SAP-Bench：多模态大型语言模型在外科动作计划中的基准测试

Mengya Xu[1,*], Zhongzhen Huang[2,*], Dillan Imans[3], Yiru Ye[4], Xiaofan Zhang[2], Qi Dou[1]
[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Sungkyunkwan University, Seoul, South Korea
[4] The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[*] Equal contribution Corresponding author

Abstract

有效的评估是推动 MLLM 研究进展的关键。手术动作规划（SAP）任务旨在从视觉输入生成未来的动作序列，要求精确和复杂的分析能力。与数学推理不同的是，手术决策涉及生命关键领域，需要仔细、可验证的过程以确保可靠性和患者安全。该任务要求能够区分原子视觉动作并协调复杂的长时间实施过程，而这些能力在当前的标准中都评估得不够充分。为了解决这一问题，我们推出了 SAP-Bench，一个大规模高质量的数据集，旨在使多模态大型语言模型（MLLMs）能够进行可解释的手术动作规划。我们的 SAP-Bench 基准源自胆囊切除手术背景，平均时长为 1137.5s ，并引入了时序分段的手术动作注释，包括 1,226 经过临床验证的动作剪辑（平均时长：68.7s ）捕捉了 74 个手术中的五种基本手术动作。该数据集提供了 1,152 个战略性抽样的当前帧，每个帧都与相应的下一个动作配对，作为多模态分析的锚点。我们提出了 MLLM-SAP 框架，该框架利用 MLLMs 生成基于当前手术场景和自然语言指令的下一步动作推荐，并注入手术领域知识进行增强。为了评估我们数据集的有效性以及当前模型的广泛能力，我们评估了七个最先进的 MLLM（例如，OpenAI-o1、GPT-4o、QwenVL2.5-72B、Claude-3.5-Sonnet、GeminiPro2.5、Step-1o 和 GLM-4v），揭示了在下一步动作预测性能方面的重要缺口。数据集可在 `https://huggingface.co/datasets/SAPbench/SAPBench` 获得。

1 引言

人工智能应用日益复杂，揭示了单模态数据处理的局限性，使得研究重点转向多模态推理模型。虽然多模态大语言模型（MLLMs）在视觉-语言任务（如图像描述、视觉问答）方面取得了显著进展，但当前方法仍面临处理复杂多模态任务、适应动态开放环境以及实现类似人类的综合感知和深度推理的挑战。随着 MLLMs 能力的不断提高，迫切需要高质量、全面的数据集来更有效地评估其表现。

一个健全的基准框架不仅量化模型性能，还揭示了关键的优点和弱点。例如，[7] 表明虽然现代 MLLMs 展现出强大的整体图像理解能力，但它们对局部区域进行推理的能力仍然有限。此外，[14] 强调了细粒度关系推理中的挑战，表明在细节场景理解上仍有改进空间。MLLMs 的持续进步取决于开发更严格和系统的评估框架。随着这些模型在复杂性上的增长，它们需要越来越细致和高质量的基准来准确衡量其能力。这种共生互动，即 MLLMs 的进步推动更好评估的需求，反过来改进评估促进模型发展，反映了一个迭代的完善过程，推动该领域向前发展。

虽然现有的基准在一般领域（例如，MMMU [26]，ScienceQA [17]）推动了多模态理解，但它们缺乏评估临床决策所需的外科手术特异性。同样，医学基准（例如，MedQA，放射学数

Preprint. Under review.

据集）也因单一模态分析（仅限于文本或图像）或被动问答形式而受到限制，未能捕捉外科动作规划所需的主动、循序渐进的推理。这在评估多模态模型如何整合多样化数据源（包括时间性视觉线索、程序性文本和外科领域知识）以支持动态临床决策方面留下了关键空白。

手术动作规划提出了独特的挑战，这需要时间、因果和多模态推理能力，而这些领域在现有基准中很少被评估。这个任务需要：(1) 精确理解原子级工具-组织交互。这些细粒度的原子动作直接改变组织状态，并构成手术工作流程的通用构建模块。(2) 长期推理以将这些原子动作协调成连贯的程序。然而，关键障碍依然存在，包括需要跟踪仪器-组织关系、监控手术进展以及跨机构差异进行泛化。所有这些困难因目前高质量手术动作规划数据集的稀缺性而加剧。

为原子动作规划开发专门的数据集对于推进该领域至关重要。这类数据集必须挑战模型：(1) 更高阶的推理能力来推断动作之间的因果关系。(2) 细粒度的视觉区分能力以识别微妙的组织状态变化。(3) 具备上下文感知的语言理解能力，以便将动作与程序目标对齐。通过对这些能力进行严格评估，我们旨在开发手术动作规划数据集，这将不仅解决当前 MLLM 评估中的局限性，还能加速临床相关多模态智能的发展，以用于真实世界的手术应用，包括增强术中决策支持和自动化程序工作流程辅助。

我们工作的贡献可以总结如下：

- 我们引入了外科手术动作计划任务，该任务从视觉输入中生成未来的外科手术动作计划。此任务要求对工具与组织的原子级互动有精确的理解，并进行长时间的推理，以弥补多模态长距离学习（MLLM）评估在临床应用中的关键空白，如术中指导和程序自动化。
- 我们引入的 SAP-Bench 数据集跨越了 $1137.50 \pm 575.61s$ 的上下文时长，并包括来自 74 名患者的 $1,226$ 个动作片段，涵盖了五种手术动作，每个片段持续 $68.66 \pm 115.99s$ 。它具有提取的 $1,152$ “当前帧”作为 MLLM 分析的关键时间锚，其中 679 来自 CholecT50，473 来自 HeiChole。
- 我们提出了一种基于 MLLM 的手术动作规划（MLLM-SAP）框架，该框架生成“下一步动作”建议。利用“当前帧”，即直接反映组织状态的视觉表示，作为输入，MLLM 模型在注入手术领域知识后处理自然语言指令。通过整体评估场景理解、过程评估和安全考虑，MLLM 为术中指导输出“下一步动作”建议。
- 我们评估了 7 种最新的（SOTA）MLLM 模型在我们提出的 SAP-Bench 数据集上的有效性，包括 OpenAI-o1, QwenVL2.5-Instruct-72B, Claude-3.5-Sonnet, GeminiPro2.5, Step-1o, 和 GLM-4v。
- 我们邀请了两位经认可的外科医生来建立对下一步首选操作的专家基准，同时通过消融研究系统地评估提示设计组件。

2 相关工作

通用多模态语言模型（MLLM）基准测试最近对 MLLM 的评估工作强调在不同领域中的广泛知识和推理能力，基准测试主要关注回顾性任务（例如，识别、感知），而不是主动决策。对于概念识别，诸如 MM-Vet [25] 、LLaVA-Bench [9] 和 Open-VQA [28] 等工作识别关键的视觉概念，如物体、实例和场景。在动作识别中，诸如 MMBench [10] 、SEED-Bench [6] 、EQBEN [21] 和 Visual CoT [18] 等基准测试识别主体执行的动作。对于基于属性的任务，CV-Bench [19] 和 Q-Bench [22] 评估模型识别视觉属性的能力（例如，风格、情感、质量），而物体定位则由 MDVP-Bench [8] 和 VL-Checklist [29] 负责。空间关系理解（例如，" 左侧"，“之前”）的基准测试在 VSR [9] 和 GQA [4] 中进行，而物体交互，对于现实世界推理至关重要，则在 SEED-Bench [6] 、VL-Checklist [29] 、ARO [27] 和 CODIS [12] 中进行评估。虽然这些基准测试已经推进了通用视觉和常识性推理，但对于程序性规划或领域特定推理（特别是我们工作所针对的手术中与工具-组织交互和长时间因果动态的主动规划）的评估仍显不足。将 MLLM 应用于顺序任务规划展示了显著的潜力。这些模型结合了视觉-语言的理解与连续决策制定，使其能够处理环境上下文、建立任务目标并生成可执行的计划。在日常活动规划场景中的有效性已经在最近的研究中被充分记录 [2, 3, 5] 。

医学专用的多模态大语言模型 (Medical-specific MLLMs) 在临床领域中，诸如 Med-Flamingo [13] 和 Rad-DINO [16] 这样的模型在诊断任务（例如，放射报告生成或医学图像问答）中表现出色。然而，这些系统专为被动观察（例如，解释扫描结果或笔记）而设计，而非主动决策，如手术规划或术中推理。尽管像 MedReason [23] 这样的数据集引入了结构化医学推理，但它们仍局限于诊断情境，缺乏手术中所需的动态、工具意识的推理。在外科人工智能的早期工作中，主要集中在回顾性分析，例如手术阶段识别 [1] 、器械分

割 [24] 和交互检测 [15] 。这些努力改善了对手术场景的理解，但未能对更高级的认知过程进行建模，例如手术程序规划或术中决策。最近的项目如 SurgRaw [11] 试图通过捕捉原始的手术工作流程来弥补这一差距，但它们缺乏对推理或长周期任务分解的明确标注。我们的工作通过引入一个手术动作规划的基准，统一了这些视角，结合了一般多模态大语言模型评估的多模态广度、医学多模态大语言模型的领域特异性，以及现有外科手术人工智能数据集中缺乏的程序严谨性。

3 方法

3.1 手术行动计划 (SAP) 任务定义

SAP 任务将复杂的工序分解为通用且基础的动作，并从视觉输入中生成长远的连续动作计划，达成用户定义的目标。具体来说，手术动作规划器通过整合两个主要输入来制定动作计划 $\mathcal{A} = \{a_1, \ldots, a_T\}$：（1）视觉历史 $\mathcal{H} = \{v_1, \ldots, v_t\}$ ，将过去的观察编码为视频剪辑或帧的序列，以及（2）目标规范 G ，以自然语言指令形式提供（例如，“腹腔镜胆囊切除术中的下一步动作是什么”）。规划器生成一系列离散动作 $\{a_t\}_{t=1}^{T}$ ，每个 $a_t \in \{1, \ldots, C\}$ 属于预定义的 C 手术动作集合，确保在 T 步骤范围内从当前状态进展到目标。

3.2 SAP-Bench 数据集基准

来源：我们的 SAP-Bench 数据集取自三个外科视频数据集：CholecT50 [15] 和 HeiChole [20] ，分别包含 50 和 24 个胆囊切除术。标准的胆囊切除术包括七个阶段：(1) 准备阶段，(2) 胆囊三角切除，(3) 夹闭和切割，(4) 胆囊切除，(5) 胆囊打包，(6) 清理和凝固，(7) 胆囊牵引。在我们的研究中，我们专注于处理来自第 2 到第 4 阶段的视频片段作为外科动作规划的背景，包括胆囊三角切除、夹闭和切割、胆囊切除，因为这些阶段代表了手术中最关键和技术要求最高的部分，其中精确的解剖操作和决策是至关重要的。

手动动作标注和动作片段提取动作标签包括吸引、凝固、解剖、组织牵引和血管夹闭。我们制定了这些 5 外科动作的标注标准。吸引：通过吸引从手术部位去除液体或组织碎片。凝固：通过热能来使蛋白质变性，形成稳定的血块并封闭组织，以促进止血。解剖：在手术过程中分离或切割组织层以暴露底层结构并进入更深的解剖层。夹闭：应用外科夹子以控制出血或暂时闭塞血管或组织。组织牵引：将组织或器官移开并保持以提供更好的手术部位能见度和可及性。共享相同动作标签的连续帧被分组为动作片段。为确保数据质量，我们强制执行严格的选择标准：(a) 视觉中心性：目标动作必须占据视觉中心以保持上下文；(b) 动作纯度：每个片段必须包含唯一可辨别的动作。所有动作片段均经过严格的人工验证过程以确认标签准确性。

数据集特征如表 1 所示，我们的 SAP-Bench 数据集聚合了来自两个手术视频数据集（即 CholecT50 和 HeiChole）的数据，总共包含 74 名患者和 1,226 个动作片段及其对应的动作标签。CholecT50 数据集贡献了 50 名患者和 729 个片段，而 HeiChole 提供了 24 名患者和 497 个片段。这些 1,226 个动作片段涵盖了五种手术动作，如表 2 所示。解剖类别在集合中占主导地位，有 597 个实例（占总数的 48.7% ），平均每个片段 112.32 ± 150.34 秒，以 1 FPS 提供 67,056 帧。组织牵引在解剖之后出现频率最高（321 个片段，占 26.2% ），尽管平均持续时间显著较短（$20.65 \pm 19.84s$ ）。血管夹闭有 140 个片段，平均 31.66 ± 30.21 秒，凝固有 110 个片段，抽吸有 58 个片段。
从第二阶段到第四阶段的上下文视频段，CholecT50 的持续时间为 $1,339.06 \pm 560.63s$ ，HeiChole 的持续时间为 $717.58 \pm 325.32s$ 。动作片段的持续时间测量为 $68.66 \pm 115.99s(overall)$ ，CholecT50 的平均 $91.84 \pm 142.77s$ ，HeiChole 的平均 $34.65 \pm 36.61s$ 。
当前观察被定义为“当前帧” $F_{t+1}^{s} - 1$ ，它紧接在后续动作片段的起始帧 F_{t+1}^{s} 之前，片段跨度为 [F_{t+1}^{s} , F_{t+1}^{e}]。后续动作片段的动作标签 a_{t+1} 作为“下一个动作”的真实情况。我们提取了 1,152 个“当前帧”作为 MLLM 分析的关键时间锚点，包括来自 CholecT50 的 679 和来自 HeiChole 的 473 。每个“当前帧”都分配了一个“下一个动作”标签。

3.3 基于 MLLM 的外科手术动作规划（MLLM-SAP）框架

鉴于 MLLM 在处理图像方面的当前能力，我们仅使用当前帧来表示当前观察 $\mathcal{O}$ 作为视觉输入，而不是视频序列。MLLM 处理当前观察 $\mathcal{O}$ 和提示以生成预测的下一步动作 $\mathcal{A}$ 。这样设计的原因有：（1）MLLM 在单幅图像理解方面表现出色，（2）计算效率的考虑，以及（3）我们手术应用中对实时处理的需求。

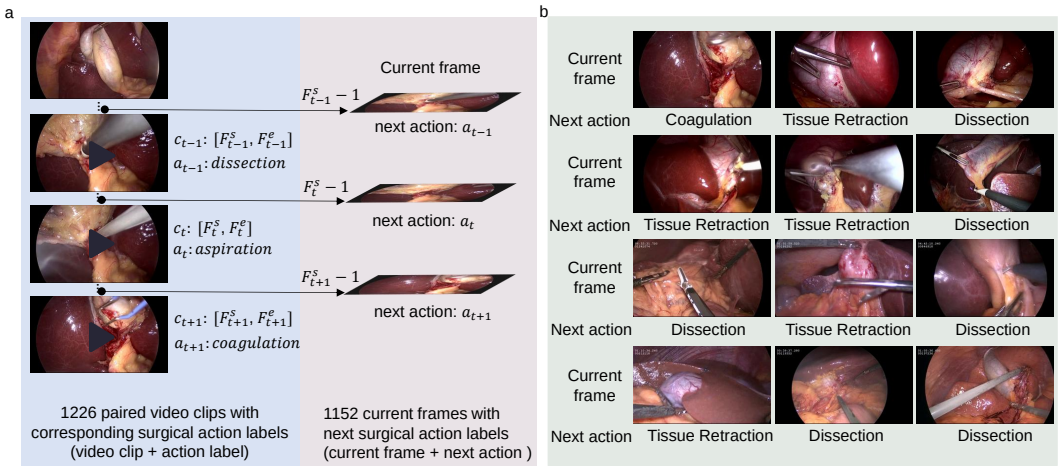

Figure 1: SAP-Bench 数据集。(a) 我们首先标注动作片段，记录每个片段的动作标签 a_{t+1} 和帧范围 [F^s_{t+1} , F^e_{t+1}]。然后，我们提取每个片段的起始帧之前的那个帧 $F^s_{t+1}-1$ 作为“当前帧”，并将后续的动作标签 a_{t+1} 标记为“下一个动作”。(b) 当前帧及其对应的下一个动作的示例。

Table 1: 数据集摘要：患者数量，总剪辑数，每个视频的上下文持续时间（平均值 ± 标准差）[秒]，动作剪辑持续时间（平均值 ± 标准差）[秒]，1 FPS 帧数和当前帧数。

Source Dataset	# Patients	# Clips	Context Duration [s]	Clip Duration [s]	1 FPS Frame Count	# Current Frame
CholecT50	50	729	1339.06 ± 560.63	91.84 ± 142.77	66 953	679
HeiChole	24	497	717.58 ± 325.32	34.65 ± 36.61	17 222	473
Overall	74	1226	1137.50 ± 575.61	68.66 ± 115.99	84 175	1152

Table 2: 动作片段计数、持续时间统计（平均值 ś 标准差），按数据集及总体的 1 FPS 帧数。

Action	CholecT50			HeiChole			Overall		
	Count	Duration [s]	1 FPS Frame Count	Count	Duration [s]	1 FPS Frame Count	Count	Duration [s]	1 FPS Frame Count
Aspiration	39	36.00 ś 47.60	1404	19	37.05 ś 20.05	704	58	36.34 ś 40.69	2108
Coagulation	89	40.62 ś 65.79	3615	21	15.95 ś 15.11	335	110	35.91 ś 60.33	3950
Dissection	298	179.75 ś 185.51	53565	299	45.12 ś 42.24	13491	597	112.32 ś 150.34	67056
Tissue Retraction	232	22.50 ś 21.35	5221	89	15.81 ś 14.13	1407	321	20.65 ś 19.84	6628
Vessel Clipping	71	44.34 ś 35.91	3148	69	18.62 ś 13.76	1285	140	31.66 ś 30.21	4433

$$\mathcal{A} = \text{MLLM}(\mathcal{O}, Prompts) \tag{1}$$

设计提示以生成行动计划医学推理任务需要严格的逻辑过程。具体来说，我们需要一个能够（1）将复杂问题分解为连贯推理步骤的模型，以及（2）遵循既定医学指南，以确保提供由循证理由支持的专业、准确的响应。与遵循明确、逐步解决方案的数学问题不同，医学问答涉及处理专业术语和从复杂的临床信息中提取关键线索以做出正确诊断，这是一个明显更具挑战性的推理过程。因此，我们的提示工程框架由两个关键组件组成：系统提示和用户提示。系统提示包含结构化的外科知识库，为 MLLM 提供领域特定的信息，包括外科流程、外科指南和行动描述（见图 2）。外科流程定义了手术中执行的标准化步骤和程序的顺序。安全协议建立了降低医疗干预风险的规则和措施。行动描述解释了外科医生执行的各个动作。用户提示捕获实时决策方面，并通过具体的查询指导系统响应，包括场景理解、进展判断、安全考虑和准备执行的动作。具体来说，场景理解分析视觉和背景元素，例如关键结构和空间关系，以确定当前状态。进展判断动态跟踪任务相对于计划工作流程的进展情况，支持实时调整。安全考虑持续评估风险，以优先考虑减少伤害的行动。准备执行的动作提供三个推荐的下一个动作。对于每个动作，提供一个解释其排序原因的理由。

由于对视觉输入的支持及其在图像理解和推理任务中的强能力，选择了

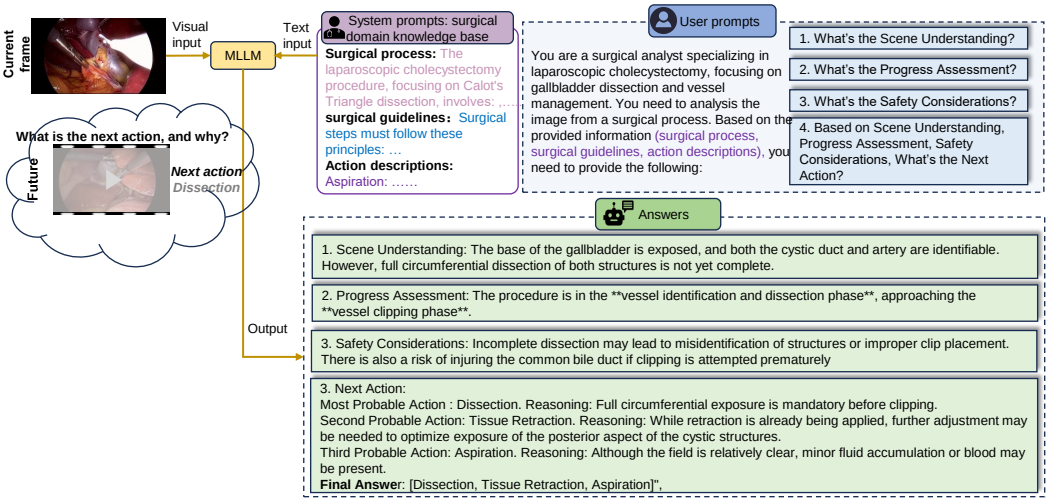

Figure 2: 我们的 MLLM 框架生成手术行动计划。给定“当前帧”作为观察输入，该输入捕获最新的组织状态，MLLM 处理系统提示和用户提供的文本指令。基于场景理解、过程进展和安全考虑的综合分析，MMLLM 模型推荐“下一步行动”。

4 实验

GPT-4o、QwenVL2.5-Instruct-72B、Claude-3.5-Sonnet、GeminiPro2.5、Step-1o 和 GLM-4v 作为基准 MLLM 模型。所有实验均在 8 NVIDIA A800 GPU 上进行。我们采用以下评价指标：样本- S： 在标准样本级准确性（样本- S ）下，预测若要被认为是正确的，模型预测的下一动作 $\hat{a}_i$ 必须在每个时间步与真实动作 a_i 精确匹配。Top-2 和 Top-3 指标提供更宽松的评估标准。在此，如果真实动作 a_i 分别出现在模型最有可能的前两个或前三个预测中，则认为预测是正确的。视频级别 S：为了评估模型在完整手术过程中的泛化性，我们通过聚合样本级别预测来计算视频级别（即病人级别）的准确性。对于每个病人案例，我们首先确定每个时间步的样本级别 S 准确性，然后在整个手术过程中对这些值求平均。这提供了模型在个体手术层面的全面表现，而不是孤立的样本。样本级别 R：为了更好地反映手术工作流程中固有的时间变异性，我们引入一个放宽的准确性指标。如果预测结果 $\hat{a}_i$ 匹配当前的真实动作 a_i 或紧随其后的动作 a_{i+1} ，则视为有效。这种 +1 步容忍度（$\hat{a}_i \in \{a_i, a_{i+1}\}$ ）适应了实际手术过程中所需的自然时间灵活性，因为在某些情况下可以安全地重新排序步骤，而不会产生临床后果。视频级别 R：为了在病人层面提供模型性能的整体评估，同时考虑到真实世界手术工作流程中固有的时间变异性，我们在视频级别（即病人级别）聚合预测结果。对于每个视频，我们使用放宽标准（样本级别 R ）计算样本级别准确性，然后通过在整个手术过程中对这些值取平均值来推导出统一的性能得分。

我们对 7 个最先进的多模态大语言模型（OpenAI-o1、GPT-4o、QwenVL2.5-72B、Claude-3.5-Sonnet、GeminiPro2.5、Step-1o 和 GLM-4v）在 CholecT50 和 HeiChole 上的评估揭示了在标准（S）和放宽（R）条件下，从样本层面和视频层面分析手术推理能力的关键差异（见表 3 和图 3 ）。主要发现包括：

1. 性能层级及模型优势。OpenAI-o1 凭借推理能力在 Cholec50 上取得了最佳的样本 S 和样本 R ，并在 HeiChole 上样本 S 中排名第二，而样本 R 中排名第一，表现出色。GPT-4o 在标准条件下持续在 Top-1 样本 S 准确率方面领先（例如，在 Cholec50 上为 34.61 %，在 HeiChole 上为 59.20 %），展现了良好的泛化能力。GLM-4v 显示最低的 Top-1 样本– S ，但最高的 Top-3 样本– S ，表明其检索增强推理能力强但精度较弱。

2. 数据集特定的挑战。CholecT50 比 HeiChole 更具挑战性，所有模型的 Top-1 得分明显较低（例如，Claude-3.5-Sonnet 在 CholecT50 上是 28.87 %，而在 HeiChole 上是 46.30 %）。这种差距在 Top-3 指标中变小，暗示了在细粒度手术步骤分类中存在更高的不确定性。

3. 评估严格性的影响。从标准（S）到宽松（R）条件的转换普遍提高了性能，Top-1 Sample- S 的增益超过 20 %（例如，GPT-4o 在 Cholec50 上从 34.61 % 跃升到 67.41 %）。这凸显了 MLLMs 对严格手术基准的敏感性。视频基础的评估（Video- S 和 Video- R ）显示了大多数模型在样本级指标上的更好表现（例如，Step-1o 在 CholecT50 上 Top-1 Video- R 是 71.47 %，而 Top-1 Sample- R 是 67.25 %）。

4. 模型的局限性。GLM-4v 和 Claude-3.5-Sonnet 在手术动作计划任务中表现不佳，分别在 Cholec50 的 Top-1 Sample- S 和 Top-3 Sample- S 中得分最低。

图 3 显示了模型对用户提示的预测文本响应，涵盖了场景理解、进度评估、安全考虑和准备执行的动作。正确的响应以绿色突出显示，而错误则以粉色标记。分析重点放在涉及 Calot 三角解剖的外科手术程序上，清楚地识别出关键的解剖结构，如胆囊管和胆囊动脉。它评估了进一步解剖的必要性，同时强调了程序的安全性和最佳可见性。

Table 3: 在标准（完全匹配）和宽松（+1 步容差）条件下，以样本级（每帧）和视频级（每患者）细度评估 CholecT50 和 HeiChole 数据集的比较性能。粗体值表示最佳性能，而 underlined 值表示每项指标的最差结果。

Dataset	Model	Sample-S			Video-S			Sample-R			Video-R		
		Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
CholecT50	OpenAI-o1	36.82	44.04	46.69	39.42	45.57	48.93	69.16	78.54	79.33	74.01	81.61	82.15
	GPT-4o	34.61	64.21	77.76	35.86	62.48	75.85	67.41	92.21	97.62	69.82	92.75	97.53
	QwenVL2.5-Instruct-72B	29.01	55.82	73.93	31.01	55.83	71.61	58.51	85.85	96.50	61.86	86.08	94.70
	Claude-3.5-Sonnet	28.87	48.31	57.44	30.87	49.62	57.82	57.23	78.38	83.78	59.23	80.07	84.58
	GeminiPro2.5	28.87	58.03	77.47	30.42	56.49	75.13	51.99	84.58	95.55	51.12	82.15	93.38
	Step-1o	34.46	53.76	69.37	35.93	55.59	69.46	67.25	85.21	95.07	71.47	87.89	95.88
	GLM-4v	18.02	51.70	81.09	21.47	55.96	81.74	36.84	86.60	96.81	41.75	89.41	97.74
HeiChole	OpenAI-o1	57.29	63.85	64.06	57.35	64.02	64.28	76.61	84.41	84.41	76.94	84.64	84.64
	GPT-4o	59.20	74.42	83.09	59.00	74.73	83.46	76.61	90.87	94.65	76.46	90.59	94.29
	QwenVL2.5-Instruct-72B	48.63	74.84	86.05	48.39	74.60	86.23	63.92	89.76	95.32	63.60	89.66	95.23
	Claude-3.5-Sonnet	46.30	64.69	66.38	45.42	64.81	66.58	65.26	82.85	83.30	64.27	82.41	82.76
	GeminiPro2.5	46.30	68.29	79.07	44.90	67.63	78.35	64.81	82.85	90.20	63.41	81.47	89.03
	Step-1o	53.91	68.50	76.74	53.63	68.59	76.68	71.94	86.41	90.65	71.73	86.78	90.49
	GLM-4v	29.66	76.69	86.02	29.58	77.22	86.12	47.77	93.97	97.10	48.17	94.09	96.98

4.1 人工评估：外科决策中的临床变异性

由于以下原因，手术流程表现出固有的变异性：(i) 技能依赖因素：外科医生的经验水平（例如，住院医生与主治医师）(ii) 风格差异：个人的手术偏好（例如，解剖技术）(iii) 情境适应性：基于组织状况的实时调整。这种变异性意味着在任何手术视频中观察到的下一个动作只是一个可能的有效方法，并不一定是唯一的最佳解决方案。

为了建立专家验证的性能基准，我们邀请了两位持有董事会认证的外科医生（A 和 B）提出他们偏好的“下一步行动”。从 CholecT50 数据集中，我们选取了 64 个样本，涉及 50 个手术视频，要求每位外科医生为每个样本指出他们偏好的下一步行动。对于这些 64 个样本，我们将评估 MLLM 模型的预测与四种不同的“下一步行动”真实值的比较：(1) 来自视频的 GT：手术视频中实际执行的“下一步行动”（来自原始外科医生示范的真实值）。(2) 外科医生 A：专家外科医生 A 所建议的偏好“下一步行动”。(3) 外科医生 B：专家外科医生 B 所建议的偏好“下一步行动”。(4) 外科医生 (A 或 B)：如果模型的预测与外科医生 A 或 B 的偏好“下一步行动”相符，即计为匹配，代表更广泛的专家共识。

如图 4 所示，我们使用 Top-k 样本 S 准确率 ($k \in 1, 2, 3$) 评估 GPT-4o 的表现，测量与以下方面的一致性：(1) 个别外科医生的偏好（执行手术的外科医生与外科医生 A 与外科医生 B），以及 (2) 两位专家判断相结合的决策。这种双重评估方法既捕捉了地方外科实践的变异，也反映了一般的专家共识。GPT-4o 在与外科医生 B 的共识（38.30 % Top-1）和综合评估（58.51 % Top-1）中表现最佳。特别值得注意的是，当结合外科医生偏好（HK 或 SG）进行评估时，相较于个别评估时的平均表现提升。

4.2 消融研究

我们进行了一项系统的消融研究，以分析不同文本组件在我们提示设计中的影响，重点关注两个关键元素：系统提示和用户提示（见表 4 ）。系统提示包括手术过程、手术指导和操作描述。用户提示包括场景理解、进度判断、安全考虑以及可执行的动作。我们评估了以下变化：（1）无 SP：完全去除所有系统提示（SP），同时保留用户提示。（2）仅 OA：用户提示仅限于可执行的动作（OA）（排除场景理解、进度判断和安全考虑），并保留系统提示。我们使用 Δ 来表示消融模型变体与其基本版本之间的性能差异（以百分点表示）。

系统提示的消融研究我们通过比较模型在有无注入外科知识情况下的表现，定量评估系统提示的影响。移除系统提示（w/o SP）在大多数模型和数据集组合中一致导致性能下降，GPT-4o 表现出最大性能下降（HeiChole Video-S 上的 $\Delta = -8.85$ ）。这突显了系统提示在维

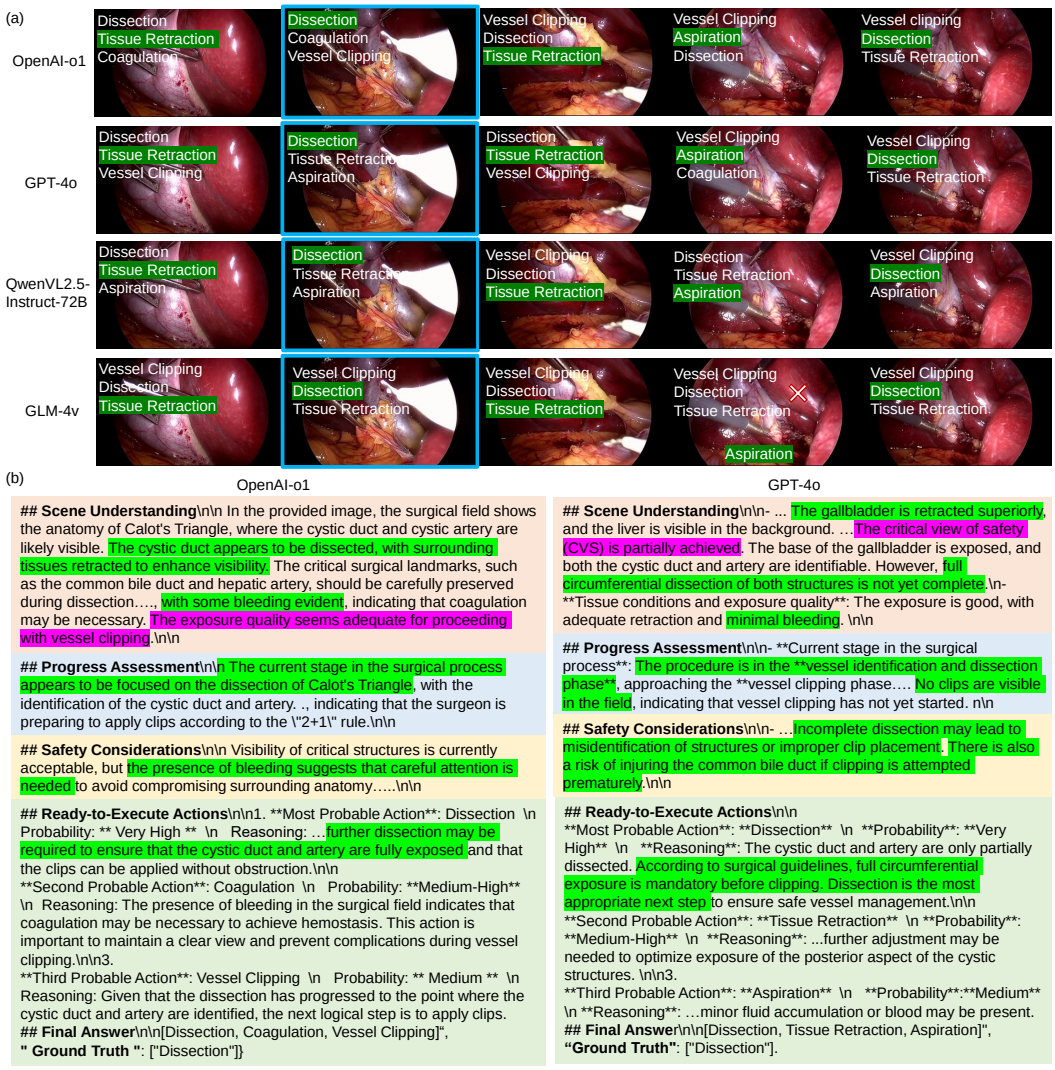

Figure 3: 结果可视化。(a) 动作规划预测：白色文本显示前三个预测的未来动作（按顺序排列），其中 × 标记错误预测。真实动作用绿色高亮显示。(b) 在 (a) 中的蓝色边框框出的示例中，MLLMs 生成预测下一个动作的文本预测。绿色高亮的文本表示正确的预测，而粉色高亮表示错误。

持上下文理解中的关键作用，特别是在需要程序知识和安全约束的复杂外科场景中。两个数据集上持续的性能下降证实了特定领域知识注入的必要性，并展示了外科 MLLMs 提示设计的普遍适用性。

关于用户提示的消融研究准备执行的用户提示（带 OA）会产生不同的效果。虽然 GPT-4o 在带 OA 模式下表现显著退化（CholecT50 Sample-S 上为 $\Delta = -8.69$，HeiChole Sample-S 上为 $\Delta = -15.86$）。GeminiPro2.5 从这种简化中受益（HeiChole Sample-S 上为 $\Delta = +6.97$），这表明模型在处理指导性上下文时存在根本性的差异。QwenVL 遭受最严重的退化（HeiChole Sample-S 上为 $\Delta = -21.36$）。这些发现表明，尽管 OA 过滤减少了幻觉风险，但它对那些推理能力较强但手术动作基础较弱的模型产生了不成比例的影响。

5 结论

有效的评估对于推动 MLLM 的研究至关重要，尤其是在手术动作计划（SAP）中，这一任务需要精确和多模态的推理，从视觉输入中生成未来的动作序列。因此，我们引入了

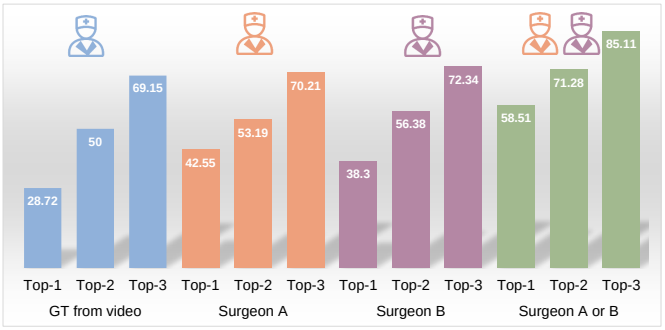

Figure 4: 模型性能评估基于四个真实参考，包括原始手术和外科医生偏好（A、B 或任意）使用 Top-k 样本 S 准确度（$k \in 1, 2, 3$ ）。

Table 4: 跨数据集的消融变体与基准模型之间的性能比较（Top-1 准确率 Δ ）。消融变体包括系统提示移除（w/o SP）和仅有可执行动作用户提示（w. OA）。Δ 表示消融模型变体与其基准版本之间的性能差异（以百分比计算）。

	CholecT50		HeiChole	
Model	Sample-S Δ	Video-S Δ	Sample-S Δ	Video-S Δ
w/o SP				
GPT-4o	-4.12	-1.87	-6.77	-6.33
GeminiPro2.5	+2.80	+1.33	-0.22	-1.15
QwenVL	-5.15	-4.78	-11.63	-11.30
w. OA				
GPT-4o	-8.69	-8.52	-15.86	-14.85
GeminiPro2.5	+5.01	+3.11	+6.97	+6.34
QwenVL	-2.06	-2.45	-21.36	-21.97

SAP-Bench 基准，这是一个从 74 例胆囊切除手术中获得的大规模数据集，包含 1,226 个临床验证的动作片段和 1,152 帧当前画面，其与下一个动作配对用于多模态分析。我们还提出了 MLLM-SAP 框架，该框架利用手术知识增强的 MLLMs 生成可操作的下一步行动建议。在评估七个领先的 MLLM（如 GPT-4o，Claude-3.5-Sonnet，QwenVL2.5-72B）时，我们发现手术动作预测中存在显著差距。为了使我们的评估具有临床现实性，我们与两位经过认证的外科医生合作建立了专家基准，并进行了消融研究以评估提示设计组件在术中决策中的影响。局限性：我们当前的框架有两个主要局限性：(1) 我们尚未使用强化学习来验证模型的有效性；(2) MLLM-SAP 架构没有纳入思维链（CoT）推理机制。未来工作：除了性能指标外，评估 MLLMs 的可信度对于确保其在医疗系统等高风险领域的可靠性至关重要，因为错误可能导致严重后果。此外，在广泛的下游应用中评估 MLLMs 有助于验证其适应性，确保其能够满足实际部署的复杂需求。

References

[1] Ayobi, N., Rodríguez, S., Pérez, A., Hernández, I., Aparicio, N., Dessevres, E., Peña, S., Santander, J., Caicedo, J.I., Fernández, N., Arbeláez, P.: Pixel-wise recognition for holistic surgical scene understanding. arXiv (2024), https://arxiv.org/abs/2401.11174

[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] 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)

[4] Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)

[5] 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)

[6] Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)

[7] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)

[8] Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. arXiv preprint arXiv:2403.20271 (2024)

[9] Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)

[10] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)

[11] Low, C.H., Wang, Z., Zhang, T., Zeng, Z., Zhuo, Z., Mazomenos, E.B., Jin, Y.: Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. arXiv preprint arXiv:2503.10265 (2025)

[12] Luo, F., Chen, C., Wan, Z., Kang, Z., Yan, Q., Li, Y., Wang, X., Wang, S., Wang, Z., Mi, X., et al.: Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. arXiv preprint arXiv:2402.13607 (2024)

[13] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)

[14] Nie, J., Zhang, G., An, W., Tan, Y.P., Kot, A.C., Lu, S.: Mmrel: A relation understanding dataset and benchmark in the mllm era. arXiv preprint arXiv:2406.09121 (2024)

[15] 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)

[16] Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Rad-dino: Exploring scalable medical image encoders beyond text supervision. arXiv preprint arXiv:2401.10815 (2024)

[17] Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: A novel resource for question answering on scholarly articles. International Journal on Digital Libraries **23**(3), 289–301 (2022)

[18] Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. arXiv e-prints pp. arXiv–2403 (2024)

[19] Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)

[20] Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical Image Analysis **86**, 102770 (2023)

[21] Wang, T., Lin, K., Li, L., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Equivariant similarity for vision-language foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11998–12008 (2023)

[22] Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. arXiv preprint arXiv:2309.14181 (2023)

[23] Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.I., et al.: Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. arXiv preprint arXiv:2504.00993 (2025)

[24] Xu, M., Islam, M., Bai, L., Ren, H.: Privacy-preserving synthetic continual semantic segmentation for robotic surgery. IEEE transactions on medical imaging (2024)

[25] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)

[26] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)

[27] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)

[28] Zeng, Y., Zhang, H., Zheng, J., Xia, J., Wei, G., Wei, Y., Zhang, Y., Kong, T., Song, R.: What matters in training a gpt4-style language model with multimodal inputs? In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7930–7957 (2024)

[29] Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221 (2022)