Memory-Enhanced SAM3 for Occlusion-Robust Surgical Instrument Segmentation

Valay Bundele[1] 
Mehran Hosseinzadeh[1]
Hendrik P. A. Lensch[1]
[1] *University of Tuebingen*

VALAY.BUNDELE@UNI-TUEBINGEN.DE
MEHRAN.HOSSEINZADEH@UNI-TUEBINGEN.DE
HENDRIK.LENSCH@UNI-TUEBINGEN.DE

Editors: Under Review for MIDL 2026

Abstract

Accurate surgical instrument segmentation in endoscopic videos is crucial for computer-assisted interventions, yet remains challenging due to frequent occlusions, rapid motion, specular artefacts, 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 memory-enhanced 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 the 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 and EndoVis18 under a zero-shot setting show absolute mcIoU improvements of around 7% and 16%, respectively, over vanilla SAM3, outperforming even prior training-based approaches.

Keywords: Video Object Segmentation, Surgical Video Analysis

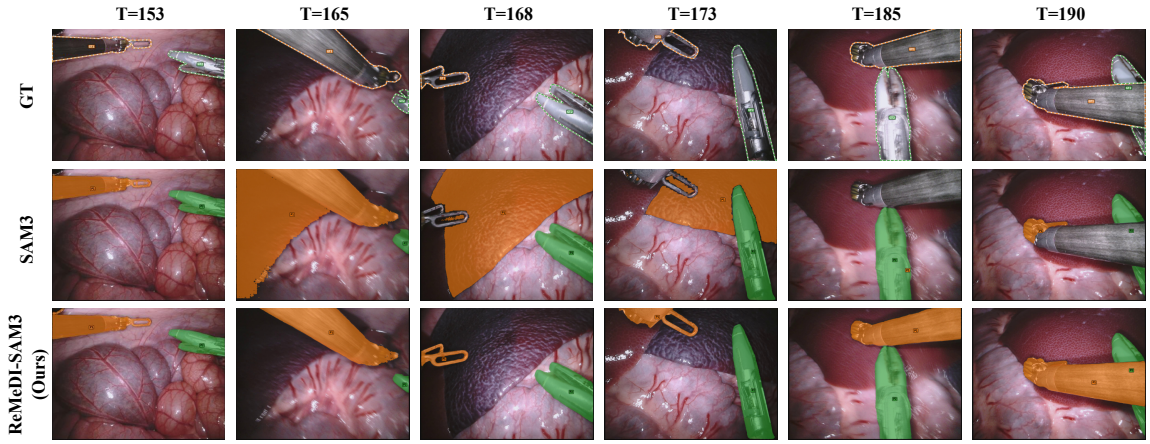

Figure 1: SAM3 vs ReMeDI-SAM3 (ours). The orange-labeled instrument gets occluded after $T = 153$ and re-appears at $T = 165$. While SAM3 produces false positives after its re-appearance, ReMeDI-SAM3 maintains consistent identities across occlusion and re-entry.

1. Introduction

Surgical instrument segmentation is central to computer-assisted interventions, supporting tasks such as tracking, workflow analysis, and intraoperative guidance. However, surgical videos feature long unstructured sequences, frequent occlusions and re-entry, and severe visual clutter from smoke, blood, and deformable tissues, making long-term temporal consistency and identity preservation challenging. Based on the recently introduced general-purpose SAM3 (Carion et al., 2025), we propose several training-free extensions to address these issues for reliable zero-shot surgical instrument tracking.

With the rise of foundation models, SAM (Kirillov et al., 2023) enabled zero-shot prompt-based segmentation, but its direct use in surgical videos is unreliable due to domain shift and prompt dependency (Yu et al., 2024a). SAM 2 (Ravi et al., 2024) introduced spatio-temporal memory for video segmentation, inspiring extensions such as SAMURAI (Yang et al., 2024) and DAM4SAM (Videnovic et al., 2025), as well as surgical adaptations including MA-SAM2 (Yin et al., 2025) and SAMed-2 (Yan et al., 2025). 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 using spatio-temporal memory, making it a strong foundation for video object segmentation. However, it allows low quality predictions to be inserted into the memory. In surgical scenes, this causes noisy masks—due to occlusion, smoke, and rapid camera motion—to be propagated, leading to progressive error accumulation. To mitigate this, we introduce *relevance-aware memory filtering* that gates memory insertion using intrinsic confidence signals, preventing unreliable predictions from contaminating memory.

Even with improved memory quality, SAM3 shows weak re-identification after long occlusions, especially when instruments reappear with large viewpoint changes or partial visibility (see Figure 1). While strict confidence thresholds stabilize general tracking, pre-occlusion frames often contain identity-critical cues despite having low confidence due to poor visibility. We therefore introduce a *occlusion–aware memory partition* that selectively retains such pre-occlusion frames under relaxed thresholds around occlusion events. While crucial for recovery, admitting lower-confidence frames also increases risk of propagating incorrect identities. To address this trade-off, we further incorporate a *feature-based re-identification module* that verifies and corrects recovered identities using multi-scale appearance features, complemented by *temporal voting* to ensure robust disambiguation when single-frame evidence is insufficient.

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 capacity. Our main contributions are as follows:

- We introduce a dual-memory design that combines relevance-aware propagation with a dedicated occlusion–aware memory for post-occlusion recovery.

- We incorporate a lightweight 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 requiring neither retraining nor architectural modifications. Our proposed approach shows zero-shot mcIoU gains of around 7% on EndoVis17 and 16% on EndoVis18 over vanilla SAM3, while also outperforming recent training-based methods.

2. Related Work

SAM-based visual tracking.

Segment Anything Model (SAM) introduced a promptable framework for general-purpose zero-shot image segmentation, later extended to video with SAM2 via a memory-based segmentation paradigm. Subsequent work focused on improving SAM2's memory reliability. DAM4SAM (Videnovic et al., 2025) proposes a distractor-aware memory that stores anchor frames to disambiguate targets while SAM2Long (Ding et al., 2025) explores multiple segmentation hypotheses to derive video-level optimal results. Motion-aware strategies such as SAMURAI (Yang et al., 2024) leverages motion modeling to re-rank mask predictions, while HiM2SAM (Chen et al., 2025) introduces motion-guided filtering to retain high-confidence frames. More recently, SAM3 unified open-vocabulary detection, segmentation, and tracking within a single spatiotemporal memory framework. Despite these advances, SAM3 still struggles with reliable re-identification and long-horizon tracking under extended occlusions and large appearance changes.

Surgical Instrument Segmentation. Surgical instrument segmentation in endoscopic videos is challenging due to sparse annotations, limited field of view, frequent occlusions, and rapid motion. ISINet (González et al., 2020) performs instance-aware detection with temporal consistency, while TAFPNet (Yuan and Ban, 2025) combines convolutional and transformer features with asymmetric temporal propagation. Text-promptable formulations have also been explored (Zhou et al., 2023), leveraging CLIP (Radford et al., 2021) for language-guided segmentation. With the emergence of SAM, several works adapted SAM and SAM2 to surgery. Yu et al. (2024b) demonstrate zero-shot SAM2 performance using point and box prompts, while domain-adapted foundation models such as BioSAM-2 (Yan et al.) improve generalization via large biomedical pretraining. Other approaches fine-tune SAM for surgical use: Yu et al. (2024a) apply LoRA adapters, SurgicalSAM (Yue et al., 2024) introduces class-level prototype prompts, and SP-SAM (Yue et al., 2023) focuses on part-level prompting. More recent works emphasize memory-based video segmentation for long-term tracking. SurgSAM2 (Liu et al., 2024) employs frame pruning for efficient memory usage, MA-SAM2 (Yin et al., 2025) introduces distractor-aware and multi-hypothesis memory mechanisms, while SAMed-2 (Yan et al., 2025) and MedSAM-2 (Zhu et al., 2024) enhance temporal memory using confidence-based replacement and dissimilarity-driven frame selection.

Unlike prior works, our approach enhances the SAM3 memory with explicit focus on post-occlusion recovery and long-horizon temporal retention. We introduce a dual-memory design for reliable propagation and re-identification, expand the temporal memory horizon via positional interpolation, and incorporate feature-based re-identification, capabilities not addressed by existing SAM-based surgical video segmentation methods.

3. Methodology

3.1. Problem Formulation

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 the number of instruments.

3.2. Preliminaries: SAM3

SAM3 is a video object segmentation model that extends SAM2 with text-promptable segmentation. Given a video $\mathcal{V} = \{I_1, \ldots, I_T\}$ and an object prompt (text, points, boxes, or masks) at a reference frame t_0, SAM3 propagates the object state to predict masks $\{P_t\}_{t=t_0}^{T}$ over time. Its architecture comprises an image encoder, detector, prompt encoder, memory bank, and mask decoder. Sparse prompts are embedded by the prompt encoder to condition the mask decoder. The image encoder and detector share a unified vision backbone (Bolya et al., 2025), and the detector follows a DETR-based open-vocabulary design (Carion et al., 2020). The mask-conditioned features from the prompted frame and the six most recent frames are stored in the 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, after which the memory is updated in a FIFO manner. While effective in general domains, this indiscriminate memory update and fixed memory capacity limit long-term consistency and identity preservation under severe occlusions and frequent re-entry in surgical videos.

3.3. 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 2. Our approach (1) restructures SAM3 memory into two components: (i) a *relevance-aware memory* that admits only high-confidence frames to stabilize temporal propagation, and (ii) an *occlusion–aware memory* that selectively retains pre-occlusion appearance cues to facilitate identity recovery. 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. Together, these components improve robustness to occlusions, tracking stability, and reliable instrument re-entry handling in surgical videos.

3.3.1. Relevance-Aware and Occlusion-Aware Memory

In SAM3, the memory bank retains a fixed number of recent frames even when their prediction reliability is low. In surgical videos, this results in frequent insertion of noisy or ambiguous masks, leading to memory contamination and error accumulation. To address this, we partition the total memory of size M into two complementary components: a *relevance-aware memory* for stable long-term tracking, and an *occlusion-aware memory* dedicated to post-occlusion identity 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

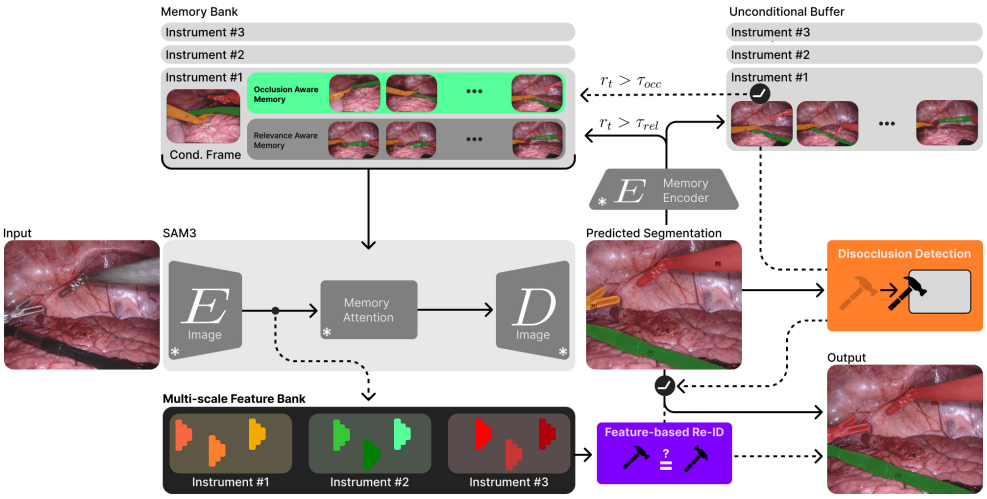

Figure 2: ReMeDI-SAM3 pipeline. We extend SAM3 with a dual-memory design and a feature-based re-identification module. For each instrument, the memory is divided into a *relevance-aware memory* that stores high-confidence entries and an *occlusion-aware memory* that is populated upon disocclusion using lower-confidence pre-occlusion frames drawn from an Unconditional Buffer that stores all past frames (3.3.1). When disocclusion is detected (tool reappears), occlusion-aware memory is first updated, after which feature-based ReID module verifies or reassigns the predicted identity using a multi-scale feature bank (3.3.3).

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 temporal 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 design preserves identity-discriminative appearance cues for post-occlusion recovery while keeping the propagation memory free of ambiguous updates.

3.3.2. Memory Capacity Expansion

Although SAM3 supports a configurable memory size, it uses only M=7 fixed temporal positional embeddings, which limits reliable indexing of larger memory. This is restrictive for long surgical videos where instruments undergo extended occlusions and re-enter with

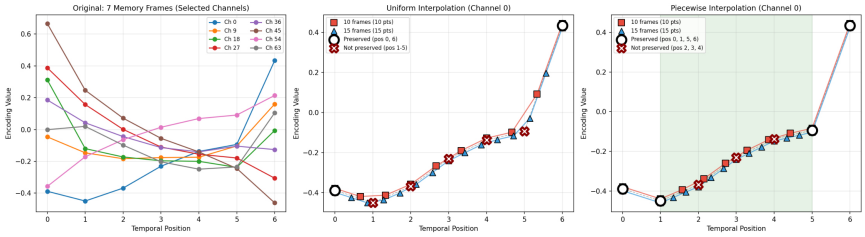

Figure 3: Visualization of temporal positional encodings and memory expansion strategies. Left: select channels of original temporal positional embeddings. Mid: uniform interpolation distributes new positions evenly over entire temporal range. Right: piecewise interpolation preserves boundary embeddings and samples new positions only in interior region.

significant pose and appearance changes. Let $\{\mathbf{p}_0, \ldots, \mathbf{p}_6\}$ denote the original temporal positional embeddings. As shown in Figure 3 (left), the boundary segments $[0, 1]$ and $[5, 6]$ exhibit different trends from the interior region $[1, 5]$, indicating stronger temporal priors at the earliest and latest memory positions. We therefore preserve the boundary segments and interpolate only within the interior interval (see Figure 3 (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 the boundary semantics while enabling denser temporal indexing in the interior region, effectively expanding the memory capacity of SAM3 without retraining.

3.3.3. Feature-Based Re-Identification post occlusions

Despite improved memory quality and capacity, prolonged occlusions can still cause identity drift upon disocclusion, particularly because occlusion-aware memory admits lower-confidence pre-occlusion frames to preserve identity cues. To address this, we introduce a *feature-based re-identification (re-ID)* mechanism that validates and corrects instrument identity upon recovery using appearance descriptors, aggregating predictions over a *temporal window* to ensure robustness as single-frame evidence can be unreliable.

Reference Feature Bank. For each instrument class i, we maintain a reference 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 using the SAM3 image encoder. At scale l, the feature is computed 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 level l and M_t^i is the predicted mask. The resulting multi-scale descriptors $\{\mathbf{f}_{t,l}^i\}_l$ are appended to $\mathcal{B}^i$ and updated online.

Occlusion Detection and Recovery Window. An occlusion event for instrument i is detected when its predicted objectness score s_t drops to zero. Once a non-zero objectness score is re-observed, a recovery phase is triggered. Identity verification from a single frame is inherently unreliable, even with strong appearance descriptors. To mitigate this, we perform re-identification over a temporal window of K consecutive frames.

Identity Verification via Temporal Voting. Let the class predicted by SAM3 for recovered instance at frame t be i. We denote by $\mathbf{f}_{t,l}^{\text{cur}}$ the appearance feature of the recovered instance at scale l, and by $\mathcal{B}^i$ the reference feature bank maintained for class i. We define *self-similarity* as the average similarity to the reference bank of class i, and *cross-instrument similarity* as the maximum average similarity to any other class.

$$s_t^{\text{self}} = \frac{1}{|\mathcal{B}^i|} \sum_{\mathbf{f} \in \mathcal{B}^i} \sum_l \cos\big(\mathbf{f}_{t,l}^{\text{cur}}, \mathbf{f}_l\big), \qquad s_t^{\text{other}} = \max_{j \neq i} \left(\frac{1}{|\mathcal{B}^j|} \sum_{\mathbf{f} \in \mathcal{B}^j} \sum_l \cos\big(\mathbf{f}_{t,l}^{\text{cur}}, \mathbf{f}_l\big) \right).$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and the summation aggregates across all scales. These frame-wise scores are temporally averaged over a recovery window of K frames: $s^{\text{self}} = \frac{1}{K} \sum_t s_t^{\text{self}}$, and $s^{\text{other}} = \frac{1}{K} \sum_t s_t^{\text{other}}$. In parallel, we compute the mean spatial overlap $\overline{\text{IoU}}$ between the recovered mask and all other tracked instruments over the same window.

The recovered identity i is accepted if $s^{\text{self}} - s^{\text{other}} \geq \delta_{\text{sim}}$ and $\overline{\text{IoU}} \leq \delta_{\text{iou}}$, where δ_{sim} and δ_{iou} are fixed verification thresholds. This enforces that the recovered instance is more similar to its own reference features than to any other instrument class by a sufficient margin, while also maintaining low spatial overlap with competing instances. If this condition is violated but $s^{\text{other}} - s^{\text{self}} \geq \delta_{\text{sim}}^-$ for some other class j, the identity is reassigned to that class j. Otherwise, the frame is rejected and excluded from future reference bank updates to prevent feature contamination.

3.3.4. 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 final segmentation map.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate on two public benchmarks, EndoVis2017 (Allan et al., 2019) and EndoVis2018 (Allan et al., 2020). EndoVis2017 contains eight annotated sequences (225 frames each), preprocessed following Shvets et al. (2018); we evaluate on all sequences and compare against prior training-based 4-fold cross-validation results. EndoVis2018 includes 11 training and

Table 1: Quantitative comparison on the EndoVis17 and EndoVis18 datasets (%). The best results are in **bold**, second-best are underlined. (Refer to Appendix B for the abbreviations)

	Category	Method	Challenge IoU	IoU	mcIoU	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 (Zero-Shot)	71.32	71.32	68.79	55.62	65.30	76.33	75.32	**90.09**	80.51	38.36	-	-
		ReMeDI-SAM3 (Ours)	**78.57**	**78.57**	**75.65**	69.07	**68.51**	**85.78**	82.52	89.90	78.68	55.06	-	-
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 (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)	**88.24**	**87.46**	**82.23**	**87.63**	**80.12**	**91.66**	90.41	**80.44**	-	-	64.33	**81.01**

4 validation videos (149 frames each); we report performance on the validation set for fair comparison. Both datasets provide pixel-wise annotations for seven instrument categories. For quantitative evaluation, we report Challenge IoU (Allan et al., 2019), which computes IoU only for instruments present in each frame, along with IoU and mean class IoU (mcIoU).

4.2. Implementation details

We set $\tau_{rel} = 0.95$ and $\tau_{occ} = 0.65$ for optimal performance. Re-identification is performed over a window of K=5 frames using multi-scale appearance features. We store up to 20 frames in the Multi-scale Feature Bank, and set the similarity margin to δ_{sim}=0.01, $\delta_{\text{sim}}^{-} = -0.01$ and δ_{iou}=0.8. All experiments are conducted on a single RTX 4090 GPU.

4.3. Results

Quantitative Comparison. We evaluate our method on the EndoVis17 and EndoVis18 benchmarks and compare against both specialist surgical segmentation models and recent SAM-based approaches, including ISINet (González et al., 2020), S3Net (Baby et al., 2023), MATIS (Ayobi et al., 2023), TP-SIS (Zhou et al., 2023), TrackAnything (Yang et al., 2023), PerSAM (Zhang et al., 2023), SurgicalSAM (Yue et al., 2024), SP-SAM (Yue et al., 2023), MA-SAM2 (Yin et al., 2025), and vanilla SAM3. In addition to global metrics, we also report class-wise IoUs. All results of our method are obtained under a training-free setting.

Table 1 summarizes the quantitative results on EndoVis17 and EndoVis18. ReMeDI-SAM3 consistently and substantially outperforms vanilla SAM3 across both benchmarks. On EndoVis17, we achieve an improvement of around 7.2% Challenge IoU and 6.9% mcIoU. On EndoVis18, we observe gains of approximately 6% IoU and 16% mcIoU. Notably, the larger improvement in mcIoU compared to Challenge IoU on EndoVis18 highlights the benefit of our identity-aware design: Challenge IoU only evaluates instruments present in the frame and does not penalize false-positive predictions of absent tools, whereas mcIoU directly captures such errors. Our re-identification mechanism effectively suppresses these

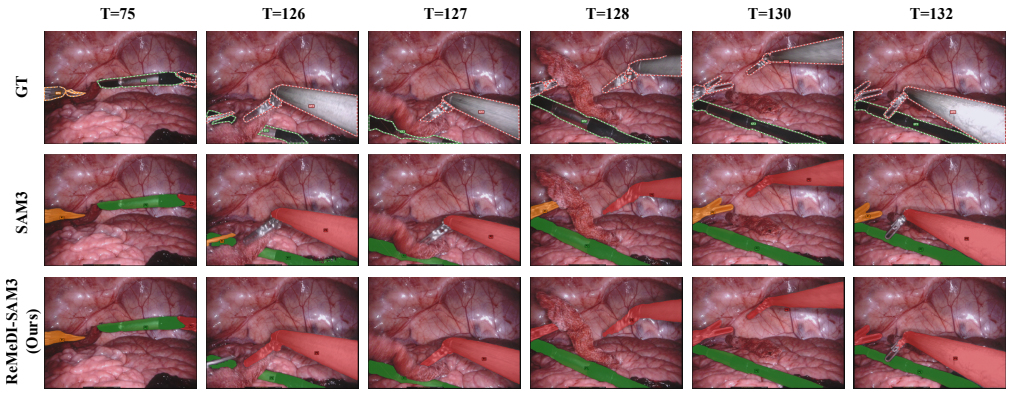

Figure 4: Qualitative comparison of SAM3 and ReMeDI-SAM3. ReMeDI-SAM3 preserves identity across long occlusions and re-entries, while SAM3 shows identity confusion.

false positives. For instance, in Sequence 2 of EndoVis18, after the Clip Applier (CA) exits the scene, SAM3 continues to hallucinate CA on tissue regions, whereas our method correctly suppresses these false positives (see Figure 7), improving the CA IoU from 32.96% to 81%. Similarly, in Sequence 5, SAM3 incorrectly predicts the Ultrasound Probe (UP) instead of the returning Prograsp Forceps (PF), whereas our approach recovers PF and suppresses UP, yielding a 36% improvement in UP IoU. Overall, our method outperforms all other zero-shot methods and even surpasses training-based approaches on both benchmarks, demonstrating the effectiveness of the proposed method.

Qualitative Comparison. Figure 4 shows a challenging occlusion and reappearance case on EndoVis17. The orange instrument (Bipolar Forceps) exits the scene after T=75, and a second red instrument (Prograsp Forceps) enters later (T=126–132). ReMeDI-SAM3 initially misses the new instrument (T=126) as it's not clearly visible but subsequently correctly recovers and assigns the red identity once sufficient evidence is available. In contrast, SAM3 incorrectly preserves orange identity after occlusion, continuing to label the new red instrument as orange due to memory drift. This shows that ReMeDI-SAM3 reliably recovers correct identities even after instrument turnover, highlighting the effectiveness of our method in maintaining identity consistency under extended occlusions and large appearance variations. Further qualitative comparisons are provided in Appendix A.

4.4. Ablation Studies

We analyze the impact of memory size, temporal interpolation scheme, and components of ReMeDI-SAM3—including relevance-aware memory (RM), occlusion-aware memory (OM), memory expansion (ME), and feature-based re-identification (Re-ID)—on EndoVis17.

Model Components. Table 4 analyzes the individual and combined contributions of different model components on EndoVis17. Introducing relevance-aware memory filtering alone yields a gain of 3.5% IoU over vanilla SAM3 by suppressing unreliable memory updates. Expanding the memory further provides an additional 1.0% IoU improvement. Incorporating the feature-based re-identification module contributes an additional 1.2% IoU, demonstrating its effectiveness for post-occlusion identity recovery. Finally, enabling the occlusion–aware memory with a relaxed occlusion threshold results in a 1.5% IoU gain,

Table 2: Effect of Memory Expansion (%) on SAM3 + RM with $\tau_{rel} = 0.95$.

Memory Size	Challenge IoU	IoU	mcIoU
7	74.81	74.81	72.46
10	74.41	74.41	72.01
15	**75.85**	**75.85**	**73.51**
20	74.64	74.64	72.09

Table 3: Uniform vs. Piecewise Temporal Positional Encoding on SAM3 + RM + ME ($\tau_{rel} = 0.95$, $M = 15$)

Method	Challenge IoU	IoU	mcIoU
Uniform	74.92	74.92	73.02
Piecewise	**75.85**	**75.85**	**73.51**

Table 4: Impact of individual components of ReMeDI-SAM3 (%).

Method	M	τ_{rel}	τ_{reID}	Challenge IoU	IoU	mcIoU
Vanilla SAM3	7	0.01	–	71.32	71.32	68.79
SAM3 + RM	7	0.95	–	74.81	74.81	72.46
SAM3 + RM + ME	15	0.95	–	75.85	75.85	73.51
SAM3 + RM + ME + Re-ID	15	0.95	–	77.01	77.01	74.41
SAM3 + RM + ME + Re-ID + OM	15	0.95	0.65	**78.57**	**78.57**	**75.65**

achieving a total improvement of 7.2% IoU and 6.9% mcIoU over vanilla SAM3. These results confirm that all components of the proposed framework are complementary and jointly critical for robust long-horizon surgical video segmentation.

Memory Expansion. Expanding memory with the proposed piecewise interpolation strategy yields clear performance gains, as shown in Table 2. Increasing the memory to 15 frames improves all metrics by around 1.0% , indicating the benefit of additional temporal context. However, further expansion to 20 frames leads to a slight performance drop, suggesting that overly large memory might introduce less informative context.

Interpolation Scheme. Replacing the proposed piecewise interpolation with uniform interpolation (Figure 3 (middle)) for memory expansion causes a drop of about 0.9% Challenge IoU and 0.5% mcIoU (M=15) on EndoVis17 (Table 3). Uniform resampling distorts the learned boundary temporal priors, whereas piecewise interpolation preserves the semantic roles of the earliest and latest memory positions while densifying the interior region.

5. 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 based on piecewise interpolation of temporal positional encodings for long-horizon reasoning. Extensive zero-shot evaluations on EndoVis17 and EndoVis18 show consistent improvements over vanilla SAM3 and prior methods, particularly in mcIoU, demonstrating superior identity preservation and false-positive suppression. A noted limitation is that ReMeDI-SAM3 may delay re-identification after occlusions due to its conservative handling of uncertain detections (Figure 9). We also plan to explore its applicability to natural-domain tracking in future work.

References

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Nicolás Ayobi, Alejandra Pérez-Rondón, Santiago Rodríguez, and Pablo Arbeláez. Matis: Masked-attention transformers for surgical instrument segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

Britty Baby, Daksh Thapar, Mustafa Chasmai, Tamajit Banerjee, Kunal Dargan, Ashish Suri, Subhashis Banerjee, and Chetan Arora. 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*, pages 6191–6201, 2023.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Ruixiang Chen, Guolei Sun, Yawei Li, Jie Qin, and Luca Benini. Him2sam: Enhancing sam2 with hierarchical motion estimation and memory optimization towards long-term tracking. *arXiv preprint arXiv:2507.07603*, 2025.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13614–13624, 2025.

Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: an instance-based approach for surgical instrument segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 595–605. Springer, 2020.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 624–628. IEEE, 2018.

Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24255–24264, 2025.

Zhiling Yan, Weixiang Sun, Rong Zhou, Zhengqing Yuan, Kai Zhang, Yiwei Li, Tianming Liu, Quanzheng Li, Xiang Li, Lifang He, et al. Biomedical sam 2: Segment anything in biomedical images and videos (2024). *ArXiv abs/2408.03286*.

Zhiling Yan, Sifan Song, Dingjie Song, Yiwei Li, Rong Zhou, Weixiang Sun, Zhennong Chen, Sekeun Kim, Hui Ren, Tianming Liu, et al. Samed-2: Selective memory enhanced medical segment anything model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–550. Springer, 2025.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

Ming Yin, Fu Wang, Xujiong Ye, Yanda Meng, and Zeyu Fu. Memory-augmented sam2 for training-free surgical video segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 328–337. Springer, 2025.

Jieming Yu, Long Bai, Guankun Wang, An Wang, Xiaoxiao Yang, Huxin Gao, and Hongliang Ren. Adapting sam for surgical instrument tracking and segmentation in endoscopic submucosal dissection videos. *arXiv preprint arXiv:2404.10640*, 2024a.

Jieming Yu, An Wang, Wenzhen Dong, Mengya Xu, Mobarakol Islam, Jie Wang, Long Bai, and Hongliang Ren. Sam 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. *arXiv preprint arXiv:2408.04593*, 2024b.

Cheng Yuan and Yutong Ban. Temporal propagation of asymmetric feature pyramid for surgical scene segmentation. *arXiv preprint arXiv:2504.13440*, 2025.

Wenxi Yue, Jing Zhang, Kun Hu, Qiuxia Wu, Zongyuan Ge, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalpart-sam: Part-to-whole collaborative prompting for surgical instrument segmentation. *arXiv preprint arXiv:2312.14481*, 2023.

Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6890–6898, 2024.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. *Advances in Neural Information Processing Systems*, 36:28611–28623, 2023.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

Appendix A. Additional Visualizations

A.1. EndoVis17

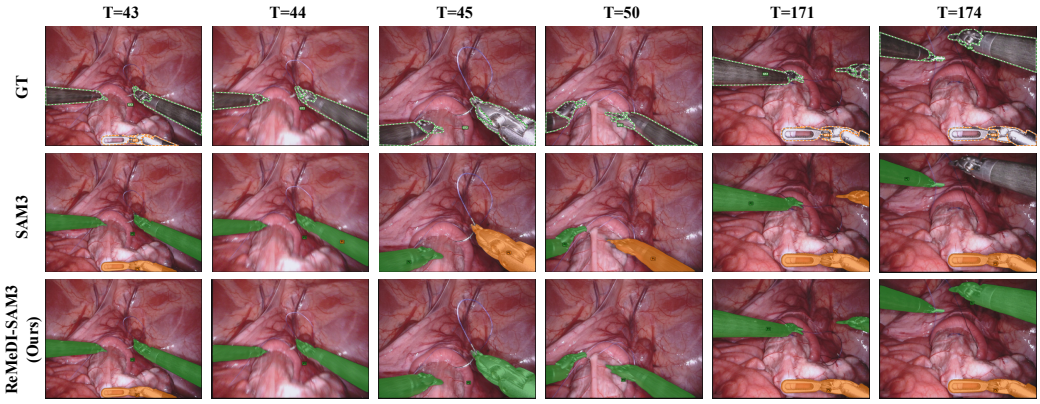

Figure 5: Qualitative comparison of SAM3 and ReMeDI-SAM3 on EndoVis17.

Qualitative Comparison. Figure 5 shows a challenging occlusion and reappearance case on EndoVis17. After the orange-labeled instrument becomes fully occluded at T=44, SAM3 exhibits identity drift and incorrectly assigns the orange identity to the visible green instrument, with this mislabeling persisting across subsequent frames. In contrast, ReMeDI-SAM3 suppresses such false-positive identity propagation during the occlusion and correctly re-identifies the true instrument upon reappearance.

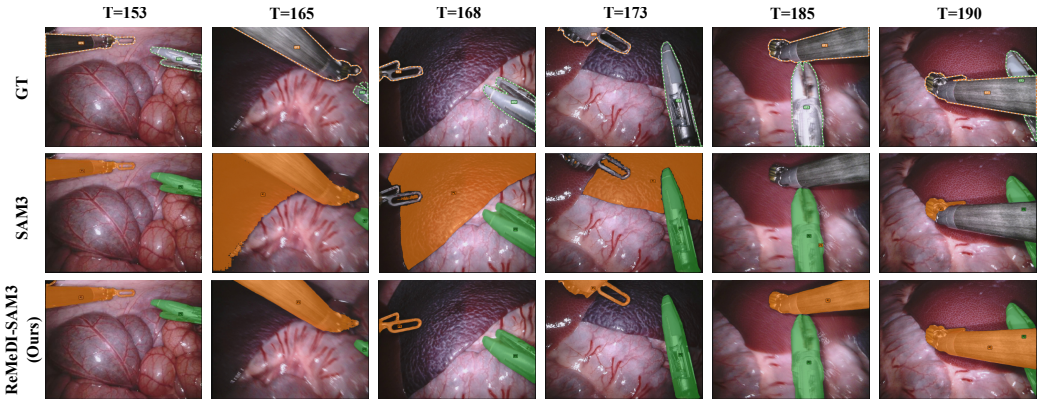

Figure 6: Qualitative comparison of SAM3 and ReMeDI-SAM3 on EndoVis17.

Qualitative Comparison. In Figure 6, after T=153, the orange-labeled instrument leaves the field of view. In its reappearance at T=165, SAM3 confuses it with surrounding tissue, failing to recover the correct instrument identity. In subsequent frames, it is either weakly detected or further confused with the green-labeled instrument, leading to persistent identity errors. In contrast, ReMeDI-SAM3 reliably recovers the orange-labeled instrument upon re-entry, maintaining consistent identity throughout. This highlights the effectiveness of our occlusion-aware memory filtering and feature-based re-identification in maintaining identity consistency under extended occlusions and large appearance variations.

A.2. EndoVis18

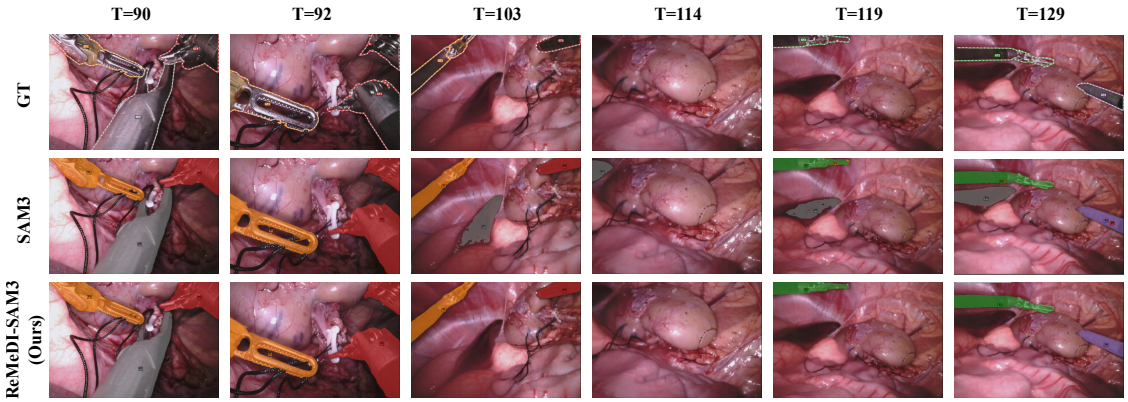

Figure 7: Qualitative comparison of SAM3 and ReMeDI-SAM3 on EndoVis18.

Qualitative Comparison. In Figure 7, after the gray instrument exits the scene, SAM3 produces false positives by incorrectly assigning tissue regions to the gray class. Following the disocclusion of other tools at T=119 and T=129, SAM3 further exhibits label confusion with high-overlap predictions. In contrast, ReMeDI-SAM3 correctly suppresses the absent gray instrument by leveraging both high-confidence and occlusion-aware memory entries. Upon tool reappearance, the feature-based re-identification module further resolves label ambiguities and overlapping predictions, yielding consistent and accurate identities.

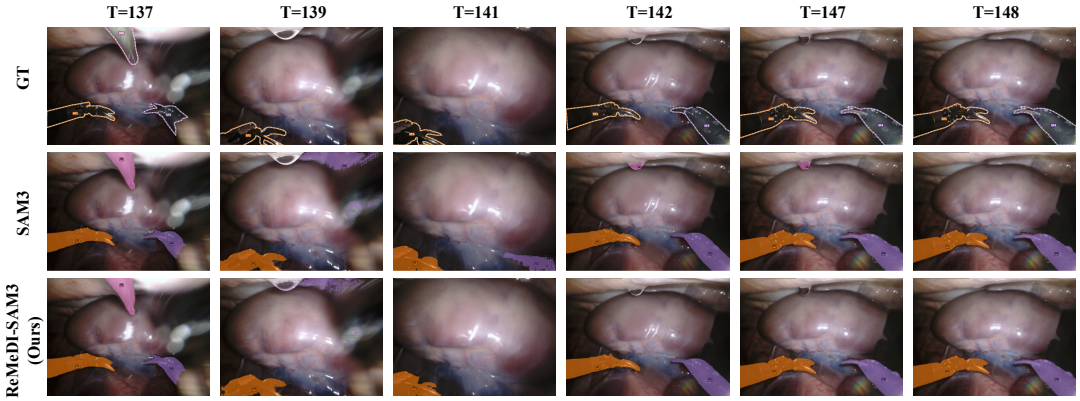

Figure 8: Qualitative comparison of SAM3 and ReMeDI-SAM3 on EndoVis18.

Qualitative Comparison. Figure 8 further demonstrates the superiority of ReMeDI-SAM3 over SAM3 under extended occlusions. After T=137, the purple- and pink-labeled instruments exit the scene. While the purple instrument is absent, SAM3 incorrectly hallucinates it on blurred background regions at T=139 and T=141. Moreover, in later frames, SAM3 falsely assigns the pink class to a water droplet, likely due to its approximate spatial location and resemblance to the instrument's last observed shape.

In contrast, ReMeDI-SAM3 effectively suppresses these post-occlusion false positives. Although it produces a weak purple prediction at T=139, this is corrected in subsequent

frames, and no false pink detections appear afterward. This robustness stems from the expanded memory capacity and the more reliable, relevance-aware memory bank maintained by ReMeDI-SAM3, which prevents spurious predictions after occlusion.

Appendix B. Detailed Instrument Categories

EndoVis17 and Endovis18 datasets provide annotated masks of 7 different surgical instrument categories each. 5 instruments are common between the two datasets, namely: Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Monopolar Curved Scissors (MCS), and Ultrasound Probe (UP). Additionally, EndoVis17 includes Vessel Sealer (VS) and Grasping Retractor (GR), while EndoVis18 also contains Suction Instrument (SI) and Clip Applier (CA). The instruments are mentioned by the abovementioned abbreviations in Table 1.

Appendix C. Limitations

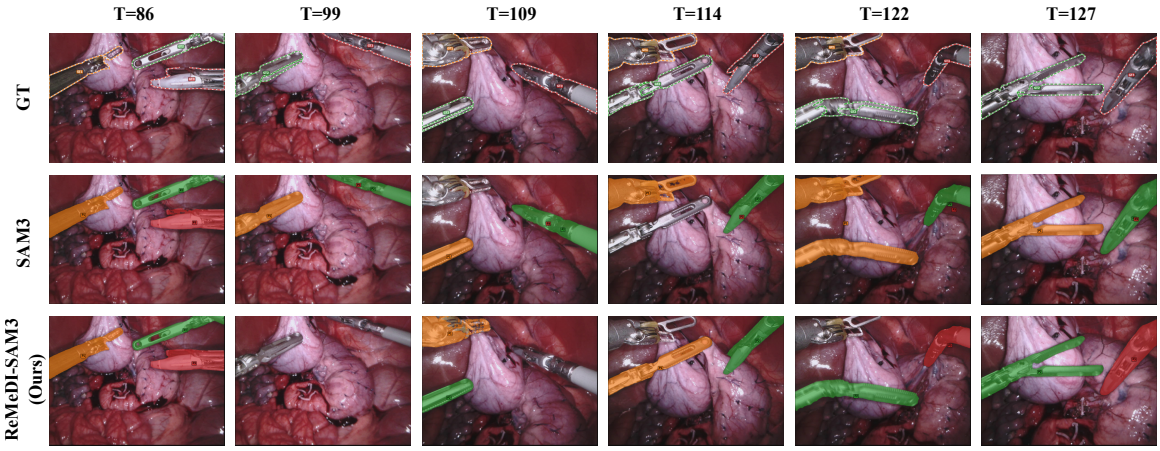

Figure 9: Qualitative comparison of SAM3 and ReMeDI-SAM3 on EndoVis17.

Qualitative Comparison. Figure 9 illustrates a limitation of ReMeDI-SAM3 in complex multi-instrument occlusion and re-entry scenarios. Although all instruments are correctly detected at T=86, they leave the field of view and later reappear with substantially different poses and spatial relationships ($T = 99$). ReMeDI-SAM3 adopts a conservative update strategy to avoid false positives, and as a result it initially refrains from predicting any instruments upon their return, even though they are present ($T = 99$). As more visual evidence accumulates, the model begins to recover a subset of instruments but remains hesitant about others, leading to delayed re-detection ($T = 109$) and a brief identity confusion ($T = 114$). Only once sufficient reliable cues become available does the model fully re-stabilize and correctly recover all instrument identities ($T = 127$).

In contrast, SAM3 continues producing predictions throughout the sequence but frequently assigns incorrect identities, showing substantial drift after occlusions. This scenario highlights an inherent trade-off in ReMeDI-SAM3: its strong emphasis on suppressing

false positives can slow down re-identification and reduce recall immediately after occlusion events, particularly when instruments re-enter with large appearance or pose changes.