk

Abstract

With the advances in surgical robotics, robot-assisted endoscopic submucosal dissection (ESD) enables rapid resection of large lesions, minimizing recurrence rates and improving long-term overall survival. Despite these advantages, ESD is technically challenging and carries high risks of complications, necessitating skilled surgeons and precise instruments. Recent advancements in Multimodal Large Language Models (MLLMs) offer promising decision support and predictive planning capabilities for robotic systems, which allow the robot to complete complex tasks in more challenging scenarios. However, the training of MLLMs requires large-scale, well-annotated datasets, and existing datasets for multi-level fine-grained ESD surgical motion reasoning are scarce and lack detailed annotations. In this paper, we design a hierarchical decomposition of ESD motion granularity and introduce a multi-level surgical motion dataset (CoPESD) for training MLLMs as the robotic **Co**-**P**ilot of **E**ndoscopic **S**ubmucosal **D**issection. CoPESD includes 17,679 images with 32,699 bounding boxes and 88,395 multi-level motions, from over 35 hours of ESD videos for both robot-assisted and conventional surgeries. Extensive experiments demonstrate the effectiveness of CoPESD in training MLLMs to comprehend surgical scenarios and reason following surgical robotic motions. As the first multimodal ESD motion dataset, CoPESD supports advanced research in ESD motion decision-making and surgical automation. The dataset is available at https://github.com/gkw0010/CoPESD.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; **Natural language generation**; • **Applied computing** → **Life and medical sciences**.

Keywords

Surgical Decision-Making, Multi-Level Motion Reasoning, Endoscopic Submucosal Dissection, Multimodal Large Language Models

ACM Reference Format:
Guankun Wang, Han Xiao, Renrui Zhang, Huxin Gao, Long Bai, Xiaoxiao Yang, Zhen Li, Hongsheng Li, and Hongliang Ren. 2025. CoPESD: A Multi-Level Surgical Motion Dataset for Training Large Vision-Language Models to Co-Pilot Endoscopic Submucosal Dissection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758200

*Both authors contributed equally to this research.
†Corresponding Author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758200

1 Introduction

Endoscopic submucosal dissection (ESD) was initially proposed in 1988 as a minimally invasive technique for the treatment of early-stage gastric neoplasms, obviating the necessity for open surgical intervention. In the following decades, its application has broadened to encompass a variety of locations within the gastrointestinal (GI) tract, extending even to the resection of deeper non-epithelial lesions [5, 10, 27]. With the development of advanced surgical robotics, robot-assisted ESD is an emerging technique to facilitate the rapid en-bloc resection of large lesions, which is crucial for

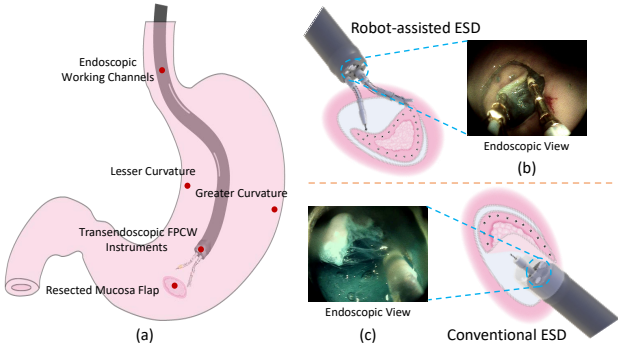

Figure 1: Illustration of Endoscopic Submucosal Dissection with different system instruments. (a) ESD surgery in the gastric body. (b) Endoscopic view of robot-assisted ESD. (c) Endoscopic view of conventional ESD.

minimizing recurrence rates and enhancing long-term survival [11, 35, 52]. Despite these benefits, ESD procedures are still technically challenging, with a high risk of complications such as perforation and bleeding, necessitating surgeons with exceptional skill and instruments with high dexterity and manipulation accuracy [42, 55].

In recent years, Multimodal Large Language Models (MLLMs) have demonstrated superior capabilities in decision support and motion planning for robotic systems [9, 23, 48, 53, 54, 56]. Fu et al. [12] have pioneered the MLLMs for surgical decision-making. The robot-assisted ESD can also benefit from MLLM's strength based on the availability of a large-scale vision-language dataset for ESD surgical motions. If there is a vision-language dataset annotated by fine-grained ESD motions, the MLLMs can learn from substantial experts' determination and serve as a co-pilot to augment surgeons' teleoperation accuracy for robot-assisted or traditional ESD and further mitigate procedural risks. Such a co-pilot is especially valuable for less experienced surgeons. However, current datasets on ESD procedures are significantly scarce. Although there have been efforts towards ESD data construction, existing datasets [3, 13, 20] remain inaccessible. The open-source dataset [7] focuses on workflow recognition and lacks fine-grained motion annotations.

To satisfy the demands for a comprehensive vision-language dataset for ESD surgical scene understanding and motion reasoning, we first conduct a hierarchical decomposition of ESD surgical motion, which is a critical task for reasoning accurate surgical motions. The granularity levels, from high to low, encompass operation, task, surgeme, and motion primitive, etc. Most studies on surgical workflow recognition primarily target the task level [7]. In contrast, our focus is on the lower motion level definition that is more closely aligned with robotic execution, which can provide more sophisticated guidance for surgeons. Consequently, we propose **CoPESD**, a multi-level surgical motion dataset for training MLLMs as the **Co**-**P**ilot of **ESD**. CoPESD comprises 17,679 images, each annotated with five levels of robotic motion instructions and corresponding bounding boxes. Figure 1 illustrates both techniques and procedures performed in the gastric stomach.

Using our proposed CoPESD dataset, we effectively adapt state-of-the-art MLLMs to function as ESD co-pilots to support surgical motion decision-making. Specifically, images are provided as inputs along with high-level language instructions. The MLLMs are guided to perceive surgical scenes, thereby outputting low-level robotic motion instructions. In summary, our contributions are threefold: First, we achieve a granular decomposition of surgical motions, providing precise motion definitions for ESD. Second, we develop CoPESD, a large-scale vision-language dataset that significantly enhances the resources available for ESD surgical scene understanding and motion reasoning. Lastly, our comprehensive evaluations of MLLMs demonstrate the significant effectiveness of CoPESD in reasoning low-level surgical motions, which allows the MLLM to offer precise guidance to surgeons. As the first multimodal ESD surgical motion dataset, CoPESD is poised to advance research in ESD motion decision-making and facilitate the development of reasoning-based models for kinematic information prediction.

2 Related Work

2.1 Language Motion Processing in Robotics

Advancements in natural language processing (NLP) have garnered significant interest in the field of robotics [31, 49], particularly in the context of learning groundings between visual and language modalities [24, 32]. Achievements in human-robot interaction encompass the development of an interactive fetching system capable of localizing objects referenced in natural language expressions [18, 40, 57]. Furthermore, there has been an increasing focus on developing language-conditioned policies for continuous visuomotor control in three-dimensional environments, employing methods such as imitation learning [21, 33, 47] and reinforcement learning [4, 39, 46]. To establish standardized benchmarks and algorithm implementations, [36] proposes CALVIN. Recent advances in MLLMs have enabled the development of Vision-Language-Action (VLA) models capable of end-to-end task execution [6, 25, 43]. These models jointly interpret natural language instructions, perform planning, and generate low-level control commands in a closed-loop manner [9]. Empirical results demonstrate that such models exhibit strong generalization to novel daily scenarios [37]. However, these models often fail to generalize to surgical domains, due to the large domain shift between surgical and general-purpose environments, and the higher complexity and precision demands of surgical tasks.

2.2 Language Motion Processing in Surgery

The concept of surgical language motion processing originated from the need to create more intuitive and efficient ways for surgeons to control robotic systems during complex procedures. This approach allows for the translation of verbal commands into precise robotic movements, streamlining the surgical work. Pioneering work [38] introduces a DVRK-based framework that integrates language motion processing with the Robot Operating System to automate surgical subtasks. Ginesi et al. [17] further contribute to this field by exploring situation awareness and autonomous task planning. They highlight the role of dynamic motion primitive and volumetric obstacle avoidance, enhancing the robot's capability to interpret and execute complex commands. Nguyen et al. [41] emphasize how AI can be used to refine the surgical pattern cutting of natural language instructions. Recently, the application of MLLMs in surgery has gained increasing attention [26, 53]. For

Level of Granularity	Definition	Complexity	Time Span	Example
Operation	The entire invasive part of Endoscopic Submucosal Dissection.	Very High	20-40 min	Endoscopic Submucosal Dissection
Task	Specific surgical activities within ESD that aims to achieve a particular goal.	High	1-15 min	Marking → ... → Submucosal Dissection
Surgeme	An elementary surgical action of determined surgical task.	Moderate	0.1-10 min	Forceps lift the mucosa flap with the assistant of Knife ⇄ Knife dissects the mucosa flap. → ...
Motion Primitive	Basic unit of motion patterns, indicating fundamental movements that the robotic system can execute.	Low	1-5 sec	Forceps reach the mucosa flap and Knife reaches the mucosa flap. → Forceps lift the mucosa flap and Knife reaches the mucosa flap. → ...
Navigating Motion Primitive	Basic unit of motion patterns with a specific direction that can be directly translated into robot commands.	Very Low	1-5 sec	Forceps lift the mucosa flap upward and Knife moves to the edge mucosa flap. ... Forceps lift the mucosa flap downward and Knife moves to the edge mucosa flap.

Figure 2: Overview of different levels of surgical motion granularity for Endoscopic Submucosal Dissection.

instance, Surgical-LLaVA [22] uses the instruction-tuning paradigm of LLaVA [29] by introducing surgical video description annotations into the MLLM training process. SCAN [19] proposes a memory-augmented querying mechanism that improves visual question answering (VQA) performance. GP-VLS [45] formulates multiple surgical tasks within a unified QA framework by representing the outputs of diverse tasks in textual form. However, they focus on describing the surgical image or video clip, lacking a deeper understanding of the surgical context and reasoning over subsequent motions, which limits their utility for precise decision-making. Therefore, we propose the CoPESD dataset to address these limitations by providing a comprehensive multimodal dataset with detailed annotations across multi-level surgical motion granularity.

3 Dataset and Benchmark

3.1 Endoscopic Submucosal Dissection Systems

ESD employs a flexible endoscope to pass through a natural orifice (mouth or anus) and GI tract. The endoscope tip is integrated with a camera to visualize the GI tract for endoscopists. Behind the endoscope tip, the endoscope features a bending section with two degrees of freedom (DoF), pitch and yaw, which are teleoperated by an endoscopist via the proximal handwheels or master console to execute the real-time lesion location [35]. Our study encompasses both traditional and robot-assisted ESD systems. The traditional ESD instruments have limited dexterity, thus their positioning process to targeted lesions generally depends on the maneuverability of the flexible endoscope. To reduce the high reliance on endoscopy skills and improve the dexterity of ESD instruments, robotic technologies are applied to ESD. Our previous work proposed DREAMS[14], a novel system that includes the transendoscopic flexible parallel continuum wrist (FPCW) with three DoFs and multi-functional instruments (such as electric knives, injection needles and forceps). During ESD procedures, the endoscopist can use the master hands to teleoperate two FPCW instruments to perform ESD bimanually.

3.2 Multiple Motion Granularity Levels

The hierarchical decomposition of surgical motion patterns is critical for advancing the ESD automation framework. Surgical interventions and the movements of the surgeon can be systematically decomposed into elements of varying granularity [16, 34, 51]. Although the literature has defined multiple granularity levels for certain surgical scenarios, a comprehensive and consistent framework has not yet been developed for ESD. In order to facilitate the decomposition and partial automation of ESD surgical motions, we define the procedure granularity levels based on prior research [38], as illustrated in Figure 2. From a hierarchical perspective, the *Operation* and *Task* delineate the surgical name and workflows. The *Surgeme* refers to the high-level language instructions. Tool motions are categorized into *Motion Primitive* and *Navigating Motion Primitive*. Previous studies have primarily focused on task level, wherein tasks are executed to achieve specific goals, aligning with the concept of partial automation. These tasks can be subdivided into several surgemes, which are typically task-specific. Within the operational context of ESD, surgemes can be constructed from a universal set of motion primitives. This prompts us to develop a motion primitive library comprising universal motion implementations, facilitating closer alignment with robotic execution.

To construct the motion primitive library, we define the execution conditions for different surgemes. The primary execution condition is the state of the mucosal flap. For instance, if the mucosal flap has been lifted by the Forceps, the Knife will initiate the dissection (refer to the flowchart in the supplementary for additional details). Another critical condition is scene corruption during the surgeme. Given the complexity of the surgical scene, lens deviation and blurring can cause scene corruption, necessitating the cessation of all motions until the scene returns to normal. Furthermore, the direction of motion is pivotal for robot-assisted surgery. It can be predetermined based on the interaction status and relative positioning of the mucosal flaps with surgical instruments. Therefore, we introduce a navigating motion primitive to enhance the motion primitive library's alignment with robotic execution, thereby providing more guidance instructions to the surgeon.

3.3 Data Construction

Our CoPESD dataset construction pipeline comprises the following steps: (1) collecting and clipping ESD videos, (2) extracting and enhancing images for bounding box annotation, (3) designing multi-level surgical motions and labeling each image, and (4) aggregating

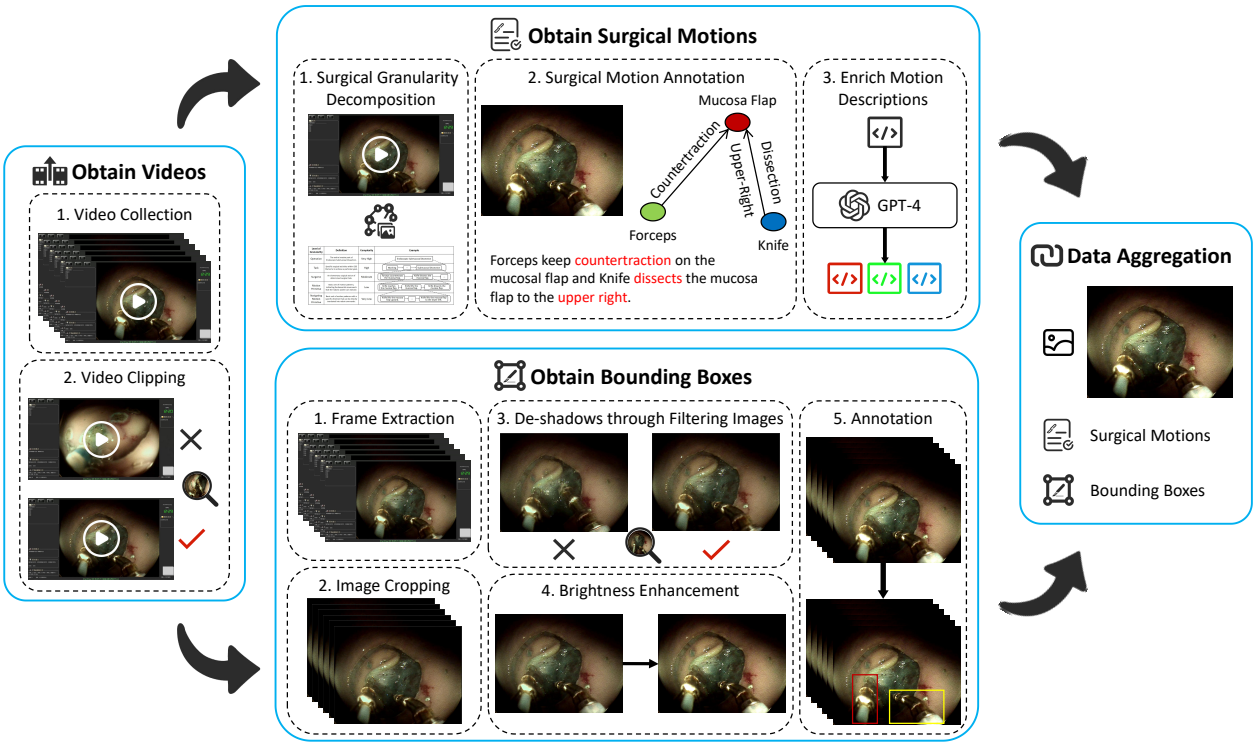

Figure 3: Overview of the construction pipeline for our CoPESD dataset, involving four key steps: video extraction, motion enrichment, bounding box annotation, and data aggregation.

image and text modality data to construct the CoPESD dataset. The detailed overview of the pipeline is presented in Figure 3.

Collection and Clipping of Representative ESD Videos. The initial step is acquiring videos utilizing the DREAMS system and traditional ESD methods. The animal study is approved by the Institutional Ethics Committee on Animal Experiments (Approval No. DWLL-2021-021). A total of 35 hours of video recordings from 40 complete ESD procedures performed on four in-vivo porcine models are collected. Based on criteria including video quality and surgical completeness, 13 procedures utilizing the DREAMS system and 6 conventional ESD procedures are selected for inclusion in the dataset. The ESD videos are recorded at 30 FPS with a resolution of 1920×1080. Following video preprocessing, expert endoscopists from Qilu Hospital provide the ESD surgical task annotations, encompassing six tasks: marking, injection, circumferential incision, subsidized injection, installation and debugging, and submucosal dissection. We specifically select and clip video sequences related to the submucosal dissection task, due to its high degree of soft tissue interaction and procedural complexity.

Image extraction and enhancement. We sample the video sequences at 1 FPS to create the CoPESD dataset, cropping the operator interface portion to achieve a final image resolution of 1306×1009. Given the constant motion of surgical instruments within the scene, the extracted images often contain motion shadows, which could obscure key features and impair model perception. We manually eliminate images with residual shadows. Additionally, due to GI anatomical constraints and hardware limitations, ESD visual signals may suffer from insufficient illumination, complicating the learning and prediction of surgical motions. To address this, we enhance image brightness using the LLCaps [2]. Following image processing, we annotate the bounding boxes of instruments present in each image, providing grounding information for MLLMs.

Multi-level surgical motion designing and annotation. As detailed in Section 3.2, we define motion granularity levels to obtain ESD surgical motions. To ensure high-quality annotations, the following steps are conducted: first, two trained medical annotators independently annotated the images. Subsequent to the initial annotation, cross-validation is performed, and any uncertainties that arose during this process are resolved through a collaborative discussion between two experienced endoscopists. Discussions occur when the surgical scene is highly complex or the direction of dissection is not clear. After completing all annotation tasks, these two endoscopists conducted quality control of the entire dataset. Annotation evaluation depends not only on visual cues but also on practical experience to decide on surgical motions and directions. Considering the inherent complexity and richness of natural language, we use the GPT-4 [1] to generate five variants for each type of motion. During the generation process, each variant retains the original semantic meaning while differing in phrasing, ensuring a diverse representation of surgical motions.

Table 1: Quantitative comparison between different MLLM models using the CoPESD dataset.

Model	Image Resolution	LLM backbone	GPT Score			mIoU (%)		
			10%shot	50%shot	100%shot	10%shot	50%shot	100%shot
LLaVA-ESD	336^2	LLaMA2-7B	83.44	84.03	83.98	30.80	59.22	60.23
		LLaMA2-13B	83.65	84.43	83.62	48.00	61.94	59.42
SPHINX-ESD	512^2	LLaMA2-7B	84.59	85.03	84.32	70.08	69.38	67.38
		LLaMA2-13B	84.53	85.12	85.39	69.24	68.63	69.18
	1024^2	LLaMA2-7B	84.69	**85.14**	85.22	67.53	70.00	69.35
		LLaMA2-13B	**84.94**	84.80	**85.63**	**71.01**	**70.02**	**70.48**

Table 2: Quantitative full-shot performance comparison of motion type and direction generation.

Model	Image Resolution	LLM backbone	Motion		Direction	
			Accuracy (%)	F-score (%)	Accuracy (%)	F-score (%)
LLaVA-ESD	336^2	LLaMA2-7B	87.16	51.43	58.33	34.29
		LLaMA2-13B	86.90	50.22	58.14	34.39
SPHINX-ESD	512^2	LLaMA2-7B	87.66	50.48	58.94	35.61
		LLaMA2-13B	88.70	61.46	60.24	50.43
	1024^2	LLaMA2-7B	88.14	55.63	63.01	38.51
		LLaMA2-13B	**89.06**	**61.98**	**63.12**	**52.95**

Finally, the images obtained through these steps, along with their corresponding multi-level ESD motions and bounding box annotations, are aggregated and formatted to contribute to the CoPESD dataset, comprising 17,679 images annotated with 32,699 bounding boxes and 88,395 multi-level motions.

4 Experiments

4.1 Training Details

To evaluate the effectiveness of our proposed CoPESD dataset, we conduct extensive experiments by adapting MLLMs to serve as an ESD co-pilot. Specifically, the input to the MLLMs consists of visual images paired with *Surgemes*, which serve as high-level motion commands. The MLLMs are prompted to comprehend the surgical scene and subsequently produce *Navigating Motion Primitives* as meticulous low-level motion instructions. We adopt SPHINX-X [15] and LLaVA-1.5 [28] as our baseline models. Open-sourced pretrained weights are fine-tuned on our CoPESD dataset, with the resulting models referred to as SPHINX-ESD and LLaVA-ESD in the experimental results. We experiment with two different model sizes, incorporating LLaMA-2-7B or LLaMA-2-13B [50] backbones. During the fine-tuning, we employ the cosine learning rate scheduler with an initial learning rate of 2e-5 and a total batch size of 64 for both SPHINX-ESD and LLaVA-ESD variants.

Moreover, we perform a thorough ablation study to assess the impact of image resolution and data proportion during fine-tuning. For the SPHINX-ESD series, we employ an input resolution of 512×512, consistent with its original input image resolution for a mixed visual encoder consisting of ConvNeXt [30] and DINO-v2 [8]. We also explore an input resolution of 1024×1024, following the image partition method in SPHINX-X to break down high-resolution input images into sequences of low-resolution image patches. For LLaVA-ESD models, we maintain the original resolution of 336×336, which aligns with the setting of its visual encoder CLIP-ViT-Large [44]. Furthermore, we utilize 10%, 50%, and 100% of the CoPESD dataset in the overall fine-tuning to study the impact of training data on the performance of MLLMs in terms of ESD surgical decision-making.

4.2 Evaluation Metrics

To thoroughly assess the performance of MLLMs on the CoPESD, we adopt three evaluation aspects: (1) Overall Response Quality Evaluation: We utilize GPT-4 to score the quality and accuracy of the responses generated by the fine-tuned models on a scale of 0-100 regarding surgical motions. (2) Grounding Evaluation: We assess the models' localization ability after fine-tuning on the CoPESD dataset. The mean Intersection over Union (mIoU) metric is employed to compare the ground-truth bounding boxes with the predicted ones. (3) Motion type and direction accuracy: We extract key items related to motion and direction from the generated responses. The evaluation metrics for this setting involve Accuracy and F-score, providing a comprehensive measure of performance.

4.3 Performance Evaluation

4.3.1 Quantitative Results. We present the quantitative performance comparisons between different MLLMs of various sizes in Table 1. Utilizing our CoPESD dataset, the MLLMs exhibit substantial capabilities for surgical motion reasoning across various model sizes and image resolution settings. When using the full dataset, SPHINX-ESD and LLaVA-ESD achieve the highest GPT scores of 83.98 and 85.63, respectively, demonstrating that the MLLMs can generate high-quality and accurate responses regarding surgical motions. These results suggest the effectiveness of the proposed dataset in adapting MLLMs for surgical decision-making. Furthermore, the models show significant grounding abilities, accurately localizing the positions of surgical instruments. With the full dataset, SPHINX-ESD and LLaVA-ESD achieve mIoU of 70.48 and 60.23,

Table 3: Quantitative comparison between different MLLM models for top-ten navigating motion primitives.

Model	Image Resolution	Model	GPT Score			mIoU (%)		
			10%shot	50%shot	100%shot	10%shot	50%shot	100%shot
LLaVA-ESD	336^2	LLaMA2-7B	86.47	88.24	89.11	31.05	60.13	63.06
	336^2	LLaMA2-13B	86.72	88.39	89.92	49.53	62.19	62.12
SPHINX-ESD	512^2	LLaMA2-7B	87.43	89.56	90.79	70.66	70.91	71.39
	512^2	LLaMA2-13B	87.22	89.68	91.94	69.41	70.29	71.59
	1024^2	LLaMA2-7B	87.31	89.85	91.49	68.74	70.15	71.01
	1024^2	LLaMA2-13B	**88.70**	**90.58**	**92.17**	**71.70**	**71.25**	**72.19**

respectively. These results highlight the enhanced surgical instrument localization and decision-making abilities of these models by training on the CoPESD. We notice that SPHINX-ESD outperforms LLaVA-ESD counterparts with the same LLM backbones. This is attributed to its higher input image resolution and the more comprehensive knowledge gained through previous training.

To further assess the surgical motions generated by MLLMs, we conduct a quantitative comparison focusing on motion and direction keywords. The experimental results are presented in Table 2. Fine-tuned on the full dataset, the MLLMs exhibit a robust capability in predicting surgical motions and orientations. Notably, the SPHINX-ESD model demonstrates significant improvements in accuracy and F-score for both key items with the increase in image resolution from 512×512 to 1024×1024. This suggests that higher image resolution enhances the model's ability to capture intricate details of surgical scenes. Additionally, utilizing a larger language model backbone generally enhanced the performance across the metrics. Furthermore, we demonstrate the capabilities of MLLMs to function as an ESD co-pilot, generating intelligent surgical robot motions, as shown in Figure 4.

4.3.2 Ablation Study. We analyze the impact of image resolution on the models' performance. As shown in Table 1, increasing the input image resolution improves both the quality of generated responses and grounding accuracy. Higher resolution allows the models to capture more details in the input images, which is crucial for generating precise surgical motions. Additionally, higher resolutions align more closely with our collected video frames, preserving fine-grained information from the visual features. Besides, we explore the effects of varying amounts of sampled data for fine-tuning. Utilizing only 10% of the dataset still produces high-quality responses, comparable to those obtained with the full dataset. This demonstrates the robustness of our CoPESD dataset, as even a small subset effectively enhances the MLLMs' capability to understand and generate detailed surgical motions.

We further explore the impact of data distribution on model performance and specifically evaluate the accuracy of the top-ten navigating motion primitives (they account for 88.32% of the distribution), and the results are shown in Table 3. We find that the accuracy of these extensively represented primitives is significantly higher than the average accuracy of all primitives in Table 1. This demonstrates that the MLLMs trained by CoPESD can indeed understand most cases well. Moreover, the accuracy of highly distributed surgical cases increases with the amount of training data, demonstrating that expanding the data volume can further improve accuracy. However, the excellent performance of MLLM on highly distributed primitives also reflects poor performance on other less distributed primitives. In the submucosal dissection task of ESD surgery, the occlusion of the operating environment leads to a high degree of similarity in dominant features, including the similarity of the target mucosal flap, surgical instruments, and surroundings. As a result, the MLLM is more likely to misinterpret less distributed primitives as other more distributed primitives.

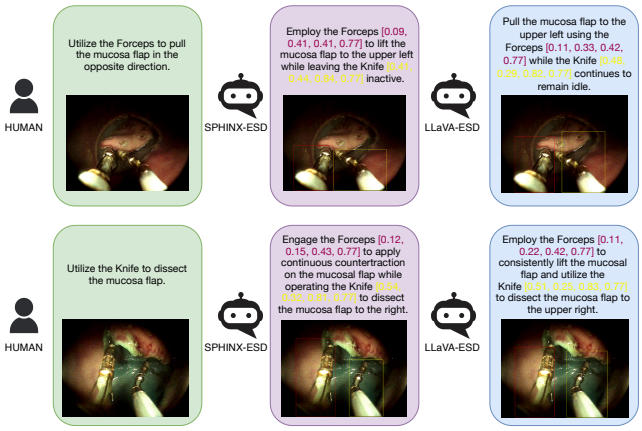

Figure 4: Demonstrations of surgical robot motion outputs from MLLMs after fine-tuning on the CoPESD dataset.

5 Conclusion

In this work, we introduce CoPESD, a comprehensive multi-level surgical motion dataset tailored for Endoscopic Submucosal Dissection. CoPESD encompasses a detailed hierarchical decomposition of ESD motions, facilitating advanced surgical motion reasoning. Through rigorous quantitative analyses on MLLMs, we demonstrate superior performance in predicting surgical motions provided by CoPESD, which makes MLLMs the ESD co-pilot. By making CoPESD publicly accessible, we aim to promote further research and development in the field of ESD motion decision-making and surgical automation, thereby advancing the integration of AI in surgical practices. In the future, our endeavors will focus on the integration of continuous multi-frame and temporal information. Furthermore, CoPESD lays a foundation for the future development of reasoning-driven surgical Visual-Language-Action models, enabling AI co-pilot systems to perform context-aware tasks in complex surgical environments. In addition, we will incorporate multi-center human data and real-world evaluations to enhance model robustness and generalizability.

Acknowledgments

We would like to thank expert endoscopists at Qilu Hospital for their help with data collection and definition of multi-granularity level motions. We also thank the help of Mr. Junyi Wang in data annotation.

References

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

[2] Long Bai, Tong Chen, Yanan Wu, An Wang, Mobarakol Islam, and Hongliang Ren. 2023. Llcaps: Learning to illuminate low-light capsule endoscopy with curved wavelet attention and reverse diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 34–44.

[3] Long Bai, Guankun Wang, Jie Wang, Xiaoxiao Yang, Huxin Gao, Xin Liang, An Wang, Mobarakol Islam, and Hongliang Ren. 2024. OSSAR: Towards Open-Set Surgical Activity Recognition in Robot-assisted Surgery. *arXiv preprint arXiv:2402.06985* (2024).

[4] Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. 2018. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Conference on Robot Learning*. PMLR, 505–518.

[5] Michael J Bourke, Horst Neuhaus, and Jacques J Bergman. 2018. Endoscopic submucosal dissection: indications and application in western endoscopy practice. *Gastroenterology* 154, 7 (2018), 1887–1900.

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[7] Jianfeng Cao, Hon-Chi Yip, Yueyao Chen, Markus Scheppach, Xiaobei Luo, Hongzheng Yang, Ming Kit Cheng, Yonghao Long, Yueming Jin, Philip Wai-Yan Chiu, et al. 2023. Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. *Nature Communications* 14, 1 (2023), 6676.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9650–9660.

[9] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. 2024. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453* (2024).

[10] Philip Wai Yan Chiu, Khek Yu Ho, and Soo Jay Phee. 2021. Colonic endoscopic submucosal dissection using a novel robotic system (with video). *Gastrointestinal Endoscopy* 93, 5 (2021), 1172–1177.

[11] YongYan Cui, Christopher C Thompson, Philip Wai Yan Chiu, and Seth A Gross. 2022. Robotics in therapeutic endoscopy (with video). *Gastrointestinal Endoscopy* 96, 3 (2022), 402–410.

[12] Jiawei Fu, Yonghao Long, Kai Chen, Wang Wei, and Qi Dou. 2024. Multi-objective Cross-task Learning via Goal-conditioned GPT-based Decision Transformers for Surgical Robot Task Automation. *arXiv preprint arXiv:2405.18757* (2024).

[13] Tasuku Furube, Masashi Takeuchi, Hirofumi Kawakubo, Yusuke Maeda, Satoru Matsuda, Kazumasa Fukuda, Rieko Nakamura, Motohiko Kato, Naohisa Yahagi, and Yuko Kitagawa. 2024. Automated artificial intelligence–based phase-recognition system for esophageal endoscopic submucosal dissection (with video). *Gastrointestinal Endoscopy* 99, 5 (2024), 830–838.

[14] Huxin Gao, Xiaoxiao Yang, Xiao Xiao, Xiaolong Zhu, Tao Zhang, Cheng Hou, Huicong Liu, Max Q-H Meng, Lining Sun, Xiuli Zuo, et al. 2024. Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection. *The International Journal of Robotics Research* 43, 3 (2024), 281–304.

[15] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935* (2024).

[16] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, Vol. 3. 3.

[17] Michele Ginesi, Daniele Meli, Andrea Roberti, Nicola Sansonetto, and Paolo Fiorini. 2021. Dynamic movement primitives: Volumetric obstacle avoidance using dynamic potential functions. *Journal of Intelligent & Robotic Systems* 101 (2021), 1–20.

[18] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3774–3781.

[19] Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. 2024. Memory-Augmented Multimodal LLMs for Surgical VQA via Self-Contained Inquiry. *arXiv preprint arXiv:2411.10937* (2024).

[20] Kaide Huang, Xianglei Yuan, Ruide Liu, Yao Zhou, Bing Hu, and Zhang Yi. 2023. An Experimental Study of nmODE in Recognizing Endoscopic Submucosal Dissection Workflow. In *2023 International Annual Conference on Complex Systems and Intelligent Science (CSIS-IAC)*. IEEE, 603–608.

[21] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2022. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*. PMLR, 991–1002.

[22] Juseong Jin and Chang Wook Jeong. 2024. Surgical-LLaVA: Toward Surgical Scenario Understanding via Large Language and Vision Models. *arXiv preprint arXiv:2410.09750* (2024).

[23] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. 2023. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766* (2023).

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 787–798.

[25] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).

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

[27] Ling Li, Xiaojian Li, Bo Ouyang, Hangjie Mo, Hongliang Ren, and Shanlin Yang. 2023. Three-dimensional collision avoidance method for robot-assisted minimally invasive surgery. *Cyborg and Bionic Systems* 4 (2023), 0042.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).

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

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986.

[31] Zirui Liu, Haichun Sun, and Deyu Yuan. 2025. Automatic analysis of alarm embedded with large language model in police robot. *Biomimetic Intelligence and Robotics* (2025), 100220.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019).

[33] Corey Lynch and Pierre Sermanet. 2020. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648* (2020).

[34] L MacKenzie, JA Ibbotson, CGL Cao, and AJ Lomax. 2001. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. *Minimally Invasive Therapy & Allied Technologies* 10, 3 (2001), 121–127.

[35] John T Maple, Barham K Abu Dayyeh, Shailendra S Chauhan, Joo Ha Hwang, Sri Komanduri, Michael Manfredi, Vani Konda, Faris M Murad, Uzma D Siddiqui, and Subhas Banerjee. 2015. Endoscopic submucosal dissection. *Gastrointestinal Endoscopy* 81, 6 (2015), 1311–1325.

[36] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7327–7334.

[37] Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. 2025. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence* (2025), 1–10.

[38] Tamás Dániel Nagy and Tamás Haidegger. 2019. A dvrk-based framework for surgical subtask automation. *Acta Polytechnica Hungarica* (2019), 61–78.

[39] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. 2022. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*. PMLR, 1303–1315.

[40] Thao Nguyen, Nakul Gopalan, Roma Patel, Matt Corsaro, Ellie Pavlick, and Stefanie Tellex. 2020. Robot object retrieval with contextual natural language queries. *arXiv preprint arXiv:2006.13253* (2020).

[41] Thanh Nguyen, Ngoc Duy Nguyen, Fernando Bello, and Saeid Nahavandi. 2019. A new tensioning method using deep reinforcement learning for surgical pattern cutting. In *2019 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 1339–1344.

[42] Hiroyuki Odagiri and Hideo Yasunaga. 2017. Complications following endoscopic submucosal dissection for gastric, esophageal, and colorectal cancer: a review

of studies based on nationwide large-scale databases. *Annals of Translational Medicine* 5, 8 (2017).

[43] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6892–6903.

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

[45] Samuel Schmidgall, Joseph Cho, Cyril Zakka, and William Hiesinger. 2024. GP-VLS: A general-purpose vision language model for surgery. *arXiv preprint arXiv:2407.19305* (2024).

[46] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. 2021. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research* 40, 12-14 (2021), 1419–1434.

[47] Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. 2020. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems* 33 (2020), 13139–13150.

[48] Shilong Sun, Chiyao Li, Zida Zhao, Haodong Huang, and Wenfu Xu. 2024. Leveraging large language models for comprehensive locomotion control in humanoid robots design. *Biomimetic Intelligence and Robotics* 4, 4 (2024), 100187.

[49] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), 25–55.

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[51] S Swaroop Vedula, Anand O Malpani, Lingling Tao, George Chen, Yixin Gao, Piyush Poddar, Narges Ahmidi, Christopher Paxton, Rene Vidal, Sanjeev Khudanpur, et al. 2016. Analysis of the structure of surgical activity for a suturing and knot-tying task. *PloS One* 11, 3 (2016), e0149174.

[52] Fanghao Wang, Youchao Zhang, Daoyuan Jin, Zhongliang Jiang, Yaqian Liu, Alois Knoll, Huanyu Jiang, Yibin Ying, and Mingchuan Zhou. 2024. Magnetic soft microrobot design for cell grasping and transportation. *Cyborg and Bionic Systems* 5 (2024), 0109.

[53] Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. 2024. Surgical-LVLM: Learning to Adapt Large Vision-Language Model for Grounded Visual Question Answering in Robotic Surgery. *arXiv preprint arXiv:2405.10948* (2024).

[54] Jiankun Wang, Weinan Chen, Xiao Xiao, Yangxin Xu, Chenming Li, Xiao Jia, and Max Q-H Meng. 2021. A survey of the development of biomimetic intelligence and robotics. *Biomimetic Intelligence and Robotics* 1 (2021), 100001.

[55] Shunsuke Yamamoto, N Uedo, R Ishihara, N Kajimoto, H Ogiyama, Y Fukushima, Sachiko Yamamoto, Y Takeuchi, K Higashino, H Iishi, et al. 2009. Endoscopic submucosal dissection for early gastric cancer performed by supervised residents: assessment of feasibility and learning curve. *Endoscopy* (2009), 923–928.

[56] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics* 3, 4 (2023), 100131.

[57] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang La, and Nanning Zheng. 2021. Invigorate: Interactive visual grounding and grasping in clutter. *arXiv preprint arXiv:2108.11092* (2021).