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], Yueming Jin[1†]

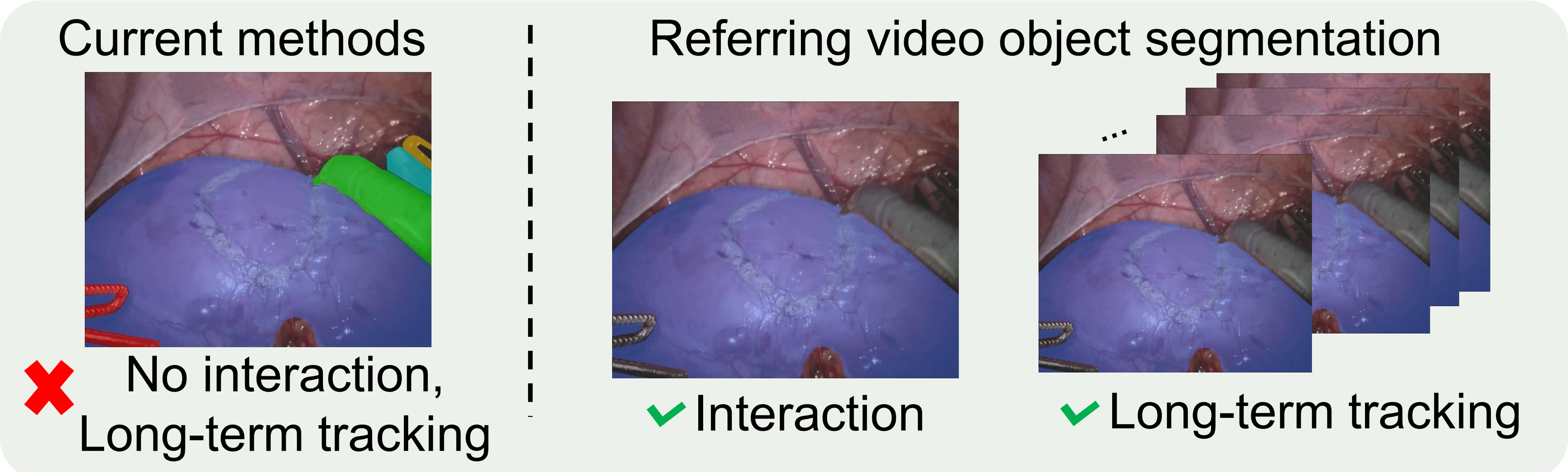
[1]National University of Singapore [2]Southern University of Science and Technology [3]University of Oxford

haofeng.liu@u.nus.edu, ymjin@nus.edu.sg

Introduction

Surgical Video Segmentation: Importance & Current Practice

- Precise instrument/tissue recognition for automation
- Real-time surgical guidance and AR-assisted surgical education
- Current methods: generate collective masks without interactivity and long-term tracking



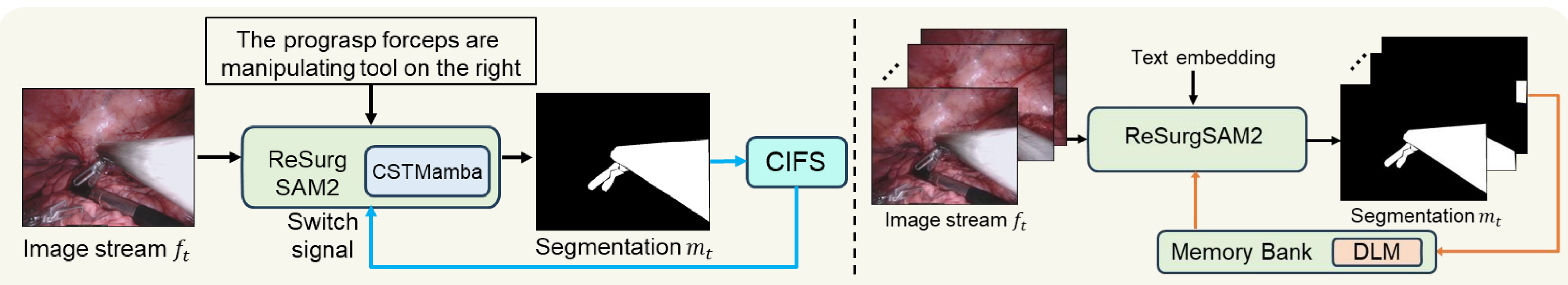
Challenges in Current Approaches

- Lack of hands-free, text-driven segmentation for specific tools/tissues
- Not designed for real-time use in operating rooms
- Fails to achieve long-term tracking for hours-long procedures

Key Question: How can we enable accurate, real-time referring segmentation with robust long-term tracking in surgical videos?

Method

ReSurgSAM2: A novel framework for accurate, real-time referring segmentation with long-term tracking in surgical videos. Seamlessly intergrade **text understanding** with **reliable long-term tracking**.



Stage 1: Text-Referred Target Detection **Stage 2**: Long-term Tracking

- **CSTMamba** – Efficient cross-modal spatio-temporal Mamba modeling: integrates STMamba with 7×7 depthwise convolution and bidirectional text–vision fusion.
- **CIFS** – Credible initial frame selection: sliding-window detection with IoU/occlusion filtering ensures robust initialization to reduce error accumulation.
- **DLM** – Diversity-driven memory: cosine-similarity selection builds a hybrid short/long-term memory, avoiding redundancy and improving long-sequence stability.

Experiment

Dataset: Ref-EndoVis17 and Ref-EndoVis18

Table 1: Dataset statistics for Ref-EndoVis17/EndoVis18 datasets.

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

Metric: J (region accuracy), F (boundary accuracy), J&F (mean), FPS.

Comparison Experiment

- **State-of-the-art Accuracy** – ReSurgSAM2 achieves the best J&F scores, with **+14.2** on Ref-EndoVis17 and clear gains on Ref-EndoVis18.
- **Robust Long-term Tracking** – Consistently stable under rapid motion and scene variations, outperforming RSVIS and RefSAM.
- **Real-time Performance** – Runs efficiently at **61.2 FPS**.

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**
		+14.17			+7.76			+3.19			+18.9

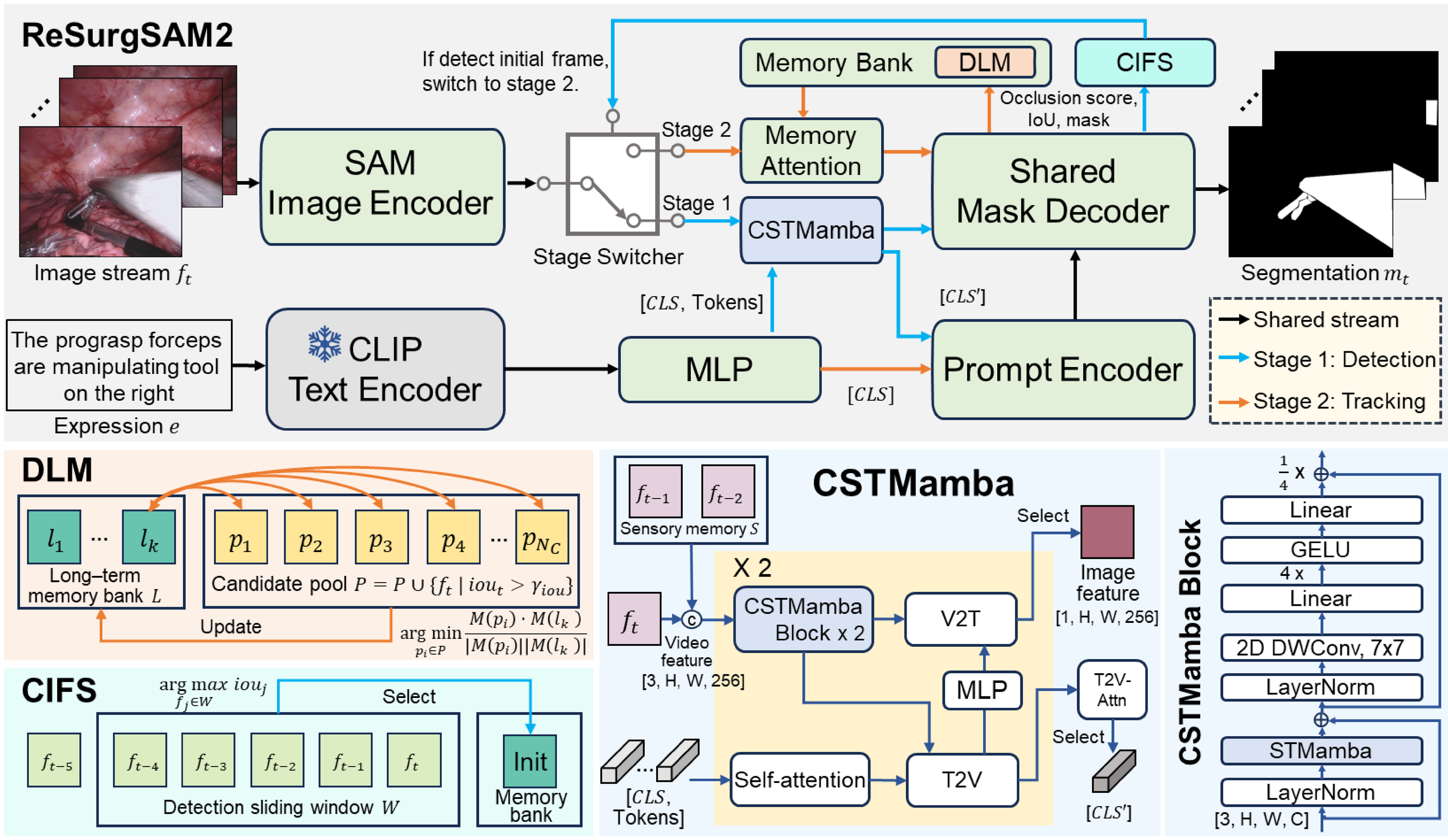


Figure 1. Overview of ReSurgSAM2

Qualitative Analysis

- **Complex scenes** – Correctly segments the **specified instrument** even when multiple similar ones appear in the same scene.
- **Clearer boundaries** – More accurate **instrument/tissue** segmentation.
- **Stable tracking** – Maintains consistency during occlusion, motion.

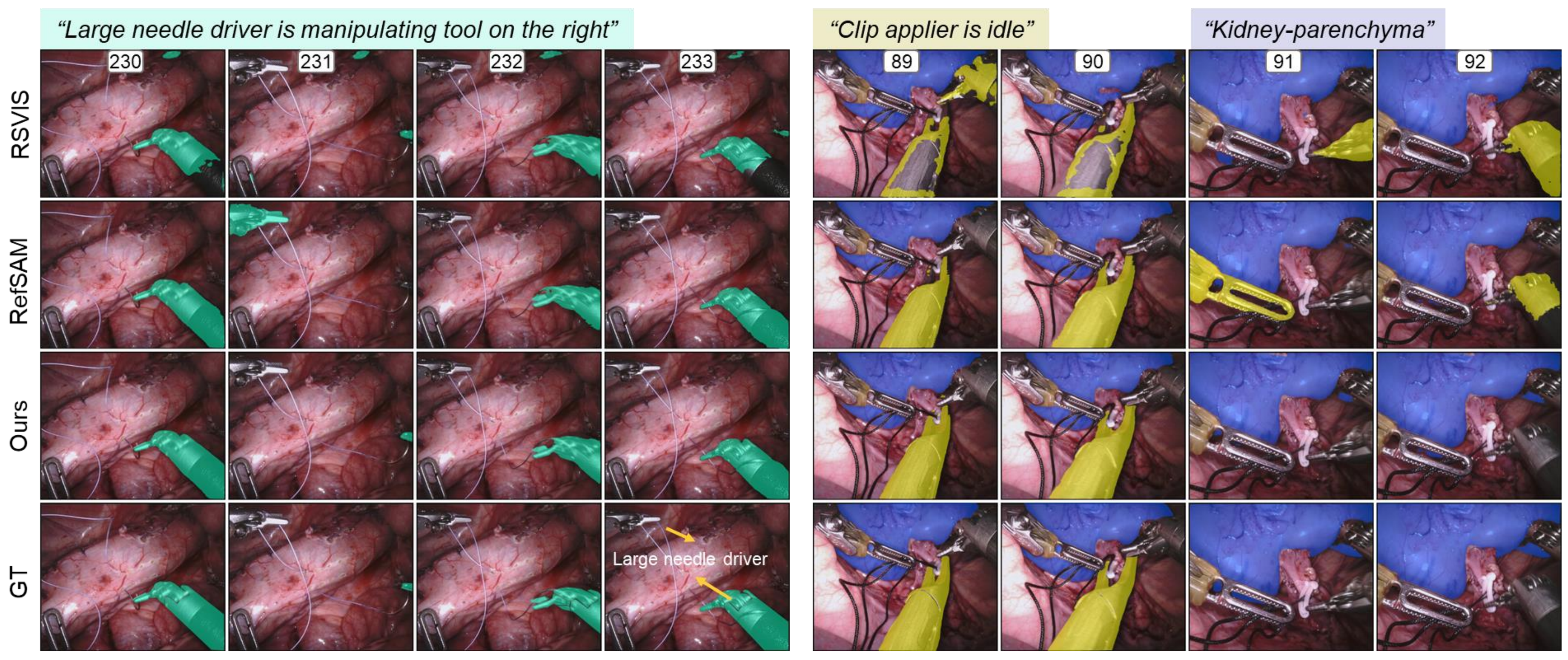


Figure 2. Visual comparison between ReSurgSAM2 and the state-of-the-art.

Ablation Study

Table 3: Component Contribution Analysis

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 4: Memory Bank Comparison

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**

- **Separation** of detection and tracking (+2.64 J&F) – improves short-term temporal modeling.
- **Cross-modal fusion** (+4.77 J&F) – enhances vision–language representation.
- **Credible initialization** (+6.14 J&F) – reduces error accumulation.
- **Diversity long-term memory** (+3.03 J&F) – strengthens long-term tracking.
- **DLM** significantly improves long-term tracking stability compared with different memory variants.

Conclusion

Problem Solved: Hands-free text-driven referring segmentation in long surgical videos.

Technical Achievements:

- 77.73 J&F on Ref-EndoVis17 (+14.17 improvement).
- Real-time performance at 61.2 FPS.
- Robust long-term tracking in hours-long procedures.

Impact: Supports intraoperative guidance, consistent analytics, and surgical training.

Project page HERE