

SurgMind-R1: Overcoming Reasoning Collapse in Surgical Vision-Language Models via Reinforcement Learning

Yifan Luo¹, Guoming Wang^{1*}, Ling Xiao², Juncheng Li¹, Siliang Tang¹

¹Zhejiang University ²Hokkaido University

{yifanluo, NB21013}@zju.edu.cn

Abstract

Vision-Language Models (VLMs) show immense promise in medicine, yet surgical scene understanding remains formidable due to complex visual obstructions and rigorous multi-step clinical logic. Standard Supervised Fine-Tuning (SFT) imparts domain knowledge, but we identify a critical flaw—the “**Reasoning Collapse**” phenomenon: during answer-only SFT, models progressively lose explicit step-by-step reasoning capabilities, regressing into “black boxes” that generate uninterpretable shortcut answers—a severe risk for clinical traceability. Constructing high-quality Chain-of-Thought (CoT) datasets to mitigate this is prohibitively expensive in the surgical domain. We propose **SurgMind-R1**, a surgical VLM with a three-stage training pipeline. We construct Cholec-Think, a reasoning-enhanced dataset combining expert and VLM-assisted CoT annotations; apply domain-adaptive SFT and reasoning-enhanced SFT on Cholec-Think to internalize surgical logic; and introduce Reinforcement Learning via Group Relative Policy Optimization (GRPO) with rule-based rewards to enforce structured thinking paths and eliminate reasoning collapse. On SurgLite-Bench, our comprehensive benchmark, SurgMind-R1 significantly outperforms state-of-the-art VLMs on complex inference tasks. It internalizes structured thinking, maintaining high accuracy in a direct *no-think* mode that resolves the high-latency bottleneck of reasoning models for real-time clinical applications, and exhibits strong zero-shot out-of-distribution (OOD) generalization across unseen surgical scenarios and novel tasks, paving the way for transparent, trustworthy computer-assisted interventions. The code, benchmark, and models will be available at <https://github.com/YifanLuo0103/SurgMind-R1>.

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable success in general-domain multimodal understanding through large-scale pretraining and Supervised Fine-Tuning (SFT) (OpenAI et al., 2024; Team et al., 2024; Bai et al., 2023). Extending these models to medicine is highly promising, yet clinical decision-making requires not only perception but also reliable multi-step reasoning. The central challenge is therefore to adapt VLMs from shallow visual recognition to trustworthy medical inference.

Among medical applications, surgical scene understanding is especially challenging because procedures unfold in dynamic environments with ambiguous anatomies, unstable illumination, smoke, blood, and severe tool–tissue occlusion. As a result, tasks such as surgical phase recognition, instrument–tissue interaction modeling, and critical view of safety assessment cannot be solved by simple visual matching alone, but require robust spatiotemporal reasoning over the surgical workflow.

SFT is the dominant paradigm for adapting foundation models to medical imaging (Moor et al., 2023; Zhang et al., 2023; Li et al., 2025), and recent surgical VLMs largely follow this strategy. However, answer-only SFT encourages shortcut learning, causing the model to align inputs directly to labels instead of learning transferable multi-step logic. Although Chain-of-Thought (CoT) prompting can improve reasoning (Wei et al., 2022), constructing high-quality surgical CoT data is prohibitively expensive because it requires clinically valid, expert-level annotation.

More critically, we observe that standard surgical SFT can induce “**Reasoning Collapse**”: powerful general-domain VLMs progressively lose their explicit step-by-step reasoning ability and regress to shortcut-style answer generation. The resulting

* Corresponding author.

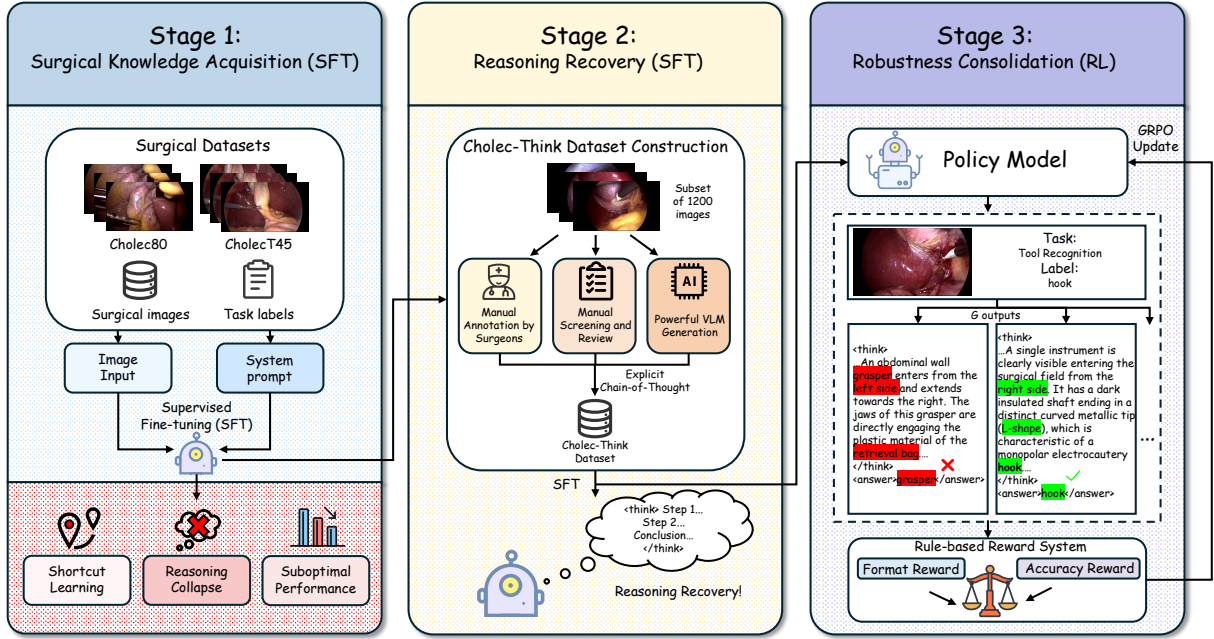


Figure 1: Overview of SurgMind-R1. The proposed three-stage pipeline, consisting of two SFT stages and one RL stage, progressively acquires surgical knowledge, restores explicit reasoning, and consolidates reasoning robustness.

outputs are often empty, truncated, or structurally unstable, undermining interpretability and clinical traceability. In surgery, where transparency and decision justification are essential, such collapse makes the model difficult to trust.

To address these challenges, we propose **SurgMind-R1**, a surgical VLM trained with a three-stage pipeline, shown in Fig. 1, to acquire domain knowledge while preserving explicit reasoning. We first build **Cholec-Think** from Cholec80 (Twinanda et al., 2017) and CholecT45 (Nwoye et al., 2022) by combining expert annotations with rigorously filtered VLM-assisted CoT generation. Stage 1 performs domain-adaptive SFT for surgical knowledge acquisition, Stage 2 uses reasoning-enhanced SFT to recover explicit surgical reasoning, and Stage 3 applies GRPO (Shao et al., 2024) with rule-based rewards to stabilize reasoning quality and improve answer accuracy.

To evaluate SurgMind-R1, we construct **SurgLite-Bench**, a multimodal surgical benchmark spanning 4 surgical scenarios and 5 tasks from tool recognition to critical view of safety assessment. In the hardest task, strict surgical triplet recognition, SurgMind-R1 outperforms the strongest zero-shot general-domain baseline by 22.82% and the strongest fine-tuned model by 8.53%. In other tasks, SurgMind-R1 outperforms the strongest zero-shot baseline by 12.49%–51.02% and the strongest fine-tuned

model by 9.79%–15.26%. It also demonstrates robust zero-shot OOD generalization, and, through ablation studies, shows that the three-stage pipeline effectively prevents reasoning collapse.

In summary, our main contributions are as follows:

1. We propose SurgMind-R1, a novel vision-language model for the surgical domain. Evaluated on our challenging surgical benchmark, SurgMind-R1 achieves state-of-the-art performance across multiple complex inference tasks. Notably, it internalizes surgical logic to maintain high accuracy in a low-latency *no-think* mode and demonstrates strong zero-shot out-of-distribution generalization.
2. We establish SurgLite-Bench, a multimodal surgical benchmark including 4 surgical scenarios and 5 tasks, ranging from fundamental visual perception and temporal analysis to high-level reasoning.
3. We identify a critical limitation in traditional SFT for VLMs: the “Reasoning Collapse” phenomenon, where models progressively lose their explicit step-by-step reasoning capabilities and regress to generating uninterpretable shortcut answers. We define indicators to quantify reasoning collapse and demonstrate through experiments that our training pipeline effectively addresses this issue.

2 Related Work

2.1 Vision-Language Models in Surgery

General Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023b), have achieved advanced natural image-text understanding through large-scale pretraining. However, they often struggle with domain-specific tasks like medical reasoning due to the significant domain gap. The rise of large-scale VLMs has driven numerous domain-specific adaptations in the medical field. Systems like LLaVA-Med (Li et al., 2023a) and others (Chen et al., 2024c) have achieved significant results in radiology visual question answering (VQA) and related diagnostic tasks. In the surgical domain, recent studies have systematically evaluated the potential of large VLMs for surgical artificial intelligence (Rau et al., 2025). Furthermore, comprehensive efforts have been made to construct large-scale surgical multimodal databases, such as SurgVLM-DB, alongside systematic evaluation benchmarks like SurgVLM-Bench (Zeng et al., 2025), to facilitate the development of surgical intelligence.

2.2 Surgical Scene Understanding

Surgical scene understanding is crucial for computer-assisted interventions. Recent advancements have demonstrated excellent performance in endoscopic scenes using grounded multimodal large language models (Wang et al., 2026) and comprehensive evaluation frameworks (Liu et al., 2025b). However, these models often lack an explicit reasoning process, acting as “black boxes” that directly output answers. While effective for basic perception, using Supervised Fine-Tuning (SFT) on final answer labels remains the dominant strategy for tailoring large models to the medical domain (Zambrano Chaves et al., 2025; Chen et al., 2024a). Recent research highlights a critical limitation of this approach: while SFT primarily encourages “memorization” of the training data, reinforcement learning (RL) can significantly help models “generalize” to unseen scenarios (Chu et al., 2025). This motivates the need to move beyond simple SFT for complex surgical reasoning.

2.3 Reasoning and RL in Foundation Models

Enhancing the reasoning capabilities of foundation models has become a focal point of recent research. Techniques such as Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2024) and

Direct Preference Optimization (DPO) (Rafailov et al., 2023) have been widely adopted to align models with human intent. In the medical domain, complex reasoning has been shown to improve medical problem-solving capabilities, which can significantly benefit from reinforcement learning (Chen et al., 2024b). Consequently, recent works have introduced RL-driven medical VLMs, such as Med-R1 (Lai et al., 2026) and MedVLM-R1 (Pan et al., 2025), using GRPO with rule-based accuracy rewards to incentivize diagnostic reasoning in general medical and radiology VQA. In the surgical domain, Surgery-R1 (Hao et al., 2025) applies GRPO to Surgical Visual Question Localized Answering (Surgical-VQLA), coupling format constraints with spatial bounding-box grounding and answer rewards. Distinct from these approaches, our work targets fine-grained surgical scene comprehension—encompassing instrument identification, workflow phase recognition, and structured action triplet deduction—via a multi-component reward. Crucially, beyond reward-driven policy optimization, we identify and formulate the critical “Reasoning Collapse” phenomenon in surgical VLMs under standard SFT, resolving it via our three-stage training pipeline to ensure verified clinical reasoning paths while supporting low-latency internalized logic.

3 Methods

3.1 Overview

SurgMind-R1 follows a three-stage training pipeline that first adapts a general VLM to surgical visual understanding, then restores explicit reasoning, and finally consolidates reasoning robustness. In Stage 1, the model is fine-tuned on standard surgical question-answer data from Cholec80 (Twinanda et al., 2017) and CholecT45 (Nwoye et al., 2022) to acquire domain knowledge. Since this answer-only SFT induces *Reasoning Collapse*, Stage 2 further trains the model on Cholec-Think, a compact Chain-of-Thought (CoT) (Wei et al., 2022) dataset designed to recover step-by-step surgical reasoning. Stage 3 applies Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with rule-based format and accuracy rewards, encouraging the model to preserve structured reasoning while improving clinically grounded predictions.

3.2 Dataset

3.2.1 Train Set

Our experimental framework is built upon two widely recognized endoscopic video datasets for laparoscopic cholecystectomy: Cholec80 (Twinanda et al., 2017) and CholecT45 (Nwoye et al., 2022). The Cholec80 dataset consists of 80 surgical videos performed by 13 surgeons. We utilize frames extracted at 1 fps, which are annotated for two primary tasks: surgical instrument presence detection (7 instrument classes) and surgical phase recognition (7 phase classes). The CholecT45 dataset, a subset of Cholec80 comprising 45 videos, is dedicated to the more complex task of surgical action triplet recognition. It provides fine-grained annotations for $\langle \text{instrument}, \text{verb}, \text{target} \rangle$ relations at 1 fps, encompassing 6 instruments, 10 verbs, and 15 anatomical targets, which combine into 100 unique action triplets. This task demands a comprehensive spatiotemporal observation of the surgical scene to locate active instruments and correctly associate their actions with specific anatomical targets.

We apply video-level de-duplication and splitting to avoid leakage between overlapping Cholec80 and CholecT45 videos. More details on the dataset split are provided in Appendix A.1.

3.2.2 Cholec-Think Construction

To endow the model with explicit reasoning capabilities, we constructed a novel reasoning-enhanced dataset, Cholec-Think. Building upon the raw frames and annotations from the training splits of Cholec80 and CholecT45, Cholec-Think bridges the semantic gap between the task query and the final answer by introducing structured Chain-of-Thought (CoT) (Wei et al., 2022) reasoning steps. It combines expert-curated reasoning paths with VLM-assisted generation and manual clinical screening, yielding 1,200 balanced reasoning samples for tool & phase and triplet tasks. More details on the construction process are provided in Appendix A.2.

3.3 SurgLite-Bench

To comprehensively evaluate the capabilities of SurgMind-R1 and provide an efficient, lightweight evaluation tool for surgical vision-language models, we propose SurgLite-Bench, a compact, multimodal surgical benchmark characterized by its cross-domain and cross-task nature. Existing medical VLM benchmarks predominantly focus on

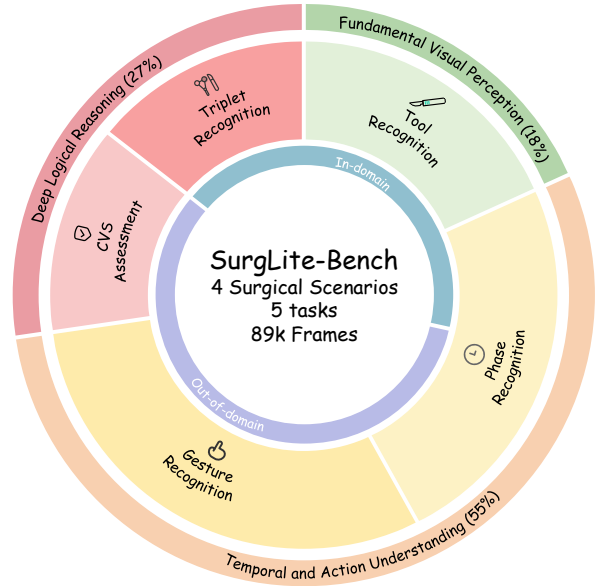


Figure 2: Statistics of SurgLite-Bench. The benchmark covers 4 surgical scenarios and 5 tasks spanning hierarchical levels from fundamental visual perception to deep logical reasoning. Its distribution across datasets and tasks provides a compact yet diverse evaluation testbed for both in-domain performance and OOD generalization.

static radiological images (e.g., X-rays, MRI) or simple visual question answering, lacking the dynamic complexity and rigorous logical deduction required in surgical environments. SurgLite-Bench addresses this gap by integrating 6 publicly available datasets, encompassing 4 distinct surgical scenarios and 5 diverse tasks. With a moderate scale of 89,659 annotated frames, it is specifically designed to facilitate lightweight deployment and efficient evaluation of surgical VLM capabilities. The benchmark statistics are summarized in Fig. 2.

To systematically assess the model’s cognitive capabilities, the tasks in SurgLite-Bench are hierarchically structured into three levels of increasing complexity. 1. Fundamental Visual Perception evaluates the model’s ability to identify basic surgical elements, specifically through Tool Recognition on the Cholec80 dataset. 2. Temporal and Action Understanding assesses the capability to analyze surgical workflows and fine-grained actions, encompassing Phase Recognition on Cholec80 (Twinanda et al., 2017) (Laparoscopic Cholecystectomy) and AutoLaparo (Wang et al., 2022) (Laparoscopic Hysterectomy), as well as Gesture Recognition on the JIGSAWS (Yixin Gao and Hager, 2014) (Robot-Assisted Surgical Skill Assessment) and

SAR-RARP50 (Psychogyios et al., 2024) (Robot-Assisted Radical Prostatectomy) datasets. 3. Deep Logical Reasoning challenges the model with tasks requiring multi-step deduction and spatiotemporal association, including Surgical Triplet Recognition on the CholecT45 (Nwoye et al., 2022) dataset and Critical View of Safety (CVS) Assessment on the Endoscapes (Mascagni et al., 2025) dataset.

By spanning across familiar domains (cholecystectomy) and entirely unseen domains (hysterec-tomy, prostatectomy, robotic surgery), SurgLite-Bench provides a rigorous testbed not only for in-domain task accuracy but also for zero-shot out-of-distribution (OOD) generalization, ensuring that our model is evaluated on its true clinical reasoning capabilities rather than mere dataset memorization.

3.4 Training Pipeline

3.4.1 Stage 1 & 2: Supervised Fine-Tuning (SFT)

The foundation of SurgMind-R1 is built upon the Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) architecture. The SFT process is divided into two sequential phases to progressively inject domain knowledge and reasoning capabilities.

Stage 1: Surgical Knowledge Acquisition (SFT) fine-tunes the base model using standard question-answer pairs derived from the Cholec80 and CholecT45 datasets. The primary objective here is to adapt the general-purpose VLM to the specific visual and semantic distribution of the surgical domain, enabling it to accurately recognize basic instruments, phases, and actions.

However, as demonstrated in our ablation studies, models fine-tuned solely via this standard SFT suffer from a severe reasoning collapse—they progressively lose their explicit thinking capabilities, reverting to direct, uninterpretable answer generation. To counteract this, we introduce Stage 2: Reasoning Recovery (SFT), where the model is fine-tuned using the 1,200 instances from the CholecThink dataset. By learning from explicit CoT data, the VLM successfully reactivates its surgical logical thinking, outputting a step-by-step reasoning process before arriving at the final diagnosis. Notably, our empirical results reveal that fine-tuning with a small number of high-quality CoT samples is sufficient to fully restore the model’s structural reasoning capabilities. This finding is particularly significant, as it provides a practical solution to the severe scarcity and prohibitive annotation costs of

medical CoT datasets.

Despite these advancements, SFT possesses inherent limitations. Even with CoT data, SFT models tend to memorize task-specific shortcuts and superficial correlations rather than acquiring robust, generalizable reasoning skills. This restricts their domain transferability and leaves the reasoning process susceptible to hallucinations when faced with edge cases.

3.4.2 Stage 3: Reinforcement Learning

To overcome the limitations of SFT, enhance reasoning robustness, and ensure the generation of explicit thought processes, we introduce Stage 3: Robustness Consolidation (RL) based on Reinforcement Learning, utilizing a subset of the surgical data (Cholec-GRPO).

1) Group Relative Policy Optimization (GRPO)

We adopt the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm to fine-tune the model. Unlike Proximal Policy Optimization (PPO) (Schulman et al., 2017), which requires an auxiliary value model (critic) that is often the same size as the policy model, GRPO estimates the baseline directly from a group of sampled outputs. This eliminates the memory overhead of the critic model, making it highly efficient for large-scale VLMs.

Specifically, for each input prompt x , GRPO samples a group of G outputs $\{y_1, y_2, \dots, y_G\}$ from the current policy π_θ . The objective function is formulated to maximize the expected advantage while penalizing deviations from the reference policy π_{ref} (which is initialized as the Stage 2 SFT model) via Kullback-Leibler (KL) divergence. The GRPO objective is defined as:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{x, \{y_i\}_{i=1}^G} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) \hat{A}_i \right) - \beta \mathbb{D}_{KL}(\pi_\theta \parallel \pi_{\text{ref}}) \right) \right], \quad \rho_i = \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)}. \quad (1)$$

where the expectation is taken over $x \sim P(x)$ and $\{y_i\}_{i=1}^G \sim \pi_{\text{old}}(y|x)$, ρ_i is the importance ratio, ϵ is the clipping hyperparameter, β is the KL penalty coefficient, and $\hat{A}_i$ is the relative advantage of the i -th output within the group. The advantage $\hat{A}_i$ is computed by standardizing the rewards r_i obtained by each output in the group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\})} \quad (2)$$

This group-based standardization ensures that the model learns to favor the relatively better reasoning paths among its own generated candidates, promoting continuous self-improvement without relying on an external critic.

2) Reward Function

Format Reward. We design a rule-based format reward to enforce strict adherence to the predefined response structure. A response is rewarded only when it follows the required `<think> ... </think> <answer> ... </answer>` template, ensuring that the model consistently exposes an explicit reasoning path before producing the final answer.

Accuracy Reward. We further introduce an accuracy reward to evaluate the factual correctness of the prediction extracted from the `<answer>` tag. This reward is computed against the ground-truth surgical labels and provides task-aware supervision for both tool & phase recognition and surgical triplet recognition, allowing GRPO to optimize not only structural compliance but also clinical prediction quality.

The specific reward function is provided in Appendix B.

4 Experiments

4.1 Experimental Setup

Our SFT training is based on the LLaMA-Factory (Zheng et al., 2024) framework. For the Supervised Fine-Tuning (SFT) stages (Stage 1 and Stage 2), we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022). For the Reinforcement Learning stage (Stage 3), we applied Group Relative Policy Optimization (GRPO) (Shao et al., 2024) also using LoRA. More details on compute and hyperparameters are provided in Appendix D.1 and Appendix D.3.

4.2 Main Results: Comparison with State-of-the-Art

Experimental Setup. To rigorously assess the model’s performance on in-domain surgical scene understanding, we conducted evaluations on the designated test splits of the Cholec80 and CholecT45 datasets, which constitute the in-domain subset of our proposed SurgLite-Bench. The evaluation encompasses both observational tasks (Tool Recognition) and comprehension tasks (Phase Recognition, Surgical Triplet Recognition). The test sets comprise a total of 20,098 frames

for the Cholec80 tasks and 19,105 frames for the CholecT45 triplet task.

Evaluation Metrics. We designed strict evaluation criteria to ensure clinical reliability. For the Tool and Phase Recognition tasks, we use Accuracy as the primary metric. A prediction is considered correct only if it achieves an exact match with the ground truth: the predicted list of tools must contain the exact same items as the label, and the predicted surgical phase must be identical to the ground truth.

For the highly complex Surgical Triplet Recognition task, we employ five distinct granularities of evaluation to comprehensively measure the model’s reasoning capabilities:

1. **Strict Triplet Accuracy:** The prediction is correct only if the number of predicted triplets matches the ground truth, and the content of all triplets is completely identical. This is an extremely challenging metric that demands perfect scene comprehension.
2. **Relaxed Triplet Accuracy:** The prediction is considered correct if at least one predicted triplet exactly matches a ground truth triplet.
3. **Component Accuracy (Instrument / Verb / Target):** We independently evaluate the three components of the triplets, contributing three distinct metrics. A prediction is correct if the number of predicted triplets matches the ground truth, and the set of the specific component exactly matches the ground truth set.

Results Analysis. As shown in Table 1, SurgMind-R1 consistently outperforms both zero-shot general-domain VLMs and the fine-tuned 7B baseline across surgical scene understanding tasks. On the core triplet metrics, it improves over the strongest zero-shot baseline by 22.82% in strict accuracy and 51.02% in relaxed accuracy, while also surpassing the SFT baseline by 8.53% and 15.06%, respectively. The gains extend to tool and phase recognition, indicating that the proposed training pipeline substantially improves both fine-grained interaction reasoning and broader surgical workflow comprehension.

Performance: Think vs. No-Think. As observed in Table 1, both the explicit reasoning (*think*) mode and direct (*no-think*) mode achieve competitive state-of-the-art results across all tasks. We emphasize that the core clinical value of the *think*

Model	Strict Triplet	Relaxed Triplet	Instrument	Verb	Target	Tool	Phase
Zero-shot VLMs							
LLaVA-1.5-7B (Liu et al., 2023)	3.21	4.70	11.15	5.78	11.10	0.01	3.58
Llama-4-Scout-17B-16E	1.17	2.91	9.17	4.17	3.20	2.04	1.52
InternVL3-8B (Zhu et al., 2025)	6.82	14.16	24.78	10.95	21.33	25.08	30.25
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	2.85	5.70	3.35	3.76	3.93	5.60	28.44
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	4.45	13.81	14.07	9.18	8.13	21.24	38.28
Qwen2.5-VL-72B-Instruct (Bai et al., 2025b)	7.30	21.60	19.45	12.76	9.23	12.24	35.72
Fine-tuned VLM							
Qwen2.5-VL-7B-Instruct (SFT)	21.59	57.56	47.35	39.21	23.90	57.86	56.71
SurgMind-R1 (think)	29.83	72.62	59.11	48.94	33.08	67.54	69.21
SurgMind-R1 (no-think)	30.12	72.38	62.25	49.00	33.82	73.12	70.88

Table 1: Performance comparison of different models on surgical video understanding tasks. All metrics are reported as percentages (%).

Mode	Avg. Inference Time (ms)	Avg. Output Tokens
Think	166.03	359.21
No-Think	88.53	17.30

Table 2: Inference efficiency comparison between think and no-think modes.

mode lies in its explicit transparency, interpretability, and decision traceability—providing clinicians with verifiable step-by-step rationales behind high-stakes intraoperative deductions rather than merely aiming to dominate every metric. Conversely, the high accuracy of the *no-think* mode reveals that SurgMind-R1 successfully internalizes structured clinical logic into its model weights through our three-stage pipeline. Across most surgical tasks, *no-think* mode exhibits slightly higher or comparable accuracy compared to *think* mode. This is primarily because direct answer generation effectively bypasses the potential variance and minor error accumulation inherent in multi-token autoregressive CoT verbalization, relying instead on the internalized clinical logic embedded within the model weights. This observation is consistent with recent findings showing that explicit Chain-of-Thought prompting can degrade accuracy or introduce reasoning noise (Liu et al., 2025a; Sadanandan and Behzadan, 2026). Meanwhile, *think* mode maintains strong performance across all evaluations while providing explicit deductive paths.

Efficiency and Latency: Think vs. No-Think. We measured the average inference latency per sample and the average generated token count across the full test set under identical hardware conditions. As shown in Table 2, compared with the explicit think mode, the *no-think* mode reduces the aver-

age inference time by 46.68% (from 166.03 ms to 88.53 ms per sample) and decreases the average output length by 95.18% (from 359.21 to 17.30 tokens). Combined with its comparable task performance in Table 1, this efficiency gain confirms that SurgMind-R1 internalizes surgical reasoning logic, substantially reducing generation overhead and making it suitable for latency-sensitive surgical deployment.

4.3 Out-of-Distribution Generalization

Experimental Setup. To rigorously evaluate the zero-shot out-of-distribution (OOD) generalization capabilities of SurgMind-R1, we leverage the OOD subset of our proposed SurgLite-Bench. This subset spans entirely unseen surgical scenarios and novel task definitions that the model never encountered during its training on laparoscopic cholecystectomy data. Specifically, the OOD evaluation includes four distinct datasets: (1) AutoLaparo (Wang et al., 2022), which tests Phase Recognition in a different surgical domain (Laparoscopic Hysterectomy); (2) SAR-RARP50 (Psychogyios et al., 2024), which introduces a novel, fine-grained Gesture Recognition task in Robot-Assisted Radical Prostatectomy; (3) JIGSAWS (Yixin Gao and Hager, 2014), which tests Gesture Recognition in Robot-Assisted Surgical Skill Assessment; and (4) Endoscapes (Mascagni et al., 2025), which challenges the model with a highly complex clinical reasoning task: Critical View of Safety (CVS) Assessment.

Evaluation Metrics. Given the inherent class imbalance prevalent in surgical datasets, we adopt metrics that provide a balanced view of model performance across all classes. For AutoLaparo, SAR-

Model	AutoLaparo			SAR-RARP50			JIGSAWS			Endo CVS
	Acc	Macro Acc	Macro F1	Acc	Macro Acc	Macro F1	Acc	Macro Acc	Macro F1	Mean Acc
LLaVA-1.5-7B	8.89	13.24	0.0699	16.23	12.61	0.0938	5.23	7.54	0.0209	35.45
Llama-4-Scout-17B-16E	4.85	13.04	0.0325	29.12	14.54	0.0668	10.57	7.63	0.0306	0.00
InternVL3-8B	23.65	14.43	0.0963	26.90	15.06	0.0987	6.73	8.72	0.0448	79.16
Qwen2.5-VL-3B-Instruct	18.96	13.86	0.1049	21.76	14.14	0.1107	9.58	6.63	0.0377	43.55
Qwen2.5-VL-7B-Instruct	18.08	11.51	0.0791	27.89	15.03	0.1160	11.14	7.93	0.0597	60.41
Qwen2.5-VL-72B-Instruct	16.78	10.68	0.0803	30.29	17.09	0.1243	12.48	7.15	0.0563	78.12
SurgMind-R1	24.99	15.89	0.1502	31.87	20.28	0.1965	20.31	13.37	0.1163	76.40

Table 3: Zero-shot OOD evaluation on unseen surgical scenarios and novel tasks across different baselines.

RARP50, and JIGSAWS, we report Overall Accuracy, Macro Accuracy, and Macro F1 score. For the Endoscapes CVS assessment, which evaluates three specific safety criteria, we report the Mean Criterion Accuracy across the three criteria.

Results Analysis. As shown in Table 3, SurgMind-R1 shows strong zero-shot generalization across unseen surgical domains and task definitions. It achieves the best overall results on AutoLaparo, SAR-RARP50, and JIGSAWS, with especially clear gains on JIGSAWS where accuracy improves by 7.83% over the strongest baseline. On the more demanding Endoscapes CVS task, SurgMind-R1 remains competitive despite being slightly below InternVL3-8B and Qwen2.5-VL-72B, suggesting that the proposed reasoning-enhanced training improves transferability while preserving robust performance on complex clinical reasoning tasks.

4.4 Ablation Study

To validate the effectiveness of our proposed three-stage training pipeline, we conducted an ablation study comparing models at different developmental stages: the base Qwen2.5-VL-7B model (Bai et al., 2025b), Stage 1 (Domain-adaptive SFT), Stage 2 (Reasoning-enhanced SFT), and Stage 3 (GRPO). Crucially, to observe the phenomenon of reasoning collapse and recovery, all models in this ablation study were explicitly prompted to output their step-by-step reasoning process before providing the final answer.

Reasoning Collapse Quantitative Indicators. To make this phenomenon measurable, we quantify reasoning collapse with three complementary indicators:

1. Empty Reasoning Rate (**ERR**) measures the proportion of outputs that preserve the `<think></think>` structure but contain no effective reasoning content inside the tags.
2. Incomplete Think-Tag Rate (**ITR**) measures

the proportion of outputs from which a complete `<think>...</think>` span cannot be extracted due to missing or malformed tags.

3. Out-of-Tag Residue Rate (**OTR**) measures the proportion of outputs that still contain non-empty residual text after removing complete `<think>` and `<answer>` blocks, indicating that the answer or intermediate content has escaped the prescribed structure. Together, these three indicators characterize reasoning collapse from the perspectives of content emptiness, structural corruption, and format instability.

Results Analysis: Performance Evolution. As shown in Table 4, each stage contributes to the final performance. Stage 1 provides the main domain adaptation gain, improving strict and relaxed triplet accuracy by 15.55% and 41.82% over the base model, while Stage 2 further improves these metrics by 7.58% and 12.90% through explicit CoT supervision. Stage 3 brings the best overall reasoning performance, ultimately outperforming the base model by 23.79% in strict triplet accuracy and 56.88% in relaxed triplet accuracy. The slight drop in tool recognition from Stage 2 to Stage 3 suggests that excessive reasoning may be less beneficial for purely perceptual tasks, but the overall trend confirms the complementary roles of domain adaptation, reasoning recovery, and RL-based consolidation.

Reasoning Collapse Across VLM Backbones. To investigate whether reasoning collapse is specific to a certain model or represents a broader failure mode in surgical adaptation, we evaluated additional VLM backbones alongside their answer-only SFT variants. As shown in Table 5, all evaluated architectures exhibit marked reasoning degradation following answer-only SFT. In particular, smaller-scale models (e.g., Qwen2.5-VL-3B) and certain architectures (e.g., LLaVA-1.5-7B) suffer a com-

Method	Strict Triplet	Relaxed Triplet	Instrument	Verb	Target	Tool	Phase
Qwen2.5-VL-7B (Baseline)	6.04	15.74	24.17	11.73	13.84	25.41	24.29
Stage 1 (SFT)	21.59	57.56	47.35	39.21	23.90	57.86	56.71
Stage 2 (+ CoT SFT)	29.17	70.46	57.75	48.44	32.35	68.70	65.52
Stage 3 (+ RL), SurgMind-R1	29.83	72.62	59.11	48.94	33.08	67.54	69.21

Table 4: Ablation study of different training stages on the 7B model. All metrics are reported as percentages (%).

Model	ERR↓	ITR↓	OTR↓	Avg. Words
InternVL3-8B	0.00	0.00	2.99	114.6
InternVL3-8B (SFT)	7.98	0.00	5.64	74.7
LLaVA-1.5-7B	0.00	76.93	83.43	23.4
LLaVA-1.5-7B (SFT)	0.00	100.00	100.00	4.3
Qwen2.5-VL-3B	0.00	6.25	6.40	119.1
Qwen2.5-VL-3B (SFT)	0.00	100.00	100.00	4.4
Qwen2.5-VL-7B	0.00	2.79	2.97	127.1
Qwen2.5-VL-7B (SFT)	19.93	3.69	28.79	59.3

Table 5: Quantitative analysis of reasoning collapse across different VLM baselines and corresponding answer-only SFT versions. ERR, ITR, OTR are reported as percentages (%).

plete loss of explicit reasoning, where ITR and OTR reach 100.00% and the average output length drops to ~ 4 words (collapsing entirely into direct shortcut answers). Larger models also show severe structural degradation. These findings substantiate that reasoning collapse is a cross-backbone vulnerability under standard answer-only SFT.

Results Analysis: Resolving Reasoning Collapse.

As shown in Table 6, answer-only SFT clearly induces reasoning collapse: Stage 1 sharply increases empty reasoning and out-of-tag residue, with OTR reaching 28.79%. Stage 2 largely repairs this behavior, reducing ERR and ITR to 0.00% and OTR to 1.09%, indicating that a compact set of high-quality CoT samples can restore stable reasoning structure. Stage 3 further eliminates the remaining residue and keeps all collapse indicators at 0.00%, showing that GRPO consolidates both format compliance and robust reasoning behavior. Furthermore, beyond format compliance, we conducted a manual evaluation by a medical expert on generated rationale quality, confirming substantial gains in clinical correctness, structural completeness, and hallucination reduction (detailed in Appendix C.3).

5 Conclusion

In this work, we introduce SurgMind-R1, a novel vision-language model designed to tackle the complexities of surgical scene understanding. We iden-

Method	ERR↓	ITR↓	OTR↓
Qwen2.5-VL-7B (Baseline)	0.00	2.79	2.97
Stage 1 (SFT)	19.93	3.69	28.79
Stage 2 (+ CoT SFT)	0.00	0.00	1.09
Stage 3 (+ RL), SurgMind-R1	0.00	0.00	0.00

Table 6: Quantitative analysis of reasoning collapse across different training stages on the 7B model. All metrics are reported as percentages (%).

tify a critical bottleneck in adapting foundational VLMs to the medical domain: the ‘‘Reasoning Collapse’’ phenomenon induced by standard SFT. To overcome this, we propose a robust three-stage training pipeline. By first adapting the model to the surgical domain, then explicitly enforcing structural reasoning through our newly curated Cholec-Think dataset, and finally elevating the logical deduction quality with GRPO, we effectively address reasoning collapse. Evaluated on our newly established multimodal surgical benchmark, SurgLite-Bench, SurgMind-R1 achieves state-of-the-art performance on complex inference tasks such as surgical triplet recognition. Crucially, the model demonstrates the ability to deeply internalize structured surgical thinking, maintaining high accuracy even in the low-latency *no-think* mode without explicit reasoning outputs. Furthermore, it exhibits strong zero-shot out-of-distribution generalization across entirely unseen surgical scenarios and novel tasks. Our work provides a practical paradigm for developing transparent, reliable, and interpretable medical AI, bridging the gap between foundation models and trustworthy computer-assisted surgical interventions.

Limitations

This study has several limitations. The training data is primarily derived from laparoscopic cholecystectomy; although SurgLite-Bench includes OOD evaluation across other surgical scenarios, broader validation on more diverse procedures, institutions, imaging devices, and patient populations is still

required. Cholec-Think is a compact reasoning dataset constructed through a hybrid expert and VLM-assisted annotation pipeline. Although all reasoning chains are manually screened, the dataset may still contain annotation bias or omit rare but clinically important reasoning patterns. The current formulation mainly evaluates frame-level or sparsely sampled visual understanding, and therefore does not fully capture long-horizon temporal dependencies in continuous surgical videos. In addition, the rule-based GRPO rewards and exact-match evaluation protocols improve controllability and reproducibility, but they may under-reward semantically correct answers with alternative wording or equivalent clinical interpretations. Accordingly, SurgMind-R1 should be regarded as a research prototype for transparent surgical reasoning rather than a deployable clinical decision-making system without further prospective validation.

Ethical Considerations

This work is intended for research on surgical vision-language reasoning and is not designed or validated for direct clinical decision-making. All visual data used in this study are drawn from publicly available surgical datasets and are used in accordance with their original access conditions, anonymization procedures, licenses, and terms of use. We do not collect new patient-identifying information. The physician-provided reasoning annotations were created with informed consent for research use. Although SurgMind-R1 improves transparent surgical reasoning, its predictions may still be affected by dataset bias, limited procedural diversity, and domain shift; therefore, any clinical use would require prospective validation, institutional approval, and oversight by qualified medical professionals.

Acknowledgments

This work was supported by the NSFC (62436007), the Zhejiang NSF (LRG25F020001), the Key R&D Program in Zhejiang Province (2025C01030), and the Ministry of Culture and Tourism (No. 2023DMKLB002).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu,

Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. *Qwen2.5-vl technical report*. *Preprint*, arXiv:2502.13923.

Jiawei Chen, Dingkan Yang, Yue Jiang, Mingcheng Li, Jinjie Wei, Xiaolu Hou, and Lihua Zhang. 2024a. Efficiency in focus: Layernorm as a catalyst for fine-tuning medical visual language pre-trained models. *arXiv preprint arXiv:2404.16385*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuoqpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024c. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 10818–10838. PMLR.

Pengfei Hao, Shuaibo Li, Hongqiu Wang, Zhizhuo Kou, Junhang Zhang, Guang Yang, and Lei Zhu. 2025. Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. *arXiv preprint arXiv:2506.19469*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*.

- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Yuheng Li, Yuxiang Lai, Maria Thor, Deborah Marshall, Zachary Buchwald, David S Yu, and Xiaofeng Yang. 2025. Towards universal text-driven ct image segmentation. *arXiv preprint arXiv:2503.06030*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2025a. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 38489–38517.
- Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. 2025b. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*.
- Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. 2025. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *Scientific Data*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.
- Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 2022. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.
- Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, Maxence Boels, Jiayu Huo, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, Sebastien Ourselin, Mengya Xu, An Wang, Yanan Wu, and 31 others. 2024. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Anita Rau, Mark Endo, Josiah Aklilu, Jaewoo Heo, Khaled Saab, Alberto Paderno, Jeffrey Jopling, F Christopher Holsinger, and Serena Yeung-Levy. 2025. Systematic evaluation of large vision-language models for surgical artificial intelligence. *arXiv preprint arXiv:2504.02799*.
- Binesh Sadanandan and Vahid Behzadan. 2026. [When chain-of-thought backfires: Evaluating prompt sensitivity in medical language models.](#) *Preprint*, arXiv:2603.25960.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer,

- and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. 2017. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97.
- Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, Hongbin Liu, Jiazheng Wang, Fan Zhang, Nicolas Padoy, Nassir Navab, and Hongliang Ren. 2026. Endochat: Grounded multimodal large language model for endoscopic surgery. *Medical Image Analysis*, page 103789.
- Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. 2022. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Carol E. Reiley Narges Ahmadi Balakrishnan Varadarajan Henry C. Lin Lingling Tao Luca Zappella Benjamin B ejar David D. Yuh Chi Chiung Grace Chen Ren e Vidal Sanjeev Khudanpur Yixin Gao, S. Swaroop Vedula and Gregory D. Hager. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*.
- Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*.
- Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, Xiaochun Cao, Yutong Ban, Qi Dou, Yang Liu, and Yueming Jin. 2025. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Dataset

A.1 Dataset Split

Since the visual data of CholecT45 originates from a subset of Cholec80, we implemented a strict de-duplication protocol to prevent data leakage and redundant training. Specifically, we removed the 45 overlapping videos from Cholec80, resulting in a disjoint set of 35 videos exclusively for Cholec80 tasks and 45 videos for CholecT45 tasks. The unified dataset was then partitioned at the video level following a 7 : 1 : 2 ratio, as summarized in Table 7.

Split	CholecT45	Cholec80	Total
Train	31	25	56
Validation	5	3	8
Test	9	7	16
Total	45	35	80

Table 7: Video-level dataset split after de-duplication.

A.2 Cholec-Think Construction

To overcome the prohibitively high cost of dense medical annotations, Cholec-Think was developed using a scalable hybrid pipeline: a small seed set of expert-curated reasoning paths was combined with large-scale VLM-assisted generation using

Task	Stratum	Samples
Tool & phase	Video ID	600
Triplet	0 triplets	100
Triplet	1 triplet	250
Triplet	2 triplets	150
Triplet	≥ 3 triplets	100
Total	–	1,200

Table 8: Sampling strategy for Cholec-Think construction.

gemini-3.1-pro-preview. Three professional physicians from collaborating hospitals each wrote reasoning paths for 20 images in Cholec-Think. For example, in the instrument and phase identification tasks, the reasoning process includes Anatomical Identification, Instrument Identification, specific instrument localization, Action Analysis, Phase Determination, and Conclusion. In the surgical triplet (tool, action, target) identification task, the reasoning process includes Tool Identification, Target Identification, Action Identification, and Conclusion for each triplet. All generated reasoning chains were subsequently screened by physicians for clinical validity and logical coherence, and the physicians were allowed to revise the generated CoT when necessary. The physicians provided informed consent for their annotations to be used for research, and the source images were drawn only from publicly available anonymized datasets.

To ensure the reasoning dataset is representative and balanced across different levels of scene complexity, we employed a stratified sampling strategy to select 1,200 images from the training sets. The sampling design is summarized in Table 8. Within each stratum, frames are sampled via a round-robin schedule across distinct video IDs to ensure even coverage across different surgical cases and surgeons while preventing sample over-representation from longer videos.

For Stage 3 reinforcement learning, Cholec-GRPO follows the same stratified sampling strategy and contains 1,200 non-overlapping samples drawn from the training set.

B Reward Functions in GRPO

The GRPO stage uses a rule-based reward composed of a format reward and a surgical accuracy

reward:

$$R(y, y^*) = R_{\text{fmt}}(y) + R_{\text{acc}}(y, y^*), \quad (3)$$

where y is the model output and y^* is the ground-truth answer. Let $\mathcal{Y}_{\text{fmt}}$ denote the set of outputs that strictly follow the required `<think>...</think><answer>...</answer>` reasoning-answer structure. The format reward is defined as:

$$R_{\text{fmt}}(y) = \begin{cases} 1, & y \in \mathcal{Y}_{\text{fmt}}, \\ 0, & y \notin \mathcal{Y}_{\text{fmt}}. \end{cases} \quad (4)$$

Let $F1(A, B) = 2|A \cap B|/(|A| + |B|)$ denote the set-level F1 score between the predicted set A and the ground-truth set B . For the tool & phase task, the answer is parsed as `[tool1, ...]; [phase]`. Given predicted tools $\hat{T}$, ground-truth tools T , predicted phase $\hat{p}$, and ground-truth phase p , the accuracy reward is:

$$R_{\text{acc}}^{\text{tool\&phase}} = \frac{1}{2} F1(\hat{T}, T) + \frac{1}{2} \mathbb{I}[\hat{p} = p]. \quad (5)$$

For the triplet task, the answer is parsed as a set of triplets `[instrument, verb, target]`. Let $\hat{G}$ and G be the predicted and ground-truth triplet sets, and let $\hat{I}, \hat{V}, \hat{C}$ and I, V, C denote the corresponding instrument, verb, and target component sets. The accuracy reward is:

$$R_{\text{acc}}^{\text{triplet}} = \frac{1}{2} F1(\hat{G}, G) + \frac{1}{6} F1(\hat{I}, I) + \frac{1}{6} F1(\hat{V}, V) + \frac{1}{6} F1(\hat{C}, C). \quad (6)$$

Thus, the triplet reward jointly measures exact triplet-level correctness and partial component-level correctness, while the tool & phase reward balances multi-label tool recognition with exact phase matching.

C Additional Results

C.1 Extended Baseline and Variants Comparisons

To provide a comprehensive benchmark comparison, we extended our evaluation to include the recent foundation model Qwen3-VL-8B-Instruct, as well as CoT-prompted (think mode) variants of general zero-shot VLMs (LLaVA-1.5-7B, InternVL3-8B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B). In the think mode setting, zero-shot models were explicitly instructed to produce step-by-step reasoning prior to generating answers.

Model	Strict Triplet	Relaxed Triplet	Instrument	Verb	Target	Tool	Phase
Zero-shot VLMs							
LLaVA-1.5-7B	3.21	4.70	11.15	5.78	11.10	0.01	3.58
LLaVA-1.5-7B (think)	3.83	5.62	10.30	6.28	10.24	0.00	2.21
Llama-4-Scout-17B-16E	1.17	2.91	9.17	4.17	3.20	2.04	1.52
InternVL3-8B	6.82	14.16	24.78	10.95	21.33	25.08	30.25
InternVL3-8B (think)	2.45	6.96	21.63	5.23	12.08	19.15	19.92
Qwen2.5-VL-3B-Instruct	2.85	5.70	3.35	3.76	3.93	5.60	28.44
Qwen2.5-VL-3B-Instruct (think)	2.36	4.31	11.33	6.60	4.56	11.06	26.92
Qwen2.5-VL-7B-Instruct	4.45	13.81	14.07	9.18	8.13	21.24	38.28
Qwen2.5-VL-7B-Instruct (think)	6.04	15.74	24.17	11.73	13.84	25.41	24.29
Qwen2.5-VL-72B-Instruct	7.30	21.60	19.45	12.76	9.23	12.24	35.72
Qwen3-VL-8B-Instruct	8.45	19.64	20.26	13.63	11.78	16.16	18.40
Fine-tuned VLM							
Qwen2.5-VL-7B-Instruct (SFT)	21.59	57.56	47.35	39.21	23.90	57.86	56.71
SurgMind-R1 (think)	29.83	72.62	59.11	48.94	33.08	67.54	69.21
SurgMind-R1 (no-think)	30.12	72.38	62.25	49.00	33.82	73.12	70.88

Table 9: Extended performance comparison including recent SOTA models and zero-shot CoT-enabled variants on surgical video understanding tasks. All metrics are reported as percentages (%).

Model	Structure Completeness $\uparrow$	Anatomical Accuracy $\uparrow$	Action Accuracy $\uparrow$	Coherence $\uparrow$	Hallucination Rate (%) $\downarrow$
Qwen2.5-VL-7B (Baseline)	2.76	1.50	1.72	2.96	18.0
Stage 1 (SFT)	1.76	2.85	2.93	2.54	28.0
Stage 2 (+ CoT SFT)	4.28	3.44	3.62	3.88	16.0
Stage 3 (+ RL), SurgMind-R1	4.33	3.40	3.62	4.13	5.0

Table 10: Manual evaluation of surgical rationale quality across training stages.

As summarized in Table 9, Qwen3-VL-8B-Instruct (Bai et al., 2025a) achieves the highest strict triplet accuracy (8.45%) among all zero-shot baselines, yet still falls far short of domain-adapted SurgMind-R1 (30.12%), highlighting that scaling general-domain pretraining alone is insufficient for nuanced surgical spatiotemporal reasoning. Moreover, prompting general-purpose VLMs for explicit CoT does not consistently yield improvements; on certain fine-grained tasks (e.g., InternVL3-8B and Qwen2.5-VL-3B on phase and triplet recognition), prompting for explicit rationales even leads to slight performance drops due to domain mismatch and visual hallucinations. In contrast, SurgMind-R1 achieves consistent state-of-the-art accuracy across both *think* and *no-think* modes, demonstrating the critical efficacy of our structured three-stage training pipeline.

C.2 Qualitative Analysis of Reasoning Collapse

Figure 3 provides qualitative examples of reasoning collapse and its resolution across different training stages.

The base Qwen2.5-VL model often produces incorrect reasoning because it lacks sufficient surgical domain knowledge. After Stage 1 domain-adaptive SFT, the model becomes better aligned with surgical labels but tends to rely on shortcut patterns, resulting in empty, truncated, or structurally unstable reasoning. Stage 2 reasoning-enhanced SFT restores explicit step-by-step reasoning, but the generated rationales may still contain hallucinated or clinically inconsistent intermediate steps. In contrast, SurgMind-R1 produces accurate and clinically coherent reasoning after GRPO, showing that the final stage not only improves format compliance but also consolidates the quality of surgical reasoning.

C.3 Manual Evaluation of Rationale Quality

To verify that the recovery of explicit reasoning structure corresponds to genuine clinical reasoning quality rather than superficial template matching, we conducted a blind evaluation of generated rationales across different training stages by one medical expert with clinical domain backgrounds.

Evaluation Protocol. We randomly sampled

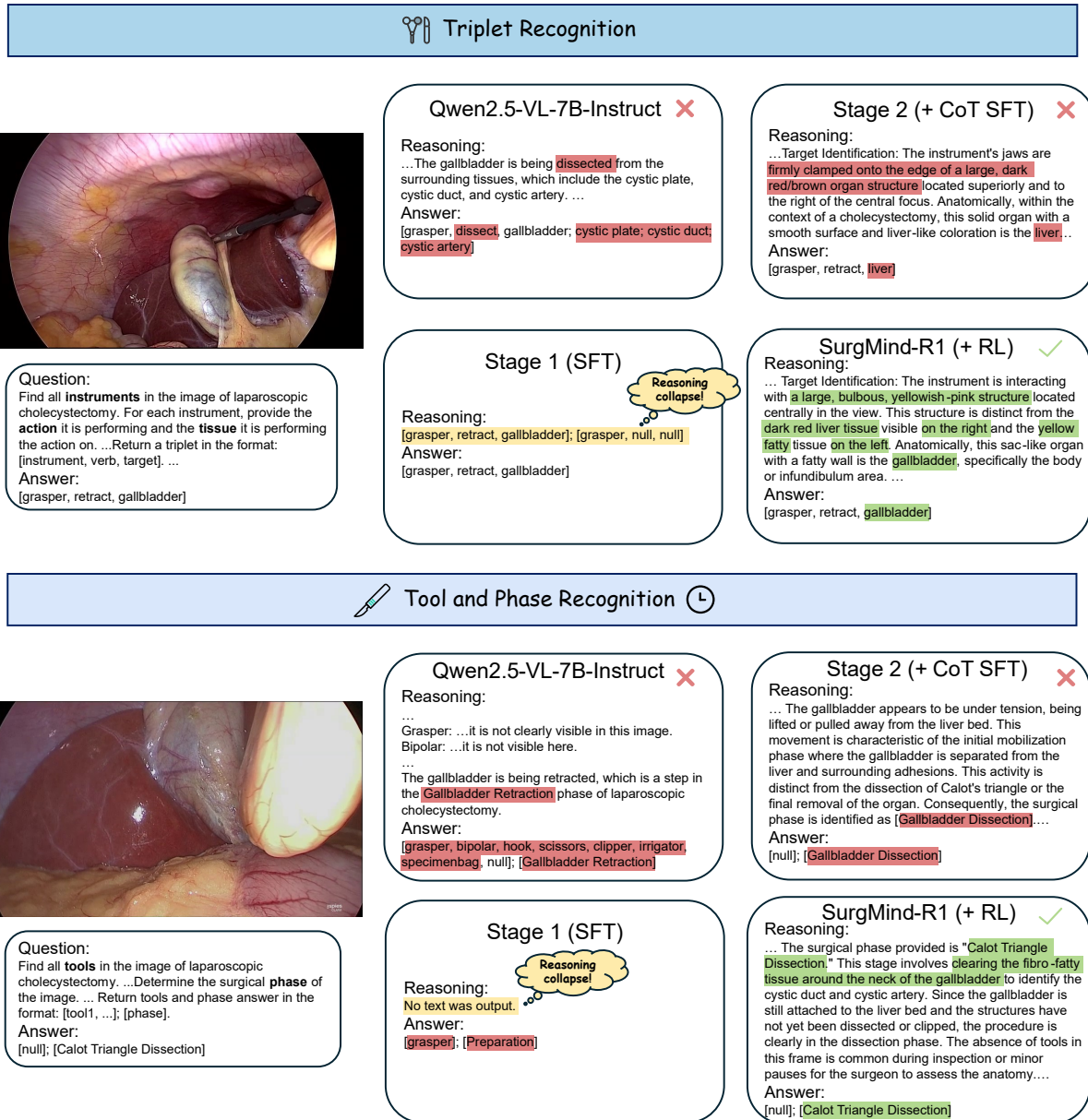


Figure 3: Qualitative examples of reasoning collapse and recovery across training stages.

100 cases from the in-domain test set. For each case, the rationales generated by four models—Baseline (Qwen2.5-VL-7B), Stage 1 (Domain-adaptive SFT), Stage 2 (Reasoning-enhanced SFT), and SurgMind-R1 (Stage 3 RL)—were evaluated across five clinically motivated dimensions:

- Reasoning Structure Completeness** (1–5 scale): Evaluates whether the generated rationale systematically covers the necessary clinical deduction steps (e.g., anatomical identification and action analysis).
- Anatomical Reasoning Accuracy** (1–5 scale): Measures the factual correctness of anatomical identification and spatial refer-

ences.

- Action Reasoning Accuracy** (1–5 scale): Assesses the validity of instrument-tissue interaction modeling and procedural action deduction.
- Reasoning Coherence** (1–5 scale): Assesses the internal logical consistency between intermediate observational deductions and the final conclusion.
- Hallucination Rate (%)**: Measures the proportion of generated rationales that contain non-existent anatomical structures, phantom surgical instruments, or fabricated procedural

events.

Results Analysis. As reported in Table 10, although Stage 1 SFT improves basic anatomical and action accuracy over the zero-shot baseline, its reasoning structure completeness and coherence drop noticeably, accompanied by an elevated hallucination rate of 28.0% (consistent with the reasoning collapse). Stage 2 reasoning-enhanced SFT restores explicit structural completeness to 4.28 and boosts anatomical and action reasoning accuracy to 3.44 and 3.62, respectively. Finally, Stage 3 GRPO consolidates reasoning robustness by improving logical coherence to 4.13 and drastically reducing the hallucination rate from 16.0% to 5.0%. These findings provide empirical confirmation that our training pipeline restores meaningful clinical reasoning capability rather than superficial format adherence.

D Reproducibility

D.1 Compute

All experiments were conducted on a workstation equipped with two NVIDIA A800 (80GB) GPUs. The approximate computational budget was 15 GPU-hours for Stage 1 SFT, 1 GPU-hour for Stage 2 CoT SFT, and 24 GPU-hours for Stage 3 GRPO, resulting in a total training budget of 40 GPU-hours.

D.2 Framework Versions

The main software dependencies include PEFT 0.17.1, Transformers 4.57.1, PyTorch 2.4.0+cu124, Datasets 4.0.0, and Tokenizers 0.22.2. The GRPO training stage was implemented using the open-r1-multimodal framework.

D.3 Hyperparameters

The detailed parameters and hyperparameters for the three training stages are summarized in Tables 11–13.

E Artifact Licenses and Privacy

E.1 Artifact Licenses and Terms of Use

All public datasets utilized in this work are subject to their respective original licenses and terms of use. Our derived artifacts, including SurgLite-Bench and Cholec-Think, are developed strictly for academic research purposes and are distributed in full compliance with the access and licensing conditions of the underlying datasets. The accompanying codebase, model checkpoints, and annotation

files are provided with detailed documentation and appropriate licensing terms specifying permissible non-commercial research use.

E.2 Privacy and Data Anonymization

This study uses publicly available endoscopic and surgical video datasets and does not collect new patient-identifying information. The original datasets are used in accordance with the anonymization procedures, access conditions, and terms of use specified by their respective providers. The data used in this work consist of surgical visual scenes and task annotations; they do not include user-generated text or offensive content.

Parameter	Value	Parameter	Value
Cutoff length	1024	Max image pixels	589,824
LoRA rank	16	LoRA alpha	32
LoRA dropout	0.05	Learning rate	5e-5
LR scheduler	Cosine	Warmup ratio	0.03
Epochs	1	Train batch size / device	8
Gradient accumulation	4	Eval batch size / device	2
Dataloader workers	4	Precision	bfloat16

Table 11: Parameters and hyperparameters for Stage 1: domain-adaptive SFT.

Parameter	Value	Parameter	Value
Validation split	0.1	Seed	42
Cutoff length	2048	Max image pixels	589,824
LoRA rank	32	LoRA alpha	64
LoRA dropout	0.05	Learning rate	1e-4
LR scheduler	Cosine	Warmup ratio	0.03
Epochs	6	Train batch size / device	4
Eval batch size / device	4	Gradient accumulation	4
Dataloader workers	4	Precision	bfloat16

Table 12: Parameters and hyperparameters for Stage 2: reasoning-enhanced CoT SFT.

Parameter	Value	Parameter	Value
Max prompt length	1024	Max completion length	1024
Train batch size / device	4	Gradient accumulation	1
Training steps	300	Generations per prompt	4
Max image pixels	401,408	Min image pixels	3,136
LoRA rank	16	LoRA alpha	32
LoRA dropout	0.05	Learning rate	1e-6
Optimizer	AdamW	Precision	bfloat16

Table 13: Parameters and hyperparameters for Stage 3: GRPO training.