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

SIRA: Reasoning-Aware Surgical Instrument Segmentation via Query-Anchored Alignment

Zhibo Zhang[1], Qijie Wang[1], and Zengqiang Yan[1(✉)]

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology
{zzb226,wangqijie,z_yan}@hust.edu.cn

Abstract. Surgical instrument segmentation (SIS) plays a critical role in robotic assistance and surgical workflow analysis. However, most existing SIS methods formulate segmentation as a category-driven localization problem, limiting their ability to capture procedural context and task-dependent semantics in surgical workflows. We introduce **Reasoning-Aware Surgical Instrument Segmentation (RA-SIS)**, a task formulation that frames segmentation as query-conditioned inference under surgical context. To benchmark this setting, we construct **SurgRS**, a surgical reasoning segmentation dataset consisting of 41,000 image–text pairs, which aligns instance-level masks with structured query–answer supervision to enable semantic grounding at the pixel level. Based on SurgRS, we propose **Surgical Instrument Reasoning and Segmentation Assistant (SIRA)**, a multimodal framework that disentangles target-level and query-level semantics and integrates them with visual features through query-anchored dual alignment. By aligning query semantics with spatial features and segmentation prompts, SIRA enhances semantic-visual consistency in mask prediction. Extensive experiments on SurgRS demonstrate improvements over existing reasoning-aware baselines. Code is available at https://github.com/linxir226/SIRA.

Keywords: Surgical instrument segmentation · Reasoning-aware segmentation · Multimodal alignment.

1 Introduction

Surgical Instrument Segmentation (SIS) aims to localize surgical instruments in intraoperative scenes, enabling applications such as navigation, robotic assistance, workflow analysis, and skill assessment [1]. Most existing SIS methods formulate segmentation as a category-driven localization problem, relying primarily on visual features. Early approaches adopt CNNs [2, 3], transformers [4], or fine-tuned large segmentation models such as Mask2Former [5] and SAM [1, 6, 7]. Text-prompted extensions, including SP-SAM [8] and TP-SIS [9], introduce language guidance but still treat textual input as auxiliary conditioning, with segmentation largely dominated by appearance-based matching. However, they do not explicitly ground segmentation to procedural roles or task-dependent contextual cues in surgical workflows.

Recently, reasoning-aware segmentation has emerged as a paradigm for handling context-dependent and semantically complex queries [10]. Instead of predicting masks solely from predefined labels, models interpret structured textual instructions to infer pixel-level targets. Representative methods include LISA [11], GSVA [12], PathMR [13], and GLaMM [14]. VRS-HQ [15] pioneered the use of SAM2 [16] as the segmentation model, establishing itself as a robust benchmark. However, these approaches are primarily developed for natural scenes or relatively simple medical scenarios.

In real surgical workflows, instrument localization is inherently task-driven. Surgeons identify instruments by their functional role in a procedural stage, such as enabling suturing or maintaining exposure, rather than by category names [17]. Accurate segmentation therefore requires interpreting procedural intent and contextual relationships before pixel-level prediction [18]. This motivates a task-driven segmentation formulation.

Motivated by these observations, we introduce **Reasoning-Aware Surgical Instrument Segmentation (RA-SIS)**, which, to the best of our knowledge, is the first formulation that explicitly models surgical instrument segmentation under task-driven and context-dependent queries instead of explicit category prompts. In RA-SIS, segmentation is framed as a query-conditioned inference process, where models must first infer procedural intent and contextual semantics before producing pixel-level masks.

To benchmark RA-SIS, we construct **SurgRS**, a surgical reasoning segmentation dataset consisting of 41,000 image–text pairs that explicitly aligns refined instance-level masks with structured query–answer supervision, enabling query-conditioned pixel-level inference.

Based on SurgRS, we introduce **Surgical Instrument Reasoning and Segmentation Assistant (SIRA)**, a multimodal framework that disentangles target-level and query-level semantics and integrates them with visual features through query-anchored dual alignment. Extensive experiments on SurgRS demonstrate state-of-the-art performance.

2 Method

2.1 Dataset Construction

Surgical Instrument Segmentation Dataset. Early benchmarks such as EndoVis-17 [19] and EndoVis-18 [20] provide pixel-level masks for surgical instruments, primarily supporting object localization tasks. Subsequent datasets, including SAR-RARP50 [21], PhaKIR [22], and SurgVU [23], introduce additional supervisory signals such as action labels and phase annotations to enrich surgical context. However, these contextual cues are not explicitly grounded to instance-level segmentation masks and therefore cannot directly support RA-SIS, which requires mapping query semantics to precise pixel-level targets. To support RA-SIS, a dedicated dataset should satisfy the following requirements: (i) large-scale surgical images covering realistic procedural variations; (ii) instance-level segmentation masks distinguishing multiple instruments and components;

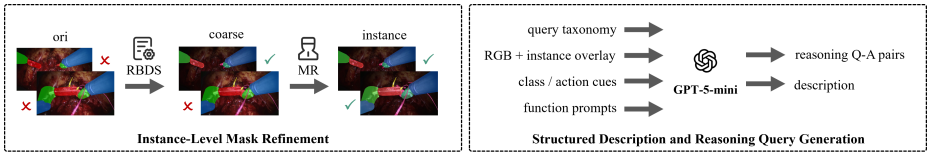

Fig. 1. Overall construction pipeline of SurgRS.

(iii) expert-driven, task-oriented query-answer pairs reflecting surgical reasoning; and (iv) explicit alignment between textual supervision and corresponding segmentation targets. Motivated by these principles, we construct the **Surgical Reasoning Segmentation Dataset (SurgRS)**, a benchmark specifically designed for task-driven reasoning-aware surgical instrument segmentation.

SurgRS Construction Pipeline. SurgRS is constructed based on the SAR-RARP50 dataset [21], which provides action labels offering task-driven procedural context conducive to reasoning-aware segmentation, as well as diverse and richly annotated surgical frames that support large-scale instance refinement. The overall construction pipeline of SurgRS is illustrated in Fig. 1 and consists of two major stages.

(1) Instance-Level Mask Refinement. Following the preprocessing protocol of SAR-RARP50, we construct SurgRS from its RGB frames, semantic masks, and action annotations, and extract high-quality mask-annotated frames as the foundation of SurgRS. We remove 201 frames without valid targets and convert the remaining semantic masks into instance-level masks. Since semantic labels in SAR-RARP50 often merge multiple instrument instances, we refine them into instance-level masks.

Specifically, we design rule-based decomposition strategies (RBDS) to split separable semantic masks into coarse instance masks, particularly for instruments with clear spatial separation and stable positional cues. RBDS performs connected-component analysis for each semantic tool region: large components are kept as instance candidates, while small regions caused by occlusion or thin structures are merged into the semantically closest component. For multi-instance tools, roles are assigned according to relative position and temporal continuity. Since RBDS only produces coarse masks, we then perform manual refinement (MR) in Supervisely to correct merged regions, resolve boundary ambiguities, and eliminate labeling errors. All refined masks are further verified to ensure structural correctness and annotation consistency.

(2) Structured Description and Reasoning Query Generation. As shown in Table 1, we define nine representative reasoning query types arising in robotic-assisted prostate suturing, with an additional *Other* (OT) category for rare but clinically valuable cases. This taxonomy constrains query design to clinically meaningful reasoning skills with strong intraoperative relevance.

Table 1. SurgRS reasoning query taxonomy.

Code	Type	Definition
LT	Leading Tool	Identify the instrument executing the maneuver.
SE	Support Exposure	Analyze exposure or counter-traction roles.
SGT	Suture Guide/Tension	Identify instruments guiding or tensioning sutures.
PAI	Procedural Action Inference	Infer the current action from visible instruments.
NSI	Next-Step Importance	Identify instruments important for the next action.
NDL	Near/Deep Layout	Analyze spatial distribution in surgical field.
OR	Occlusion Relation	Identify occlusion or overlap relationships.
PL	Position Locate	Localize instruments using positional cues.
CR	Commonsense Reference	Localize instruments via commonsense functional descriptions.
OT	Other	Cover rare clinically meaningful cases.

To enhance clinical relevance and contextual accuracy, for each frame we use the vision-capable GPT-5-mini conditioned on the RGB image, labeled instance overlay, visible instrument categories, action annotation, and commonsense prompts about instrument functions. The model jointly generates a clinically natural frame description and 1-3 reasoning query-answer pairs according to the interaction complexity, consistent with prior reasoning segmentation frameworks [13, 15]. To enforce explicit reasoning-mask alignment, each answer must reference the corresponding segmentation targets using exact class names and attach a `<SEG>` marker. All query-answer pairs are manually checked with the RGB image and labeled overlay to ensure that the query is clinically reasonable and the target answer is correct.

2.2 SIRA Architecture

Overall Architecture. As illustrated in Fig. 2, SIRA consists of a multimodal large language model (MLLM), a query-anchored dual alignment module, and a SAM2-based segmentation model [16].

The MLLM adopts Chat-UniVi [24] with LoRA [25] adapters for surgical domain adaptation. Given a textual query and an input image, it produces (i) target-level embeddings from `<SEG>` tokens, (ii) query-level embeddings from `<QUERY>` tokens, and (iii) explanatory text. The segmentation model follows SAM2, comprising a vision backbone and a mask decoder.

Since the MLLM and SAM2 operate in different representation spaces, and `<SEG>` prompts alone cannot encode full reasoning semantics, we introduce a query-anchored dual alignment module to bridge them. Query-level embeddings are projected into the visual feature space and injected into spatial features via cross-attention to produce query-conditioned representations. Target-level embeddings are projected into the SAM2 prompt space and further aligned with query semantics to enhance consistency between target activation and global reasoning context. The mask decoder then integrates aligned visual features and prompt embeddings to generate one or multiple masks under task-driven specifications. The entire framework is trained end-to-end with segmentation and text supervision.

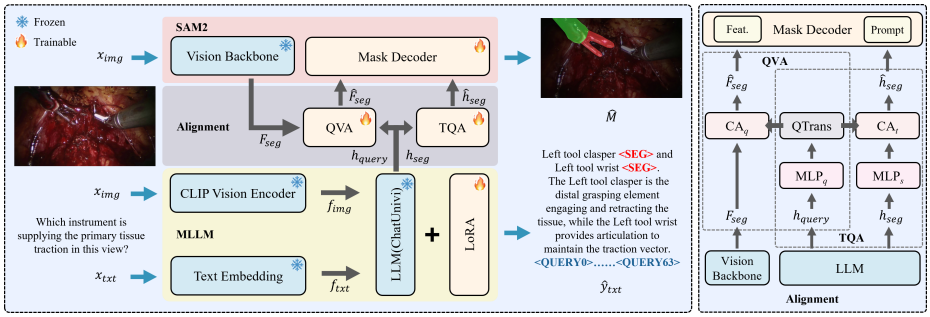

Fig. 2. Overview of the proposed SIRA framework.

Disentangled Semantic Encoding. Existing reasoning-aware segmentation methods typically rely on a single dedicated `<SEG>` token to encode target-specific semantics from the MLLM, which entangles object activation and query-level reasoning within a unified representation. Such coupled encoding is insufficient for capturing the rich global context embedded in complex reasoning queries. To explicitly disentangle target-level activation from query-level semantics, we introduce a structured semantic encoding scheme that separates instance-oriented signals and coarse reasoning context in the latent space.

In addition to conventional `<SEG>` tokens indicating candidate segmentation targets, we incorporate a fixed-length set of N_q learnable `<QUERY>` tokens to model global reasoning semantics independent of explicit instance activation.

Given a tokenized textual expression $x_{txt} = \{w_1, \ldots, w_L\}$ and an image input x_{img}, the MLLM processes the multimodal input and autoregressively generates a response sequence $\hat{y}_{txt} = \{\hat{y}_1, \ldots, \hat{y}_N\}$. The sequence $\hat{y}_{txt}$ includes explanatory tokens, multiple `<SEG>` tokens corresponding to potential targets, and a contiguous block of N_q `<QUERY>` tokens representing query-level reasoning semantics. Let $\mathbf{H} \in \mathbb{R}^{N \times d}$ denote the hidden states of the final MLLM layer, where d is the embedding dimension. We extract the representations associated with the `<SEG>` and `<QUERY>` tokens as

$$\boldsymbol{h}_{seg} \in \mathbb{R}^{K \times d}, \quad \boldsymbol{h}_{query} \in \mathbb{R}^{N_q \times d}, \tag{1}$$

where K denotes the number of `<SEG>` tokens. Here, $\boldsymbol{h}_{seg}$ encodes instance-specific activation cues, while $\boldsymbol{h}_{query}$ captures disentangled coarse-grained reasoning context derived from x_{txt} and x_{img}. This structured semantic decomposition establishes complementary instance-level and query-level representations for subsequent alignment.

Query-Anchored Dual Alignment. Although $\boldsymbol{h}_{seg}$ provides instance-level activation cues, it does not fully preserve the coarse query semantics encoded in $\boldsymbol{h}_{query}$, which are essential for precise localization. As illustrated in Fig. 2, we introduce a query-anchored dual alignment mechanism to propagate disentangled query representations into spatial features and prompt embeddings.

(1) Query-to-Visual Alignment (QVA). Let $\boldsymbol{h}_{query} \in \mathbb{R}^{N_q \times d_m}$ denote the query-level embeddings extracted from the MLLM. Here, N_q is the number of `<QUERY>` tokens and d_m is the embedding dimension. Let $\boldsymbol{F}_{seg} \in \mathbb{R}^{HW \times d_v}$ denote the fine-grained spatial features extracted by the SAM2 backbone. We first project $\boldsymbol{h}_{query}$ into the vision feature space:

$$\boldsymbol{E}_{query} = \mathrm{QTrans}\big(\mathrm{MLP}_q(\boldsymbol{h}_{query})\big), \tag{2}$$

where $\boldsymbol{E}_{query} \in \mathbb{R}^{N_q \times d_v}$ encodes query semantics in the visual feature dimension, and $\mathrm{QTrans}(\cdot)$ denotes a lightweight transformer block for modeling inter-query dependencies. We then align query semantics with spatial representations via cross-attention:

$$\hat{\boldsymbol{F}}_{seg} = \mathrm{CrossAttention}_q(\boldsymbol{F}^Q_{seg}, \boldsymbol{E}^K_{query}, \boldsymbol{E}^V_{query}), \tag{3}$$

where spatial features act as queries, and query embeddings serve as keys and values. The resulting $\hat{\boldsymbol{F}}_{seg} \in \mathbb{R}^{HW \times d_v}$ represents query-conditioned visual features.

(2) Target-to-Query Alignment (TQA). The instance-level embeddings $\boldsymbol{h}_{seg} \in \mathbb{R}^{K \times d_m}$ are first projected into the segmentation prompt embedding space:

$$\tilde{\boldsymbol{h}}_{seg} = \mathrm{MLP}_s(\boldsymbol{h}_{seg}), \tag{4}$$

where $\tilde{\boldsymbol{h}}_{seg} \in \mathbb{R}^{K \times d_p}$ and d_p denotes the SAM2 prompt embedding dimension. To anchor target activation within shared query semantics, we further align the projected target embeddings with the query representations:

$$\hat{\boldsymbol{h}}_{seg} = \mathrm{CrossAttention}_t(\tilde{\boldsymbol{h}}^Q_{seg}, \boldsymbol{E}^K_{query}, \boldsymbol{E}^V_{query}). \tag{5}$$

Through QVA and TQA, both spatial features and target prompts are conditioned on the same disentangled query representation, establishing a shared semantic anchor across modalities.

Training Objective. The aligned visual features $\hat{\boldsymbol{F}}_{seg}$ and target prompt embeddings $\hat{\boldsymbol{h}}_{seg}$ are fed into the SAM2 mask decoder to produce one binary mask per target, naturally supporting multi-target prediction under a single query.

SIRA is trained end-to-end with both segmentation and text generation supervision. The overall objective is defined as

$$\mathcal{L}_{total} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{txt}\mathcal{L}_{txt}, \tag{6}$$

where $\mathcal{L}_{bce}$ and $\mathcal{L}_{dice}$ denote the binary cross-entropy and Dice losses for mask prediction, respectively, and $\mathcal{L}_{txt}$ is the cross-entropy loss for autoregressive text generation.

Table 2. Quantitative comparison of reasoning segmentation performance on SurgRS (% for gIoU, cIoU, and Dice).

Method	Overall			Reasoning Query Type									
	gIoU	cIoU	Dice	LT	SE	SGT	PAI	NSI	NDL	OR	PL	CR	OT
LISA	65.05	74.32	74.21	79.34	71.14	69.24	73.23	74.55	74.43	68.30	70.86	58.81	54.35
GSVA	65.30	75.08	74.39	79.14	72.20	69.52	73.03	74.52	74.89	67.21	69.64	60.68	53.94
PathMR	66.42	75.14	76.25	80.18	74.11	72.39	75.91	76.40	75.69	71.15	73.37	63.20	61.10
VRS-HQ	70.79	80.60	79.61	83.53	78.57	75.40	78.53	78.92	79.78	72.31	76.29	64.97	59.82
SIRA	**72.24**	**81.78**	**80.93**	**86.82**	**80.05**	**76.76**	**79.53**	**80.02**	**81.05**	**74.42**	**76.78**	**65.58**	**61.61**

3 Experiments

Dataset. We evaluate SIRA on the proposed SurgRS dataset. SurgRS is constructed based on SAR-RARP50 [21] and focuses exclusively on radical prostatectomy procedures. The dataset contains approximately 16,000 surgical images (resolution 1920×1080) and around 41,000 structured image-text pairs, covering 20 instrument categories and over 127,000 instance-level segmentation targets. Following the original SAR-RARP50 protocol, we split the dataset into training and testing sets with a ratio of 4:1.

Implementation Details. We fine-tune the MLLM (Chat-UniVi-7B) using LoRA with rank 8 while freezing the remaining language model parameters. The number of query tokens is set to $N_q = 64$. The alignment module (ViT, cross-attention, and MLP projections) and the SAM2 mask decoder are trained, while freezing the SAM2 vision backbone and other components. Training is performed using the AdamW optimizer with a learning rate of 3e-4 and no weight decay. We adopt a WarmupDecayLR scheduler with 100 warm-up iterations. The model is trained for 10 epochs (13,000 iterations per epoch) on the SurgRS training set. The segmentation loss weights are set to $\lambda_{bce} = 2$ and $\lambda_{dice} = 0.5$, and the reasoning loss weight is $\lambda_{txt} = 1$. Experiments are conducted on two NVIDIA GeForce RTX 3090 GPUs using DeepSpeed. The per-GPU batch size is 1 with gradient accumulation set to 1, resulting in a total batch size of 2. The training process consumes approximately 24GB of GPU memory, while inference requires about 16GB.

Baselines and Evaluation Metrics. We compare SIRA with four state-of-the-art reasoning segmentation models, LISA [11], GSVA [12], PathMR [13], and VRS-HQ [15], all implemented with 7B-scale language backbones for fair comparison. Following prior work, we adopt generalized IoU (gIoU) and complete IoU (cIoU) for evaluation. In addition, we report the Dice coefficient, which is widely used in medical image segmentation.

Comparison Results. Table 2 reports the reasoning segmentation performance on SurgRS. SIRA achieves the best results across all overall metrics and reason-

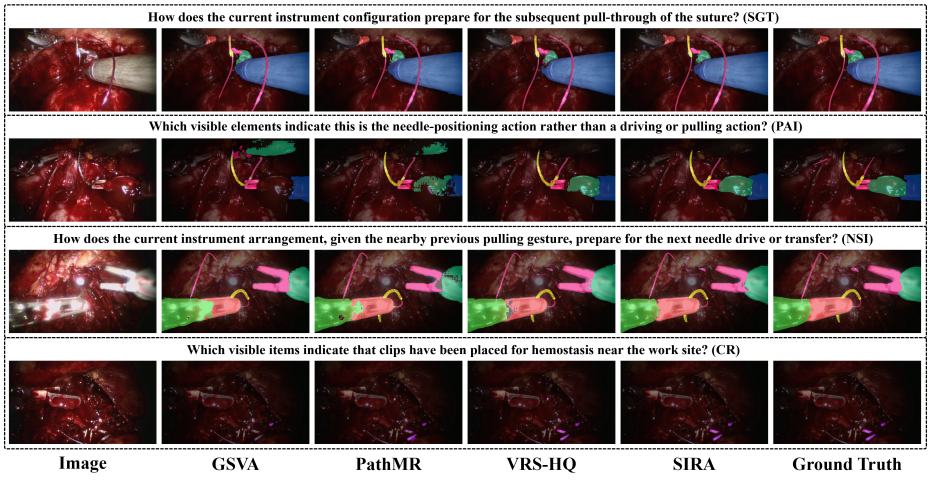

Fig. 3. Exemplar qualitative results of different approaches on SurgRS.

Table 3. Ablation study of QVA and TQA on SurgRS (%).

Baseline	QVA	TQA	gIoU	cIoU	Dice
•	○	○	70.79	80.60	79.61
•	○	•	71.40	81.23	80.09
•	•	○	71.93	81.62	80.59
•	•	•	**72.24**	**81.78**	**80.93**

ing query types. Fig. 3 illustrates qualitative comparisons under several challenging surgical conditions. In scenarios involving ambiguous boundaries and small-scale instruments, SIRA produces more complete and spatially coherent masks, with clearer delineation of fine structures. Under low-illumination environments and for infrequently occurring targets, the proposed method maintains stable localization, whereas competing methods exhibit partial activation or omission. Furthermore, in cases affected by motion blur due to rapid instrument movement, SIRA preserves consistent mask prediction and reduces fragmentation artifacts. These observations indicate that the proposed query-anchored dual alignment mechanism enhances semantic-visual consistency and improves robustness across diverse intraoperative challenges.

Ablation Analysis. Table 3 reports ablations of QVA and TQA. Enabling either module consistently improves over the baseline, with QVA providing slightly larger gains. Combining QVA and TQA yields the best results, indicating that the two alignments are complementary for query-conditioned surgical instrument segmentation.

4 Conclusion

We introduce RA-SIS, a reasoning-aware surgical instrument segmentation formulation that models segmentation as query-conditioned inference under procedural context. To benchmark this setting, we construct SurgRS, a dataset aligning instance-level masks with structured query-answer supervision for pixel-level semantic grounding. We further propose SIRA, a multimodal framework that disentangles query- and target-level semantics and aligns them with visual features via query-anchored dual alignment. Experiments demonstrate consistent improvements over state-of-the-art reasoning segmentation methods on SurgRS. Future work will extend RA-SIS to video-based surgical understanding, where temporal consistency and motion dynamics introduce additional challenges.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. W. Yue, J. Zhang, K. Hu, Y. Xia, J. Luo, and Z. Wang. SurgicalSAM: Efficient class promptable surgical instrument segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6890–6898, 2024.
2. O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.
3. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
4. J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. *Med. Image Anal.*, 85:102762, 2023.
5. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
6. A. Kirillov et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
7. N. Ayobi, A. Pérez-Rondón, S. Rodríguez, and P. Arbeláez. MATIS: Masked-attention transformers for surgical instrument segmentation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023.
8. W. Yue, J. Zhang, K. Hu, Q. Wu, Z. Ge, Y. Xia, J. Luo, and Z. Wang. SurgicalPart-SAM: Part-to-whole collaborative prompting for surgical instrument segmentation. *arXiv preprint arXiv:2312.14481*, 2023.
9. Z. Zhou, O. Alabi, M. Wei, T. Vercauteren, and M. Shi. Text promptable surgical instrument segmentation with vision-language models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 28611–28623, 2023.

10. Y. Liu, B. Peng, Z. Zhong, Z. Yue, F. Lu, B. Yu, and J. Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.
11. X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. 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. Z. Xia, D. Han, Y. Han, X. Pan, S. Song, and G. Huang. GSVA: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.
13. Y. Zhang et al. PathMR: Multimodal visual reasoning for interpretable pathology diagnosis. *arXiv preprint arXiv:2508.20851*, 2025.
14. H. Rasheed et al. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.
15. S. Gong, Y. Zhuge, L. Zhang, Z. Yang, P. Zhang, and H. Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29183–29192, 2025.
16. N. Ravi et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
17. Z. Zeng et al. SurgVLM: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025.
18. Y. Long et al. Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Sci. Robot.*, 10(104).
19. M. Allan et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.
20. M. Allan et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.
21. D. Psychogyios et al. SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.
22. T. Rueckert et al. Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (PhaKIR). *arXiv preprint arXiv:2511.06549*, 2025.
23. A. Zia et al. Surgical visual understanding (SurgVU) dataset. *arXiv preprint arXiv:2501.09209*, 2025.
24. P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.
25. E. J. Hu et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, volume 1, page 3, 2022.