https://doi.org/10.1109/LRA.2025.3535184

From Decision to Action in Surgical Autonomy: Multi-Modal Large Language Models for Robot-Assisted Blood Suction

Sadra Zargarzadeh[1], Maryam Mirzaei[1], Yafei Ou[1], and Mahdi Tavakoli[1,2], *Senior Member, IEEE*

***Abstract*—The rise of Large Language Models (LLMs) has impacted research in robotics and automation. While progress has been made in integrating LLMs into general robotics tasks, a noticeable void persists in their adoption in more specific domains such as surgery, where critical factors such as reasoning, explainability, and safety are paramount. Achieving autonomy in robotic surgery, which entails the ability to reason and adapt to changes in the environment, remains a significant challenge. In this work, we propose a multi-modal LLM integration in robot-assisted surgery for autonomous blood suction. The reasoning and prioritization are delegated to the higher-level task-planning LLM, and the motion planning and execution are handled by the lower-level deep reinforcement learning model, creating a distributed agency between the two components. As surgical operations are highly dynamic and may encounter unforeseen circumstances, blood clots and active bleeding were introduced to influence decision-making. Results showed that using a multi-modal LLM as a higher-level reasoning unit can account for these surgical complexities to achieve a level of reasoning previously unattainable in robot-assisted surgeries. These findings demonstrate the potential of multi-modal LLMs to significantly enhance contextual understanding and decision-making in robotic-assisted surgeries, marking a step toward autonomous surgical systems.**

***Index Terms*—Medical robots and systems, multi-modal large language models, surgical robot, planning, laparoscopy.**

I. Introduction

ROBOT-assisted surgery (RAS) has enormously changed the way many surgeons operate. Surgical robots can enhance accuracy and dexterity, provide better anatomical access, and minimize invasiveness, surgery time, and the need for revision surgery [1]. With the development of surgical robots and the da Vinci Research Kit (dVRK) [2], along with realistic surgical simulation environments [3]–[5], the automation of surgical sub-tasks such as tissue retraction [6], suturing [7], endoscopic camera control [8], cutting [9], and body fluid removal [10], has been an area of research in the past few years. These are the building blocks of surgeries that form the foundation for enhancing bottom-up surgical autonomy [11], [12], and automating these commonly faced sub-tasks provides the basic robot skills necessary for reaching a more advanced level of autonomy, including the ability to reason and plan tasks.

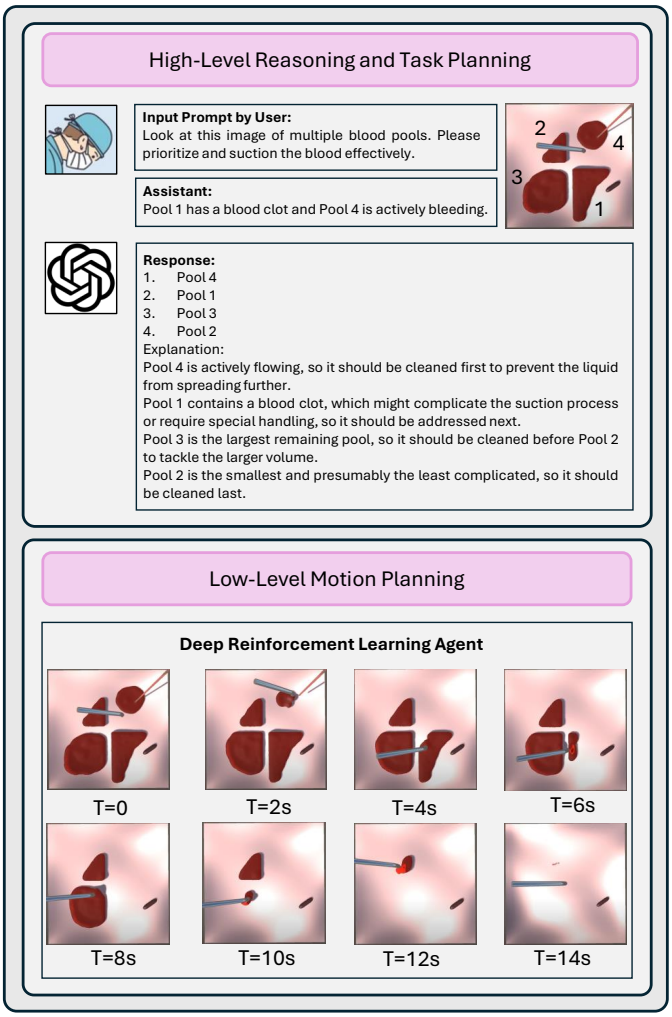


Fig. 1: The high-level task reasoning and planning for the blood suction task is performed by the LLM, and the low-level motion planning and execution is done by the DRL agent.

Autonomous execution of surgical sub-tasks is typically learned through model-based methods [13] or data-driven approaches like deep reinforcement learning (DRL) [14] and

Manuscript received: August, 10, 2024; Revised November, 5, 2024; Accepted January, 19, 2025.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. This research was supported by the Canada Foundation for Innovation (CFI), the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), and Alberta Innovates. (Corresponding author: Sadra Zargarzadeh)

[1]Sadra Zargarzadeh, Maryam Mirzaei, Yafei Ou, and Mahdi Tavakoli are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. {sadra.zar, mmirzaei, yafei.ou, mahdi.tavakoli}@ualberta.ca

[2]Mahdi Tavakoli is with the Department of Biomedical Engineering, University of Alberta, Edmonton, AB, Canada.

Digital Object Identifier (DOI): see top of this page.

imitation learning [15], [16]. However, these methods often lead to behavior that lacks human interpretability, explainability, and adaptability in decision-making. They prioritize maximizing cumulative rewards and efficient exploration of the state space but do not implicitly account for the risks of the actions taken, thus failing to assure safety standards [17]–[19]. Moreover, DRL and imitation learning struggle with dynamic adaptability, particularly in scenarios where surgical conditions deviate from the norm due to unexpected patient anatomy or sudden complications such as bleeding. This inability to reason and act in unforeseen circumstances underscores the need for a reasoning framework that can adapt to challenges, ensuring efficacy and transparent decision-making in autonomous surgical tasks. As surgical sub-tasks involve physical interaction with patients, any automation and decision-making by the robot must be clear and understandable by the operator, resembling human-like behavior to ensure safety and reliability.

Autonomous surgical planning in robot-assisted sub-tasks demands a human-like reasoning unit capable of pre-operatively planning task execution and intraoperatively modifying the plan to accommodate unforeseen circumstances. This approach would enhance the explainability of the robot's decisions and minimize associated risks. Large Language Models (LLMs), trained on vast amounts of text data, have revolutionized natural language understanding and have been adopted in various domains beyond NLP, such as planning and interaction for robots. They can serve as a high-level reasoning unit, breaking down given commands into smaller subtasks to be executed by the robot's lower-level systems, including motion planning and control.

Integrating LLMs into robots poses challenges due to their struggle with real-world complexities. While LLMs offer general knowledge and expertise, they lack a connection to tangible reality, leading to errors and potentially unsafe recommendations. Extracting information from LLMs for robots requires balancing theoretical knowledge with practical understanding to navigate dynamic environments and facilitate effective human-robot interaction. An agent needs to comprehend semantic aspects of the world, the range of available skills, how these skills influence the environment, and how changes in the world map back to language [20]. The integration of robotics and LLMs, as introduced through Google's SayCan [21], PaLM-E [22], and more recent methods [23]–[30], presents immense opportunities for exploration in domains such as surgical robotics.

By leveraging multi-modal LLMs in a zero-shot manner, in this work we aim to surmount the limitations of current autonomous systems, introducing a level of reasoning and adaptability previously unattainable in robot-assisted surgeries. Integrating images with text allows the multi-modal LLM to capture important visual nuances, such as spatial relationships and the presence of surgical tools near blood pools, which may not be fully conveyed through text alone. This integration is pioneering, as it combines the theoretical knowledge of LLMs with the practical demands of surgical environments as identified and communicated by the medical staff. We propose an LLM-powered framework capable of high-level reasoning, mid-level motion planning, and execution for autonomous blood suctioning in robot-assisted surgeries. The reasoning and prioritization responsibility is delegated to the higher-level task planning LLM, and the motion planning and execution are delegated to the DRL model, leading to a distributed agency between the two components. Expanding on the foundation laid with autonomous blood suctioning, our vision extends beyond this surgical subtask and sets a precedent for the future of autonomous systems in the surgical field.

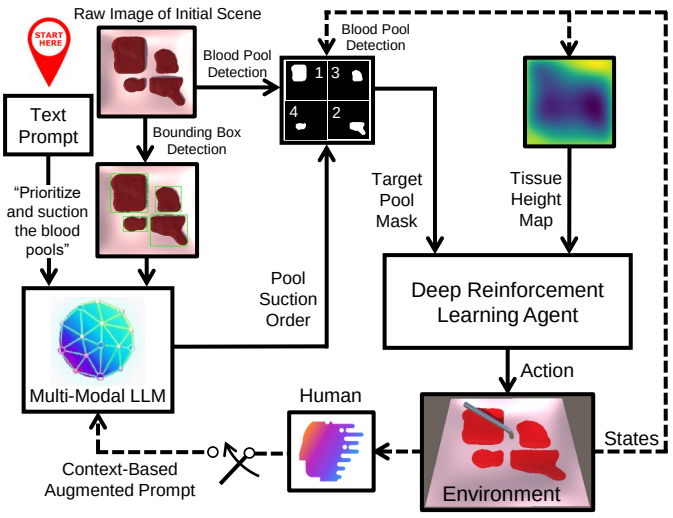


Fig. 2: System architecture.

The main contributions of this work are as follows:

1) We propose an LLM-powered framework for autonomous robot-assisted blood suction, where task reasoning and planning are managed by the LLM, while motion planning and execution are handled by a DRL-trained agent.
2) We compare the performance of LLM reasoning to random reasoning and no reasoning modules in terms of blood removal time and tool movement.
3) We analyze how augmenting the prompts with context and expert-defined guidelines influences the reasoning capabilities of the LLM in zero-shot prompting. A user study is also conducted to assess the similarity to human-like behavior across the three modules.

The paper is structured as follows. Section II reviews the integration of LLMs in robotics and surgery and their impact on these fields. Section III introduces the methods used, including the system architecture, multi-modal LLMs, prompt augmentation, the simulation environment, and the deep reinforcement learning module. Section IV presents and discusses the experiments and results. Section V outlines the limitations and future work, while Section VI concludes the paper.

II. Related Work

A. Large Language Models in Robotics

In a pioneering work in LLM-Robotics integration, Google introduces SayCan [21], a method that aims to extract knowledge stored within LLMs for physically grounded tasks. The

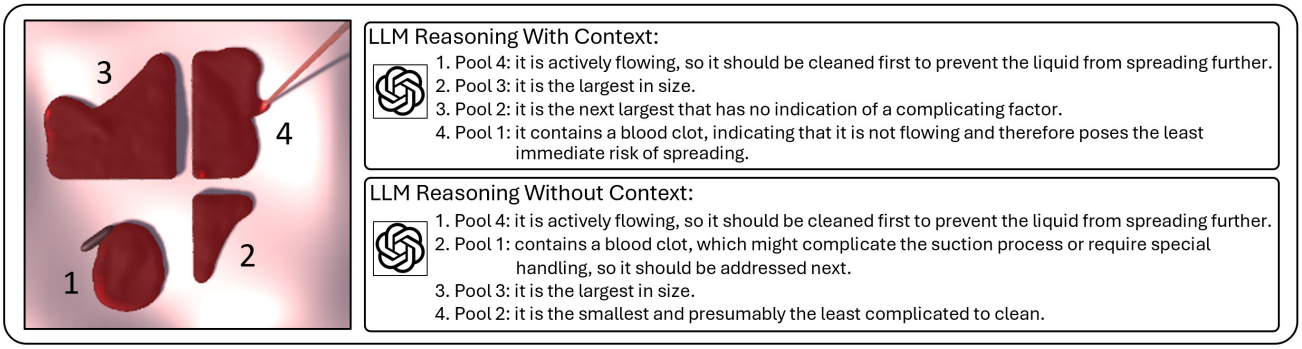


Fig. 3: An example of LLM reasoning with (LRWC) and without (LRWOC) context-based prompt augmentation. The guideline provided to the LLM is as context is as follows: Address active bleeding first, consider pool size next, and address the blood clot pool last, as coagulation ensures that flow in this pool has ceased and will not propagate further.

LLM, 'Say', breaks down instructions into subtasks and evaluates each skill's contribution likelihood. Affordance functions, 'Can', assess each skill's success potential in the environment. This approach combines these factors to determine the effectiveness of each skill in fulfilling the instruction. In their later work, PaLM-E [22], they propose an embodied language model that processes multi-modal sentences, blending visual, continuous state estimation, and textual inputs, leading to the integration of real-world continuous sensor data into language models and creating a direct connection between words and sensory experiences.

While these recent works have shown great advancement in the integration of LLMs into robotics, they mainly focus on general tasks such as object manipulation. More specific application domains, such as robot-assisted surgery, pose new and important challenges that need to be addressed to ensure the safety of patients in surgical operations. For instance, despite advancements, existing LLM applications have yet to fully tackle the real-time adaptation and decision-making required in dynamic surgical settings. Nevertheless, the success of LLMs in broader robotics tasks can lay a foundational understanding that is crucial when approaching more specialized domains such as robot-assisted surgery where data is scarce.

B. Large Language Models in Surgery and Surgical Planning

LLMs can assist in surgical planning by analyzing vast amounts of medical literature, patient data, and clinical guidelines to suggest the most appropriate surgical approaches. This includes evaluating the risks and benefits of different surgical options and considering patient-specific factors such as age and previous surgical history and can be used in various surgeries such as joint arthroplasty [31], oral and maxillofacial surgery [32], and orthopedics [33]. It has the ability to process and generate complex language-based instructions, enabling bidirectional communication with the medical team in real-time in an intuitive, fast, and natural way. This capability is essential for improved decision-making in dynamic surgical environments, where rapid and accurate communication is crucial for patient safety and successful outcomes [34].

Although an initial step towards integrating LLMs in robotic surgery has been made in SUIFA [35] through tasks such as needle lift and shunt insertion, the current gaps in the literature include the adoption of LLMs as reasoning units in the planning process of surgical tasks where decision-making is crucial. In this work, we investigate the reasoning capability of LLMs in prioritizing the suction of multiple blood pools under different circumstances and integrate them into robot-assisted surgery for autonomous blood suctioning.

III. METHODS

A. Blood Suctioning Task and System Architecture

Blood is among the most frequent types of fluids encountered in surgical settings, as bleeding is a common and sometimes unpredictable occurrence during operations. Surgeons typically need to address bleeding promptly by clearing the area and pinpointing its source before proceeding with other tasks. Suction of the blood with the proper tool becomes necessary to extract blood from the surgical site. Consequently, this task is indispensable and often consumes significant time and effort, and automating this process reasonably and safely can alleviate the burden on surgeons. In a dynamic environment, such as the human body, fluids such as blood move around, forming pools with different conditions. For instance, characteristics such as active bleeding, blood clots, variations in blood pool sizes, and the closeness of blood pools to critical organs or surgical instruments affect the priority with which they need to be suctioned.

As shown in Fig. 1, an image of the tissue scene containing multiple blood pools is annotated with bounding boxes around each pool and passed to the multi-modal large language model, along with the text prompt, "Look at this image of multiple blood pools. Please prioritize and suction the blood effectively." The proposed system architecture, illustrated in Fig. 2, depicts how the multi-modal LLM, as the high-level reasoning unit, uses its reasoning capability to prioritize the order in which the blood pools need to be suctioned and informs the masking sensor accordingly. The pools are masked in the order that they need to be suctioned and then fed into the trained DRL agent, which acts as the lower-level motion planner, along with the tissue height map, and leads to an action taken by the suction tool. In complex scenarios where additional information is needed for the LLM to prioritize suction, human input can augment the initial prompt with context.

B. Multi-Modal Large Language Models

Our methodology leverages multi-modal LLMs to allow for information from diverse modalities, including text and images. In multi-modal LLMs, textual data undergoes processing through a standard language model architecture. Feature embeddings extracted from image inputs are concatenated with the textual embeddings, yielding a multi-input representation. This fused representation is then fed into a multi-layer neural network, facilitating joint learning across modalities. In this work, we employed the pre-trained OpenAI GPT-4V model [36] for image understanding and reasoning.

As shown in Fig. 1 and explained in the system architecture section, the multi-modal LLM accepts a text prompt and an image of the labeled blood pools. The prompt outlines essential details about blood pools, including the presence of clots, and signs of active bleeding. Each of these aspects is crucial in deciding the urgency of suctioning blood pools during medical procedures. Size indicates the amount of blood, while clots indicate coagulation and cessation of blood flow. Active bleeding is a significant signal that could suggest the potential for blood to spread more extensively. By incorporating these factors into the prompt, we equip the multi-modal LLM with essential context for making informed decisions regarding blood pool suctioning. The LLM conducts reasoning based on the image and the prompt, yielding the proper priority for suctioning the blood, and explains why this priority is chosen in a zero-shot manner. Our zero-shot approach allows the LLM to generate relevant responses for each prompt without specific training examples, relying on its general understanding of language semantics and prompt cues to address previously unseen conditions effectively.

In scenarios where we encounter a combination of factors concerning blood pools simultaneously, such as when both active bleeding and a blood clot exist in two of the pools, the LLM may lack consistency in reasoning due to the lack of training on medical data. To address this, we activate the switch as seen in Fig. 2 and exemplified in Fig. 3, and augment the prompt with additional contexts providing a guideline for the model in generating a consistent priority in complex scenarios. This approach remains in a zero-shot manner as we refrain from furnishing the model with specific examples. Instead, we enhance the prompt through context augmentation as part of a prompt engineering process.

C. Blood Suction Environment and Mask Sensor Mechanism

In our recent work [10], a blood suction simulation environment for RL was built using position-based fluids (PBF) based on Nvidia PhysX 5, Unity, and Unity's ML-Agents toolkit. PBF is an approach that represents fluids using a large number of small particles that interact with each other. PhysX is a real-time physics engine with GPU optimization, which allows for PBF simulation. With PhysX 5 as the low-level physics engine, the main simulation environment was built in Unity. In this model, blood is simulated as particles influenced by forces like gravity and suction, with a spherical cone-shaped force field applied to particles near the suction tool. The force decays with distance to realistically simulate suction, removing particles once they reach a specified height threshold. We leverage the existing simulation environment and further build upon it in this work.

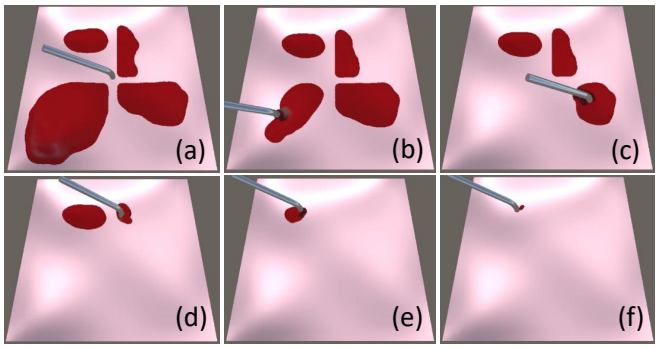


Fig. 4: Simulation Environment 1. The LLM reasoning prioritizes suctioning the pools based on their size in the absence of surgical complexities such as active bleeding and blood clots as seen in (a)-(f).

The simulation environment consists of a randomly generated tissue that contains the blood, the simulated blood, and a suction tool. We simulated a fixed amount of blood (4000 particles) through PBF and added suction force to each particle within a suction range to simulate the effect of suction. Particles that are suctioned close enough to the suction tool will be removed and marked as inactive. To introduce randomness in the shape of the tissue, Bezier surfaces with random control points were used to generate random shapes. The Bezier surfaces are represented by

$$\mathcal{S}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{m} P_{i,j} \cdot B_{n,i}(u) \cdot B_{m,j}(v),\ 0 \le u, v \le 1 \quad (1)$$

where $P_{i,j}$ are the control points, and $B_{n,i}(u)$ and $B_{m,j}(v)$ are the Bernstein basis functions defined by

$$B_{n,i}(x) = \frac{n!}{i! \cdot (n-i)!} \cdot x^i \cdot (1-x)^{n-i}. \quad (2)$$

In this work, several features are added to the simulation environment, including a module that would allow blood to continuously add to a random pool representing active bleeding, the addition of a capsule-shaped object representing a blood clot that would be randomly positioned on the tissue, and the ability to generate multiple independent blood pools. We also developed a blood pool detector algorithm that takes in the raw image of the scene along with the suction orders from the multi-modal LLM and outputs the mask of the target blood pool based on the priority reasoned by the LLM. This mechanism would allow the agent to plan the suction motion of only the target pool and would blind the agent to other existing pools unless commanded otherwise by the LLM.

We developed four simulation environments. In Environment 1, as can be seen in Fig. 4, four independent pools are randomly generated without the presence of a blood clot or active bleeding. In Environment 2, active bleeding occurs randomly in one of the pools for a fixed time interval, with no sign of a blood clot. Environment 3 features the introduction of a capsule-shaped object representing a blood clot, randomly generated near one of the pools, indicating coagulation and

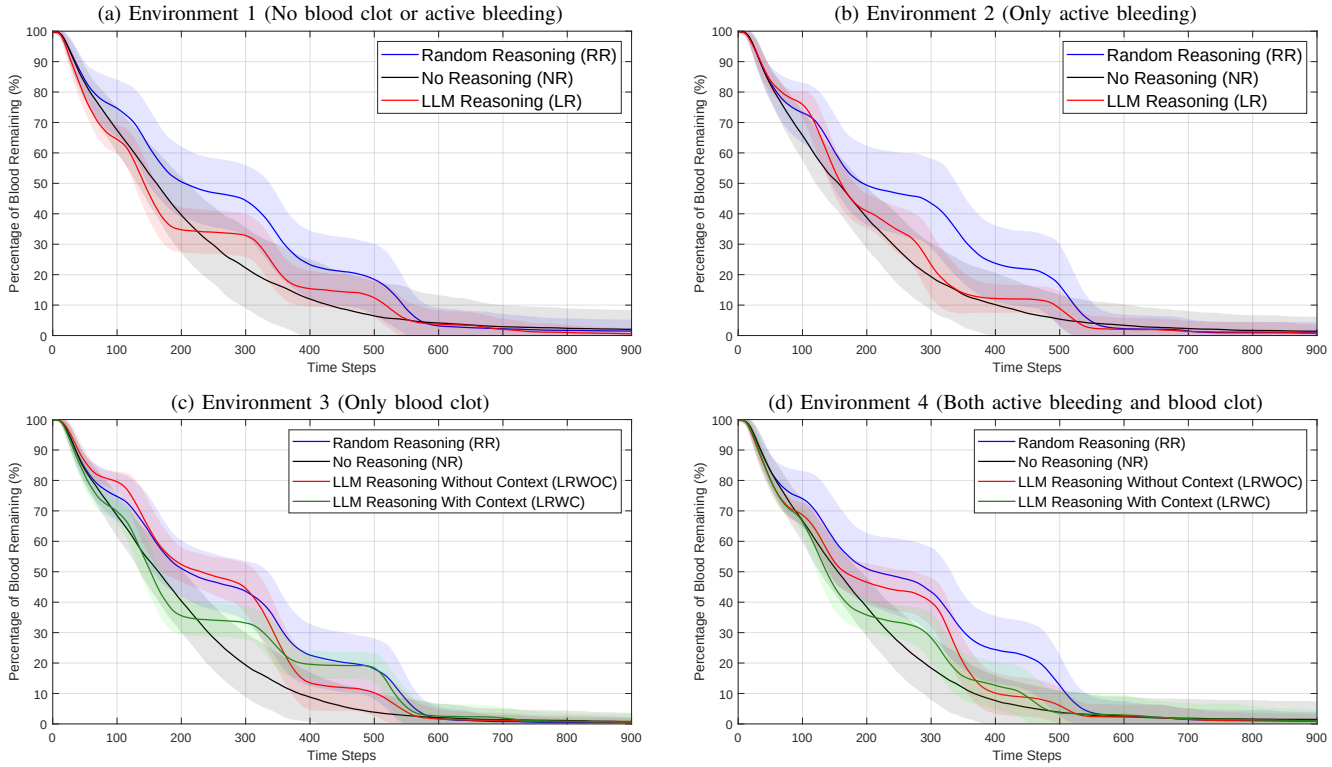


Fig. 5: Progression in blood suction in the four environments.

the cessation of potential bleeding. Finally, Environment 4 includes a pool actively bleeding, while another contains a blood clot. To test the realism of our simulation environments, visual comparisons with real surgical environments were conducted, ensuring our model visually reflects the dynamics of blood flow and bleeding during real surgical procedures.

D. Motion Planning Using Deep Reinforcement Learning

In our recent work [10], an RL agent for completing autonomous blood suction was obtained. To train this agent, we used the following reward function, which consists of a reward for the amount of blood removed during each step, an extra terminal reward for removing all blood, and an action penalty for tool movements. The number of particles being removed during each step is used to determine the amount of blood being suctioned out.

$$r(s_t, a_t, s_{t+1}) = N_p^t - N_p^{t+1} + C_1\,\delta(N_p^{t+1}) - C_2\|a_t\| \quad (3)$$

In the above equation, t is the time step, N_p^t is the number of active particles, $\delta(N_p^{t+1})$ denotes whether there are active particles remaining, and $\|a_t\|$ is the norm of the actions. The weighting factors $C_1 = 5$ and $C_2 = 0.02$ were chosen to balance task efficiency and control stability, emphasizing blood removal while discouraging excessive motion. Specifically, C_1 is set to a higher value to prioritize the reward for full blood clearance, encouraging the agent to complete the task efficiently, while C_2 is relatively small, penalizing movements without restricting necessary adjustments. The observation includes the tissue height map, the binary image mask of the blood (stacked with 3 from previous steps), and the suction tool location (stacked with 4 from previous steps). The binary image mask of the blood is synthesized from the current positions of all active particles in the blood.

IV. Experiments and Results

We investigate four reasoning modules in our experiments. When the LLM, as the high-level reasoning unit, reasons the sequence for suctioning blood pools without any additional context, it is termed LLM Reasoning Without Context (LRWOC). If additional context is provided by the assistant, leading to an augmented prompt for the LLM, we call this LLM Reasoning With Context (LRWC). When the DRL agent receives a randomly generated order based on a random permutation of the number of blood pools, this is known as Random Reasoning (RR). If the DRL agent proceeds to suction blood pools solely based on its reward function without input from a higher-level unit, it is termed No Reasoning (NR).

A. Comparison of LLM Reasoning with Random Reasoning and No Reasoning

To evaluate the performance of the RR, NR, LRWOC, and LRWC modules, we simulated the blood suction task across 400 distinct scenes (100 scenes per environment). Fig. 5 illustrates the blood remaining percentage over time, providing a comparative analysis of the different reasoning modules across different environments. Additionally, Table I presents key metrics such as the mean and standard deviations of the time to suction the actively bleeding pool in Environments 2 and 4 (T_{AB}), the time to suction 50% (T_{50}) and 95% (T_{95}) of the blood, and total tool path length (TTPL).

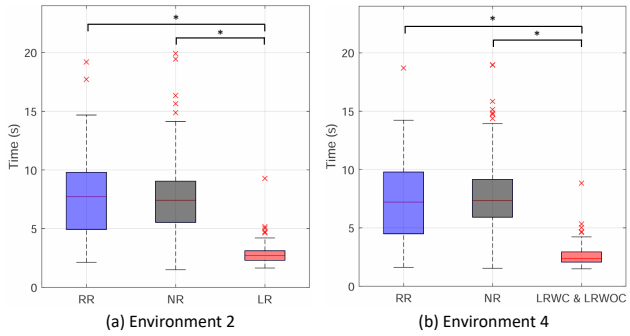


Fig. 6: Time to suction the actively bleeding pool (T_{AB}) in Environments 2 and 4 by different reasoning modules.

In Environment 1 (Fig. 5a), the LLM reasoning module reasons that larger volumes of blood must be addressed first and prioritizes suctioning the pools based on their size, resulting in a more rapid initial suction compared to the other modules leading to a faster average T_{50}. The NR module shows a gradual decrease in blood remaining by suctioning parts of pools as it moves between them, while the LR and RR modules prefer to suction one pool before moving to the next, resulting in intervals where the slope decreases indicating movement between pools.

In Environment 2 (Fig. 5b), where active bleeding is present in one of the pools, the LLM reasoning module gives priority to suctioning the pool with active bleeding first, even if it is smaller than others. This approach results in a faster T_{AB}. After addressing the pool with active bleeding, the LLM then proceeds to prioritize suctioning the pools based on their size, similar to the strategy observed in Environment 1. Providing additional instructions as the augmented prompt did not change the LLM reasoning in Environments 1 and 2 and the focus remained on prioritizing active bleeding, followed by pool size. This resulted in the same LRWOC and LRWC, denoted as LR.

In Environment 3 (Fig. 5c), the presence of a blood clot adds complexity to suctioning. LRWOC tends to prioritize the suction of the pool with the blood clot potentially due to its perceived complexity. However, the user can define rules for the LLM to follow, tailoring its reasoning to specific conditions. For this environment, we establish the following rule: 1) Address active bleeding first, 2) Consider pool size next, and 3) Address the blood clot pool last, as coagulation ensures that flow in this pool has ceased and will not propagate further. An example of this is shown in Fig. 3. This context leads to a faster average T_{50} in LRWC, prioritizing blood pool size after active bleeding. In Environment 4 (Fig. 5d), both LRWOC and LRWC modules start by suctioning actively bleeding pools. LRWC prioritizes larger pools next and saves the pool with the blood clot for last, resulting in a faster T_{50} compared to LRWOC, which targets the pool with the blood clot before proceeding based on pool size.

Fig. 6 illustrates the time to suction the actively bleeding pool (T_{AB}) in Environments 2 and 4 and shows statistically significant improvement (*) when LLM reasoning is used. Table I shows that the LRWC module results in smaller standard deviations across all metrics and environments, indicating more consistent performance. Although the NR module exhibited a marginally smaller average TTPL, its inability to reason and adapt to unforeseen circumstances makes it unreliable in highly dynamic surgical settings.

TABLE I: Key metrics in comparison of reasoning modules expressed as mean±std (Time step = 0.02 seconds).

		T_{AB}	T_{50}	T_{95}	TTPL
Environment 1	RR	–	225±82	570±128	33.9±3.2
	NR	–	164±38	557±110	32.4±4.1
	LR	–	**145±30**	**554±101**	**32.5±2.8**
Environment 2	RR	363±176	223±89	573±130	33.3±3.4
	NR	390±174	154±37	514±106	31.5±3.9
	LR	**128±50**	**161±28**	**523±95**	**33±2.8**
Environment 3	RR	–	222±71	563±124	32.4±2.9
	NR	–	166±34	466±98	31.2±4.2
	LRWOC	–	239±58	541±103	32.4±3.0
	LRWC	–	**151±23**	**553±100**	**32.7±2.9**
Environment 4	RR	359±178	227±86	535±120	33.7±3.2
	NR	393±172	155±34	462±104	30.4±3.8
	LRWOC	**130±49**	179±55	509±93	33.7±2.9
	LRWC	**130±49**	**137±18**	**470±90**	**32.9±2.8**

B. Advantage of Multi-Modal LLMs

To further demonstrate the advantage of the multi-modal LLM, we conducted an experiment to test its ability to capture contextual details that may not be explicitly provided in text. In this experiment, we presented the model with 10 images in which a surgical tool was positioned near one of the blood pools, using our original prompt without any mention of the tool in text. The multi-modal LLM correctly recognized the presence of the tool in 80% of cases (8 out of 10) and incorporated this visual information into its decision-making process, placing this priority right after the active bleeding pool. Unlike hard-coded logic, which lacks flexibility in unforeseen scenarios, the LLM leverages context awareness and visual information to adapt to these conditions interpreting nuanced visual cues that a text-only input or a hard-coded logic might overlook.

C. User Study on Closeness to Human-Like Behaviour

To assess the similarity to human-like behavior exhibited by the RR, NR, and LRWOC modules during blood suction, a survey questionnaire was conducted involving ten participants. The participants, consisting of graduate students and senior researchers with no specialized medical backgrounds, were chosen to assess how well a non-expert-defined context aligns with broader, non-specialist perspectives in evaluating the behavior of the RR, NR, and LRWOC modules. A total of thirty-six videos were collected for the survey, divided among three reasoning modules (RR, NR, LRWOC), with each module contributing twelve videos from four environments (three videos per environment), showcasing the blood suctioning task. Participants were presented with pairs of videos and asked, “If you were the human operator, which of the two videos shown below would you choose to suction the blood pools?”. This forced choice questioning process was repeated to compare all combinations of LRWOC versus NR versus RR.

A second user study survey involved collecting nine videos each of LRWOC and LRWC (in Environments 3 and 4), which were then presented to the participants with the same question. A human performance score was defined as the number of videos selected by the participants normalized by the total number of videos, as shown in Fig. 7. The objective of these surveys was to investigate whether LLM reasoning aligns more with actual human decision-making than random reasoning and no reasoning modules and also to assess how providing our user-defined context impacts the LLM's ability to mimic human decision-making. This research was approved by the University of Alberta's Research Ethics Board under approval ID Pro00139696.

A one-way analysis of variance (ANOVA) test was used in the first survey and a paired t-test in the second survey to establish statistical significance among reasoning modules, as shown in Fig. 7. The first study yielded a p-value$<$0.001, showing that LRWOC had significantly more human-like blood suction behavior. The second study also resulted in a p-value$<$0.001. The user study results, indicating a preference for the LRWC over the LRWOC module's decision-making, suggest that incorporating contextual understanding in robotic surgery could bridge the gap between automated procedures and the intuitive decision-making of humans.

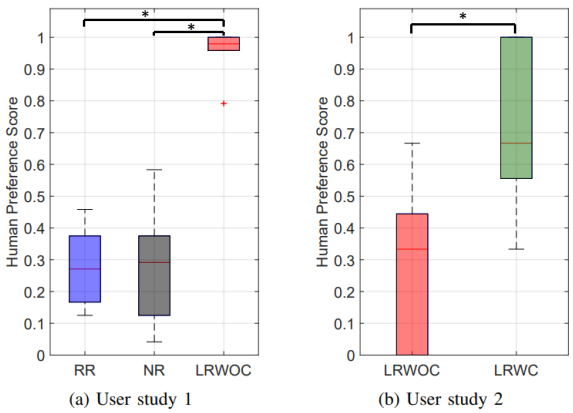
Although results show a promising stride towards explainable surgical autonomy, a thorough evaluation process including clinical trials is essential to establish the efficacy and safety of LLM-powered robotic systems in real-world surgical applications. Additionally, the development of new training protocols for surgical teams on interacting with and overseeing LLM-enabled systems, along with intuitive interfaces for surgeons to interact with and override the system's decisions when necessary, will be vital for adoption.

V. Limitations and Future Work

While the proposed method demonstrates the effectiveness of multi-modal LLMs in reasoning and decision-making within a simulated environment, several limitations and areas for future work remain, particularly for transitioning to real-world applications in surgical settings.

This study assumed that blood pools are separate and independent, simplifying interactions within the environment. Additionally, the current system operates below real-time performance due to the generation speed of OpenAI's GPT-4V, which constrains its immediate applicability in time-sensitive clinical tasks. The physical experiments demonstrated in our previous work in blood suction [10] show the real-world execution of this task and hence were not the main focus of this study.

Simulation-based environments, while inherently limited compared to real-world settings, offer the critical advantage of encompassing a wide range of scenarios, including rare but pivotal situations that are difficult to consistently reproduce in physical setups. This simulation-first approach provides a foundation to test and refine the system's decision-making capabilities across varied conditions, setting the groundwork for application in real surgical environments.



(a) User study 1 (b) User study 2

Fig. 7: Comparison on closeness to human-like behavior of different reasoning modules.

Applying this framework to real surgical contexts introduces practical challenges, such as accurate segmentation of blood from live camera images, real-time pose coordination of surgical instruments, and adapting to the dynamic and complex environment of an actual operating room. Future work will address these challenges by refining and validating the framework's applicability in real surgical tasks, exploring predictive mechanisms and pre-emptive prompts to prioritize time-sensitive actions. To enhance real-time feasibility, model distillation, quantization, and smaller, task-specific models will be investigated, enabling the system to better align with the operational demands of clinical practice.

Ensuring the accuracy and safety of LLM-based decisions is another priority. LLMs are prone to hallucinations, which could impact decision reliability. To mitigate this, future iterations will integrate rule-based checks and domain-specific constraints, as well as feedback loops from medical experts, to improve the system's robustness and decision accuracy in high-stakes environments. Gathering insights from medical professionals will also help assess the LLM's alignment with expert-defined surgical priorities, refining its adaptability to clinical needs.

Safety in surgical tasks is essential, and the current framework will be expanded to address this by incorporating force-based thresholds to control tool speed and acceleration, collision detection, and Safe Reinforcement Learning (Safe RL) techniques, such as reward shaping and risk-sensitive policies, to enhance safe tool proximity to sensitive tissues. Additionally, to better handle dynamic scenarios, future work will explore mid-sequence task re-planning to adapt to external disturbances and refine LLM reasoning capabilities through reinforcement learning from human feedback, ultimately improving the framework's scalability to other surgical and medical tasks.

VI. Conclusion

In this study, we proposed a multi-modal LLM integration in robot-assisted surgery for autonomous blood suction and investigated how the addition of a high-level reasoning unit can influence decision-making and performance. Experiments were conducted to analyze LLM reasoning in comparison to

random reasoning and no reasoning modules. Active bleeding and blood clots were introduced to influence decision-making as is also common in highly dynamic surgical settings. Results showed that the presence of a multi-modal LLM as a higher-level reasoning unit can account for these surgical complexities in decision-making and prioritization to achieve a level of reasoning and explainability previously unattainable in robot-assisted surgeries. The user study showed that incorporating contextual understanding in robotic surgery could bridge the gap between automated procedures and the intuitive decision-making of humans.

REFERENCES

[1] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," *Springer handbook of robotics*, pp. 1657–1684, 2016.

[2] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci® surgical system," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 6434–6439.

[3] Y. Ou, S. Zargarzadeh, P. Sedighi, and M. Tavakoli, "A realistic surgical simulator for non-rigid and contact-rich manipulation in surgeries with the da vinci research kit," *arXiv preprint arXiv:2404.05888*, 2024.

[4] P. M. Scheikl, B. Gyenes, R. Younis, C. Haas, G. Neumann, M. Wagner, and F. Mathis-Ullrich, "Lapgym-an open source framework for reinforcement learning in robot-assisted laparoscopic surgery," *Journal of Machine Learning Research*, vol. 24, no. 368, pp. 1–42, 2023.

[5] J. Xu, B. Li, B. Lu, Y.-H. Liu, Q. Dou, and P.-A. Heng, "Surrol: An open-source reinforcement learning centered and dvrk compatible platform for surgical robot learning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1821–1828.

[6] P. M. Scheikl, E. Tagliabue, B. Gyenes, M. Wagner, D. Dall'Alba, P. Fiorini, and F. Mathis-Ullrich, "Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 560–567, 2022.

[7] S. Sen, A. Garg, D. V. Gealy, S. McKinley, Y. Jen, and K. Goldberg, "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4178–4185.

[8] Y. Ou, S. Zargarzadeh, and M. Tavakoli, "Robot learning incorporating human interventions in the real world for autonomous surgical endoscopic camera control," *Journal of Medical Robotics Research*, vol. 8, no. 03n04, p. 2340004, 2023.

[9] N. D. Nguyen, T. Nguyen, S. Nahavandi, A. Bhatti, and G. Guest, "Manipulating soft tissues by deep reinforcement learning for autonomous robotic surgery," in *2019 IEEE International Systems Conference (SysCon)*. IEEE, 2019, pp. 1–7.

[10] Y. Ou, A. Soleymani, X. Li, and M. Tavakoli, "Autonomous blood suction for robot-assisted surgery: A sim-to-real reinforcement learning approach," *IEEE Robotics and Automation Letters*, 2024.

[11] F. Lalys and P. Jannin, "Surgical process modelling: a review," *International journal of computer assisted radiology and surgery*, vol. 9, pp. 495–511, 2014.

[12] E. Tagliabue, A. Pore, D. Dall'Alba, E. Magnabosco, M. Piccinelli, and P. Fiorini, "Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 3261–3266.

[13] J. Huang, F. Liu, F. Richter, and M. C. Yip, "Model-predictive control of blood suction for surgical hemostasis using differentiable fluid simulations," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 12 380–12 386.

[14] Y. Ou and M. Tavakoli, "Sim-to-real surgical robot learning and autonomous planning for internal tissue points manipulation using reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2502–2509, 2023.

[15] H. Su, A. Mariani, S. E. Ovur, A. Menciassi, G. Ferrigno, and E. De Momi, "Toward teaching by demonstration for robot-assisted minimally invasive surgery," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 484–494, 2021.

[16] Y. Hu and M. Tavakoli, "Autonomous ultrasound scanning towards standard plane using interval interaction probabilistic movement primitives," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3719–3727.

[17] A. Pore, D. Corsi, E. Marchesini, D. Dall'Alba, A. Casals, A. Farinelli, and P. Fiorini, "Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4025–4031.

[18] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[19] K. Fan, Z. Chen, G. Ferrigno, and E. De Momi, "Learn from safe experience: Safe reinforcement learning for task automation of surgical robot," *IEEE Transactions on Artificial Intelligence*, 2024.

[20] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[21] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[22] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[23] H. Liu, Y. Zhu, K. Kato, I. Kondo, T. Aoyama, and Y. Hasegawa, "LLM-based human-robot collaboration framework for manipulation tasks," *arXiv preprint arXiv:2308.14972*, 2023.

[24] S. S. Kannan, V. L. Venkatesh, and B.-C. Min, "Smart-LLM: Smart multi-agent robot task planning using large language models," *arXiv preprint arXiv:2309.10062*, 2023.

[25] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.

[26] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530.

[27] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "ChatGPT for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, p. 20, 2023.

[28] Z. Long, G. Killick, R. McCreadie, and G. A. Camarasa, "Robollm: Robotic vision tasks grounded on multimodal large language models," *arXiv preprint arXiv:2310.10221*, 2023.

[29] M. G. Arenas, T. Xiao, S. Singh, V. Jain, A. Z. Ren, Q. Vuong, J. Varley, A. Herzog, I. Leal, S. Kirmani *et al.*, "How to prompt your robot: A promptbook for manipulation skills with code as policies," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[30] Y. Dai, R. Peng, S. Li, and J. Chai, "Think, act, and ask: Open-world interactive personalized robot navigation," *arXiv preprint arXiv:2310.07968*, 2023.

[31] K. Cheng, Z. Li, C. Li, R. Xie, Q. Guo, Y. He, and H. Wu, "The potential of GPT-4 as an AI-powered virtual assistant for surgeons specialized in joint arthroplasty," *Annals of Biomedical Engineering*, pp. 1–5, 2023.

[32] B. Puladi, C. Gsaxner, J. Kleesiek, F. Hölzle, R. Röhrig, and J. Egger, "The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review," *International journal of oral and maxillofacial surgery*, 2023.

[33] S. Chatterjee, M. Bhattacharya, S. Pal, S.-S. Lee, and C. Chakraborty, "ChatGPT and large language models in orthopedics: from education and surgery to research," *Journal of Experimental Orthopaedics*, vol. 10, no. 1, p. 128, 2023.

[34] K. Cheng, Z. Sun, Y. He, S. Gu, and H. Wu, "The potential impact of chatgpt/gpt-4 on surgery: will it topple the profession of surgeons?" *International Journal of Surgery*, pp. 10–1097, 2023.

[35] M. Moghani, L. Doorenbos, W. C.-H. Panitch, S. Huver, M. Azizian, K. Goldberg, and A. Garg, "Sufia: Language-guided augmented dexterity for robotic surgical assistants," *arXiv preprint arXiv:2405.05226*, 2024.

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