This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

CSAP-Assist: Instrument-Agent Dialogue Empowered Vision-Language Models for Collaborative Surgical Action Planning

Jie Zhang[1,2] ⋆, Mengya Xu[1]⋆, Yiwei Wang[2], and Qi Dou[1] ✉

[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Huazhong University of Science and Technology, Wuhan, China

Abstract. Visual Planning for Assistance (VPA) in Robot-Assisted Minimally Invasive Surgery (RMIS) holds significant potential for intra-operative guidance and procedural automation. This paper presents the Collaborative Surgical Action Planning (CSAP) task, which focuses on generating cooperative action plans based on linguistic surgical goals, highlighting the crucial need for coordinated multi-tool interactions in surgical procedures. CSAP task emphasizes two core challenges: understanding tool-action interdependencies in the timeline and managing concurrent multi-tool interactions. To address these challenges, we propose CSAP-Assist, a VLM-based framework consisting of two key modules: a Recency-Centric Focus Memory Module (ReFocus-MM), which prioritizes recent surgical history while summarizing distant events to improve performance in complex scenes and long sequences; and a Hybrid Multi-Agent Module (HMM), featuring a central agent that provides an initial plan, prompting a dialogue with local agent instruments to iteratively refine their collaborative actions. We evaluated CSAP-Assist on datasets that include phantom and real surgical scenarios. Our extensive experiments show that CSAP-Assist substantially outperforms the baseline method, achieving a 15% higher planning precision for surgical action planning. The source code and dataset are available at https://github.com/einnullnull/Collaborative-Surgical-Action-Planning-Assist.

Keywords: Visual Planning for Assistance · Vision-Language Models · Surgical Video Analysis · Multi-Agent Systems.

1 Introduction

Visual Planning for Assistance (VPA) focuses on creating sequential action plans from visual inputs to achieve user-defined goals, a framework widely used in human activity planning [13]. In robot-assisted minimally invasive surgery (RMIS), VPA shows great promise, enabling real-time procedural guidance by

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

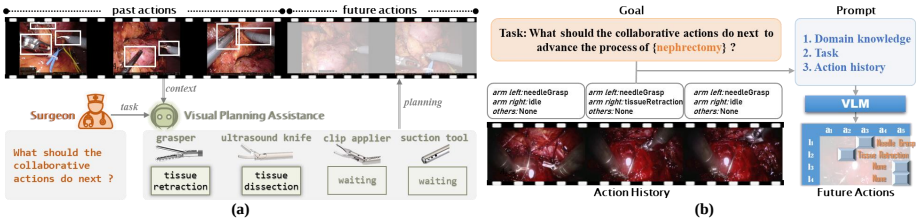

Fig. 1. (a) CSAP Task Definition: The model generates future action plans for each surgical tool based on defined goals and procedural history. (b) Baseline VLM Framework: Utilizing linguistic targets, surgical domain knowledge, and action history (frames and labels), the VLM predicts future tool actions.

translating natural language instructions (e.g., "*Suture Renal Veins*") into executable surgical actions. For example, VPA can be used to automate specific steps of robotic surgery, or provide intra-operative navigation and decision support [4]. However, deploying VPA in surgical contexts introduces unique complexities beyond conventional human activity planning. Surgical workflows require collaborative action planning, requiring synchronized execution and coordination of interdependent instrument maneuvers. These challenges distinguish Surgical VPA from traditional planning frameworks. To address these demands, we propose **Collaborative Surgical Action Planning (CSAP)** (Fig. 1 (a)), a task centered on three critical challenges: (1) modeling tool-action interdependencies, (2) handling concurrent multi-tool interactions, and (3) integrating real-time visual feedback into dynamic planning processes.

Although VPA task has significant clinical value, current research in surgical AI assistance primarily emphasizes retrospective analysis of operative videos. Dominant research directions include workflow recognition [1] and instrument segmentation [20], which, while providing insightful post-procedure analyses, are not equipped with the predictive abilities necessary for proactive decision making. Recent advances in predictive action planning aim to bridge this gap. For example, Zhao et al. [19] explore diffusion models to predict actions that achieve predetermined visual objectives. However, specifying such visual goal during surgery remains a significant challenge. Zhang et al. [17] introduce an activity grammar framework for language-based goals, yet this approach overlooks the visual state that changes over time. Additionally, both emerging approaches overlook the essential multi-tool interactions required for CSAP and fail to integrate real-time visual observations into a practical planning system, leaving a pivotal gap in intra-operative AI assistance.

Vision-Language Models (VLMs) show great promise for VPA task [9]. Their ability to jointly align visual inputs with language instructions, parse complex scenes, ground language goals in visual contexts, and generate step-by-step plans makes them well-suited for VPA, as evidenced by their success in daily planning tasks like kitchen activities [2, 5]. However, adapting current VLM-based planning frameworks for the CSAP task presents considerable challenges. Surgi-

cal procedures, such as suturing, involve lengthy, repetitive steps with intricate temporal dependencies, which are difficult for frameworks designed for short-term daily tasks to manage [2]. Furthermore, CSAP necessitates coordinating interdependent actions across multiple instruments, while current frameworks mainly focus on single-action sequence planning. To overcome these shortcomings, we propose CSAP-Assist, a VLM-based framework designed for surgical CSAP, which consists of two modules: **1) Recency-Centric Focus Memory Module (ReFocus-MM)**,which overcomes information overload in complex, long sequences by prioritizing recent history while summarizing distant history; 2) **Hybrid Multi-Agent Module (HMM)** models surgical instruments as collaborating local agents coordinated by a central planning agent. The central agent provides an initial plan, which the instrument agents then iteratively refine through communication to generate a final action plan.

We evaluated CSAP-Assist using a phantom dataset with two instruments and a real-world multi-instrument suturing dataset. Results show our approach outperforms existing baselines across key surgical action planning metrics.

The contributions of this paper are summarized as follows:

- We introduced CSAP, a novel task that explicitly addressing the underexplored challenges of domain-specific procedure understanding and multi-tool coordination in surgical VPA.
- We developed CSAP-Assist, a VLM-based planning framework that includes two innovative modules for efficient long-term memory management and coordinated multi-instrument planning.
- We demonstrated the effectiveness of CSAP-Assist through a comprehensive evaluation of both phantom and real surgical scenarios. By leveraging two reconstructed datasets from existing benchmarks, we showcased significant performance improvements over current methods.

2 Methods

2.1 Task Formulation

VPA Task Given the current observation $\mathcal{O}$ and a user-specified goal G, the model generates an action plan, a sequence of actions $A = \{a_1, \ldots, a_T\}$ to transform the current state to the goal state within a planning horizon of T steps. Each action $a_t \in \mathbb{R}^C$ is represented as a categorical label, where C is the total number of admissible actions. The observation $\mathcal{O}$ is a video history of the user's progress: $\mathcal{O} = \{v_{-H+1}, \ldots, v_0\}$, with H indicating the number of past action video clips. The user goal G is provided as a natural language description.

CASP Task As illustrated in Fig. 1(a), surgical procedures often involve the simultaneous execution of multiple actions within a single video clip to advance progress. For example, the left grasper may retract tissue while the right grasper punctures it with a needle. The observation $\mathcal{O}$ consists of the video history O_v and the action label history O_a, providing insight into the surgical progress. The goal G is specified by an instruction for the next collaborative actions: "What are

the next collaborative actions to advance progress in the {*surgery type*}?". At each timestamp, the action label O_a stores textual descriptions of the candidate tools' actions (e.g., "left grasper": "Grasp Needle", "right grasper": "Puncture", "suction tool": "Aspirate").

2.2 Baseline: VLM-based Planning Framework

Building upon prior VLM-based planning frameworks [2], we utilize GPT-4o, a powerful VLM, to predict the most probable subsequent action. While GPT-4o's strong performance across diverse tasks makes it a compelling choice, our framework remains adaptable to other VLMs. The pipeline of this baseline framework is illustrated in Fig. 1(b).

2.3 CSAP-Assist framework

Existing baseline methods struggle to effectively understand the intricate context of surgical processes, and their ability to facilitate collaborative action planning remains unclear. To address these limitations, we introduce CSAP-Assist (Fig. 2), a novel framework featuring a **Recency-Centric Focus Memory Module (ReFocus-MM)** for superior process state understanding and a **Hybrid Multi-agent Module (HMM)** for multi-action collaboration. The subsequent sections will detail these components.

Recency-Centric Focus Memory Module Prior work on context understanding for VLM-based planners (Fig. 1 (b)) typically uses the sequence of historical action labels and corresponding visual frames as contex. However, processing extensive action histories can overload planners, leading to suboptimal performance [2, 15]. In surgery, the next action depends on the dynamic progression and current tissue state, not a rigid sequence. Therefore, we propose a refined context understanding method that prioritizes a concise summary of past actions alongside detailed information about the most recent step:

$$\Omega_t = \text{VLM}(\{a_i \mid i = 1, \ldots, t-1\}, \langle v_t, a_t \rangle, \text{Instruction}) \tag{1}$$

where a_i from time step 1 to $t-1$ represents the concise distant context (action labels only), $\langle v_t, a_t \rangle$ provides the detailed near context via the current video frame v_t and action label a_t, and *Instruction* denotes the overall prompt provided to VLM.

Hybrid Multi-agent Module A centralized VLM planner (Fig. 1 (b)) struggles to scale with the number of surgical tools and lacks a reflection mechanism for correcting errors [3, 8]. Inspired by multi-agent frameworks [10, 11, 16], we adopt a Hybrid Multi-agent System (HMAS) [3]. A central LLM planner generates initial actions:

$$\tilde{a}_{t+1} = \text{A}_{\mathcal{C}}(\langle f_t, a_t \rangle, \langle \Omega_t, G \rangle) \tag{2}$$

Each tool's LLM agent evaluates its assigned action and provides feedback:

$$h^k_{t+1} = \text{A}_k \left(\{a_i^{(k,:)} \mid i = 1, \ldots, t\}, \tilde{a}_{t+1}, u_{t,\mathcal{K}} \right) \tag{3}$$

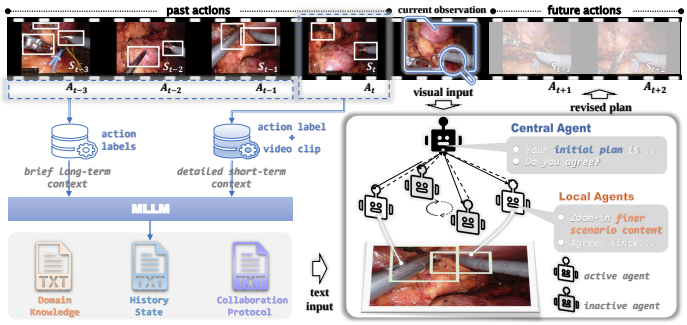

Fig. 2. CSAP-Assist Framework Pipeline: The ReFocus-MM first summarizes historical information by fusing concise distant context (historical action labels) with detailed near context (recent video content). Next, in the HMM, the central agent formulates initial plans, integrating the summarized history and the current scene. Local agents then engage in dialogue with the central agent, providing feedback and iteratively refining plans to reach a consensus.

where $u_{t,K} = \langle \Omega_t, G, CP \rangle$, and CP is the collaboration protocol. The central agent refines the plan based on collective feedback $h_{t+1} = [h^1_{t+1}, ..., h^K_{t+1}]$:

$$\bar{a}_{t+1} = \mathrm{A}_{\mathcal{C}}(\tilde{a}_{t+1}, h_{t+1}, u_{t,\mathcal{C}}) \tag{4}$$

This message passing continues until a consensus of all local agents is reached.

3 Experiments and results

Dataset We evaluated the CSAP-Assist framework using two datasets: a dataset of dual-instrument phantom scenarios for validation, and a dataset of multi-instrument real-world scenarios to assess its application potential.

The MISAW dataset [6] comprises 27 micro-surgical anastomosis sequences performed by six participants on artificial blood vessels, employing bilateral needle holders. We annotated CASP labels by refining the original activity taxonomy into six distinct classes: idle, catch needle, hold needle, pull needle, hold artificial vessel, and insert artificial vessel. While the original dataset provides per-frame action labels for individual instruments, we consolidated consecutive frames with identical actions into temporally coherent action clips. Each clip encapsulates synchronized bilateral interactions, such as {left needle holder: hold artificial vessel, right needle holder: insert artificial vessel}. Discrete action transitions (e.g., interruptions or tool shifts) were segmented into new clips to ensure temporal consistency. For evaluation, the dataset was partitioned into a training set (17 sequences, 558 clips) and a testing set (10 sequences, 169 clips), preserving inter-sequence independence and balancing procedural complexity.

The SAR-RARP50 dataset [14] provides action and surgical instrumentation labels for video segments from 50 Robot-Assisted Radical Prostatectomies.

Our CASP labels, which were derived from the coarse-grained labels of the SAR-RARP50 dataset, were designed to facilitate the collaborative use of six instruments. They include needle grasping, needle puncture, needle handover, needle holding, suture pulling, thread catch, idle, tissue retraction, fluid aspiration, thread dropping, knotting, as well as endoscopic movements like endoscopy down, endoscopy left, endoscopy right, endoscopy zoom in, and endoscopy zoom out. From three long sequences performed at an experienced consultant skill level, we extracted 213 action clips.

Metrics Traditional activity planning metrics like mAcc and ED [18], designed for rigid human activities, poorly reflect surgical realities where context-dependent criticality governs action significance. For instance, non-critical actions (e.g., Idle) tolerate temporal misalignment, while procedural steps (e.g., suture insertion) demand strict spatiotemporal precision.

To bridge this gap, we introduce **Context-Aware mAcc (CA-mAcc)** and **Context-Aware ED (CA-ED)**, which incorporate context-aware tolerance into surgical evaluation. CA-mAcc relaxes temporal constraints for non-critical actions (treat them correct if matched within a 3-step window) while requiring exact matches for critical steps, reflecting the priorities of intra-operative decision-making. CA-ED enhances alignment flexibility for non-critical actions by using minimal Damerau-Levenshtein distance to future ground truth, while maintaining strict sequence fidelity for critical actions. Both metrics offer dual granularity: evaluating action steps at the sample level and assessing sequences case-wise with mean ± standard deviation across sequences.

Comparison Experiments
Baselines: As a zero-shot framework, CSAP-Assist was evaluated against several baselines in two main categories. First, its effectiveness was compared to ***fully-supervised*** planning methods, including: the most probable baseline, which predicts the next action based on the immediately preceding action [13]; the probabilistic sequence baseline, which selects the most probable next action from the entire sequence of past actions [12]; and the VLaMP baseline, which treats action planning as a multimodal sequence modeling problem and utilizes pre-trained language models (GPT2) [13]. Second, to highlight the benefits of CSAP-Assist compared to other zero-shot approaches, we conducted comparisons with a random baseline that selects actions randomly at each step [7], and a VLM baseline as outlined in Section 2.2.
Comparison Results: Table 1 demonstrates that CSAP-Assist outperforms VLaMP [13], a fully supervised method for the VPA task. This enhancement likely results from CSAP-Assist's superior visual encoder and language model, which contain more extensive knowledge than VLaMP.

Furthermore, CSAP-Assist achieves significant improvements over established zero-shot baselines. Compared to the VLM baseline, our framework exhibits a 15% absolute improvement in planning precision, as measured by the CA-mAcc metric. Notably, the alignment between predicted plans and ground-truth actions

Table 1. Performance comparison across different methods on the MISAW Dataset. Best results are **bolded**.

Method	sample-level		sequence-level	
	CA-mAcc (↑)	*CA-ED* (↓)	*CA-mAcc* (↑)	*CA-ED* (↓)
Fully Supervised				
Most Probable [13]	10.61%	5.18	$10.38\% \pm 8.03\%$	5.26 ± 1.04
Probabilistic Sequence [12]	5.30%	6.11	$5.93\% \pm 1.25\%$	5.93 ± 1.25
VLaMP [13]	52.67%	2.06	$52.93\% \pm 15.11\%$	2.28 ± 0.90
Zero-Shot				
Random [7]	14.02%	4.66	$13.05\% \pm 8.74\%$	4.77 ± 1.14
VLM Baseline [18]	45.49%	3.47	$45.62\% \pm 24.56\%$	3.44 ± 1.36
CSAP-Assist(ours)	**59.84**%	**1.91**	**58.53**% ± **19.61**%	**1.99 ± 0.84**

Table 2. Performance of CSAP-Assist framework on the SAR-RARP50 dataset.

Method	sample-level		sequence-level	
	CA-mAcc (↑)	*CA-ED* (↓)	*CA-mAcc* (↑)	*CA-ED* (↓)
CSAP-Assist	49.03%	2.27	$47.60\% \pm 12.28\%$	1.95 ± 0.97

is substantially enhanced, with the CA-ED metric reflecting a 47% reduction in error. Additionally, CSAP-Assist demonstrates generalizability across diverse surgical cases, evidenced by lower standard deviation in sequence-level performance compared to baselines. These findings validate the effectiveness of the our framework in collaborative action planning.

Table 3. Ablations on proposed CSAP-Assist framework on MISAW Dataset.

Method	sample-level		sequence-level	
	CA-mAcc (↑)	*CA-ED* (↓)	*CA-mAcc* (↑)	*CA-ED* (↓)
Memory Module				
(I) w/o A	29.51%	3.09	$31.04\% \pm 16.78\%$	3.10 ± 1.08
(II) w/o O	52.87%	2.24	$52.72\% \pm 21.69\%$	2.28 ± 0.88
(III) independent frames O	57.75%	3.51	$57.36\% \pm 19.91\%$	3.62 ± 1.62
(IV) bundled frames O	55.28%	2.06	$55.56\% \pm 22.29\%$	2.06 ± 0.90
Planning Module				
(V) w/o local agents	48.11%	2.24	$52.90\% \pm 20.80\%$	2.05 ± 0.98
(VI) w/o feedback	54.08%	2.11	$56.96\% \pm 14.20\%$	2.09 ± 0.88
CSAP-Assist(ours)	59.84%	1.91	$58.53\% \pm 19.61\%$	1.99 ± 0.84

Real-scenario Testing We first validate the CSAP-Assist framework in a controlled phantom environment simulating dual-tool collaboration (Fig. 3, top). The framework is then tested on a real-world multi-instrument surgical dataset. Table 3 illustrates that although performance metrics (CA-mAcc/CA-ED) decrease in real-world scenarios due to increased procedural complexity, unclear visual cues, and noisy language instructions, the framework maintains clinically acceptable accuracy. Qualitative results (Fig. 3, bottom) show CSAP-Assist's ability to parse intricate tool-tissue interactions and iteratively refine plans via agent dialogues, even under perceptual uncertainty. For example, when visual occlusions obscure needle trajectories, the framework dynamically infers context

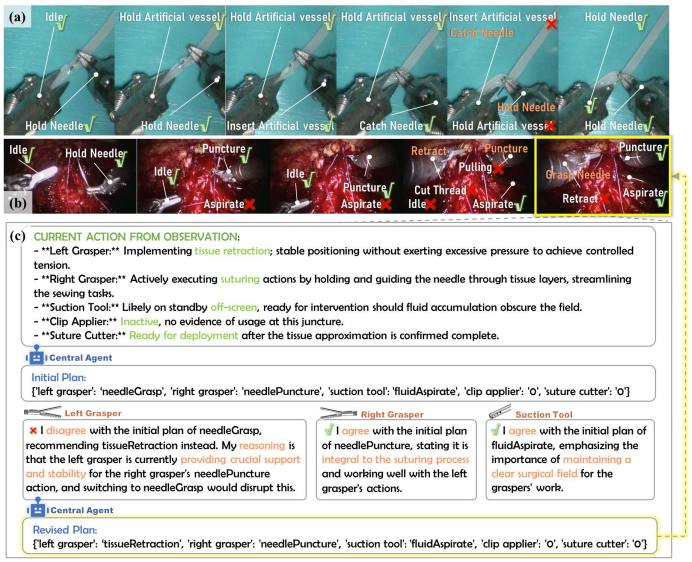

Fig. 3. Planning Result Visualizations. (a) MISAW dataset results. (b) SAR-RARP50 dataset results. White text: CSAP-Assist planning results. $\surd$: correct prediction. $\times$: incorrect prediction. Orange text: ground truth. (c) Example of HMM iteration: textual record of local tool agents refining plans through communication with the central agent.

from historical actions to maintain coherent planning. These findings underscore its adaptability to both structured and unstructured surgical workflows.

Ablation Study We systematically evaluate the contributions of two core components in CSAP-Assist: ReFocus-MM and HMM. Fisrt, we compare four configurations of ReFocus-MM. (I) **w/o A**: Relies solely on video frames, excluding historical action labels. (II) **w/o O**: Retains action labels but discards historical visual context. (III) **Independent Frames O**: Processes individual frames from historical actions as isolated visual-text pairs. (IV) **Bundled Frames O**: Aggregates frames from each action period into unified visual-text inputs. Then, we ablate HMM via two variants. (V) **w/o Local Agents**: Relies exclusively on the global agent, disabling tool-specific refinement. (VI) **w/o Feedback**: Allows local agents to generate plans independently without global coordination.

Ablation results in Table 3 highlight the importance of historical action labels in memory performance, with visual data as supplementary. Setting (III) matches ReFocus-MM's accuracy but shows weaker outcome alignment. ReFocus-MM reduces reliance on long-range visuals, improving efficiency. Removing local agents (Setting V) harms performance, and direct planning (Setting VI) is less effective than HMM with feedback. ReFocus-MM leverages recent context, while HMM balances global and local needs for surgical planning.

4 Conclusion

We propose CSAP, a novel task addressing the underexplored challenges of multi-instrument coordination and tool-action inter-dependencies in RMIS. To bridge this gap, we introduce CSAP-Assist, a vision-language framework that integrates two key innovative modules: the ReFocus-MM for context-aware temporal reasoning and the HMM for hierarchical tool-instrument collaboration. Extensive evaluations on phantom and real-world surgical procedures demonstrate CSAP-Assist's superior ability to generate precise, contextually grounded action plans, outperforming existing methods. To further enhance its clinical applicability, future work will focus on expanding CSAP-Assist's scope to encompass a broader range of surgical procedures.

Acknowledgements. This research work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Projects No. 14208424, No. A-CUHK402/23, No. AoE/E-407/24-N, in part by the National Natural Science Foundation of China under Projects No. 62322318 and 62203180.
Disclosure of Interests. The authors declare no competing interests.

References

1. Cao, J., Yip, H.C., Chen, Y., Scheppach, M., Luo, X., Yang, H., Cheng, M.K., Long, Y., Jin, Y., Chiu, P.W.Y., et al.: Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. Nature Communications **14**(1), 6676 (2023)
2. 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)
3. Chen, Y., Arkin, J., Zhang, Y., Roy, N., Fan, C.: Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4311–4317. IEEE (2024)
4. Fu, J., Long, Y., Chen, K., Wei, W., Dou, Q.: Multi-objective cross-task learning via goal-conditioned gpt-based decision transformers for surgical robot task automation. arXiv preprint arXiv:2405.18757 (2024)
5. 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)
6. Huaulmé, A., Sarikaya, D., Le Mut, K., Despinoy, F., Long, Y., Dou, Q., Chng, C.B., Lin, W., Kondo, S., Bravo-Sánchez, L., et al.: Micro-surgical anastomose workflow recognition challenge report. Computer Methods and Programs in Biomedicine **212**, 106452 (2021)
7. Islam, M.M., Nagarajan, T., Wang, H., Chu, F.J., Kitani, K., Bertasius, G., Yang, X.: Propose, assess, search: Harnessing llms for goal-oriented planning in instructional videos. arXiv preprint arXiv:2409.20557 (2024)

8. Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S.R., Rocktäschel, T., Perez, E.: Debating with more persuasive llms leads to more truthful answers. arXiv preprint arXiv:2402.06782 (2024)
9. 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)
10. Long, Y., Li, X., Cai, W., Dong, H.: Discuss before moving: Visual language navigation via multi-expert discussions. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 17380–17387. IEEE (2024)
11. Mandi, Z., Jain, S., Song, S.: Roco: Dialectic multi-robot collaboration with large language models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 286–299. IEEE (2024)
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. 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)
14. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
15. 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)
16. Zhang, C., Yang, K., Hu, S., Wang, Z., Li, G., Sun, Y., Zhang, C., Zhang, Z., Liu, A., Zhu, S.C., et al.: Proagent: building proactive cooperative agents with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17591–17599 (2024)
17. Zhang, J., Zhou, S., Wang, Y., Wan, C., Zhao, H., Cai, X., Ding, H.: Leveraging surgical activity grammar for primary intention prediction in laparoscopy procedures. arXiv preprint arXiv:2409.19579 (2024)
18. 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)
19. Zhao, Z., Fang, F., Yang, X., Xu, Q., Guan, C., Zhou, S.K.: See, predict, plan: Diffusion for procedure planning in robotic surgical videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 553–563. Springer (2024)
20. Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M.: Text promptable surgical instrument segmentation with vision-language models. Advances in Neural Information Processing Systems **36**, 28611–28623 (2023)