ReMeDI: Refined Memory for Disambiguation of Identities with SAM3 in Surgical Segmentation

Valay Bundele, Mehran Hosseinzadeh, Hendrik P.A. Lensch

University of Tübingen, Germany
{valay.bundele, mehran.hosseinzadeh, hendrik.lensch}@uni-tuebingen.de

Abstract. Accurate surgical instrument segmentation in endoscopy is crucial for computer-assisted interventions, yet remains challenging due to frequent occlusions, rapid motion, and long-term instrument re-entry. While SAM3 provides a powerful spatio-temporal framework for video object segmentation, its performance in surgical scenes is limited by indiscriminate memory updates, fixed memory capacity, and weak identity recovery after occlusions. We propose ReMeDI-SAM3, a training-free extension of SAM3, that addresses these limitations through three components: (i) relevance-aware memory filtering with a dedicated occlusion-aware memory for storing pre-occlusion frames, (ii) a piecewise interpolation scheme that expands effective memory capacity, and (iii) a feature-based re-identification module with temporal voting for reliable post-occlusion identity disambiguation. Together, these components mitigate error accumulation and enable reliable recovery after occlusions. Evaluations on EndoVis17, EndoVis18 and CholecSeg8k under a zero-shot setting show mcIoU improvements of around 5.8%, 8%, and 2% respectively, over vanilla SAM3, outperforming even prior training-based approaches. The code is available at: https://github.com/cgtuebingen/remedi-sam3.

Keywords: Video Object Segmentation, Surgical Video Analysis

1 Introduction

Surgical instrument segmentation is central to computer-assisted interventions, supporting tasks such as tracking, workflow analysis, and intraoperative guidance [8, 20, 24]. However, surgical videos feature long unstructured sequences, frequent occlusions and re-entry, making long-term temporal consistency and identity preservation challenging. Based on the recently introduced general-purpose SAM3 [6], we propose several training-free extensions to address these issues for reliable zero-shot surgical instrument tracking.

With the rise of foundation models, SAM [10] enabled zero-shot prompt-based segmentation, but its direct use in surgical videos is unreliable due to domain shift and prompt dependency [19]. SAM 2 [12] introduced spatio-temporal memory for video segmentation, inspiring extensions such as SAMURAI [16], and DAM4SAM [14], as well as surgical adaptations including MA-SAM2 [18], SAMed-2 [15] and SurgSAM2 [11]. Despite these advances, extended occlusions,

frequent re-entry, and large viewpoint changes still cause identity failures. The recently introduced SAM3 unifies open-vocabulary detection, segmentation, and tracking through spatio-temporal memory, providing a strong baseline for video object segmentation. However, its reliability-agnostic memory update allows low-quality predictions to be written into memory. In surgical videos, where occlusions and visual artifacts are frequent, this causes error accumulation and identity drift, especially after prolonged occlusions and viewpoint changes. While stricter confidence thresholds can mitigate memory contamination, they may also discard low-visibility but identity-critical pre-occlusion frames, exposing a trade-off between temporal stability and reliable identity recovery.

To address these issues, we propose a dual-partitioned memory design. First, a *relevance-aware memory partition* selectively stores high-confidence frames to prevent memory contamination. Second, an *occlusion-aware memory partition* retains pre-occlusion frames under relaxed criteria to preserve identity-critical cues for recovery. As retaining lower-confidence frames increases the risk of identity drift, we further introduce a *feature-based re-identification module* that verifies and corrects recovered identities using multi-scale appearance descriptors, combined with *temporal voting* for robust disambiguation across frames.

Long surgical procedures additionally require long-horizon temporal context, yet SAM3 is constrained by a fixed set of temporal positional encodings, which limits its effective memory capacity and causes informative early frames to be overwritten. We address this with a *novel memory expansion scheme* based on piecewise interpolation of temporal positional encodings, enabling larger memory without retraining. Together, these components form a training-free extension of SAM3 that significantly improves occlusion robustness, long-term tracking stability, and reliable instrument re-identification in videos. To the best of our knowledge, this is the first SAM-based extension that explicitly targets both accurate re-identification and scalable memory. Our main contributions are:

- We introduce a dual-memory design that combines relevance-aware propagation with a dedicated occlusion–aware memory for post-occlusion recovery.
- We incorporate a feature-based re-identification module with temporal voting for explicit identity verification and correction after occlusions.
- We propose a novel memory expansion strategy enabling long-horizon memory retention without any retraining. Overall, our approach shows zero-shot mcIoU gains of 5.8%, 8% and 2% on EndoVis17, EndoVis18 and CholecSeg8k over vanilla SAM3, while also outperforming recent training-based methods.

2 Methodology

2.1 Preliminaries

Let $\mathcal{V} = \{I_1, \ldots, I_T\}$ be an endoscopic video, where each frame $I_t \in \mathbb{R}^{H\times W\times 3}$. The task is to predict class-level segmentation maps, $\mathcal{S}_t = \{P_t^{(i)}\}_{i=1}^{N}$ for each frame, where $P_t^{(i)} \in \{0,1\}^{H\times W}$ is the binary mask of instrument i, and N is

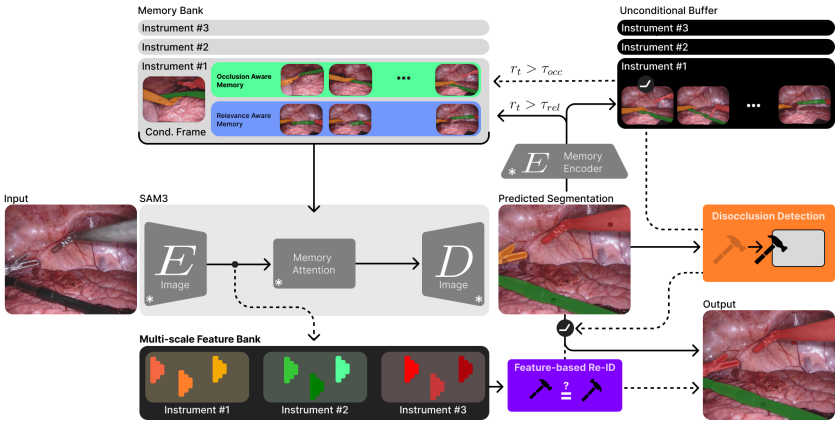

Fig. 1. ReMeDI-SAM3. We extend SAM3 with a dual-memory design and a feature-based ReID module. For each instrument, the memory is divided into a *relevance-aware memory* that stores high-confidence entries and an *occlusion-aware memory* that is populated upon disocclusion using lower-confidence pre-occlusion frames drawn from an Unconditional Buffer that stores all past frames. When disocclusion is detected (tool reappears), occlusion-aware memory is first updated, after which feature-based ReID module verifies or reassigns predicted identity using a multi-scale feature bank.

the number of instruments. SAM3 is a video object segmentation model that extends SAM2 with text-promptable segmentation. Given an object prompt (text, points, boxes, or masks) on a reference frame t_0, it propagates the object state to predict masks $\{P_t\}_{t=t_0}^{T}$. Its architecture comprises an image encoder, a detector, a prompt encoder, a memory bank, and a mask decoder. The image encoder and detector share a unified vision backbone [5], and the detector follows a DETR-based open-vocabulary design [7]. Sparse prompts are embedded by the prompt encoder to condition the mask decoder. The mask-conditioned features from the prompted frame and the six most recent frames are stored in memory. Current-frame features attend to this memory to produce temporally consistent masks. To handle ambiguity, SAM3 predicts multiple candidate masks with confidence scores and selects the highest-quality one, updating memory in a FIFO manner. While effective in general domains, this indiscriminate memory update limits long-term consistency under severe occlusions and frequent re-entry in surgery.

2.2 ReMeDI-SAM3

We propose **ReMeDI-SAM3**: **Re**fined **Me**mory for **D**isambiguation of **I**dentities with **SAM3**, a training-free extension of SAM3 to enhance temporal consistency and identity preservation in surgical videos. The pipeline is shown in Figure 1. Our approach (1) restructures SAM3 memory into two components: (i) a *relevance-aware memory* that admits only high-confidence frames, and (ii) an *occlusion–aware memory* that selectively retains pre-occlusion appearance cues.

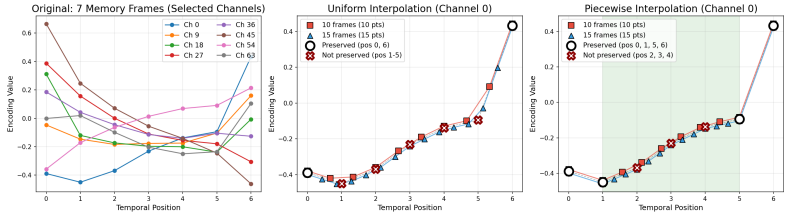

Fig. 2. Visualization of temporal positional encodings and memory expansion strategies. Left: select channels of original temporal positional embeddings. Mid: uniform interpolation distributes new positions evenly. Right: piecewise interpolation preserves boundary embeddings and samples new positions only in interior region.

Building upon this design, we further introduce (2) a *novel memory expansion scheme*, and (3) a *feature-based re-identification module with temporal voting* for robust post-occlusion identity verification and correction.

Relevance-Aware and Occlusion-Aware Memory In SAM3, a fixed-size memory bank retains recent frames regardless of prediction reliability, which in surgical videos often introduces noisy masks and causes error accumulation. To mitigate this, we split the total memory of size M into two components: a *relevance-aware memory* for stable long-term tracking and an *occlusion-aware memory* for post-occlusion recovery, each allocated $M/2$ slots.

Relevance-Aware. For each predicted mask P_t, SAM3 outputs a quality score c_t and an objectness score s_t, from which a reliability score is defined as, $r_t = s_t \cdot c_t$. The relevance-aware memory consists of recent frames whose reliability exceeds a threshold τ_{rel}:

$$\mathcal{U}_{\text{rel}} = \text{Top}_{M/2}(\{I_t \mid r_t \geq \tau_{\text{rel}}\}), \qquad |\mathcal{U}_{\text{rel}}| \leq \frac{M}{2},$$

where $\text{Top}_{M/2}(\cdot)$ denotes selecting the temporally most recent $M/2$ frames. This gating limits memory updates to reliable frames, stabilizing propagation.

Occlusion-Aware. Just before occlusion, instruments often exhibit reduced visibility and thus lower reliability scores, despite carrying critical identity cues for re-identification. To preserve this information, we maintain an *unconditional buffer* $\mathcal{U}$ that stores all past frames irrespective of r_t. Upon detecting an occlusion recovery event (i.e., when s_t transitions from zero to positive), we populate the occlusion-aware memory by selecting the $M/2$ most recent frames from the unconditional buffer that satisfy a relaxed reliability constraint:

$$\mathcal{U}_{\text{occ}} = \text{Top}_{M/2}(\{I_t \in \mathcal{U} \mid r_t \geq \tau_{\text{occ}}\}), \qquad |\mathcal{U}_{\text{occ}}| \leq \frac{M}{2},$$

where $\tau_{\text{occ}} < \tau_{\text{rel}}$. This helps preserve identity-discriminative appearance cues.

Memory Capacity Expansion Although SAM3 supports a configurable memory size, it uses only M=7 fixed temporal positional embeddings, limiting reliable indexing for long videos. This is problematic in surgical scenarios with severe occlusions and large appearance changes. Let $\{\mathbf{p}_0, \ldots, \mathbf{p}_6\}$ denote the original embeddings. As shown in Figure 2, the boundary segments $[0, 1]$ and $[5, 6]$ differ from the interior $[1, 5]$, indicating stronger temporal priors at the extremes. We therefore keep the boundary segments unchanged and interpolate only within the interior (see Figure 2 (right)). For a target memory size $M > 7$, we fix the boundary encodings as $\tilde{\mathbf{p}}_0 = \mathbf{p}_0$ and $\tilde{\mathbf{p}}_{M-1} = \mathbf{p}_6$. The remaining $M-2$ encodings are obtained by linearly resampling the interior sequence $(\mathbf{p}_1, \ldots, \mathbf{p}_5)$ to $M-2$ uniformly spaced positions. Specifically, for $k \in \{1, \ldots, M-2\}$,

$$t_k = \frac{k-1}{M-3}, \qquad u_k = 1 + 4t_k,$$

$$\tilde{\mathbf{p}}_k = (1-\alpha_k)\mathbf{p}_{\lfloor u_k \rfloor} + \alpha_k \mathbf{p}_{\lceil u_k \rceil}, \quad \alpha_k = u_k - \lfloor u_k \rfloor.$$

This corresponds to linear interpolation of $(\mathbf{p}_1, \ldots, \mathbf{p}_5)$ with aligned endpoints. The resulting encodings preserve boundary semantics while enabling denser temporal indexing in the interior region, effectively expanding the memory capacity.

Feature-Based Re-Identification Despite improved memory, long occlusions can still cause identity drift at disocclusions, particularly because the occlusion-aware memory admits lower-confidence pre-occlusion frames to preserve identity cues. To address this, we introduce a feature-based re-identification (ReID) module that validates and corrects identity upon recovery using appearance descriptors, aggregating predictions over a temporal window to ensure robustness.

For each instrument class i, we maintain a feature bank $\mathcal{B}^i$ constructed from selected frames observed up to time t. A frame contributes to $\mathcal{B}^i$ only if (i) it has a high reliability score r_t, and (ii) its prediction certainty is high, measured by bounding-box IoU agreement among three candidate masks. From each valid frame, we extract multi-scale appearance features by averaging the backbone feature map within the predicted mask: $\mathbf{f}^i_{t,l} = \frac{1}{|M^i_t|} \sum_{x \in M^i_t} \mathbf{F}_{t,l}(x)$, where $\mathbf{F}_{t,l}$ is the backbone feature map at scale l and M^i_t is the predicted mask. The resulting multi-scale descriptors $\{\mathbf{f}^i_{t,l}\}_l$ are appended to $\mathcal{B}^i$ and updated online.

An occlusion is detected when the objectness score drops to zero. When a non-zero score reappears, a recovery phase is triggered and identity is verified over the next K frames. Let i be the predicted class at recovery. We compute self-similarity to $\mathcal{B}^i$ and cross-class similarity to banks of other classes $\mathcal{B}^j$ using cosine similarity aggregated across scales and averaged over the K-frame window: $s^{\text{self}} = \frac{1}{K} \sum_t s^{\text{self}}_t$, $s^{\text{other}} = \frac{1}{K} \sum_t s^{\text{other}}_t$. The identity is accepted if $s^{\text{self}} \geq s^{\text{other}}$. If another class yields higher similarity, the label is reassigned to that class.

Final Inference Pipeline Each instrument is initialized from its first visible mask and tracked independently with a dedicated memory. Memory updates are regulated by relevance-aware filtering. Upon reappearance after occlusion,

the occlusion-aware memory is updated using pre-occlusion frames with a lower confidence threshold, followed by prediction and feature-based re-identification for identity verification and correction. Finally, predictions from all instruments are fused via quality-weighted mask fusion to obtain the final segmentation map.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate on three public benchmarks, EndoVis2017 [2], EndoVis2018 [1], and CholecSeg8k [9]. EndoVis2017 contains eight annotated sequences (225 frames each), pre-processed following [13]; we evaluate on all sequences and compare against prior training-based 4-fold cross-validation results. EndoVis2018 includes 11 training and 4 validation videos (149 frames each); we report performance on validation set for fair comparison. Both datasets provide annotations for seven instrument categories. CholecSeg8k is a surgical segmentation dataset with 8080 frames for multiple surgical structures (13 different categories). For evaluation, we report Challenge IoU (cIoU) [2], which computes IoU only for instruments present in each frame, along with IoU and mean class IoU (mcIoU).

3.2 Implementation details

We set $\tau_{rel} = 0.95$ and $\tau_{occ} = 0.75$, and perform re-identification over a window of $K = 3$ frames. Hyperparameters are selected on the EndoVis18 train set and kept fixed across datasets. All experiments are conducted on an RTX 4090 GPU.

3.3 Results

Quantitative Comparison. We evaluate our method for class-level segmentation on EndoVis17, EndoVis18, and CholecSeg8k, comparing against specialist surgical segmentation models and SAM-based approaches, including ISINet [8], S3Net [4], MATIS [3], TP-SIS [24], TrackAnything [17], PerSAM [23], Surgical-SAM [22], SP-SAM [21], MA-SAM2 [18], and SAM3. All results of our method are obtained in a training-free setting, and we report global and class-wise IoUs.

Table 1 summarizes the results. ReMeDI-SAM3 consistently outperforms vanilla SAM3 across all benchmarks. On EndoVis17, we achieve improvements of 6% IoU and 5.8% mcIoU. Across 8 sequences, ReMeDI achieves 74.50 ± 14.13 mcIoU versus 69.04 ± 17.75 for SAM3, corresponding to a 20% reduction in sequence-level standard deviation. The improvement is statistically significant (Wilcoxon p=0.023), with ReMeDI outperforming SAM3 on 7/8 sequences and a 95% cluster-bootstrap CI of [0.64, 8.13]. On EndoVis18, we obtain gains of 3.5% IoU and 8% mcIoU. The larger mcIoU gain reflects improved suppression of false positives for absent instruments via our identity-aware design. For example, in Sequence 5 of EndoVis18, SAM3 misclassifies a returning Prograsp Forceps (PF) as Ultrasound Probe (UP), whereas our method correctly restores

Table 1. Quantitative comparison on the EndoVis17 and EndoVis18 datasets (%). The best results are in **bold**, second-best are underlined. SAM-family methods use first-appearance GT-mask prompt; specialist ones follow their original settings.

	Category	Method	Challenge IoU	IoU	mcIoU	Instrument Categories								
						BF	PF	LND	MCS	UP	VS	GR	SI	CA
EndoVis17	Specialist	ISINet	55.62	52.20	28.96	38.70	38.50	50.09	28.72	12.56	27.43	2.10	-	-
		S3Net	72.54	71.99	46.55	**75.08**	54.32	61.84	43.23	28.38	35.50	27.47	-	-
		MATIS Frame	68.79	62.74	37.30	66.18	50.99	52.23	19.27	23.90	32.84	15.71	-	-
		TP-SIS	63.37	63.37	52.74	66.42	45.46	75.20	44.02	34.67	73.44	29.95	-	-
	SAM-based	TrackAnything	67.41	64.50	62.97	55.42	44.46	62.43	67.03	65.17	**83.68**	62.59	-	-
		PerSAM (Zero-Shot)	42.47	42.47	41.80	53.99	25.89	50.17	47.33	38.16	52.87	24.24	-	-
		SurgicalSAM	69.94	69.94	67.03	68.30	51.77	75.52	**86.95**	60.80	68.24	57.63	-	-
		SP-SAM	73.94	73.94	71.06	68.89	53.16	**83.80**	84.91	61.05	73.20	**72.40**	-	-
		MA-SAM2 (Zero-Shot)	62.49	62.49	59.89	54.41	50.41	64.73	72.64	70.85	73.72	32.66	-	-
		SAM3 (Mask, Zero-Shot)	78.08	70.76	68.42	55.62	65.30	76.33	75.32	**90.09**	80.51	38.36	-	-
		ReMeDI-SAM3 (Ours)	**81.34**	**76.65**	**74.29**	69.00	**68.97**	77.14	78.82	89.96	79.99	56.15	-	-
EndoVis18	Specialist	ISINet	73.03	70.94	40.21	73.83	48.61	30.98	88.16	2.16	-	-	37.68	0.00
		S3Net	75.81	74.02	42.58	77.22	50.87	19.83	92.12	7.44	-	-	50.59	0.00
		MATIS Frame	82.37	77.01	48.65	83.35	38.82	40.19	**93.18**	16.17	-	-	64.49	4.32
		TP-SIS	84.92	83.61	65.44	84.28	73.18	78.88	66.67	39.12	-	-	**92.20**	23.73
	SAM-based	TrackAnything	65.72	60.88	38.60	72.90	31.07	64.73	61.05	17.93	-	-	10.24	12.28
		PerSAM (Zero-Shot)	49.21	49.21	34.55	51.26	34.40	46.75	52.28	25.62	-	-	16.45	15.07
		SurgicalSAM	80.33	80.33	58.87	83.66	65.63	58.75	88.56	21.23	-	-	54.48	39.78
		SP-SAM	84.24	84.24	65.71	87.60	65.07	61.95	92.08	34.99	-	-	58.30	**59.96**
		SAM3 (Mask, Zero-Shot)	88.04	81.82	66.46	82.25	**74.92**	**88.14**	88.38	44.78	-	-	53.75	32.96
		ReMeDI-SAM3 (Ours)	**88.10**	**85.34**	**74.37**	**87.73**	72.65	86.60	88.80	**80.44**	-	-	65.38	39.01

PF and suppresses UP, yielding a 36% IoU improvement for UP. Although not best in every individual category, our method achieves consistently strong performance across classes, leading to superior overall metrics. On CholecSeg8k, a broader surgical semantic segmentation benchmark, we observe consistent gains of 1% IoU and 2% mcIoU as shown in Table 5. Overall, ReMeDI-SAM3 surpasses all zero-shot baselines and even outperforms several training-based approaches, demonstrating the effectiveness and robustness of our design.

Qualitative Comparison. Figure 3 illustrates a challenging occlusion case on EndoVis17. The yellow instrument (Bipolar Forceps) exits at T=75, and a second blue instrument (Prograsp Forceps) enters later (T=126–132). ReMeDI-SAM3 briefly misses the new instrument at T=126 but correctly assigns the blue identity once sufficient evidence accumulates. In contrast, SAM3 preserves the old yellow identity after occlusion. This shows that ReMeDI-SAM3 reliably recovers correct identities even after instrument turnover, highlighting the effectiveness of our method in maintaining identity consistency under occlusions.

Efficiency/Scalability. SAM3 already maintains per-object memory by fusing per-object masks with image features; ReMeDI differs in using a larger memory bank and a different frame-selection strategy. On EndoVis18 with a single RTX 4090, ReMeDI incurs modest overhead compared to SAM3: 6,299 vs 6,010 MB peak GPU memory (+4.8%), 4,820 vs 4,767 MB average GPU memory (+1.1%), and 351.1 vs 329.7 ms/frame latency (+6.5%). The expanded memory bank (M=15 vs 7) adds only 5% to memory-attention cost since compute is dominated by image encoder. To ensure scalability, unconditional buffer has a maximum size of 15 while ReID bank is capped at 15 entries/object.

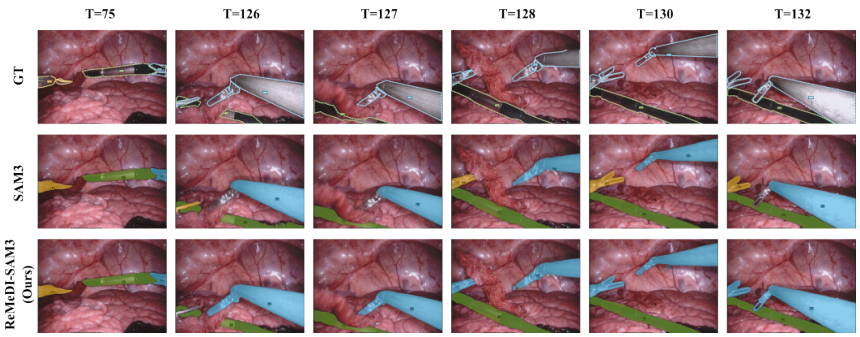

Fig. 3. ReMeDI-SAM3 vs SAM3. ReMeDI-SAM3 correctly detects the second blue instance entering later while SAM3 confuses it with the occluded yellow instrument.

Table 2. Effect of Memory Expansion (%) on SAM3 + RM + OM + ReID.

Memory Size	Challenge IoU	IoU	mcIoU
7	80.58	75.84	73.59
15	**81.34**	**76.65**	**74.29**
20	80.77	76.11	73.57

Table 3. Uniform vs. Piecewise Temporal Positional Encoding on ReMeDI-SAM3 (M=15).

Method	Challenge IoU	IoU	mcIoU
Uniform	80.13	75.66	74.20
Piecewise	**81.34**	**76.65**	**74.29**

3.4 Ablation Studies

We analyze the impact of memory size, temporal interpolation, and components of ReMeDI-SAM3: relevance-aware memory (RM), occlusion-aware memory (OM), memory expansion (ME), and feature-based re-identification (ReID).

Model Components. Table 4 analyzes the individual and combined contributions of different model components on EndoVis17. Relevance-aware memory filtering provides a 3.5% mcIoU gain over SAM3 by suppressing noisy updates. Adding the occlusion-aware memory partition yields a further 0.5% mcIoU. The feature-based re-identification module contributes 1.4% IoU by correcting post-occlusion identity drift, while expanding memory adds 0.8% IoU. Overall, the final model achieves a total improvement of 6.0% IoU and 5.9% mcIoU, confirming that the components are jointly essential for robust long-horizon segmentation.

Memory Expansion. Expanding memory with our proposed interpolation strategy yields clear performance gains on EndoVis17 (Table 2). Increasing the memory to 15 frames improves all metrics by around 0.8% , indicating the benefit of additional temporal context. Further expansion slightly degrades performance, suggesting that overly large memory might introduce less informative context.

Interpolation Scheme. Replacing the proposed piecewise interpolation with uniform interpolation for memory expansion causes a drop of about 1.2% cIoU and 1% IoU on EndoVis17 (Table 3). Uniform resampling distorts learned boundary temporal priors whereas piecewise interpolation preserves the semantic roles of the earliest and latest memory positions while densifying the interior.

Table 4. Impact of individual components of ReMeDI-SAM3 on EndoVis17.

Method	Challenge IoU	IoU	mcIoU
Vanilla SAM3	78.08	70.76	68.42
+ RM	79.94	73.74	72.03
+ RM + OM	80.93	74.46	72.53
+ RM + OM + ReID	80.58	75.84	73.58
+ RM + OM + ReID + ME	**81.34**	**76.65**	**74.29**

Table 5. Zero-shot performance comparison on CholecSeg8k.

Method	Challenge IoU	IoU	mcIoU
PerSAM (Mask)	33.67	33.54	25.88
SAM3 (Mask)	88.33	87.81	79.37
ReMeDI-SAM3 (Ours)	**89.68**	**89.07**	**81.25**

4 Conclusion

We present ReMeDI-SAM3, a training-free extension of SAM3 for robust surgical instrument segmentation. Our framework integrates relevance-aware memory filtering for stable long-term propagation, occlusion-aware memory and feature-based re-identification for reliable post-occlusion recovery, and a memory expansion strategy for long-horizon reasoning. Extensive evaluations on EndoVis17, EndoVis18 and CholecSeg8k show consistent improvements over SAM3 and prior methods, showing superior identity preservation and false-positive suppression.

Acknowledgments. The work described in this paper was conducted in the framework of the Graduate School 2543/1 “Intraoperative Multi-Sensory Tissue Differentiation in Oncology" (project ID 40947457) funded by the German Research Foundation (DFG - Deutsche Forschungsgemeinschaft). This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) – EXC number 2064/1 – Project number 390727645 and SFB 1233, TP 2, Project number 276693517. The authors thank International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting V. Bundele and M. Hosseinzadeh.

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. Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: Matis: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
4. Baby, B., Thapar, D., Chasmai, M., Banerjee, T., Dargan, K., Suri, A., Banerjee, S., Arora, C.: From forks to forceps: A new framework for instance segmentation of surgical instruments. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6191–6201 (2023)
5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual em-

beddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
8. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 595–605. Springer (2020)
9. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
10. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint arXiv:2408.07931 (2024)
12. 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)
13. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). pp. 624–628. IEEE (2018)
14. Videnovic, J., Lukezic, A., Kristan, M.: A distractor-aware memory for visual object tracking with sam2. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24255–24264 (2025)
15. Yan, Z., Song, S., Song, D., Li, Y., Zhou, R., Sun, W., Chen, Z., Kim, S., Ren, H., Liu, T., et al.: Samed-2: Selective memory enhanced medical segment anything model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–550. Springer (2025)
16. Yang, C.Y., Huang, H.W., Chai, W., Jiang, Z., Hwang, J.N.: Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv preprint arXiv:2411.11922 (2024)
17. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
18. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented sam2 for training-free surgical video segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–337. Springer (2025)
19. Yu, J., Bai, L., Wang, G., Wang, A., Yang, X., Gao, H., Ren, H.: Adapting sam for surgical instrument tracking and segmentation in endoscopic submucosal dissection videos. arXiv preprint arXiv:2404.10640 (2024)
20. Yuan, C., Ban, Y.: Temporal propagation of asymmetric feature pyramid for surgical scene segmentation. arXiv preprint arXiv:2504.13440 (2025)
21. Yue, W., Zhang, J., Hu, K., Wu, Q., Ge, Z., Xia, Y., Luo, J., Wang, Z.: Surgicalpart-sam: Part-to-whole collaborative prompting for surgical instrument segmentation. arXiv preprint arXiv:2312.14481 (2023)

22. 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)
23. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
24. 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)