ReSurgSAM2: Referring Segment Anything in Surgical Video via Credible Long-term Tracking

Haofeng Liu[1], Mingqi Gao[2], Xuxiao Luo[1], Ziyue Wang[1], Guanyi Qin[1], Junde Wu[3], and Yueming Jin[1(✉)]

[1] National University of Singapore, Singapore, Singapore
[2] Southern University of Science and Technology, Shenzhen, China
[3] University of Oxford, Oxford, United Kingdom
haofeng.liu@u.nus.edu, ymjin@nus.edu.sg

Abstract. Surgical scene segmentation is critical in computer-assisted surgery and is vital for enhancing surgical quality and patient outcomes. Recently, referring surgical segmentation is emerging, given its advantage of providing surgeons with an interactive experience to segment the target object. However, existing methods are limited by low efficiency and short-term tracking, hindering their applicability in complex real-world surgical scenarios. In this paper, we introduce ReSurgSAM2, a two-stage surgical referring segmentation framework that leverages Segment Anything Model 2 to perform text-referred target detection, followed by tracking with reliable initial frame identification and diversity-driven long-term memory. For the detection stage, we propose a cross-modal spatial-temporal Mamba to generate precise detection and segmentation results. Based on these results, our credible initial frame selection strategy identifies the reliable frame for the subsequent tracking. Upon selecting the initial frame, our method transitions to the tracking stage, where it incorporates a diversity-driven memory mechanism that maintains a credible and diverse memory bank, ensuring consistent long-term tracking. Extensive experiments demonstrate that ReSurgSAM2 achieves substantial improvements in accuracy and efficiency compared to existing methods, operating in real-time at 61.2 FPS. Our code and datasets are available at https://github.com/jinlab-imvr/ReSurgSAM2.

Keywords: Robotic-assisted surgery · Referring segmentation · Long-range video object tracking · Video-language learning.

1 Introduction

Surgical video scene segmentation is vital in computer-assisted surgery [5, 10]. By precise identification and differentiation of surgical instruments and tissues, this technology provides cognitive assistance to surgeons for decision-making [9, 14, 15, 25, 31]. Most existing methods in surgical scene segmentation rely only on video data [16, 30, 32]. Despite achieving real-time and satisfactory performance, these systems generate semantic masks for all instruments and tissues collectively, without providing surgeons the ability to interactively identify and track

specific objects of interest. Referring segmentation, which automatically identifies and segments the specific object according to a textual expression, benefits various surgical applications [24, 32]. Integrated with AR technology, it could enhance surgical education by enabling trainees to interactively explore specific instruments through AR overlays [8, 21]. During intraoperative surgery, such ability enables the system to focus on regions of interest specified by surgeons, which can optimize workflow, providing accurate and personalized navigation support for safer and higher-quality outcomes of patient care.

To enable this capability, RSVIS [24] takes the first step to investigate referring instrument segmentation in the surgical domain. They propose a video-instrument synergistic network to learn both video-level and instrument-level knowledge to improve performance. However, this method relies only on short-term information of three consecutive frames, leading to inherent challenges in long-term tracking. Referring segmentation in surgery still remains under-explored. In natural domains, the typical task is referring video object segmentation (RVOS). The online RVOS methods [11, 26] can offer real-time processing; however, they lack robust long-term tracking essential for surgical scenarios, as these methods are typically developed for short videos (under 10 seconds). This limitation is particularly critical in surgical procedures, which generally last long durations in hours, with dynamic scene variations and instrument movement. In contrast, offline RVOS alternatives [27, 28] achieve better performance through extended temporal integration but require future frame information, making them unsuitable for intraoperative applications.

Recently, Segment Anything Model 2 (SAM2) [18] has gained attention for its interactive framework with satisfactory tracking capabilities, presenting a promising potential to enhance the referring segmentation task. By providing visual prompts on initial frames as permanent memory, it can perform consistent tracking with memory attention. However, visual prompts (e.g., bounding boxes or multi-points) rely on the object's presence in the first few frames and impose labeling burdens on surgeons during operations. Since the target object may not be present at the beginning of surgery, such interactions are suboptimal [15]. Instead, textual expressions offer greater flexibility and are the closest form to audio. Integrating textual prompts into SAM2 marks a crucial step toward hands-free interaction in surgery [12, 24]. Furthermore, identifying the reliable initial frame for tracking is critical for RVOS, as inaccurate segmentation masks can introduce error accumulation, significantly degrading overall performance [3]. Finally, SAM2 employs a greedy strategy that selects the nearest frames as memory without assessing their reliability, limiting long-term temporal modeling [13].

To address these challenges, we propose ReSurgSAM2, a novel two-stage framework built on SAM2 that sequentially performs text-referred target detection followed by tracking, enabling efficient and accurate RVOS in long-range surgical videos. During the detection stage, we identify the reliable initial frame using our Cross-Modal Spatial-Temporal Mamba (CSTMamba) and Credible Initial Frame Selection (CIFS) strategy. CSTMamba efficiently captures dedi-

cated spatial-temporal dependencies across video frames while integrating multi-modal features, enabling precise specified object detection and segmentation for robust initial frame selection. Leveraging these accurate detection results, CIFS selects the optimal frame for tracking initialization based on confidence. During tracking, our Diversity-Driven Long-term Memory (DLM) mechanism empowers SAM2 to track the object reliably throughout the entire surgical video by conditioning on a diverse and dependable memory bank. Extensive experiments on datasets containing instruments and tissues demonstrate significant performance improvements over existing methods, maintaining real-time at 61.2 FPS.

2 Method

2.1 ReSurgSAM2 Framework

SAM2 [18] extends SAM [7] to video domains by incorporating temporal memory attention while preserving its segmentation capabilities. It implements a short-term memory mechanism through a queue, which conditions the current frame features on both the permanent initial frame and nearest predictions through greedy selection. For each frame, the mask decoder generates predictions with two key scores: the intersection-over-union (IoU) score and the occlusion score. The IoU score estimates the alignment between the prediction with ground truth, while the occlusion score employs a signed confidence scheme - positive values signify object presence, negative values indicate absence, and the magnitude reflects confidence. This dual-scoring system enables object tracking even during occlusions. While SAM2 demonstrates promising performance in general video domains, it faces limitations in adapting to surgical RVOS, including vision-language integration, reliable initial frame identification, and long-term tracking.

To overcome the above limitations, we propose ReSurgSAM2, a specialized two-stage framework for surgical applications that seamlessly integrates text-referred target detection with tracking, as illustrated in Fig. 1. Given the t-th frame $f_t \in \mathbb{R}^{3\times H\times W}$ from an image stream and its linguistic expression e, we separately employ SAM image encoder and the frozen CLIP text encoder with a trainable MLP to extract features, with our goal of obtaining the segmentation mask $m_t \in \mathbb{R}^{H\times W}$ for the text-referred target. In the first stage, our model enhances detection reliability by utilizing CSTMamba and a mask decoder to generate a high-fidelity segmentation mask. The scores of the mask are subsequently fed into the CIFS, which performs credible frame selection prepared for the tracking phase to mitigate error accumulation. Upon identification of the optimal initial frame by CIFS, the model switches to the tracking stage, where the prompt encoder accepts the CLS token from text features, and the model ensures reliable and consistent object tracking throughout the video by integrating vanilla short-term memory with long-term memory using the DLM.

2.2 Identify the Reliable Initial Frame with CSTMamba and CIFS

Cross-Modal Spatial-Temporal Mamba (CSTMamba): Referring segmentation using a single frame often yields a suboptimal result, and propa-

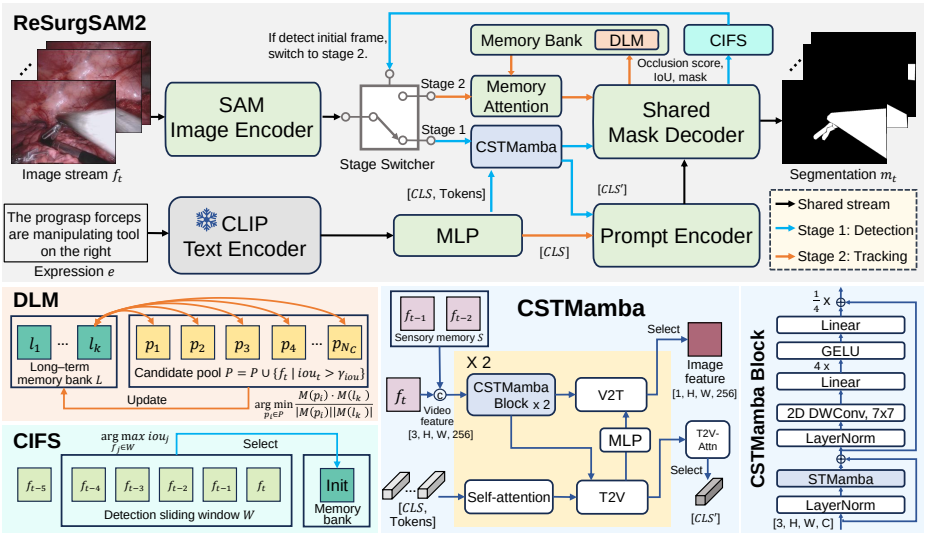

Fig. 1: Overview of ReSurgSAM2. The model begins with the text-referred target detection using CSTMamba to provide credible frames for selection using CIFS. Upon detecting the initial frame, CIFS activates the tracking stage, in which DLM offers diverse and reliable memory for consistent long-term tracking.

gating this result as an initial reference for tracking potentially causes error accumulation. Therefore, precise target detection is crucial for our two-stage framework. While transformer-based approaches can enhance segmentation by leveraging self-attention, this mechanism suffers from quadratic complexity [22], limiting the real-time surgical applications. Recently, Mamba [4] has emerged as a promising alternative with linear complexity and selective information propagation. STMamba [29] leverages these advantages for video segmentation; however, it lacks cross-modal capabilities for RVOS and is constrained by Mamba's linear scanning mechanism, which restricts its ability to capture fine-grained pixel-level information.

To address these limitations, we propose the CSTMamba, which integrates a sensory memory bank, the CSTMamba block, cross-modal attention mechanism to facilitate comprehensive cross-modal spatial-temporal modeling. To enable temporal interaction, the sensory memory bank S stores the most recent two frame features extracted from the image encoder. Then, the CSTMamba takes the language features and video features from the current frame and sensory memory as input for cross-model spatial-temporal modeling, and outputs the fused image feature $E(f_t)'$ and CLS' token. The video input can be written as $[E(f_t), E(f_{t-1}), E(f_{t-2})] \in \mathbb{R}^{3\times H'\times W'\times 256}$, where $E(\cdot)$ is the image encoder and $[\cdot]$ is concatenation, H' and W' are the feature height and width.

The CSTMamba block, designed for comprehensive spatial-temporal modeling, is illustrated in Fig. 1. It integrates STMamba with a 7×7 2D depth-wise

convolution (DWConv) to capture global features with selective scanning and fine-grained local features with an expanded receptive field. Building upon this foundation, an inverted bottleneck [20] expands the MLP block to four times the input dimension, effectively enhancing feature representation by utilizing the enlarged middle layer to improve spatial-temporal interactions. Additionally, bidirectional cross-modal attention mechanisms—text-to-vision (T2V) and vision-to-text (V2T) [23]—facilitate cross-modal interactions. The CSTMamba processes visual-temporal information and cross-modal signals simultaneously, generating enriched spatial-temporal features and a fused CLS' token with strong cross-modal representations for final mask prediction.

Credible Initial Frame Selection (CIFS): When automatically selecting an initial reference frame for tracking, it is essential to identify a highly reliable frame to mitigate error accumulation [13]. However, in surgical environments, the small inter-class similarity among instruments and among tissues increases the risk of false detections, leading to unreliable segmentation. To address this challenge, we implement CIFS that requires the model to predict the object as present with high confidence based on IoU score and occlusion score across N_w consecutive frames before frame selection begins. Before selection, it uses the sliding window to detect the qualified frames, which can be formulated as:

$$W = \{f_j \mid j \in [t - N_w + 1, t] \wedge iou_j > \delta_{iou} \wedge sigmoid(o_j) > \delta_o\}, \tag{1}$$

where W is the sliding window, iou_j and o_j are the IoU score and predicted occlusion score for the t-th frame, δ_{iou} and δ_o are their respective thresholds, and $sigmoid(\cdot)$ maps the score to range $[0, 1]$. Upon $|W| = N_w$, the one with the highest IoU score is selected as the initial reference among these N_w qualifying frames. With the selection of an optimal and credible initial frame, ReSurgSAM2 enters its second stage, where it performs robust tracking by leveraging this reference frame to propagate predictions throughout the remaining sequence, ensuring both semantic fidelity and temporal consistency.

2.3 Diversity-driven Long-term Memory

In surgical environments, videos typically have long durations with dynamic scene variations and instrument movement. However, SAM2's greedy strategy of selecting only the most recent frames as memory hinders effective long-term tracking, leading to redundancy and potential viewpoint overfitting, thereby limiting its ability to capture anatomical changes. To overcome this challenge, we propose the DLM mechanism that enhances SAM2's vanilla memory bank by strategically selecting frames from a candidate pool. This mechanism enriches the memory bank by using the candidate pool to extend the temporal coverage range and collecting the frames that capture diverse spatial-temporal information to mitigate viewpoint overfitting. Additionally, it ensures the inclusion of high-confidence frames to minimize error accumulation.

The proposed DLM mechanism updates the candidate pool in inference as:

$$P = P \cup \{f_t \mid iou_t > \gamma_{iou}\}, \tag{2}$$

where P is the candidate pool, f_t is the t-th frame, iou_t is its predicted IoU score, and γ_{iou} is a confidence threshold. Each element in P is indexed by p_i, where p_i refers to the i-th candidate. This mechanism selects high-confidence frames as memory candidates for mitigating error propagation. When the candidate pool reaches its capacity N_p, we store the most diverse candidate based on its cosine similarity to the latest long-term memory frame:

$$p^* = \arg\min_{p_i \in P} \frac{M(p_i) \cdot M(l_k)}{|M(p_i)||M(l_k)|}, \tag{3}$$

where $M(\cdot)$ is the memory encoder, p_i is i-th candidate frame, l_k is the latest frame in long-term memory bank L, and p^* is the selected frame. After selection, the pool P is cleared to extend the temporal coverage range of the long-term memory, and the updated memory bank is utilized for the $(t+1)$-th frame's memory attention mechanism. To enhance efficiency, we maintain a queue with capacity N_l for long-term memory, keeping the initial frame permanently in the long-term memory. By concatenating SAM2's vanilla short-term memory with our long-term memory using the DLM mechanism, ReSurgSAM2 maintains a memory bank that is reliable and diverse, boosting consistent long-term tracking.

3 Experiment

3.1 Dataset and Implementation Details

The experiments were conducted with Ref-EndoVis17 and Ref-EndoVis17 building upon EndoVis17 [2], EndoVis18 dataset [1] and RSVIS [24]. The EndoVis17 comprises 3000 frames across 10 sequences, including eight training sequences, eight test sequences from identical scenes, and two additional test sequences, with instrument labels. The EndoVis18 dataset consists of 15 sequences with comprehensive scene segmentation annotations. As introduced in RSVIS, both datasets were reannotated with consistent instance-specific labels to facilitate RVOS. Building upon RSVIS [24], we performed meticulous refinement to address inconsistencies and omissions in instrument labeling that would otherwise compromise experimental validity. We further enriched the datasets by incorporating tissue-specific annotations from EndoVis18, including kidney parenchyma, covered kidney, and small intestine. For Ref-EndoVis17, we merged sequences from the identical scenes to prevent cross-contamination between training and test sets, with sequences 2, 5, and 6 designated as the test set. Following RSVIS [24],

Table 1: Dataset statistics for Ref-EndoVis17 and Ref-EndoVis18 datasets. "Pair" denotes text-mask pairs.

Dataset	Training				Testing			
	Sequence	Frame	Object	Pair	Sequence	Frame	Object	Pair
Ref-EndoVis17(tool)	7	2100	20	4873	3	900	10	2265
Ref-EndoVis18(tool)	11	1639	34	3787	4	596	15	1384
Ref-EndoVis18(tissue)	11	1639	25	2995	4	596	7	807

we allocated sequences 2, 5, 9, and 15 as the test set for Ref-EndoVis18. This division ensures balanced object distribution across training and testing. Table 1 presents statistics for both datasets.

ReSurgSAM2 employs the Hiera-small backbone [19] initialized with SAM2 pre-trained weights [18], with an image input size of 512. During training, we follow SAM2 to conduct prompt segmentation by loading three frames for text-referred object detection, followed by seven frames for tracking. The model is trained for 30 epochs utilizing the same training strategies as SAM2. For inference, unlike RSVIS, we generate text expressions based on the first appearance of each object to accommodate the whole surgical video. The hyperparameters are set as follows: $\delta_o = 0.9$, $\delta_{iou} = 0.7$, $\gamma_{iou} = 0.95$, $N_w = 5$, $N_p = 5$ and $N_l = 4$. Unlike semantic segmentation using challenge IoU [1, 2], RVOS tracks the object across the whole video, including when it is occluded. Therefore, for evaluation metrics, we adopt $\mathcal{J}$ and $\mathcal{F}$ [17] for accuracy, where $\mathcal{J}$ assesses region accuracy and $\mathcal{F}$ evaluates boundary accuracy following RVOS [6], with $\mathcal{J}\&\mathcal{F}$ representing their mean, and the frame-per-second (FPS) for efficiency. All metrics are higher-is-better. All experiments utilized the same training data on an NVIDIA A6000 GPU.

3.2 Comparison and Ablation Study

Comparison. To validate the effectiveness of ReSurgSAM2, we conducted comprehensive comparisons against state-of-the-art methods summarized in Table 2. The comparison methods include offline methods (ReferFormer and MUTR) and online methods (RSVIS, OnlineRefer, and RefSAM). Offline methods achieve stable performance by processing 64 frames concurrently during inference, thus reducing false detections. In contrast, RSVIS only relies on short-term information, leading to suboptimal long-term tracking. While OnlineRefer and RefSAM exhibit modest long-term tracking capabilities by query propagation, the performance remains suboptimal on long-range sequences in Ref-EndoVis17. In comparison, ReSurgSAM2 shows superior cross-modal and long-term tracking capabilities in RVOS, with substantial $\mathcal{J}\&\mathcal{F}$ improvements: 14.17 on Ref-EndoVis17, 7.76 on Ref-EndoVis18 tool, and 3.19 on Ref-EndoVis18 tissue datasets.

The qualitative comparison with RSVIS and RefSAM is illustrated in Fig. 2. RSVIS lacks robust instrument discrimination in complex scenarios, causing in-

Table 2: Quantitative comparison with state-of-the-art methods.

Method	Setting	Ref-EndoVis17(tool)			Ref-EndoVis18(tool)			Ref-EndoVis18(tissue)			FPS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	
ReferFormer [27]	Offline	62.41	62.28	62.55	71.09	70.96	71.23	61.84	69.9	53.78	42.3
MUTR [28]	Offline	60.97	60.76	61.18	67.56	67.79	67.33	63.53	71.48	55.58	32.3
RSVIS [24]	Online	61.22	61.37	61.07	68.35	68.55	68.15	65.69	72.91	58.47	22.1
OnlineRefer [26]	Online	60.32	60.29	60.34	72.19	71.88	72.50	70.56	77.58	63.55	25.6
RefSAM [11]	Online	63.56	63.77	63.35	72.86	73.40	72.31	71.90	77.66	66.14	25.4
ReSurgSAM2	Online	**77.73**	**77.77**	**77.69**	**80.62**	**80.94**	**80.31**	**75.09**	**80.93**	**69.25**	**61.2**

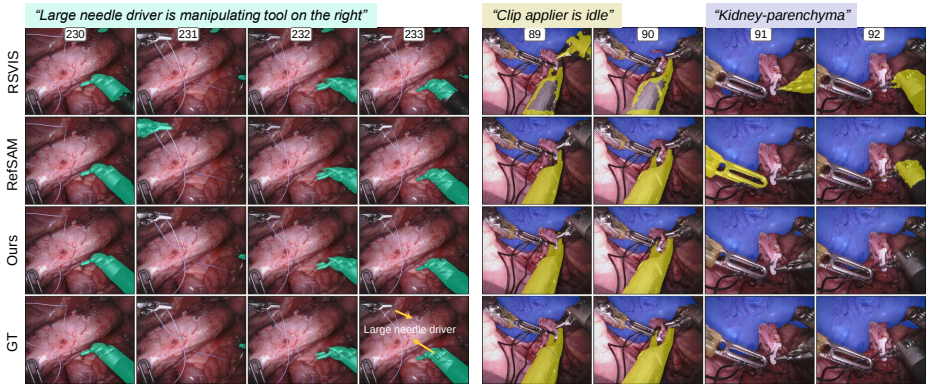

Fig. 2: Qualitative comparison in Ref-EndoVis17(left) and Ref-EndoVis18(right).

complete segmentation. Both RSVIS and RefSAM exhibit limited tracking stability during rapid object movements and scene variations, due to their limited long-term modeling. In contrast, ReSurgSAM2, equipped with a robust initialization and diverse long-term memory, performs reliable and consistent tracking.

Ablation Study. We conduct comprehensive ablation studies on Ref-EndoVis17 to evaluate the effectiveness of each proposed component, with results presented in Table 3. For ablation, the two-stage RVOS framework uses the first detected frame with $iou_t > 0.7$ and $sigmoid(o_t) > 0.9$ as a reference to activate tracking of stage 2. This design helps capture short-term temporal dependencies, leading to a $\mathcal{J}\&\mathcal{F}$ gain of 2.64. Integrating CSTMamba strengthens spatial-temporal referring segmentation to generate a more reliable reference for tracking, achieving a 4.77 improvement in $\mathcal{J}\&\mathcal{F}$. Furthermore, the CIFS strategy further enhances $\mathcal{J}\&\mathcal{F}$ by 6.14, as it selects a more reliable reference as memory, mitigating error accumulation. DLM further enhances vanilla memory mechanism with long-term temporal modeling by maintaining a diverse and long-range memory bank, contributing a 3.03 boost in $\mathcal{J}\&\mathcal{F}$. To validate DLM's effectiveness, we explored various memory bank designs: an extended short-term memory based on the vanilla memory bank, and a long-term memory with interval sampling every five frames (storing three frames each), as shown in Table 4. By integrating all proposed components, ReSurgSAM2 ultimately achieves 77.73 in $\mathcal{J}\&\mathcal{F}$ while maintaining real-time performance at 61.2 FPS.

Stage 2	CSTMamba	CIFS	DLM	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
				61.15	61.46	60.84	**70.1**
√				63.79	63.77	63.82	68.2
√	√			68.56	68.51	68.61	67.5
√	√	√		74.70	74.67	74.72	63.1
√	√	√	√	**77.73**	**77.77**	**77.69**	61.2

Table 3: Ablation studies on Ref-EndoVis17.

Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Vanilla	74.70	74.67	74.72
Extended	74.68	74.64	74.72
Interval	75.32	75.27	75.37
DLM	**77.73**	**77.77**	**77.69**

Table 4: Ablation on memory bank design.

4 Conclusion

This paper presents ReSurgSAM2, a two-stage framework via credible long-term tracking for surgical referring segmentation. While existing methods are limited in efficiency and long-term tracking, our approach addresses these limitations through reliable initial frame identification and a long-term memory mechanism building upon SAM2. Extensive experiments on surgical datasets demonstrate that ReSurgSAM2 significantly outperforms existing methods, offering a practical and efficient solution for real-time surgical video analysis.

Acknowledgments. This work was supported by Ministry of Education Tier 1 Start up grant, NUS, Singapore (A-8001267-01-00); Ministry of Education Tier 1 grant, NUS, Singapore (A-8001946-00-00).

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

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. arXiv preprint arXiv:2411.17646 (2024)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
5. Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.A., Stoyanov, D.: Exploring intra-and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging **41**(11), 2991–3002 (2022)
6. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Asian Conference on Computer Vision (ACCV). pp. 123–141 (2019)
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: International Conference on Computer Vision. pp. 4015–4026 (2023)
8. Kovoor, J.G., Gupta, A.K., Gladman, M.A.: Validity and effectiveness of augmented reality in surgical education: a systematic review. Surgery **170**(1), 88–98 (2021)
9. Li, H., Li, H., Chen, J., Zhong, R., Niu, K., Fu, H., Liu, J.: Aif-sfda: Autonomous information filter-driven source-free domain adaptation for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
10. Li, H., Ou, M., Li, H., Qiu, Z., Niu, K., Fu, H., Liu, J.: Multi-view test-time adaptation for semantic segmentation in clinical cataract surgery. IEEE Transactions on Medical Imaging (2025)

11. Li, Y., Zhang, J., Teng, X., Lan, L., Liu, X.: Refsam: Efficiently adapting segmenting anything model for referring video object segmentation. arXiv preprint arXiv:2307.00997 (2024)
12. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. In: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond (2024)
13. Liu, Y., Yu, R., Yin, F., Zhao, X., Zhao, W., Xia, W., Yang, Y.: Learning quality-aware dynamic memory for video object segmentation. In: European Conference on Computer Vision. pp. 468–486 (2022)
14. Low, C.H., Wang, Z., Zhang, T., Zeng, Z., Zhuo, Z., Mazomenos, E.B., Jin, Y.: Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. arXiv preprint arXiv:2503.10265 (2025)
15. Moglia, A., Georgiou, K., Georgiou, E., Satava, R.M., Cuschieri, A.: A systematic review on artificial intelligence in robot-assisted surgery. International Journal of Surgery **95**, 106151 (2021)
16. Ou, M., Li, H., Liu, H., Wang, X., Yi, C., Hao, L., Hu, Y., Liu, J.: Mvd-net: Semantic segmentation of cataract surgery using multi-view learning. In: IEEE EMBC. pp. 5035–5038 (2022)
17. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
18. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
19. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: International conference on machine learning. pp. 29441–29454 (2023)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
21. Sheik-Ali, S., Edgcombe, H., Paton, C.: Next-generation virtual and augmented reality in surgical education: a narrative review. Surgical technology international **33** (2019)
22. Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. ACM Computing Surveys **55**(6), 1–28 (2022)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Wang, H., Yang, G., Zhang, S., Qin, J., Guo, Y., Xu, B., Jin, Y., Zhu, L.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. IEEE Transactions on Medical Imaging **43**(12), 4457–4469 (2024)
25. Wang, Z., Wu, J., Low, C.H., Jin, Y.: Medagent-pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)
26. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2761–2770 (2023)

27. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)
28. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., Dong, H., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6449–6457 (2024)
29. Yang, Y., Xing, Z., Yu, L., Huang, C., Fu, H., Zhu, L.: Vivim: A video vision mamba for medical video segmentation. arXiv preprint arXiv:2401.14168 (2024)
30. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: Surgicalsam: Efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6890–6898 (2024)
31. Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., et al.: Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv preprint arXiv:2506.02555 (2025)
32. Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M.: Text promptable surgical instrument segmentation with vision-language models. Advances in Neural Information Processing Systems **36**, 28611–28623 (2023)