SurgReasoner: Surgical Reasoning Segmentation with Dynamic Difficulty-Aware Reinforcement Learning

Dong Zhang[1,2] **Cheng Xue**[1,2,†*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology
and Its Interdisciplinary Applications,
Ministry of Education, Nanjing, China
{dongz, cxue}@seu.edu.cn

Abstract

Surgical reasoning segmentation requires a model to infer the target instrument or anatomical structure from an implicit clinical query and produce a pixel-level mask. This setting is far more challenging than category-level surgical segmentation, as targets are specified by functional role, spatial relation, or anatomical interaction rather than explicit class names. It also exposes a key limitation of existing reinforcement learning-based visual grounding methods: standard reward normalization tends to under-optimize clinically important hard samples, such as small instruments, occluded organs. To address this, we introduce SurgSeg-14K, a surgical reasoning segmentation benchmark with over 14K image-mask pairs across seven surgical scenarios and 30 fine-grained categories. Each sample links an implicit clinical query to a binary mask, spatial annotations, and a reasoning trajectory. We further propose SurgReasoner, which decomposes the task into two stages: clinical query interpretation for spatial prompt generation, and prompt-based mask generation with a frozen segmentation model. To improve reinforcement learning for spatial reasoning, we introduce a dynamic difficulty-aware reweighting strategy that combines intrinsic target difficulty with rollout correctness, enabling training to focus on challenging targets. Experiments on SurgSeg-14K show that SurgReasoner outperforms the strongest grounding baseline by large margins, with consistent gains especially on hard samples, suggesting a practical path toward reasoning-driven surgical visual perception.

1 Introduction

Surgical scene understanding increasingly requires models to answer natural language queries with precise pixel-level grounding. Recent Multimodal Large Language Models (MLLMs) link visual perception with language interaction, enabling semantic understanding and preliminary localization in medical and surgical images[1–3]. However, surgical assistance demands more than textual responses or coarse localization: the model must infer a specific instrument or anatomical structure from an implicit clinical query and generate a category-level binary mask. Unlike category-level segmentation, where the target class is explicit, this setting requires reasoning from functional roles, spatial relationships, and anatomical interactions. For example, a query may refer to "the instrument retracting tissue to maintain a clear surgical field" rather than directly naming a surgical tool.

Existing surgical vision benchmarks are poorly aligned with this requirement. Surgical segmentation methods mainly perform category-level parsing of instruments or organs [4–6], thus lacking support for language-conditioned instance disambiguation and implicit clinical queries [7, 8]. Surgical visual

*Corresponding author. †These authors contributed equally to this work.

question answering (VQA) methods support semantic interaction but output text rather than pixel-level masks [9, 10]. Recent reasoning segmentation methods combine multimodal reasoning with visual grounding [11, 12], but mostly target general or broad medical scenarios. Applying them to surgical scenes remains challenging, as surgical targets are often small, occluded, visually similar, or described by functional cues rather than explicit names.

To address this gap, we introduce SurgSeg-14K, a surgical reasoning segmentation benchmark linking implicit clinical queries to category-level binary masks. The dataset draws on eight open-source surgical datasets and one private dataset, covering seven procedure types and diverse instruments and anatomical structures. For each target object, we construct an implicit clinical query and a reasoning trajectory from visual, functional, and spatial cues. Bounding boxes and centroids derived from binary masks support spatial prompt generation and reward computation without extra annotation. Using GPT-4o [13] with a structured prompting and filtering protocol, we generate clinically inspired queries and chain-of-thought trajectories [14], followed by consistency checks against the corresponding masks. A subset of generated samples is further reviewed by clinical experts for query plausibility, target uniqueness, and mask-reasoning consistency.

We further propose SurgReasoner, a reinforcement learning based framework for surgical reasoning segmentation [15]. SurgReasoner decomposes the task into clinical query interpretation for spatial prompt generation and prompt-based mask generation with a frozen segmentation model. This design enables the reasoning module to be optimized using mask-derived spatial rewards without jointly retraining the segmentation model. Compared with SFT-based localization schemes that often rely on special tokens and dense annotations [11, 12], SurgReasoner trains the reasoning module through rule-based reinforcement learning, using mask-derived spatial signals as rewards.

A key challenge in this RL formulation is the difficulty imbalance of surgical targets. Standard reward normalization may under-optimize clinically important hard cases, such as tiny instruments and partially occluded organs. To address this issue, we introduce dynamic difficulty-aware advantage reweighting for GRPO. It combines objective difficulty from mask-derived target properties with subjective difficulty from rollout correctness, allowing training to progressively focus on under-mastered targets that are small and occluded while maintaining stability.

To summarize, our main contributions are as follows:

- We introduce surgical reasoning segmentation and construct SurgSeg-14K, a clinically informed benchmark linking implicit clinical queries to category-level binary masks, mask-derived spatial annotations, and reasoning trajectories.
- We propose SurgReasoner, a reinforcement-optimized spatial prompting framework that decomposes the task into clinical query interpretation and prompt-based mask generation, enabling mask-derived reward optimization without retraining the frozen segmentation model.
- We introduce dynamic difficulty-aware advantage reweighting for GRPO, which adaptively focuses training on small, occluded, and under-mastered surgical targets by combining intrinsic objective and subjective target difficulty with rollout-based correctness.
- We conduct comprehensive experiments against general-purpose, medical-specific, and grounding-oriented MLLMs, demonstrating the effectiveness of SurgReasoner in surgical visual grounding and pixel-level segmentation.

2 Related Work

2.1 Post-Training and Reinforcement Learning for MLLMs

Instruction tuning and RL are key post-training strategies for adapting MLLMs to downstream tasks [16, 17]. Instruction tuning boosts task following via image-text instructions, as in LLaVA [18] and LLaVA-OV [19]. Recent RL-based post-training enhances multimodal reasoning, visual understanding and response quality [20–22]. Most methods combine supervised fine-tuning with RL but focus on general reasoning and response quality. In contrast, surgical reasoning segmentation requires spatially precise prompts with semantic and geometric fidelity. Small, occluded or visually similar surgical targets produce weak rewards, complicating the task. To address this, we propose difficulty-aware advantage reweighting for RL-based spatial prompt optimization.

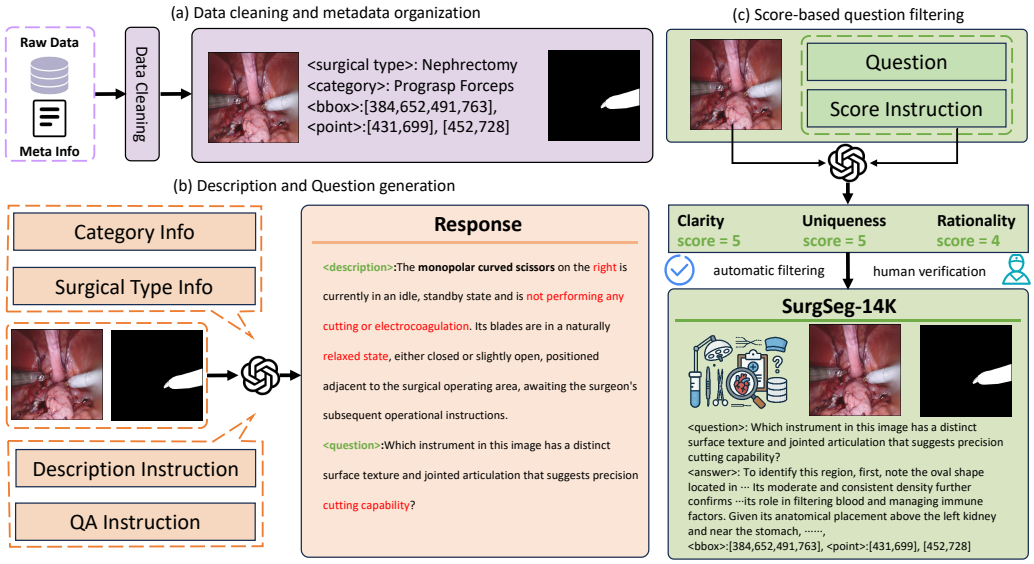


Figure 1: Overview of the SurgSeg-14K construction pipeline: (1) Data cleaning and metadata organization, (2)Description and Question generation, (3) Score-based question filtering.

2.2 Reasoning Segmentation and Visual Grounding

Visual grounding aims to localize image regions referred to by natural language expressions, while reasoning segmentation further requires the model to infer the target through semantic reasoning before generating a mask. Recent MLLMs exhibit strong joint visual-language reasoning capabilities [23–25], and foundation models like SAM enable effective prompt-based segmentation [26]. Several recent works integrate MLLMs with segmentation models to achieve reasoning-aware grounding [11, 27]. Seg-Zero explores reinforcement learning for reasoning-guided segmentation [28], and MedReasoner further adapts such pipelines to medical images [29]. However, these methods are primarily designed for general or broad medical scenes, with explicit descriptions and clearly defined object boundaries. Surgical scenarios present unique challenges: tiny, deformable, occluded targets, and queries based on functional roles rather than category names. Thus, directly applying existing methods is largely inadequate for surgical grounding. We bridge this gap by proposing a dedicated surgical reasoning benchmark and optimizing spatial prompting under target difficulty imbalance.

2.3 Surgical Scene Understanding

Surgical scene understanding has been widely studied through various tasks such as instrument and organ segmentation, surgical workflow recognition, and surgical visual question answering (VQA) [30–33]. Existing surgical segmentation methods mainly focus on category-level parsing of instruments or anatomical structures, providing pixel-level masks for predefined classes but offering only limited support for language-conditioned instance disambiguation [5, 6]. Surgical VQA methods enable flexible natural language interaction with surgical images or videos, but their outputs are typically purely textual answers and do not provide precise pixel-level grounding [9, 10, 34]. As a result, current surgical benchmarks still do not fully support the more challenging task of inferring a specific surgical target from an implicit clinical query and producing its category-level binary masks.

Our work differs from prior surgical segmentation and VQA studies by formulating surgical reasoning segmentation, a new task where implicit clinical queries are linked to category-level binary masks through reasoning trajectories. SurgSeg-14K provides a dedicated benchmark for this task, while SurgReasoner decomposes the problem into two stages: clinical query interpretation and prompt-based mask generation. This formulation allows the reasoning module to be trained with mask-derived spatial rewards while keeping the segmentation model fully frozen.

3 SurgSeg-14K Dataset

3.1 Construction of SurgSeg-14K

Each sample in SurgSeg-14K consists of four components: a surgical image, a category-specific binary target mask, an implicit clinical query, and a structured multi-step reasoning trajectory. The dataset is constructed through a three-stage pipeline, as shown in Figure 1: mask standardization, implicit query and rationale generation, and expert-based quality control.

Mask standardization. We collect surgical semantic segmentation data from multiple public datasets and one private dataset, covering seven distinct surgical scenarios. Since the source datasets follow different label taxonomies, we first standardize the surgical types and category names for anatomical structures and surgical instruments. We then convert each semantic label map into category-specific binary target masks by extracting one binary mask for each annotated category. These target masks serve as the primary spatial supervision for downstream query generation, derived spatial annotations, and reward computation. Bounding boxes and centroid coordinates are automatically computed from the target masks and thus do not require any additional manual annotation.

Query and rationale generation. For each target mask, we use GPT-4o under a structured prompting protocol to generate a target description based on visual appearance, spatial location, functional state, and potential surgical role. The resulting description is then converted into an implicit clinical query, in which the target is referred to by functional or spatial cues rather than by a direct category name. We also generate a structured reasoning trajectory that explains step by step how the target can be localized from the query and the surrounding image context.

Quality control. We apply automatic filtering and sampled clinical expert review. First, GPT-4o scores each generated query on semantic clarity, target uniqueness, and clinical plausibility (1–5 scale). Samples with any score < 4 are removed, as are those with mismatches between the reasoning trajectory and target mask. Second, clinical experts review a random subset to assess query plausibility, target uniqueness, and mask-rationale consistency; ambiguous or inconsistent samples are revised or removed. Details of dataset composition, statistics and examples are provided in Appendix A.

4 Methodology

4.1 Problem Definition

Reasoning segmentation in surgical scenes aims to generate accurate pixel-level segmentation masks given a surgical image I and an implicit clinical referring expression E. Unlike standard referring segmentation, our task requires the model to infer an explicit intermediate reasoning chain R to enhance interpretability and clinical reliability. Formally, given surgical image I and language query E, the task is defined as:

$$\{R, M\} = \mathcal{F}(I, E) \tag{1}$$

where $\mathcal{F}$ denotes the overall framework, R is the explicit reasoning trajectory, and M is the final pixel-level segmentation mask for the target surgical instrument or organ.

4.2 SurgReasoner Architecture

SurgReasoner decomposes surgical reasoning segmentation into two stages: spatial prompt generation and prompt-based mask prediction. Given a surgical image I and an implicit clinical query Q, the Surgical Reasoning Module (SRM) first interprets the query and generates a reasoning trajectory R together with a spatial prompt S. The Spatial Segmentation Module (SSM) then takes I and S as input and produces the final pixel-level target mask.

The SRM is built on Qwen2.5-VL and is denoted as f_θ. After supervised fine-tuning on SurgSeg-14K, it learns to understand surgical terminology, instrument functions, anatomical structures, and functional relations in implicit clinical queries. Formally, the SRM predicts

$$R, S = f_\theta(I, Q), \tag{2}$$

where the spatial prompt is defined as $S = \{B, P_1, P_2\}$. Here, B denotes the predicted bounding box of the referred target, P_1 and P_2 denotes the target key points used to provide additional spatial

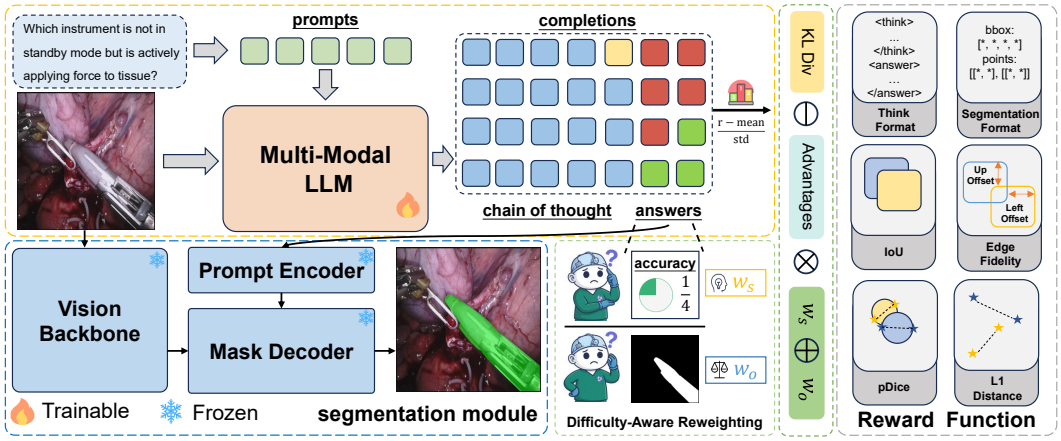


Figure 2: Overview of the proposed SurgReasoner architecture. SRM generates a reasoning chain and provides segmentation prompts. SSM produces pixel-wise mask.

guidance for segmentation. During training, the reference bounding box and centroid coordinates are automatically derived from the ground-truth target mask and used for reward computation.

The SSM is implemented with a frozen SAM2 model[35], denoted as g_ϕ. Given the image and the spatial prompt generated by the SRM, the SSM predicts the target mask:

$$M = g_\phi(I, S). \tag{3}$$

The segmentation module remains entirely frozen throughout training, and only the reasoning module is optimized through reinforcement learning. This decoupled design allows the reward signal to be computed directly from the final mask quality while updating only the SRM through policy optimization. It also avoids expensive joint retraining of the segmentation model and enables the segmentation backbone to be upgraded or replaced independently.

4.3 Reward Design

We design reward functions to optimize the Surgical Reasoning Module (SRM) for both valid output formatting and accurate spatial prompt generation. Given a generated response, the total reward is defined as

$$r = \lambda_f r_{\text{fmt}} + \lambda_s r_{\text{sp}},$$

where r_{fmt} denotes the format reward, r_{sp} denotes the spatial accuracy reward, and λ_f and λ_s are weighting coefficients. The reference bounding box and centroid coordinates used in the reward computation are automatically derived from the ground-truth target mask.

Format reward. The format reward encourages the SRM to produce parseable reasoning and spatial prompts. It consists of two terms:

$$r_{\text{fmt}} = r_{\text{think}} + r_{\text{seg-format}}.$$

The thinking format reward r_{think} evaluates the structural validity of the model's output, focusing on the formatting of the reasoning components. The segmentation format reward $r_{\text{seg-format}}$ verifies whether the predicted spatial prompt follows a valid structure, including a bounding box and point coordinates within the image range. This reward ensures that the output can be reliably parsed and passed to the segmentation module.

Spatial accuracy reward. The spatial accuracy reward measures whether the predicted prompt correctly localizes the referred target:

$$r_{\text{sp}} = \lambda_{\text{iou}} r_{\text{iou}} + \lambda_{\text{edge}} r_{\text{edge}} + \lambda_{\text{pdice}} r_{\text{pdice}} + \lambda_{\text{distance}} r_{\text{distance}}.$$

The bounding-box IoU reward r_{iou} encourages overlap between the predicted bounding box and the mask-derived reference bounding box. The edge fidelity reward r_{edge} penalizes deviations between

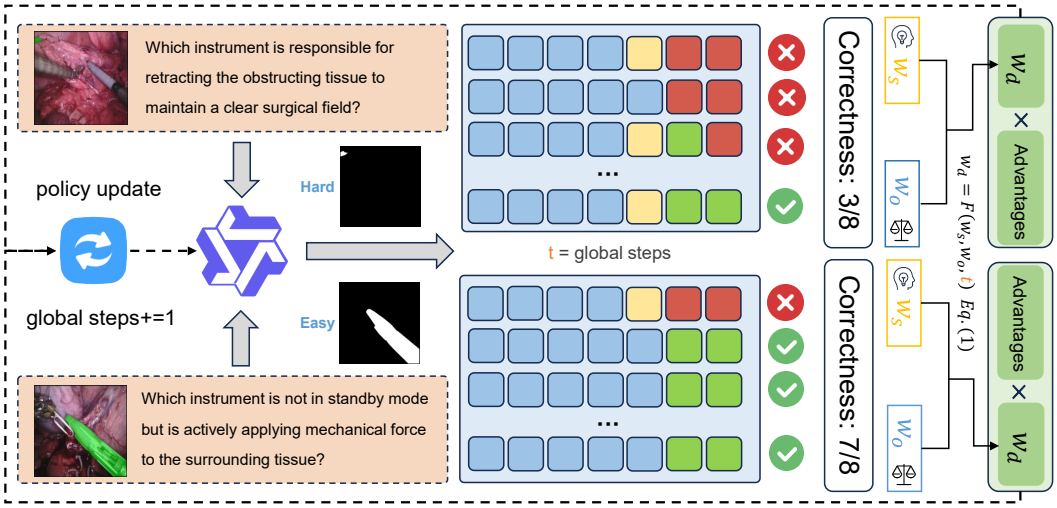


Figure 3: Illustration of our dynamic difficulty-aware advantage reweighting mechanism for GRPO training.

the predicted and reference box boundaries. The point reward r_{distance} measures the distance between the predicted point and the mask-derived centroid or reference point. Finally, the mask reward r_{pdice} evaluates the spatial correspondence between a predicted keypoint pair and a groundtruth keypoint pair. Together, these terms encourage the SRM to generate spatial prompts that are both valid and geometrically aligned with the target. Detailed formulations and calculations of each reward component are provided in Appendix B.1.

4.4 Dynamic Difficulty-Aware Advantage Reweighting

Surgical reasoning segmentation exhibits substantial sample-level difficulty variation. Large and visually salient targets are often easy to localize, whereas small, occluded, or visually ambiguous targets may produce weaker and less stable reward signals during GRPO training. Treating all samples uniformly can therefore under-emphasize these hard cases during policy updates. To address this issue, we introduce dynamic difficulty-aware advantage reweighting, which estimates sample difficulty from two complementary sources: objective difficulty, derived from mask-level target properties such as relative target area, and subjective difficulty, estimated from the model's rollout correctness during training. The two difficulty estimates are dynamically fused according to the training progress and used to reweight the GRPO group-normalized advantage, encouraging the policy to focus more on under-mastered surgical targets with small and occluded regions.

Objective difficulty. We first define an objective difficulty weight from the target mask size. Let $\lambda_q \in (0, 1]$ be the foreground area ratio of the target mask in sample q. Smaller targets are assigned larger weights by normalizing the reciprocal area ratio:

$$w_o(q) = 0.5 + \frac{\frac{1}{\lambda_q} - \min_i \frac{1}{\lambda_i}}{\max_i \frac{1}{\lambda_i} - \min_i \frac{1}{\lambda_i}}, \tag{4}$$

where the minimum and maximum are computed over the training set. Thus, $w_o(q) \in [0.5, 1.5]$.

Subjective difficulty. Objective difficulty is fixed throughout training and cannot reflect whether the current model has already mastered a sample. We therefore define a subjective difficulty weight based on rollout correctness. Let ρ_q be the correctness rate among the GRPO rollouts for sample q, where a rollout is considered correct if its output satisfies the predefined format and localization criteria. A lower ρ_q indicates that the sample is harder for the current policy. We map ρ_q to a difficulty weight using

$$w_s(q) = A + \frac{B - A}{1 + \exp(k(\rho_q - \rho_0))}, \tag{5}$$

where A and B define the weight range, ρ_0 is the midpoint, and k controls the slope.

Dynamic fusion. At early training stages, rollout correctness can be noisy, so we rely more on objective difficulty. As training progresses, rollout correctness becomes more informative of model-specific difficulty. We therefore combine the two weights with a cosine schedule. Let t be the current training step, T be the total number of steps, and $\tau = t/T$. We define

$$\alpha_s(t) = \frac{1 - \cos(\pi\tau)}{2}, \quad \alpha_o(t) = 1 - \alpha_s(t), \tag{6}$$

and compute the final difficulty weight as

$$w_d(q) = \alpha_o(t) w_o(q) + \alpha_s(t) w_s(q). \tag{7}$$

Finally, for each rollout of sample q, we reweight the GRPO advantage as

$$\tilde{A}_i = w_d(q) \hat{A}_i, \tag{8}$$

where $\hat{A}_i$ denotes the group-normalized advantage. The standard GRPO optimization objective is otherwise unchanged. This reweighting increases the contribution of small or under-mastered samples during policy updates while preserving the original relative ranking of rollouts within each group.

5 Experiments

5.1 Experimental Settings

Dataset We train SurgReasoner on SurgSeg-14K, using the data preparation strategy mentioned in Section 3.1. We hold out 3K samples as the test set, and use the remaining data for training. All quantitative results reported in this paper are obtained on the test set.

Implementation Details. We employ Qwen2.5-VL-7B and SAM2-Large as our default reasoning model and segmentation model, respectively.(more details in Appendix C.2).

Evaluation Metrics We compute three evaluation metrics: IoU, pDice, and Dice to assess model performance. IoU measures the bounding box localization accuracy predicted by MLLMs. pDice quantifies keypoint pair matching quality by calculating the spatial overlap of circles, where each circle takes the predicted keypoint pair as diameter endpoints. Dice assesses segmentation quality based on masks generated by downstream models conditioned on MLLM outputs.

5.2 Main Results

We compare SurgReasoner against a range of representative general-purpose, medical-specific, and grounding-oriented multimodal models under a unified evaluation setting, using the same fixed SAM2 segmentation backbone and prompt templates. We note that most baseline models are evaluated in an off-the-shelf prompting setting without additional training, whereas SurgReasoner is trained on the SurgSeg-14K training split. Therefore, this comparison serves two complementary purposes: to assess the difficulty of SurgSeg-14K for existing MLLMs, and to show the performance achieved by a model explicitly optimized for surgical reasoning segmentation.

As shown in Table 1, SurgReasoner achieves state-of-the-art results across all evaluation metrics, outperforming general-purpose models such as GPT-4o and Qwen2.5VL-72B[36], medical-domain models including HuatuoGPT-7B[37] and Lingshu-7B[38], as well as grounding-specific methods like Seg-Zero and VLMR1-REC-3B[39]. While general MLLMs like GPT-4o (IoU 5.43) and Gemini2.5-flash (IoU 10.56) exhibit only basic cross-modal understanding, they lack the fine-grained spatial precision required for surgical grounding. Medical-specific models such as HuatuoGPT-7B (IoU 4.83) and Lingshu-7B (IoU 15.78) show only modest improvements but still fail to match the demands of clinical reasoning and localization. Among grounding-specific models, SurgReasoner establishes a clear lead, surpassing Seg-Zero (IoU 31.04) by over 13 IoU points, clearly validating the effectiveness of our RL-driven reasoning strategy for accurate spatial prompting. This superiority holds across most surgical types, with particularly strong performance in nephrectomy, ophthalmic surgery, and gastrointestinal endoscopy. All models, including ours, still face persistent challenges in more complex scenarios such as myotomy and hysterectomy, highlighting remaining difficulties in highly occluded or low-contrast intraoperative environments.

Table 1: Performance comparison of different MLLMs on surgical image grounding and segmentation tasks. Best results are bolded, second-best are underlined. Abbreviations: Myo. (Myotomy), Nephr. (Nephrectomy), Chole. (Cholecystectomy), Prosta. (Prostatectomy), Hyster. (Hysterectomy), Ophth. (Ophthalmic Surgery), GI Endo. (Gastrointestinal Endoscopy).

Method	IoU %	pDice %	Dice %	Surgical Type (IoU %)						
				Myo.	Nephr.	Chole.	Prosta.	Hyster.	Ophth.	GI Endo.
General MLLMs										
GPT-4o	5.43	2.01	3.75	4.70	4.76	5.69	3.80	2.76	10.02	27.64
Gemini2.5-flash	10.56	4.95	5.43	11.04	7.15	8.68	4.11	10.28	21.89	22.69
Qwen2.5VL-7B	26.61	6.67	26.98	15.07	26.47	32.66	38.15	10.35	39.29	55.84
InternVL3-8B	5.87	1.49	5.11	4.13	6.08	13.65	4.82	1.95	4.08	21.23
Qwen2.5VL-72B	28.28	9.29	30.02	13.0	29.95	34.48	30.75	30.61	29.78	62.70
LLava-1.5-7B	12.88	2.73	16.65	11.20	13.43	10.03	15.91	10.21	16.33	16.50
Medical-Specific MLLMs										
MedR1-2B	2.43	2.14	5.80	1.12	2.45	1.85	7.74	0.96	0.37	6.35
MiniInternVL-4B	2.42	0.89	2.55	1.49	2.76	6.23	1.93	0.98	0.91	4.51
HuatuoGPT-7B	4.83	1.89	10.77	4.05	5.06	8.53	4.21	2.53	4.45	6.02
Lingshu-7B	15.78	3.47	18.06	11.87	13.72	16.47	14.83	14.40	24.93	15.42
Grounding-Specific MLLMs										
VLMR1-REC-3B	23.21	6.05	25.65	12.35	28.00	25.47	25.23	25.71	19.69	37.74
UniVG-R1	14.53	–	14.52	11.19	9.10	9.04	24.84	14.70	22.44	15.44
Seg-Zero	31.04	7.53	31.88	14.24	32.42	38.40	34.02	32.24	33.97	71.61
MedReasoner	34.23	8.12	34.95	16.85	38.76	40.12	36.89	35.47	41.23	73.85
SAM4MLLM-8B	13.60	–	12.45	14.02	9.78	12.62	17.52	16.94	13.29	12.49
SurgReasoner(Ours)	**44.22**	**11.66**	**44.78**	**20.52**	**53.55**	**45.03**	**40.45**	**41.23**	**57.26**	**76.16**

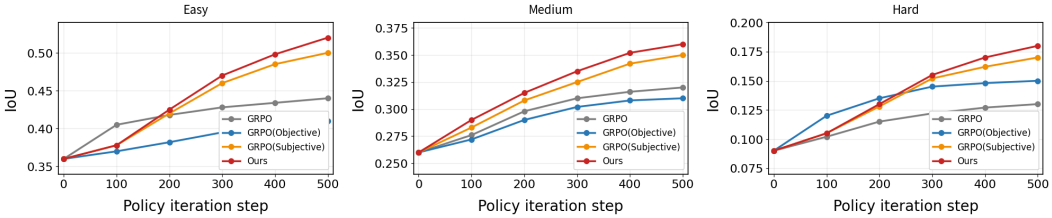


Figure 4: IoU as a function of policy iteration steps on three test subsets stratified by *objective* difficulty (easy, medium, and hard). We compare GRPO, GRPO (objective), GRPO (subjective), and our method.

5.3 Ablation on Training Strategies

To isolate the effect of our training strategy, we compare variants built on the same Qwen2.5-VL-7B reasoning backbone and the same frozen SAM2-Large segmentation module. As shown in Table 2, supervised fine-tuning brings only a modest improvement over the zero-shot Qwen2.5-VL baseline. GRPO substantially improves over supervised fine-tuning, showing the benefit of optimizing the reasoning module with spatial rewards. Objective difficulty alone improves Dice but does not consistently improve IoU, suggesting that static mask-size based weighting is insufficient by itself. Subjective difficulty yields stronger gains by capturing the model's current mastery of each sample. We further adopt a static fusion strategy with equal weighting (0.5:0.5) of objective and subjective difficulty, which achieves stable gains but is still limited by fixed weights. By contrast, our dynamic fusion achieves the best overall performance, indicating that mask-derived target properties and rollout correctness provide complementary signals for spatial prompt optimization.

5.4 Performance on Hard Targets

Motivated by the challenge of difficulty imbalance in surgical scenes, we further evaluate model performance across three test subsets stratified by objective difficulty: easy, medium, and hard. The specific stratification criteria for objective difficulty are provided in detail in Appendix A.2. As shown in Figure 4, our method consistently outperforms all GRPO variants (vanilla, objective-only,

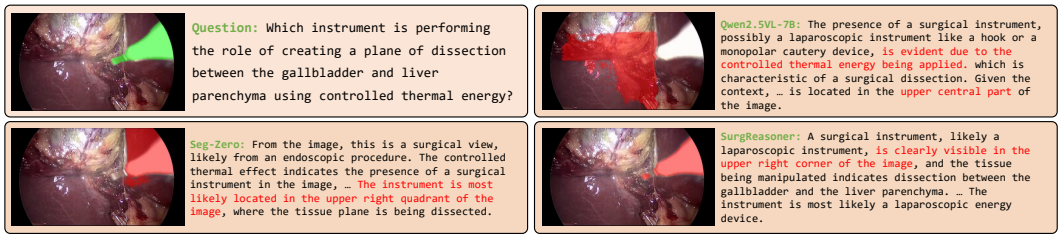


Figure 5: Qualitative comparisons between Qwen2.5-VL-7B, Seg-Zero, and SurgReasoner.

Table 2: Ablation of training strategies using the same Qwen2.5-VL-7B reasoning backbone and frozen SAM2 segmentation module.

Method	IoU %	pDice %	Dice %
Qwen2.5-VL-7B zero-shot	26.61	6.67	26.98
Qwen2.5-VL-7B / SFT	28.13	5.45	29.19
Qwen2.5-VL-7B / GRPO	40.83	9.85	41.07
Qwen2.5-VL-7B / GRPO + objective difficulty	39.88	9.34	42.00
Qwen2.5-VL-7B / GRPO + subjective difficulty	43.20	10.78	43.99
Qwen2.5-VL-7B / GRPO + *static* objective-subjective fusion	42.31	10.02	43.15
SurgReasoner / dynamic objective-subjective fusion	**44.22**	**11.66**	**44.78**

and subjective-only) across all difficulty levels, with both faster convergence and higher final IoU. Notably, the performance gap between our method and the baselines grows significantly on the hardest difficulty subset, where vanilla GRPO and its variants struggle to produce stable and precise spatial prompts. This result confirms that our dynamic difficulty-aware reweighting provides the strongest benefits for the most challenging surgical targets, validating our core motivation to address hard cases with explicit difficulty-aware optimization.

5.5 Qualitative Results

Figure 5 presents representative qualitative comparisons on a laparoscopic cholecystectomy example, where the query refers to the instrument creating a dissection plane using controlled thermal energy. Qwen2.5-VL-7B produces an imprecise spatial description and generates a mask that does not tightly localize the intended target. Seg-Zero correctly identifies the presence of the surgical instrument but outputs an overly expanded mask, lacking fine-grained precision and failing to distinguish the instrument from the surrounding tissue. In contrast, SurgReasoner generates a more accurate spatial prompt and produces a mask that is better aligned with the referred instrument. Additional qualitative examples and failure cases are provided for reference in Appendix C.4.

6 Conclusion

In this work, we systematically revisit the research paradigm of surgical visual perception for reasoning-segmentation from dataset construction, model design, and GRPO optimization. We first demonstrate that existing surgical datasets and models lack tight integration of semantic reasoning and precise pixel-level segmentation, and thus construct SurgSeg-14K for model training. We then analyze the limitations of conventional GRPO in surgical reasoning, especially the neglect of clinically critical hard samples, and propose a dynamic difficulty-aware advantage reweighting mechanism to explicitly alleviate this issue. Finally, we address GRPO's inherent biases in surgical grounding via the proposed reweighting and improved reward design. Trained on SurgSeg-14K, our SurgReasoner surpasses existing surgical perception models across all evaluation metrics, further strengthening the integration capability of semantic reasoning and pixel-level segmentation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62401143), the Natural Science Foundation of Jiangsu Province (BK20241301), and the Big Data Computing Center of Southeast University.

References

[1] Long Bai, Mobarakol Islam, and Hongliang Ren. Cat-vil: Co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 397–407. Springer, 2023.

[2] Hongqiu Wang, Yueming Jin, and Lei Zhu. Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation. *arXiv preprint arXiv:2306.02651*, 2023.

[3] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[4] Bhuvan Sachdeva, Naren Akash, Tajamul Ashraf, Simon Müller, Thomas Schultz, Maximilian WM Wintergerst, Niharika Singri, Kaushik Murali, and Mohit Jain. Phase-informed tool segmentation for manual small-incision cataract surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 446–455. Springer, 2025.

[5] Qiyuan Wang, Shang Zhao, Zikang Xu, and S Kevin Zhou. Lacoste: Exploiting stereo and temporal contexts for surgical instrument segmentation. *Medical Image Analysis*, 99:103387, 2025.

[6] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

[7] Liping Sun and Xiong Chen. Pixel-wise contrastive learning for multi-class instrument segmentation in endoscopic robotic surgery videos using dataset-wide sample queues. *IEEE Access*, 12:156867–156877, 2024.

[8] Markus Fox, Mario Taschwer, and Klaus Schoeffmann. Pixel-based tool segmentation in cataract surgery videos with mask r-cnn. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 565–568. IEEE, 2020.

[9] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7):1409–1417, 2024.

[10] Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*, 2023.

[11] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

[12] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.

[13] OpenAI. Gpt-4o (gpt-4 omni). Online, 2024. `https://openai.com/research/gpt-4o/`.

[14] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. arxiv 2022. *arXiv preprint arXiv:2210.03493*, 2022.

[15] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*, 2024.

[16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[17] Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, et al. Reverse-engineered reasoning for open-ended generation. *arXiv preprint arXiv:2509.06160*, 2025.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[20] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *arXiv preprint arXiv:2503.17352*, 2025.

[21] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[22] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[24] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[25] Lu Niu and Cheng Xue. Noise-aware few-shot learning through bi-directional multi-view prompt alignment. *arXiv preprint arXiv:2603.11617*, 2026.

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[27] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[28] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[29] Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11577–11585, 2026.

[30] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43. Springer, 2022.

[31] Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*, 2025.

[32] Jian Jiang, Chenxi Lin, Yiming Gu, Zengyi Qin, Zhitao Zeng, Kun Yuan, Yonghao Long, Xiang Xia, Cheng Yuan, Yuqi Wang, et al. Surg-r1: A hierarchical reasoning foundation model for scalable and interpretable surgical decision support with multi-center clinical validation. *arXiv preprint arXiv:2603.12430*, 2026.

[33] Yuxin He, An Li, and Cheng Xue. Cauclip: Bridging the sim-to-real gap in surgical video understanding via causality-inspired vision-language modeling. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6431–6435. IEEE, 2026.

[34] Cheng Xue, Shiyu Zhao, Danqiong Wang, Cheng Chen, Guanyu Yang, and Yang Chen. Td-sam: Temporal and distance-guided adaptations of sam for accurate surgical instrument segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[36] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[37] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. 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, 2024.

[38] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[39] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[40] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023.

[41] Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025.

[42] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[43] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[44] Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, page 103726, 2025.

[45] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. 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, 2022.

[46] N Ghamsarian et al. Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. sci. data 11 (1), 373 (2024).

[47] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas De Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *International Conference on Multimedia Modeling*, pages 218–229. Springer, 2021.

[48] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[49] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

[51] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):32, 2024.

[52] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025.

[53] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.

[54] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[55] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

A More Details on SurgSeg-14K Dataset

A.1 Dataset Composition

As shown in Table 3,, our SurgSeg-14K contains 7 surgical types with their corresponding publicly available datasets: Myotomy (DSAD [40]), Nephrectomy (EndoVis2017 [41],EndoVis2018[42]), Cholecystectomy (CholecSeg8k[43]), Prostatectomy (GraSP [44]), Hysterectomy (AutoLaparo[45]), Ophthalmic Surgery (Cataract-1K[46]), and Gastrointestinal Endoscopy (Kvasir-Instrument[47]). This selection covers a diverse range of surgical procedures, spanning abdominal, urological, ophthalmic, and gastrointestinal domains, thereby highlighting the breadth of anatomical regions and tasks addressed in our dataset. The details of each dataset and how we employ them are shown below.

AutoLaparo is collected from Prince of Wales Hospital in Hong Kong, which contains laparoscopic hysterectomy videos with manual annotations for surgical phase recognition, surgical instrument detection and spatial localization.

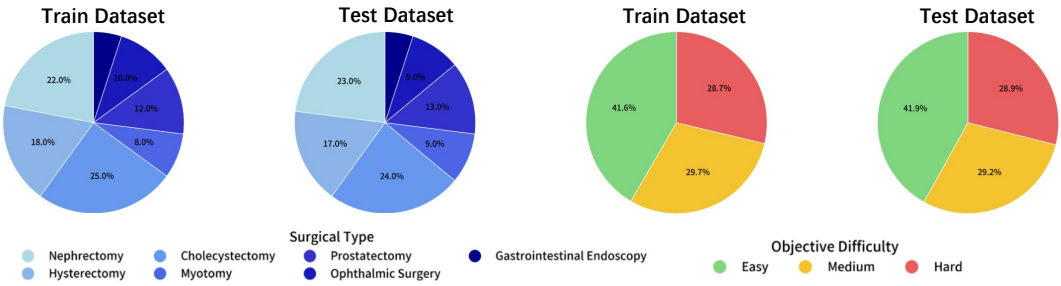


Figure 6: Distribution of SurgSeg-14K across surgical types and objective difficulty levels for the training and test splits.

Cataract-1K consists of a large collection of cataract surgical videos acquired from eye clinic scenarios, providing fine-grained surgical phase labels and pixel-level anatomical and instrument segmentation annotations for selected key frames.

DSAD (Dresden Surgical Anatomy Dataset) provides abundant laparoscopic abdominal images with high-quality semantic segmentation masks for multiple abdominal organs, abdominal wall and vascular structures, supporting anatomical structure recognition and segmentation tasks.

EndoVis2017 is recorded with the da Vinci robotic surgical system, which provides abdominal surgical sequences annotated with multiple types of robotic instruments and is suitable for surgical instrument recognition and localization research.

EndoVis2018 consists of 19 sequences of porcine training cases recorded using the da Vinci X or Xi surgical system. It provides annotations for 7 types of robotic surgical instruments as well as segmentation masks for anatomical objects and medical devices, supporting instrument recognition, tissue recognition, and instrument localization.

CholecSeg8K contains standard laparoscopic cholecystectomy videos captured from multiple surgeons. It provides hierarchical annotations of surgical phases and presence labels of common surgical instruments, serving as a fundamental benchmark for abdominal surgical vision tasks.

GraSP includes robot-assisted prostatectomy surgical videos, annotated with fine-grained surgical phases, operation steps, atomic actions and surgical instrument categories.

Kvasir-Instrument provides endoscopic gastrointestinal surgical frames with pixel-level instrument and tissue annotations, covering typical clinical scenarios of gastrointestinal endoscopy surgery.

A.2 Dataset Statistic

SurgSeg-14K contains over 14K image-mask-query pairs spanning seven distinct surgical scenarios: nephrectomy, hysterectomy, cholecystectomy, myotomy, prostatectomy, ophthalmic surgery, and gastrointestinal endoscopy. These samples cover 30 fine-grained categories of surgical instruments and anatomical structures, offering diverse intraoperative visual patterns for reasoning-driven segmentation. Each entry includes an implicit clinical query, a pixel-level binary target mask, and a structured reasoning trajectory, with bounding boxes and centroids automatically derived from the masks.

As shown in Figure 6, we provide both surgical-type and objective-difficulty distributions for the train and test splits. The sample difficulty is quantitatively divided into three levels based on the computed objective difficulty weight w_o, which ranges from 0.5 to 1.5. Specifically, samples with $w_o \in [0.5, 0.8]$ are defined as easy cases, $w_o \in (0.8, 1.2]$ as medium cases, and $w_o \in (1.2, 1.5]$ as hard cases. The dataset maintains balanced proportions across surgical procedures, and similarly balanced ratios of easy, medium, and hard samples in both splits. Hard cases typically involve small instruments, partially occluded organs, or instruments hidden behind tissue, and account for approximately 29% of the test set. The resulting benchmark supports comprehensive evaluation of language-guided surgical segmentation under varying clinical and visual conditions.

Table 3: Overview of surgical types and corresponding datasets in SurgSeg-14K

Types	Datasets
Hysterectomy	AutoLaparo [45]
Ophthalmic Surgery	Cataract-1K [46]
Cholecystectomy	CholecSeg8k[43]
Myotomy	DSAD [40]
Nephrectomy	EndoVis2017 [41], EndoVis2018 [42]
Prostatectomy	GraSP [44]
Gastrointestinal Endoscopy	Kvasir-Instrument[47]

A.3 Reasoning Trajectory Generation and Quality Control

We treat the reasoning trajectory as a structured localization rationale rather than authoritative clinical narration, and apply a three-stage quality control pipeline.

Constrained generation. Prompts restrict trajectories to observable visual cues (appearance, spatial relations, tissue contact) and forbid unverifiable functional or procedural claims.

Mask-consistency checks. We automatically verify that the bounding box and centroid referenced in each trajectory match the geometry derived from the ground-truth mask; mismatches beyond a tolerance are rejected and regenerated.

Sampled human review. A random 50% subset is reviewed by clinical experts for clinical logic, target uniqueness, and rationale–mask consistency; ambiguous cases are revised or removed.

Common failure modes. We observe three recurring failure modes, targeted by the above checks: (i) **hallucinated functional states**, e.g., describing an idle instrument as "cutting tissue"; (ii) **confused visually similar instruments** under heavy occlusion; and (iii) **over-specific anatomical assertions** beyond the visible field.

A.4 Illustration of Dataset Visual Samples

Figures 8–Figures 11 illustrate representative visual samples and annotation formats selected from our SurgSeg-14K dataset, covering several typical surgical scenarios. For each illustrated case, we present the raw intraoperative view, pixel-level segmentation mask, implicit clinical reasoning query, and corresponding reasoning trajectory.

B More Details on SurgReasoner Framework

B.1 Group Relative Policy Optimization

In this work, we adopt Group Relative Policy Optimization (GRPO) [48] as the core reinforcement learning algorithm. GRPO optimizes the policy π_θ by sampling a set of candidate responses $\{\mathbf{o}_i\}_{i=1}^{N}$ from the old policy $\pi_{\theta_{\text{old}}}$ for each input query $\mathbf{q}$, where each response $\mathbf{o}_i$ has length $|\mathbf{o}_i|$. The policy is updated based on a normalized advantage $\hat{A}_{i,t}$ computed for each token, forming the objective $\mathcal{J}_{\text{GRPO}}(\pi_\theta)$:

$$\mathcal{J}_{\text{GRPO}}(\pi_\theta) = \mathbb{E}_{\mathbf{q}\sim p_Q,\{\mathbf{o}_i\}_{i=1}^{N}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{q})}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbf{o}_i|}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}\mid \mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid \mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t},\ \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid \mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid \mathbf{q},\mathbf{o}_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right]\right\}\right], \tag{9}$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}$ denotes the token-level importance ratio, and $\hat{A}_{i,t}$ is the normalized group advantage derived from either of two supervision strategies:

Outcome Supervision. The reward model assigns a single scalar reward r_i to each output sequence o_i. The normalized group advantage is defined as:

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \mu_r}{\sigma_r}, \quad \forall t,$$

where μ_r and σ_r represent the mean and standard deviation of rewards $\{r_1, \ldots, r_G\}$. **Process Supervision.** Step-wise rewards $r_i^{(j)}$ are assigned to each intermediate reasoning step. The normalized advantage is accumulated as:

$$\hat{A}_{i,t} = \sum_{j:\text{index}(j)\geq t} \tilde{r}_i^{(j)}, \quad \text{where } \tilde{r}_i^{(j)} = \frac{r_i^{(j)} - \mu_R}{\sigma_R}.$$

GRPO is highly compatible with reward models trained from pairwise preference data, as it naturally exploits relative reward comparisons. Moreover, it reduces training overhead substantially by removing the need for an independent value network.

B.2 Reward Functions

Reasoning Formats Reward. We define two sub-reward components, namely R_{think} and $R_{\text{seg-format}}$. Specifically, R_{think} is assigned a fixed score of 1 once the model generates a complete and well-structured reasoning block. As for R_{answer}, it consists of two equally weighted parts: a score of 0.5 is awarded if the bounding box field conforms to the specification, and another 0.5 is given when the key point coordinate field satisfies the format constraints, where both fields must comply with the predefined JSON structure. Notably, this reward only supervises the format integrity and structural legitimacy of model outputs, without evaluating the factual correctness or logical quality of the reasoning content.

Grounding Box Reward. This reward evaluates the predicted box $B_p = [x_1^p, y_1^p, x_2^p, y_2^p]$ against the ground-truth box $B_g = [x_1^g, y_1^g, x_2^g, y_2^g]$, where all coordinates are normalized to the range $[0, 1]$. Specifically, the IoU reward is utilized to measure the spatial overlap between the predicted box and the ground-truth box, which is calculated as:

$$r_{\text{iou}} = \frac{\text{Area}(B_p \cap B_g)}{\text{Area}(B_p \cup B_g)}. \tag{10}$$

Second, the Edge reward quantifies the alignment of four bounding box edges (left, right, top, bottom), with errors normalized by the diagonal length of B_g:

$$r_{\text{edge}} = 1 - \frac{1}{4}\sum_{k=1}^{4} |B_p^{(k)} - B_g^{(k)}|, \tag{11}$$

where $B^{(1)}, B^{(2)}, B^{(3)}, B^{(4)}$ denote the left, right, top, and bottom edges of the box, respectively.

Grounding Points Reward. This reward evaluates the quality of the predicted key point pair $P_p = \{p_1^p, p_2^p\}$ by comparing it with the ground-truth pair $P_g = \{p_1^g, p_2^g\}$, where each point $p = (x, y)$ is normalized to $[0, 1]$. First, the pDice reward estimates the spatial overlap between circles formed by each point pair. Each uses the two points as diameter endpoints, and the Dice score is computed as:

$$r_{\text{pdice}} = \frac{2 \cdot \text{Area}(O_p \cap O_g)}{\text{Area}(O_p) + \text{Area}(O_g)}, \tag{12}$$

where O_p and O_g are the circles constructed from P_p and P_g, respectively. Second, the L1 distance reward computes the between corresponding points:

$$r_{\text{distance}} = 1 - \frac{1}{2}\sum_{i=1}^{2} \left\| p_p^{(i)} - p_g^{(i)} \right\|_1. \tag{13}$$

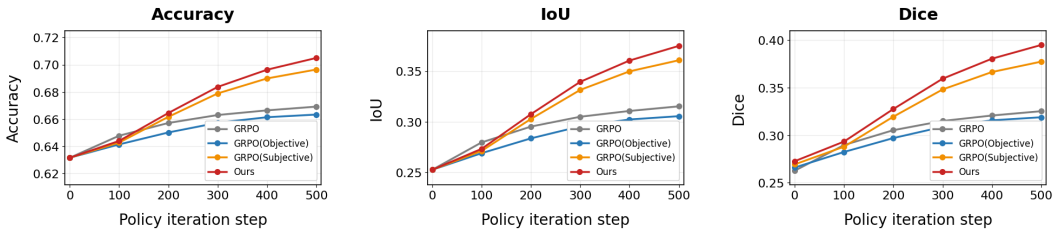


Figure 7: Changes in accuracy, IoU, and Dice over training steps for four configurations: GRPO, GRPO(objective), GRPO(subjective) and ours.

C More Details on Experiments

C.1 Models

We conduct a comprehensive comparison across a wide range of models. For general MLLMs, we utilized GPT-4o, Gemini-2.5-flash[49], Qwen2.5VL-7B/72B[36], LLava-7B[18] and InternVL3-8B[24]. For medical-specific MLLMs, we selected MedR1-2B[50], MiniInternVL-4B[51], HuatuoGPT-7B, Lingshu-7B, and UniVG-R1[52]. For grounding-specific models, we included SAM4MLLM[12], VLMR1-REC-3B[39] and SegZero-7B and MedReasoner.

C.2 Implementation Details

We employ Qwen2.5-VL-7B and SAM2-Large as our default reasoning model and segmentation model, respectively. SurgReasoner is trained on an 8 NVIDIA H100-80G GPUs with the veRL . During training, we use a total batch size of 16 with a sampling number of 8 per training step. The initial learning rate is set to 1e-6 and the weight decay is 0.01.

C.3 Step-wise Performance Analysis

To further investigate the effectiveness of our improved GRPO framework, we perform a step-wise analysis of model performance throughout reinforcement learning. As shown in Figure 7, we compare the evolution of accuracy, IoU, and Dice scores under four configurations: the vanilla GRPO baseline, GRPO weighted solely by objective difficulty (GRPO (objective)), GRPO weighted solely by subjective difficulty (GRPO (subjective)), and our full method with dynamic difficulty weighting. Their performance is identical at the beginning of training. As training progresses, clear divergence emerges across variants. The vanilla GRPO baseline exhibits the slowest improvement, while the two single-difficulty weighting variants achieve only moderate gains. In contrast, our full method consistently outperforms all baselines across all metrics, attaining the highest accuracy, IoU, and Dice scores at every policy iteration step. This persistent performance advantage demonstrates that our improved weighting strategy provides more effective learning guidance, accelerating convergence and leading to superior segmentation performance.

C.4 Ablation on Segmentation Backbones

C.5 Additional Qualitative Results

We provide additional qualitative visualization results for various surgical scenarios in Figure 12, and conduct qualitative comparisons of reasoning trajectories, spatial prompts, and segmentation masks produced by SurgReasoner and baseline methods.

C.6 Cross-Dataset Generalization

To assess whether SurgReasoner learns transferable spatial reasoning rather than memorizing training patterns, we conduct an out-of-distribution (OOD) evaluation. We train on the eight public source datasets and evaluate on our private clinical subset, which contains surgical scenarios and instrument distributions absent from training.

To evaluate the impact of different segmentation backbones on surgical reasoning segmentation, we conduct a controlled ablation study. We keep the reasoning outputs fixed and replace the prompt-based segmentation backbone with four representative medical segmentation models: SurgSAM2[53], SAM-Med2D [54], MedSAM2 [55], and SAM2-Large. As shown in Table 4, combining points and bounding boxes yields better results than using either prompt type alone across all backbones. SAM2-Large achieves the best performance and is therefore adopted as our default segmentation module.

Table 4: Dice scores for different segmentation backbones under four prompt types.

Model	Dice %		
	Pts	BBox	P+B
SurgSAM2	35.6	38.6	40.9
SAM-Med2D	37.3	41.2	42.0
MedSAM2	38.8	43.6	43.9
SAM2-Large	**39.5**	**44.3**	**44.7**

Table 5: Cross-dataset OOD evaluation: models are trained on the eight public datasets and tested on the held-out private clinical subset.

Method	IoU %	pDice %	Dice %
SFT	25.03	5.52	27.17
GRPO	32.21	7.23	34.12
SurgReasoner (Ours)	**36.57**	**9.76**	**37.89**

As shown in Table 5, the SFT model generalizes poorly (IoU 25.03), while RL-optimized variants remain substantially stronger. SurgReasoner achieves 36.57 IoU, outperforming both SFT and vanilla GRPO, confirming that the difficulty-aware RL objective learns transferable spatial reasoning rather than fitting public-scenario idiosyncrasies.

C.7 Matched-Training Comparison

To validate the effectiveness of our training strategy, we fine-tune representative baselines on the same SurgSeg-14K training split under identical protocols, covering three categories: SFT-tuned MLLM backbones (Qwen2.5-VL-7B, InternVL3-8B, LLaVA-1.5-7B with frozen SAM2), the referring segmentation method LISA-7B, and the reasoning-aware method MedReasoner.

As shown in Table 6, SurgReasoner outperforms all baselines. With the same backbone and RL recipe, RL improves over SFT (40.83 vs. 28.13 IoU), and our dynamic reweighting further raises IoU to 44.22, confirming that the gains come from the proposed training strategy rather than data or protocol differences.

D Limitations

Potential limitations of our method should be properly noted. The framework relies on a frozen SAM2 segmentation backbone, and still suffers from localization errors on extremely small instruments and heavily occluded targets. In addition, its effectiveness is validated on seven typical surgical scenarios, which may limit its generalization to other surgical procedures with distinct imaging conditions.

E Broader Impact

E.1 Societal Impact

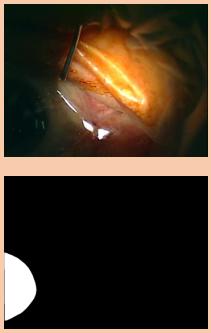
Our proposed SurgReasoner framework for surgical scene reasoning and segmentation brings tangible positive values to intelligent endoscopic surgery and medical image analysis research. The method provides a reliable technical paradigm for fine-grained perception and understanding of complex surgical scenarios, which can promote the development of vision-based intraoperative auxiliary diagnosis systems.The model is capable of accurate reasoning segmentation of surgical tissues and instruments, helping clinicians quickly locate key anatomical structures and surgical targets, and effectively assisting intraoperative decision-making and surgical workflow analysis.

Table 6: Matched-training comparison on the SurgSeg-14K test set. All methods are trained on the same training split and evaluated with identical protocols.

Method	Training	IoU %	pDice %	Dice %
Qwen2.5-VL-7B+SAM2	SFT	28.13	5.45	29.19
LLaVA-1.5-7B+SAM2	SFT	19.21	4.76	21.03
InternVL3-8B+SAM2	SFT	12.12	3.45	13.64
Qwen2.5-VL-7B+SAM2	RL	40.83	9.85	41.07
LISA-7B-LLaVA1.5	SFT	22.74	5.21	24.56
MedReasoner	SFT+RL	41.42	9.61	41.73
SurgReasoner (Ours)	RL	**44.22**	**11.66**	**44.78**

E.2 Ethics Statement

All surgical-related images used in this study come from ethically approved public medical datasets with patients' informed consent. All data are fully anonymized to remove personally identifiable information and protect patient privacy. Our SurgReasoner model and findings are only for academic and non-commercial research in medical computer vision. Unauthorized commercial use and unvalidated clinical adaptation are strictly prohibited. Researchers citing this work shall abide by academic and medical research ethics.

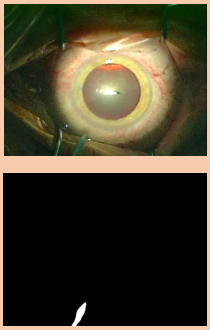


Question: Which dark circular opening, located in the lower-central portion of the exposed anterior eye and surrounded by the textured iris, is the aperture that regulates light entry into the eye?

Think: In this ophthalmic surgical view, the dark circular opening located in the lower-central area of the exposed anterior segment is clearly visible … dark opening and the iris, along with its location at the anatomical center of the iris, confirms that this structure is the pupil, the aperture responsible for regulating light entry into the eye.

Surgical Type: Ophthalmic Surgery **Category:** pupil

Answer: Bbox: [1,383,162,726] **Points:** [[75,431],[26,646]]



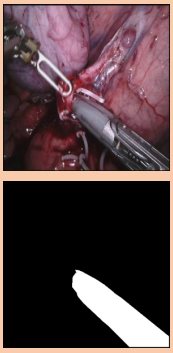
Question: Which instrument is likely being used to initiate a controlled corneal entry point during this anterior segment procedure?

Think: This is a typical anterior segment view of cataract surgery, with the eye stabilized by a … where the primary surgical entry would begin. This role — creating the initial corneal access point to enter the anterior chamber — is consistent with the function of a SlitKnife. Therefore, the instrument at the inferior limbus is the SlitKnife.

Surgical Type: Ophthalmic Surgery **Category:** SlitKnife

Answer: Bbox: [343,648,417,767] **Points:** [[370,724],[403,692]]

Figure 8: Example of Ophthalmic Surgery from our SurgSeg-14K dataset.

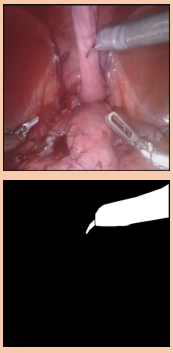


Question: Which structure has its jaws partially open and appears to be in the process of releasing after completing a sealing maneuver?

Think: Examine the state of all instruments. One device's jaws are slightly parted but still in close contact with tissue, suggesting recent release … other instrument in the lower right has closed jaws and is not engaged with vessels. The transitional state partially open, post-sealing — is specific to the Vessel Sealer's operational cycle.

Surgical Type: Nephrectomy **Category:** Vessel Sealer

Answer: Bbox: [212,272,508,508] **Points:** [[283,315],[265,300]]



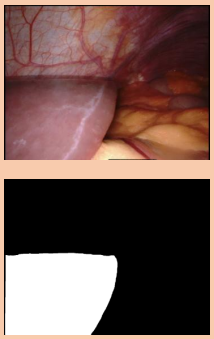
Question: Which device is currently positioned at the central tubular structure, with its curved metallic tip exposed and ready for activation to perform dissection or cutting?

Think: All instruments in the field belong to a robotic surgical system. The lower instrument is engaged with tissue in a working state, while the left instrument is also positioned … perform dissection or cutting. The exposed metallic blade confirms it is not a passive tool. This instrument corresponds to the monopolar curved scissors.

Surgical Type: Nephrectomy **Category:** Monopolar Curved Scissors

Answer: Bbox: [254,12,507,165] **Points:** [[473,115],[490,53]]

Figure 9: Example of Nephrectomy from our SurgSeg-14K dataset.

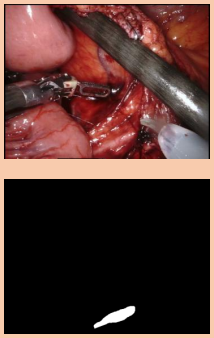


Question: What structure must be gently displaced or avoided during this step to prevent obscuring the surgical target, given its bulk and fixed position in the upper abdomen?

Think: The procedure requires clear visualization of the esophageal-gastric junction, which lies posterior and inferior … fixed anatomical attachment via the falciform ligament and its substantial mass make it a critical obstacle to overcome. Recognizing its role as a necessary retraction target during exposure confirms this is the liver.

Surgical Type: Myotomy **Category:** liver

Answer: Bbox: [8,320,1269,966] **Points:** [[141,751], [788,713]]



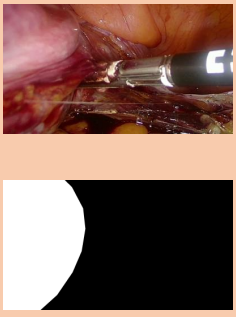
Question: Which structure, running along the posterior abdominal wall and appearing as a smooth, cord-like tube, connects the kidney to the bladder and transports urine?

Think: In this surgical view, we first note that the described structure is the muscular tube connecting the kidney and bladder to transport urine, characterized by a retroperitoneal course along the posterior … cord-like structure in the lower-central right portion of the view, which corresponds to the target ureter.

Surgical Type: Myotomy **Category:** ureter

Answer: Bbox: [418,590,610,694] **Points:** [[565,618],[522,635]]

Figure 10: Example of Myotomy from our SurgSeg-14K dataset.

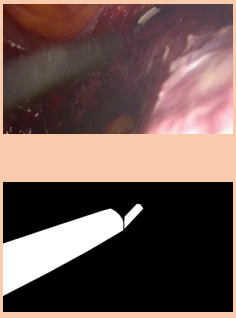


Question: Which structure is currently being mobilized to allow access to the cardinal ligaments and uterine vessels during this step of the procedure?

Think: In a hysterectomy, the surgeon must first free the uterus from surrounding attachments to expose … connective tissue strands indicates active mobilization of supporting ligaments. Since the goal of this maneuver is to isolate the uterine vasculature and ligamentous supports, the structure being freed must be the uterus.

Surgical Type: Hysterectomy **Category:** uterus

Answer: Bbox: [0, 0, 284, 793] **Points:** [[123,228], [225,503]]



Question: Which energized instrument is being used to dissect the vesicouterine fold from the underlying uterine structures using thermal energy during hysterectomy?

Think: The image shows a laparoscopic view of a hysterectomy, with dissection occurring in the plane between the uterus and bladder, consistent with opening the vesicouterine fold. On the left side of the … fold using thermal energy while avoiding excessive bleeding or injury to adjacent structures such as the bladder.

Surgical Type: Hysterectomy **Category:** Electric hook

Answer: Bbox: [0,127,827,668] **Points:** [[649,301], [676,288]]

Figure 11: Example of Hysterectomy from our SurgSeg-14K dataset.

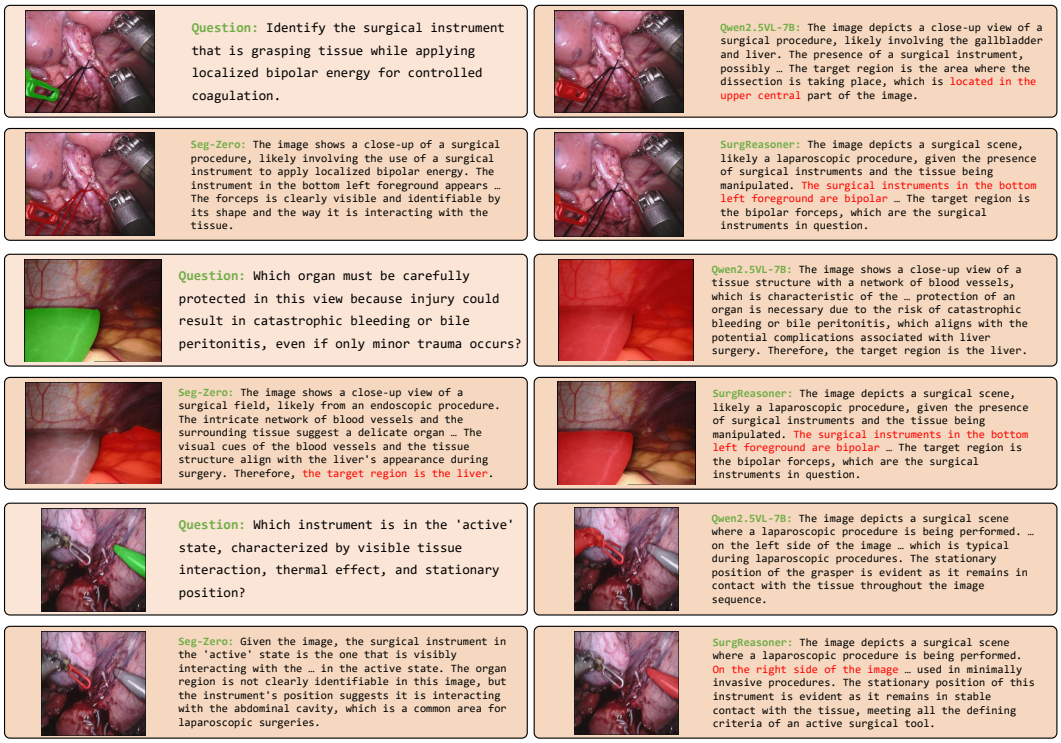


Figure 12: Qualitative comparisons between SurgReasoner and baseline methods.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's core contributions, technical innovations and research scope, which are consistent with experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper analyzes the limitations and generalization boundary of the proposed method in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include the oretical results

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The complete details of the implementation can be found in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: Our datasets and code will be publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All key experimental details including data split, optimizer and hyperparameters are clearly presented in the paper and appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Detailed training/test splits, hyperparameters, prompt settings, and optimizer configurations are reported in paper and the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides the type of compute workers(GPU) in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All surgical image data used are fully anonymized and comply with the NeurIPS Code of Ethics without privacy risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are cited and comply with the original licensing terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code introduced in the paper is well documented

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study uses de-identified surgical image data collected under institutional ethical approval and/or patient informed consent. No participant interaction or intervention was conducted for this research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.