Surgical Action Planning with Large Language Models

Mengya Xu[1⋆], Zhongzhen Huang[2,3⋆], Jie Zhang[1,4], Xiaofan Zhang[2,3], and Qi Dou[1] ✉

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai AI Laboratory, Shanghai, China
[4] Huazhong University of Science and Technology, Wuhan, China

Abstract. We introduce the Surgical Action Planning (SAP) task for cholecystectomy procedures, which generates future action plans from visual inputs to address the absence of intraoperative predictive planning in current intelligent applications. SAP shows great potential for enhancing intraoperative guidance and automating procedures. However, it faces challenges such as understanding instrument-tissue relationships and tracking surgical progress. Large Language Models (LLMs) show promise in understanding surgical video content but remain underexplored for predictive decision-making in SAP, as they focus mainly on retrospective analysis. Challenges like data privacy, computational demands, and modality-specific constraints further highlight significant research gaps. To tackle these challenges, we introduce **LLM-SAP**, a Large Language Model-based Surgical Action Planning framework that predicts future actions and generates text responses by interpreting natural language prompts of surgical goals. The text responses potentially support surgical education, intraoperative decision-making, procedure documentation, and skill analysis. LLM-SAP integrates two novel modules: the Near-History Focus Memory Module (NHF-MM) for modeling historical states and the prompts factory for action planning. We evaluate LLM-SAP on our constructed CholecT50-SAP dataset using models like Qwen2.5 and Qwen2-VL, demonstrating its effectiveness in next-action prediction. Pre-trained LLMs are tested in a zero-shot setting, and supervised fine-tuning (SFT) with LoRA is implemented. Our experiments show that Qwen2.5-72B-SFT surpasses Qwen2.5-72B with a 19.3% higher accuracy. The source code and dataset are available at https://github.com/XuMengyaAmy/SAP.

Keywords: Surgical Action Planning · Large-Language Models · Surgical Video Analysis

⋆ Equal contribution.
Jie Zhang conducted this work during her research internship at The Chinese University of Hong Kong.

1 Introduction

In Robot-assisted Minimally Invasive Surgery (RMIS), current intelligent applications, such as surgical workflow recognition [9], instrument segmentation [1,20], action recognition [3, 11], and medical question-answering [2, 16], primarily focus on retrospective analysis rather than future decision-making. While valuable for procedure analysis, they lack the ability to support intraoperative future decision-making, underscoring the need for our Surgical Action Planning (SAP) task, a forward-looking framework. Our SAP streamlines complex procedures by decomposing them into discrete actions, and generates long-horizon sequential action plans from visual inputs, enabling the achievement of user-defined objectives.

Large language models (LLMs) [23] have shown impressive reasoning capabilities, enabling applications such as question answering [7], machine translation [17], and information extraction [8]. Building on this foundation, recent advancements in Large Language Models (LLMs) and Vision Language Models (VLMs) have demonstrated significant potential for understanding surgical video content, yet their application in future decision-making remains underexplored. For instance, while LLMs have been leveraged to refine and enrich surgical concepts through hierarchical knowledge augmentation [21], these approaches primarily focus on retrospective analysis and comprehension rather than predictive planning. LLMs show great promise for the SPA task. Their ability to jointly align visual inputs with language prompts, parse complex scenes, ground language goals in visual contexts, and generate step-by-step plans makes them well-suited for SPA, as evidenced by their success in daily planning tasks like daily activities [4, 6, 10]. Meanwhile, interpretable text analysis from LLMs can be utilized for surgical education prior to operations, providing decision support and assistance during critical intraoperative phases. It offers recommendations to surgeons, helps document and summarize surgical procedures, and serves as a valuable tool for postoperative skill analysis and improvement.

Additionally, fine-tuning LLMs on private data addresses data privacy concerns and enables task-specific customization, unlike zero-shot inference which relies on pre-trained language models and risks exposing sensitive information. However, fine-tuning large models is challenging due to their parameter sizes and resource demands. To address these limitations, efficient fine-tuning methods, such as supervised fine-tuning (SFT) [18], have been developed to reduce computational costs while maintaining performance.

The contributions of our work can be summarized as follows:

- We introduced the Surgical Action Planning (SAP) task for cholecystectomy procedures, which generates future surgical action plans from visual inputs, focusing on forward-looking decision-making rather than retrospective analysis.
- We developed the Large Language Models-based Surgical Action Planning framework (LLM-SAP) that predicts future actions by integrating two innovative modules: the Near-History Focus Memory module (NHFM) for modeling historical states, and the prompts factory for generating action plans.

- We offered a flexible solution for customizing the zero-shot and fine-tuning capabilities of both LLMs and VLMs by thoughtfully designing the processing of visual observations, enabling the adaptation to both modalities.
- We introduce Relaxed Accuracy (ReAcc), a novel evaluation metric that adopts a flexible approach to assess action forecasting. By flexibly considering action forecasting successful if they occur within the current or subsequent one step, ReAcc accounts for the dynamic and adaptive nature of surgical actions.
- We evaluate the effectiveness of the LLM-SAP framework on our constructed CholecT50-SAP dataset, derived from the CholecT50 dataset [13], for surgical action planning. The evaluation involves testing with state-of-the-art (SOTA) LLM and VLM models, including Qwen2.5 and Qwen2-VL, under both zero-shot and fine-tuning experimental settings.

2 Methods

2.1 Surgical action planning (SAP) task definition

In SAP, the model generates an action plan $\mathcal{A} = \{a_1, \ldots, a_t\}$ by leveraging two key inputs: a visual history $\mathcal{H}$ and a user-defined goal G, aiming to transition the current state to the desired goal within a planning horizon of T steps. The visual history $\mathcal{H}$, represented as a sequence of video clips $\{v_1, \ldots, v_t\}$, captures the progression toward the goal over time. Meanwhile, the goal G is expressed as a natural language description, such as *"Provide analysis and the next action for laparoscopic cholecystectomy."* Each action a_t in the plan corresponds to a categorical label within a set of C possible actions.

2.2 Our LLM-SAP

By breaking down complex surgical procedures into a fixed closed set of actions, we develop the LLM-SAP to predict the next action and link these actions into long-horizon action chains, powered by advanced LLMs (see Fig. 1). LLM-SAP has versions: one processes text descriptions of visual history, while the other directly analyzes visual history. Both are guided by action-planning prompts. By leveraging domain knowledge and prompts, the system generates a response in a structured format. The following sections will provide further details.

Near-history focus memory module (NHFM) Previous studies on history understanding for VLM-based planners generally rely on a combination of historical action labels and their associated frames to understand the historical state. $\mathcal{H}_t = \text{VLM}(\{\langle f_i, a_i \rangle \mid i = 1, \ldots, t\}, \text{Prompt})$. However, incorporating lengthy action histories can overwhelm the planning process, often resulting in reduced performance [4,19]. In surgical settings, the next action is determined by the current tissue state and the procedure's dynamic progression, rather than adhering to a rigid sequence of past actions. To address this limitation, we introduce an

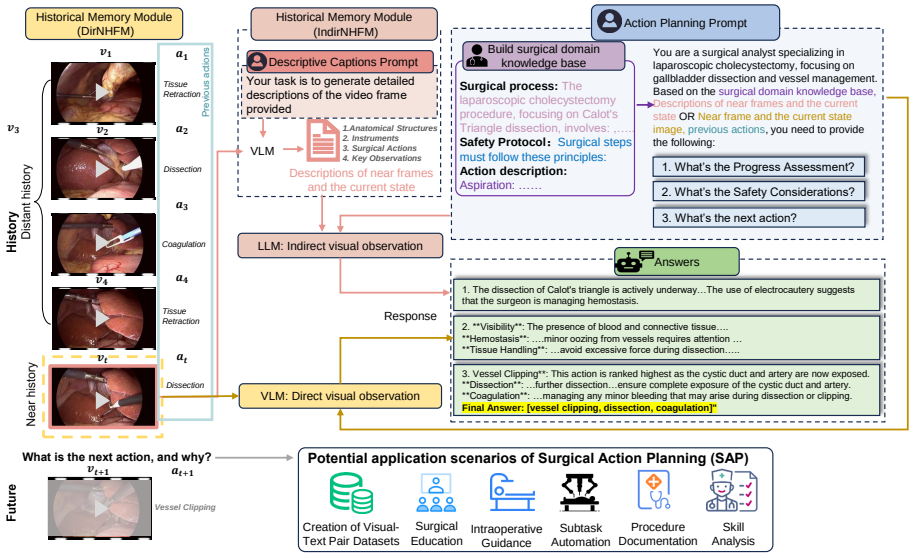

Fig. 1. The architecture of our LLM-SAP. It is developed to predict the next surgical action and form long-horizon action chains by breaking procedures into a fixed set of actions, powered by advanced LLMs. LLM-SAP has two versions: a text-based LLM planning model that utilizes the text descriptions of the visual history in IndirNHFM, and a VLM planning model that uses visual history directly in DirNHFM, both guided by the action planning prompt. The data flow of these two versions is indicated by the pink and yellow lines, respectively. Given the surgical domain knowledge base and action planning prompts like *"What's the next action?"*, LLMs analyze the history memory state to understand the surgical progress and generate structured responses, including progress assessment, safety considerations, and future action recommendations.

enhanced history state understanding approach to build the Historical Memory Module (HMM) that emphasizes a compact summary of past actions for the distant history while focusing on detailed information from the near history.

Direct visual observation for creating NHFM (DirNHFM): a_i for time steps 1 to $t-1$ encapsulates a compact representation of distant historical actions (action labels only), while $\langle v_t, a_t \rangle$ captures the detailed near history through the current video frame v_t and its associated action label a_t. Additionally, *Prompt* refers to the overarching prompt supplied to the VLM. It can be formulated as Equation 1. The details of the *Prompt* design will be introduced in the next section.

$$\mathcal{A}_{\text{DirNHFM}} = \text{VLM}(\{a_i \mid i = 1, \ldots, t-1\}, \langle v_t, a_t \rangle, \mathit{Prompt}) \tag{1}$$

Indirect visual observation for creating NHFM (IndirNHFM): To address the limitation of LLMs that do not support visual frames as input, we propose an indirect approach leveraging visual observations for HMM creation. Specifically, we use the VLM to process visual frames and generate descriptive

captions. These captions serve as text-based input for LLMs unable to process visual data directly, enabling our framework to adapt to such LLMs (see Equation 2 and 3).

$$C_t = \text{VLM}(v_t, \textit{DCPrompts}) \tag{2}$$

$$\mathcal{A}_{\text{IndirNHFM}} = \text{LLM}(\{a_i \mid i = 1, \ldots, t-1\}, \langle C_t, a_t \rangle, \textit{Prompt}) \tag{3}$$

Prompts factory for generating action plans Fig. 1) shows the prompts factory, including (1) *Descriptive captions prompts (DCPrompts)* To generate descriptive captions using the VLM , we employ the following prompt: *"You are a professional surgical analysis assistant specializing in laparoscopic cholecystectomy. Your task is to generate detailed descriptions of the video frame provided, focusing on anatomical structures, tool manipulation, key surgical steps, and environmental features."* (2) *Action planning prompts (APPrompts)*: Firstly, *surgical domain knowledge base* is build based on *surgical process*, *safety protocol*, and *action description*. Next, the APPrompts include *"Based on the provided Surgical Domain Knowledge Base, Descriptions of near frames and the current state (IndirNHFM) OR Video frames provided (DirNHFM), Previous actions, Last action, you need to provide the following:* ***progress assessment***, ***safety considerations***, ***ready-to-execute actions****: Provide three actions. For each action, provide a rationale explaining why it is ranked in that order."*

Zero-shot and supervised fine-tuning Building on our carefully designed historical memory module, we begin by evaluating the performance of open-source models in surgical action planning. Given that even current advanced models struggle with such tasks, our goal is to enhance the performance of open-source models in this domain. Based on prior efforts, we adopt a distillation-based approach to generate high-quality data from stronger models. In particular, we leverage GPT-4o, which has demonstrated exceptional performance, to generate fine-tuning data. In total, we obtain 118 samples for fine-tuning data (63 for IndirNHFM and 55 for DirNHFM).

2.3 Implementation details

We used Llama-Factory [24] to fine-tune the large language models (LLM and VLM) with LoRA [5]. we conducted training over 50 epochs. The fine-tuning of all LLMs was performed on 8 NVIDIA A800 GPUs, using a learning rate of $1e-4$ and a batch size of 8.

3 Experiments and results

Dataset Our constructed CholecT50-SAP dataset is derived from the CholecT50 dataset [13], which consists of 50 cholecystectomy surgical procedures. We grouped consecutive frames sharing the same action into action clips in CholecT50-SAP

(see Fig. 2). The action labels include {Aspiration, Coagulation, Dissection, Tissue Retraction, Vessel Clipping}. We focus on the video segments encompassing *"Calot's triangle dissection, duct and vessel clipping, and dissection from the liver bed"* as the context for implementing action planning. The 50 chosen video segments have an average duration of 6.36 minutes each. Including 5 historical action clips, these 50 video segments from cholecystectomy procedures comprise a total of 225 samples. The dataset was split into training (35 videos, 168 samples) and testing (15 videos, 57 samples) sets, following [14].

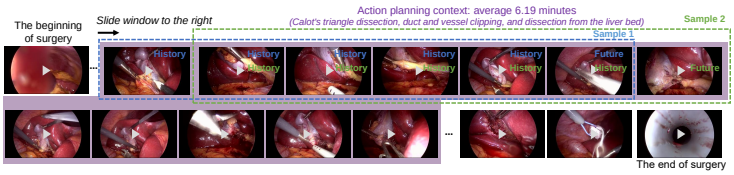

Fig. 2. Example of the CholecT50-SAP dataset we constructed.

Metrics We report the sample and video-level accuracy under standard and relaxed conditions. **Sample-level accuracy (SLAcc):** The standard SLAcc requires the predicted next action $\hat{a}_i$ to exactly match the ground truth action a_i at each time step. On the other hand, Top-2 and Top-3 SLAcc evaluate the proportion of samples in which the true action label a_i for the future video clip is included among the top two or three predictions, respectively. **Video-level accuracy (VLAcc):** We report VLAcc to evaluate the model's performance on individual surgical patients. This is calculated by averaging the mean of SLAcc across all surgical patients. **Relaxed accuracy (ReAcc):** Our proposed ReAcc adopts a more flexible evaluation approach, motivated by the need to account for the dynamic nature of surgical actions. It considers future action recommendations successful if the model's suggested actions $\hat{a}_i$ occur within the current step or the subsequent one-step $\{a_i, a_{i+1}\}$ ($\hat{a}_i \in \{a_i, a_{i+1}\}$), aligning with the variability and adaptability required in surgical procedures.

Action planning results Building on an LLM-based planning framework, we employ multiple state-of-the-art (SOTA) LLM (Qwen2.5) and VLM (Qwen2-VL) to predict the probable next action. In addition to the zero-shot experiments, we also conduct the supervised fine-tuning (SFT) approach [18] (see Table 1). We also provided comparative results from existing models under the supervised fine-tuning, such as AntGPT [22], Most Prob [15], and Probabilistic Sequence (PS) [12]. In the zero-shot setting, these baseline models fail to generate valid outputs because their predictions cannot be properly mapped back to the action labels.

The analysis of the experimental results can be summarized as follows: (1) **InDirNHFM vs. DirNHFM:** In zero-shot experiments, Qwen2.5-72B achieves a

Table 1. Comparison of performance across different methods on the CholecT50-SAP dataset, with the best results highlighted in bold.

Models	HMM	Standard Condition						Relaxed Condition					
		SLAcc			VLAcc			Re SLAcc			Re VLAcc		
		Top1	Top2	Top3	Top1	Top2	Top3	Top1	Top2	Top3	Top1	Top2	Top3
Zero-Shot													
Qwen2.5-32B	IndirNHFM	45.61	63.16	66.67	53.42	62.48	64.65	67.44	**95.35**	**97.67**	78.97	**97.38**	98.33
Qwen2.5-72B	IndirNHFM	45.61	59.65	66.67	53.42	59.31	64.65	67.44	83.72	**97.67**	78.97	89.21	**99.05**
Qwen2-VL	DirNHFM	40.35	57.89	68.42	48.37	56.37	66.32	**72.09**	88.37	90.70	**82.48**	93.45	95.24
Supervised Fine-Tuning													
AntGPT [22]	-	19.10	-	-	19.20	-	-	33.17	-	-	29.76	-	-
Most Prob [15]	-	**49.25**	-	-	44.86	-	-	62.81	-	-	59.33	-	-
PS [12]	-	48.15	-	-	46.70	-	-	65.28	-	-	60.26	-	-
Qwen2.5-72B-SFT	IndirNHFM	45.61	**78.95**	80.70	53.42	80.05	82.27	67.44	93.02	**97.67**	78.97	93.33	98.33
Qwen2.5-32B-SFT	IndirNHFM	45.61	**78.95**	**85.96**	53.42	**83.05**	**87.89**	67.44	**95.35**	**97.67**	78.97	96.67	98.33
Qwen2-VL-72B-SFT	DirNHFM	47.37	54.39	68.42	**54.31**	58.55	65.98	60.47	74.42	90.70	67.30	83.25	94.76

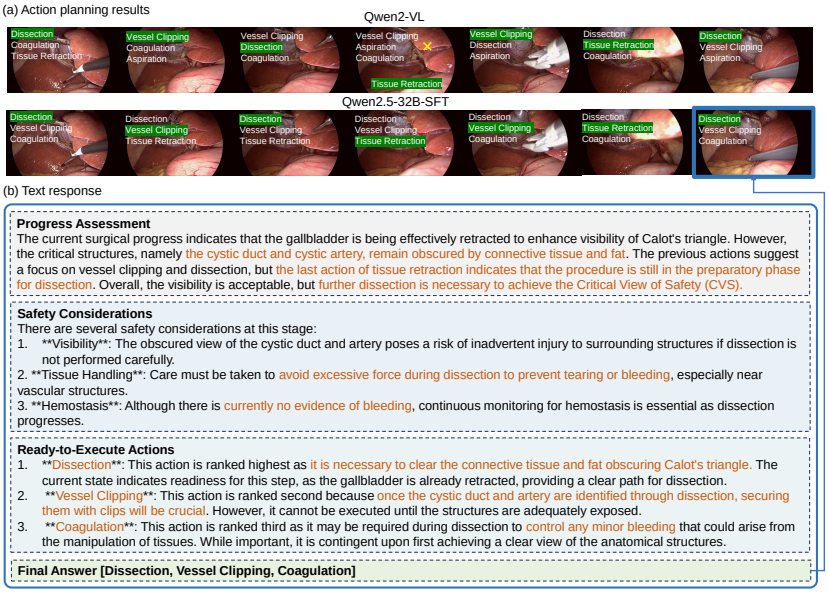

Fig. 3. (a) Action planning results visualization. White text displays the planning results, showing the Top 3 future action predictions in order. × indicates incorrect prediction. The text with a green background color indicates the ground truth actions. (b) Example of the text response used to derive the action planning answer.

standard top-1 SLAcc of 45.61% and VLAcc of 53.42% with IndirNHFM, outperforming Qwen2-VL with DirNHFM by 5.26% (45.61% vs. 40.35%) in SLAcc and 5.05% (53.42% vs. 48.37%) in VLAcc. This indicates that indirect visual observation better captures contextual nuances, potentially due to richer textual representations. (2) **Zero-Shot vs. SFT:** Qwen2.5-72B-SFT outperforms Qwen2.5-72B, achieving 19.3% higher standard top-2 SLAcc (78.95% vs. 59.65%), 20.74%

higher standard top-2 VLAcc (80.05% vs. 59.31%), 9.3% higher relaxed top-2 SLAcc (93.02% vs. 83.72%), and 4.12% higher relaxed top-2 VLAcc (93.33% vs. 89.21%). This demonstrates that SFT significantly boosts model performance for SAP by leveraging task-specific data. (3) **Qwen2.5-32B-SFT *(best model)* vs others:** Although Qwen2-VL-72B-SFT achieves the best standard top-1 SLAcc of 47.37 and VLAcc of 54.31, Qwen2.5-32B-SFT demonstrates strong overall performance across all metrics, with 44.3% higher standard top-2 SLAcc (78.95% vs. 54.39%), 41.8% higher standard top-2 VLAcc (83.05% vs. 58.55%), 28.1% higher relaxed top-2 SLAcc (95.35% vs. 74.42%), and 16.2% higher relaxed top-2 VLAcc (96.67% vs. 83.25%). (4) Our LLM-based methods yield more flexible outputs and text-based justifications than existing models. Fig. 3(a) presents the action planning results for different LLMs and Fig. 3(b) shows the interpretable reasoning processes for a single sample.

Ablation experiments on HMM creation We conducted ablation studies on the Historical Memory Module (HMM), as shown in Fig. 4. (i) **with an image from near history, along with the last previous action**: $\{f_t, a_t\}$; (ii) **with only near history**: C_t or v_t (iii) **with near history and its associated action**: $\langle v_t, a_t \rangle$; (iv) **with previous action labels and near history**: $\{a_1, \ldots, a_{t-1}\}, \langle v_t, a_t \rangle$. The comparison between settings (ii) and (iv) highlights the importance of previous action labels in modeling historical states. Among the four HMM creation methods, setting (iv) demonstrates the best results. Therefore, we adopt setting (iv) as the default method for creating the HMM.

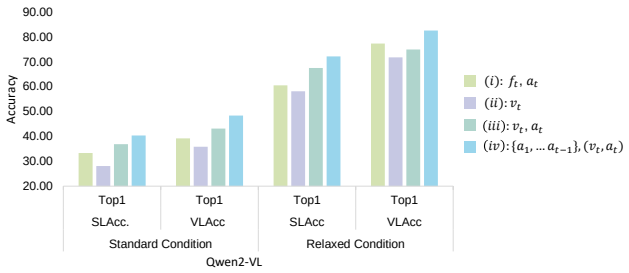

Fig. 4. Ablation experiments on HMM creation based on Qwen2-VL.

4 Conclusion

We introduce the Surgical Action Planning (SAP) task in computer-assisted surgery, generating future action plans from visual inputs with a focus on forward-looking decision-making. This task has immense potential to enhance intraoperative guidance and procedural automation. Key challenges such as complex instrument-tissue relationships, temporal dependencies, progress tracking, and data privacy concerns have hindered progress in the SAP task. To address these

challenges, we propose the LLM-SAP framework leveraging large language models to predict actions and provide interpretable responses by integrating our proposed NHF-MM and prompts factory. Evaluated on our constructed CholecT50-SAP dataset using state-of-the-art models such as Qwen2.5 and QwenVL, LLM-SAP demonstrated effectiveness in recommending future actions while effectively addressing data privacy through SFT. **Future work:** To enhance LLM-SAP, we will leverage reasoning-based LLMs to use text responses from earlier steps as input for subsequent steps, enabling prior-conditioned and continuous planning while highlighting differences in structured outputs. Additionally, we aim to expand the framework to support more procedures and robotic integration.

Acknowledgements. This research work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Projects No. 24209223, No. N_CUHK410/23, No. T45-401/22-N, and in part by the National Natural Science Foundation of China under Project No. 62322318. We appreciate Dr. Hongyu Wang for providing support in the code implementation.
Disclosure of Interests. The authors declare no competing interests.

References

1. Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: Matis: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
2. Bai, L., Wang, G., Islam, M., Seenivasan, L., Wang, A., Ren, H.: Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. Information Fusion **113**, 102602 (2025)
3. Bai, L., Wang, G., Wang, J., Yang, X., Gao, H., Liang, X., Wang, A., Islam, M., Ren, H.: Ossar: Towards open-set surgical activity recognition in robot-assisted surgery. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 14622–14629. IEEE (2024)
4. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking multimodal large language models for human-level planning. arXiv preprint arXiv:2312.06722 (2023)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
6. Huang, D., Hilliges, O., Van Gool, L., Wang, X.: Palm: Predicting actions through language models @ ego4d long-term action anticipation challenge. arXiv preprint arXiv:2306.16545 (2023)
7. Jiang, J., Zhou, K., Zhao, W.X., Li, Y., Wen, J.R.: Reasoninglm: Enabling structural subgraph reasoning in pre-trained language models for question answering over knowledge graph. arXiv preprint arXiv:2401.00158 (2023)
8. Jiao, Y., Zhong, M., Li, S., Zhao, R., Ouyang, S., Ji, H., Han, J.: Instruct and extract: Instruction tuning for on-demand information extraction. arXiv preprint arXiv:2310.16040 (2023)
9. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. IEEE Transactions on Medical Imaging **40**(7), 1911–1923 (2021)

10. Kim, S., Huang, D., Xian, Y., Hilliges, O., Van Gool, L., Wang, X.: Palm: Predicting actions through language models. In: European Conference on Computer Vision. pp. 140–158. Springer (2024)
11. Kiyasseh, D., Ma, R., Haque, T.F., Miles, B.J., Wagner, C., Donoho, D.A., Anandkumar, A., Hung, A.J.: A vision transformer for decoding surgeon activity from surgical videos. Nature biomedical engineering **7**(6), 780–796 (2023)
12. Mutegeki, R., Han, D.S.: A cnn-lstm approach to human activity recognition. In: 2020 international conference on artificial intelligence in information and communication (ICAIIC). pp. 362–366. IEEE (2020)
13. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
14. Nwoye, C.I., Yu, T., Sharma, S., Murali, A., Alapatt, D., Vardazaryan, A., Yuan, K., Hajek, J., Reiter, W., Yamlahi, A., et al.: Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. Medical Image Analysis **89**, 102888 (2023)
15. Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M.L., Jain, U., Desai, R.: Pretrained language models as visual planners for human assistance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15302–15314 (2023)
16. 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. pp. 281–290. Springer (2023)
17. Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., Tu, Z.: Document-level machine translation with large language models. arXiv preprint arXiv:2304.02210 (2023)
18. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
19. Xie, J., Zhang, K., Chen, J., Yuan, S., Zhang, K., Zhang, Y., Li, L., Xiao, Y.: Revealing the barriers of language agents in planning. arXiv preprint arXiv:2410.12409 (2024)
20. Yu, J., Wang, A., Dong, W., Xu, M., Islam, M., Wang, J., Bai, L., Ren, H.: Sam 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. arXiv preprint arXiv:2408.04593 (2024)
21. Yuan, K., Srivastav, V., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. arXiv preprint arXiv:2410.00263 **2** (2024)
22. Zhao, Q., Wang, S., Zhang, C., Fu, C., Do, M.Q., Agarwal, N., Lee, K., Sun, C.: Antgpt: Can large language models help long-term action anticipation from videos? arXiv preprint arXiv:2307.16368 (2023)
23. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 **1**(2) (2023)
24. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372 (2024)