DENSE-TEMPORAL SURGICAL VIDEO OBJECT SEGMENTATION VIA AGENTIC TRAJECTORY RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

High-frame-rate surgical video object segmentation (VOS) exposes finer-grained intraoperative dynamics over densely sampled sequences. Existing VOS methods often fail to detect unreliable propagation and recover once tracking drift contaminates memory. We propose **SurgATR**, an Agentic Trajectory Recovery framework for dense-temporal surgical VOS. SurgATR first performs trajectory reflection to identify propagation failures from recent target evolution. Once triggered, an agent directly intervenes in both spatial and temporal tracker states: spatially, it re-anchors the target with corrective prompts; temporally, it rolls back and reconstructs tracker memory from a reliable state to prevent error accumulation. To avoid incorrect recovery toward visually similar distractors, cycle validation further filters identity-inconsistent recoveries, while global Viterbi selection determines a globally coherent trajectory. To support systematic study of dense-temporal surgical VOS, we introduce **SurgDVOS**, the largest high-frame-rate surgical VOS benchmark, with 330,246 annotated frames and 869,922 pixel-level masks, supporting both mask-guided and referring VOS. Experiments on SurgDVOS and three public natural-video benchmarks show consistent improvements, with SurgATR achieving state-of-the-art performance in mask-guided and referring VOS while demonstrating strong cross-domain generalization.

1 INTRODUCTION

Surgical video object segmentation (VOS) plays a pivotal role in surgical scene understanding by enabling persistent perception of surgical instruments and tissues over time. Through continuous segmentation and tracking while preserving physical identity, surgical VOS recovers spatiotemporal trajectories that capture the evolution of surgical procedures (Perazzi et al., 2016; Pont-Tuset et al., 2017; Xu et al., 2018). High-frame-rate (HFR) surgical VOS further resolves finer-grained motion variations, instrument dynamics, and transient interaction patterns through densely sampled observations. Such temporally resolved trajectories enable precise analysis of surgical activities, including motion pattern characterization, operative behavior assessment, and tool-tissue interaction modeling, providing richer dynamic cues for surgical scene understanding (Carstens et al., 2023).

Despite these advantages, HFR surgical VOS poses distinctive challenges. Dense temporal sampling can resolve finer-scale transitions involving rapid motion, occlusion, instrument exchange, and reappearance. However, these transitions challenge identity propagation and can introduce correspondence errors among visually similar targets. As illustrated in Fig. 1, such errors can rapidly accumulate in temporal memory, leading to persistent identity drift or tracking failure in modern VOS systems (Oh et al., 2019; Cheng et al., 2024; Ravi et al., 2025). Recent agent-assisted methods can re-ground lost targets, but rarely restore the corrupted tracker state itself, causing erroneous tracker prompts and temporal memory to persist into subsequent propagation (Jiang et al., 2026; Jin et al., 2026). Moreover, pinpointing the onset of propagation failure requires temporally dense observations, whereas the sparse sampling of existing surgical VOS datasets obscures the fine-grained progression from failure onset through drift to recovery (Allan et al., 2019; 2020; Ross et al., 2020).

To bridge this gap, we introduce SurgDVOS, the largest HFR surgical VOS benchmark to date. SurgDVOS comprises 558 clips drawn from 50 laparoscopic liver resection videos, totaling 330,246 annotated frames and 869,922 pixel-level masks. Through dense temporal sampling and frame-level annotation, SurgDVOS preserves transient changes associated with rapid motion, occlusion, disap-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

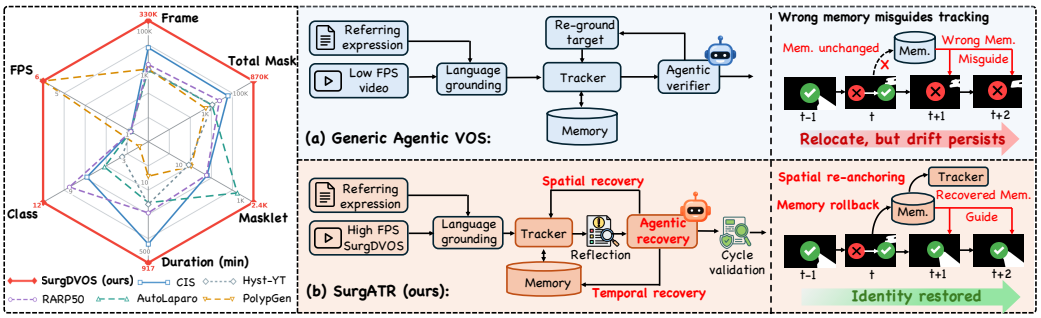


Figure 1: **Overview of SurgDVOS and agentic recovery for surgical VOS. Left:** SurgDVOS provides large-scale, densely sampled annotations. **Right:** (a) Generic agentic VOS re-localizes the target but leaves tracker memory unchanged, causing persistent drift. (b) SurgATR directly controls tracker and memory states for spatial re-anchoring and temporal rollback, restoring target identity.

pearance, and reentry, thereby capturing fine-grained temporal dynamics. Built on the same target trajectories, SurgDVOS supports two practically meaningful settings, *mask-guided* and *referring VOS*, enabling systematic evaluation of temporal consistency, failure localization, identity recovery, and language-guided segmentation in HFR surgical videos.

To actively detect and recover unreliable propagation, we further propose **SurgATR**, an Agentic Trajectory Recovery framework that supports both *mask-guided* and *referring VOS*. As illustrated in Fig. 1, trajectory reflection exploits fine-grained trajectory observations to assess temporal consistency and identify deviations indicative of propagation errors or tracking failures. Once a failure is detected, SurgATR performs coordinated agentic recovery from two complementary perspectives. Spatially, it adaptively re-anchors the target according to the failure pattern; temporally, it restores tracker memory from a reliable state, preventing errors from persisting in subsequent frames. Then, cycle validation compares recovery hypotheses with historical target observations to reject identity-inconsistent interventions. Additionally, global Viterbi selection jointly considers the original propagation and validated recovery candidates to select a globally coherent trajectory. Experiments on SurgDVOS and three public natural-video benchmarks (Pont-Tuset et al., 2017; Xu et al., 2018; Hong et al., 2023) demonstrate consistent gains across mask-guided and referring VOS, with particularly strong gains in scenarios involving disappearance, reappearance, and identity ambiguity.

Our main contributions are summarized as follows:

- We introduce **SurgDVOS**, the largest high-frame-rate surgical VOS benchmark, supporting both *mask-guided* and *referring VOS* for comprehensive evaluation of temporal consistency, failure localization, identity recovery, and language grounding.
- We propose **SurgATR**, a unified framework for mask-guided and referring VOS that leverages fine-grained temporal evidence for identity propagation under dense dynamics.
- We develop a coordinated spatiotemporal agentic recovery mechanism for SurgATR, where trajectory reflection identifies unreliable propagation, spatial re-anchoring and temporal memory restoration recover corrupted target states, and cycle validation filters recovery candidates for Viterbi selection of globally coherent trajectories.
- Extensive experiments on SurgDVOS and three public natural-video benchmarks show consistent gains, with SurgATR achieving 68.2 and 65.6 $\mathcal{J}\&\mathcal{F}$ on mask-guided and referring VOS, respectively, while demonstrating strong cross-domain generalization.

2 Related Work

Surgical Video Object Segmentation. Surgical video object segmentation (VOS) continuously segments surgical instruments and tissues over time. EndoVis 2017/2018 (Allan et al., 2019; 2020), ROBUST-MIS (Ross et al., 2020), and CholecInstanceSeg (Alabi et al., 2025) provide scene annotations across diverse laparoscopic procedures. More recent work extends surgical segmentation toward target-conditioned video tracking. RSVIS (Wang et al., 2024) introduces language-referred

Table 1: Comparison with representative surgical video segmentation datasets.

Source Dataset	Procedure	Sequence	Frame	Total Mask	Masklet	Duration (min)	Object Classes	Label FPS
CIS (Alabi et al., 2025)	Chole.	10	19,029	31,416	77	317	7	1
ClinicDB (Bernal et al., 2015)	Colon.	29	612	647	32	10	1	1
AutoLaparo (Wang et al., 2022)	Gyn.	300	1,800	3,963	738	30	5	1
DSAD-V (Carstens et al., 2023)	Rect.	76	4,390	4,248	91	73	11	1
PolypGen (Ali et al., 2023)	Colon.	21	2,037	1,760	23	6	1	**6**
Hyst-YT (Fang et al., 2026)	Gyn.	6	1,973	4,113	19	33	3	1
SurgAI3.8K (Madad Zadeh et al., 2023)	Gyn.	51	3,817	3,817	51	64	1	1
RARP50 (Psychogyios et al., 2024)	Prost.	10	3,252	10,503	74	54	9	1
EndoVis17 (Allan et al., 2019)	Neph.	3	900	2,265	10	15	7	1
EndoVis18 (Allan et al., 2020)	Neph.	4	596	2,191	22	10	11	1
SurgDVOS (ours)	Liver	**558**	**330,246**	**869,922**	**2,448**	**917.4**	**12**	**6**

instrument segmentation, while Surgical SAM2 (Liu et al., 2024) and ReSurgSAM2 (Liu et al., 2025a) adapt promptable VOS models to surgical scenes. Despite this progress, existing benchmarks remain dominated by sparsely sampled sequences, limiting the study of fine-grained temporal dynamics and failure evolution. SurgDVOS addresses this gap with high-frame-rate, densely annotated trajectories for systematic evaluation of identity recovery under fine-grained surgical dynamics.

Referring Video Object Segmentation. Referring VOS grounds language expressions into object trajectories over time. MTTR (Botach et al., 2022), ReferFormer (Wu et al., 2022), and OnlineRefer (Wu et al., 2023a) establish transformer-based and online language-conditioned propagation, while SgMg (Miao et al., 2023) and MeViS (Ding et al., 2025) strengthen cross-modal interaction and motion-aware target grounding. Recent benchmarks further extend referring VOS toward more challenging temporal settings, with MeViS emphasizing motion-centric expressions and LongRVOS (Liang et al., 2025a) focusing on long-duration target propagation. Building on foundation models, MPG-SAM2 (Rong et al., 2025) further integrates language grounding with SAM2-based video propagation. In contrast, SurgATR re-engages language guidance upon propagation failure and directly intervenes in both tracker prompts and temporal memory for identity recovery.

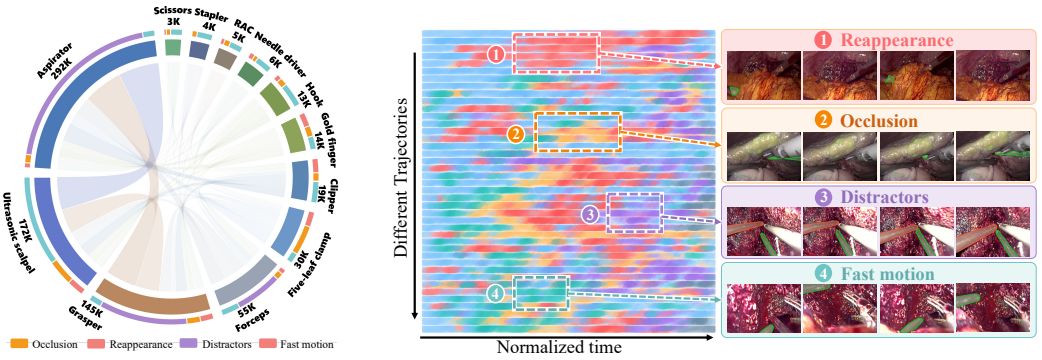
3 SurgDVOS Benchmark

We analyze the construction, annotations, and spatiotemporal characteristics of SurgDVOS in this section. Further details and additional analyses are provided in Appendix A.

Benchmark Construction. SurgDVOS is curated from 50 laparoscopic liver resection procedures collected at partner hospitals. The source videos were recorded at 25 fps with a resolution of 1920×1080 and divided into representative clips capturing rapid motion, occlusion, and prolonged disappearance. Six surgeons generated initial dense masks using interactive SAM2 (Ravi et al., 2025), followed by frame-wise inspection and correction by 12 additional surgical interns. Each clip underwent three independent rounds of review for mask accuracy and temporal identity consistency, yielding high-quality pixel-level annotations over extended trajectories.

Data Annotations. As shown in Table 1, SurgDVOS contains 558 sequences, 330,246 annotated frames, 869,922 pixel-level masks, and 2,448 masklets, establishing the largest annotation corpus among existing surgical VOS benchmarks. Its dense 6 fps annotation resolves fine-scale temporal transitions that are often missed under sparse sampling, including rapid motion, occlusion onset, reappearance, and frequent target interactions. Built upon the same densely annotated trajectories of surgical instruments, SurgDVOS supports both *mask-guided* and *referring VOS*, enabling evaluation under both mask- and language-based settings. As illustrated in Fig. 2, these densely observed dynamics, together with frequent instrument co-occurrence and visually similar distractors, create challenging conditions for temporal consistency, failure localization, and identity recovery.

Spatiotemporal Challenge Analysis. Beyond annotation scale, SurgDVOS exhibits substantial diversity in both temporal dynamics and spatial interaction patterns. As shown in Fig. 2b, instrument trajectories frequently involve reappearance, occlusion, distractor interference, and rapid motion, with multiple challenge types often interleaved within the same trajectory. The trajectory atlas further reveals their dense and heterogeneous temporal evolution over extended sequences. Fig. 3 complements this analysis spatially, showing distinct instrument-specific occupancy patterns and



(a) (b)

Figure 2: **Data characteristics of SurgDVOS. (a)** Instrument co-occurrence patterns across annotated trajectories. **(b)** Temporal challenge visualization with a trajectory atlas and representative cases of reappearance, occlusion, distractor ambiguity, and fast motion.

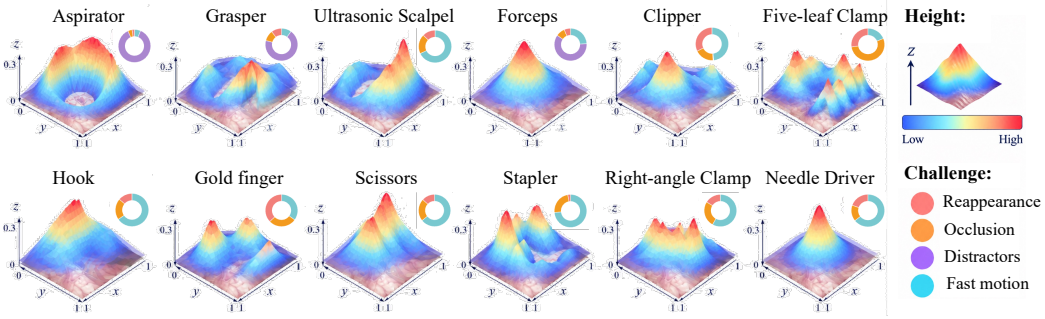


Figure 3: **Spatial challenge landscapes in SurgDVOS.** For each instrument category, surface height indicates spatial occupancy probability and color indicates challenge likelihood. Donut charts summarize the relative composition of challenge types.

challenge hotspots with varying spatial entropy. Together, these characteristics expose a broad range of spatiotemporal conditions that are largely obscured under sparse temporal sampling.

4 METHOD

4.1 FRAMEWORK OVERVIEW

Figure 4 illustrates the overall pipeline of SurgATR. Given a video and task-specific input, a vision-language model (VLM) first grounds the specified target and generates prompts for tracker propagation. Then, trajectory reflection monitors recent target evolution and triggers recovery when propagation becomes unreliable. Upon failure, agentic recovery constructs alternative trajectories through complementary spatial re-anchoring and temporal memory rollback, with cycle validation retaining only consistent recovery hypotheses. Finally, we adopt global Viterbi selection to jointly optimize over the validated recovery trajectories and obtain a globally coherent propagation path. The complete inference procedure of SurgATR is summarized in Algorithm 1 in Appendix C.

4.2 AGENTIC TRAJECTORY RECOVERY

Dense surgical video sequences expose rapid target transitions, introducing correspondence errors that accumulate in tracker memory and progressively destabilize subsequent propagation. SurgATR addresses this failure through three stages: trajectory reflection detects erroneous propagation from dense temporal evidence, agentic recovery adaptively intervenes in spatial prompts and temporal memory, and cycle validation rejects inconsistent recovery hypotheses.

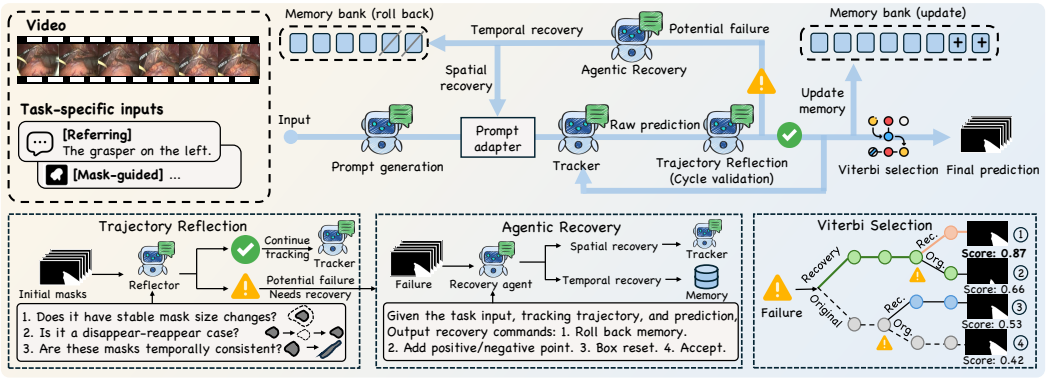


Figure 4: **Overview of the SurgATR framework.** SurgATR uses a VLM to interpret task-specific inputs and generate target cues, which are adapted into compatible prompts for propagation. Trajectory reflection monitors propagation reliability, while agentic recovery performs complementary spatial re-anchoring and temporal memory rollback upon failure. Recovery trajectories are filtered by cycle validation and globally resolved through Viterbi selection.

Trajectory Reflection. Given a video $\mathcal{V} = \{I_t\}_{t=1}^{T}$ and task-specific input q, SurgATR initializes the tracker at frame t_0 with prompt p_{t_0}, yielding the initial state $(\hat{Y}_{t_0}^0, M_{t_0}^0, c_{t_0}^0, o_{t_0}^0)$. For subsequent frames $t = t_0 + 1, \ldots, T$, baseline propagation is performed as

$$(\hat{Y}_t^0, M_t^0, c_t^0, o_t^0) = \mathcal{T}(I_t, M_{t-1}^0), \tag{1}$$

where $\mathcal{T}$ denotes the video object tracker, $\hat{Y}_t^0$ and M_t^0 denote the predicted mask and updated tracker memory, while c_t^0 and o_t^0 denote the mask confidence and predicted target presence, respectively. To avoid unnecessary agent intervention, SurgATR first evaluates propagation reliability by monitoring recent tracking trajectories. Specifically, we characterize target evolution through complementary temporal cues $h_t = [h_t^{\text{area}}, h_t^{\text{overlap}}, h_t^{\text{motion}}, h_t^{\text{presence}}]$, capturing target area variation, temporal mask overlap, target displacement, and predicted target presence, respectively. These cues are aggregated over a local temporal window by the reflection function $\mathcal{R}$, with recovery triggered when $z_t = \mathbb{1}[\mathcal{R}(h_{t-w:t}) < \tau] = 1$. Since the reflection cues are derived directly from tracker predictions, recovery is invoked only upon detecting erroneous trajectory evolution.

Spatiotemporal Agentic Recovery. Once unreliable propagation is detected, SurgATR invokes a vision-language recovery agent to interpret the failure pattern and determine the appropriate intervention. The agent operates on both spatial localization and temporal memory, enabling recovery from target mislocalization as well as corrupted propagation states.

For temporal recovery, the agent first determines whether the current memory state has been affected by erroneous propagation. Let $t_a < t$ denote the most recent reliable anchor frame, and let M_{t-1}^k denote the input memory used by recovery branch k to process frame t. When memory rollback is selected, SurgATR restores this memory to the reliable anchor state by setting $M_{t-1}^k \leftarrow M_{t_a}^0$. Memory entries accumulated after t_a are discarded, and the intervening frames are not replayed. Otherwise, the branch retains the current baseline memory by setting $M_{t-1}^k \leftarrow M_t^0$. The resulting memory is then used to process the current frame, with spatial re-anchoring applied when selected.

For spatial recovery, the agent generates a corrective prompt conditioned on the task input, current observation, and recent target evolution. For the k-th recovery hypothesis, we have

$$p_t^k = \pi_{\text{spa}}\left(q,\, I_t,\, \hat{Y}_{t-w:t}^0,\, h_{t-w:t}\right) \in \left\{\mathcal{P}_t^k,\, \mathcal{B}_t^k\right\}, \tag{2}$$

where π_{spa} denotes the spatial recovery policy, $\mathcal{P}_t^k = \{(x_i, y_i, \ell_i)\}_{i=1}^{N_k}$ is a point prompt with $\ell_i \in \{+1, -1\}$ indicating positive or negative guidance, and $\mathcal{B}_t^k = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denotes box guidance. The resulting prompt re-anchors the target at frame t using the selected input memory:

$$(\hat{Y}_t^k, M_t^k) = \mathcal{T}\left(I_t, M_{t-1}^k; p_t^k\right). \tag{3}$$

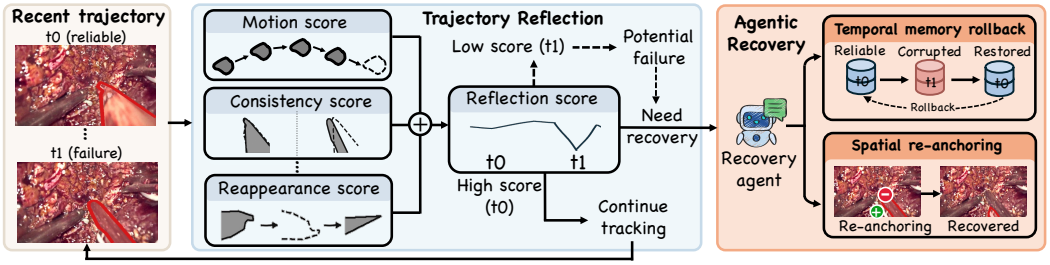


Figure 5: **Trajectory reflection and agentic recovery in SurgATR.** Trajectory reflection exploits dense temporal cues to detect abnormal propagation. Once triggered, SurgATR adaptively re-anchors the target spatially and restores reliable tracker memory temporally, while cycle validation filters identity-inconsistent recovery hypotheses.

When both temporal and spatial interventions are selected, spatial re-anchoring therefore operates on the restored memory state. Each recovery branch maintains an independent memory state, allowing alternative corrections to be explored without contaminating the baseline propagation.

Cycle Validation. Spatial re-anchoring may recover visually plausible targets with incorrect identity, particularly in the presence of similar instruments. We therefore validate each recovery branch by evaluating the reverse trajectory with trajectory reflection. Starting from the recovered target at frame t, the branch is traced back toward the reliable anchor t_a, where the same temporal cues are evaluated along the reverse trajectory. Denoting the resulting reverse cues by $h^{k,\mathrm{rev}}_{t:t_a}$, a recovery is retained only if the reverse trajectory remains reliable, *i.e.*, $\mathcal{R}(h^{k,\mathrm{rev}}_{t:t_a}) \geq \tau$. This cycle check rejects recovery hypotheses that fail to retrace a reliable trajectory toward the established target identity.

4.3 Global Viterbi Selection

To identify a globally optimal propagation path, SurgATR jointly evaluates all validated recovery trajectories using Viterbi selection, balancing local recovery quality with temporal continuity.

Trajectory Hypotheses. Let state 0 denote the baseline trajectory and states $1, \ldots, K$ the validated recovery trajectories. For each hypothesis k, SurgATR aggregates propagation reliability, semantic consistency, and identity consistency into a frame-level score $U_t(k) = \lambda_h H^k_t + \lambda_s S^k_t + \lambda_c C_k$, where H^k_t, S^k_t, and C_k represent propagation reliability, semantic consistency, and identity consistency from cycle validation, respectively. Unavailable or identity-inconsistent hypotheses are excluded from temporal optimization, leaving only valid recovery branches for global selection.

Global Temporal Selection. Then, SurgATR selects the globally most reliable and temporally coherent propagation path from the baseline and validated recovery hypotheses. Let $s_t \in \{0, 1, \ldots, K\}$ denote the selected hypothesis at frame t. The optimal sequence is defined as

$$s^*_{t_0:T} = \arg\max_{s_{t_0:T}} \sum_{t=t_0}^{T} U_t(s_t) - \lambda_{\mathrm{sw}} \sum_{t=t_0+1}^{T} \mathbb{1}(s_t \neq s_{t-1}), \tag{4}$$

where λ_{sw} controls the transition penalty between competing trajectory hypotheses. The corresponding Viterbi recursion is initialized as $V_{t_0}(k) = U_{t_0}(k)$ and proceeds as

$$V_t(k) = U_t(k) + \max_j \left[V_{t-1}(j) - \lambda_{\mathrm{sw}} \mathbb{1}(j \neq k) \right], \quad t = t_0 + 1, \ldots, T. \tag{5}$$

Backtracking recovers the optimal hypothesis sequence $s^*_{t_0:T}$ and yields the final trajectory $\hat{Y}_t = \hat{Y}_t^{\,s^*_t}$. By jointly balancing trajectory reliability and consistency, SurgATR adaptively integrates recovery branches into propagation, yielding a globally coherent trajectory with stable target identity.

5 Experiments

Implementation Details. We instantiate SurgATR with SAM2-Small (Ravi et al., 2025) as the frozen tracker and Qwen3.5-9B (Qwen Team, 2026) as the vision-language recovery agent. The

Referring VOS	Ref-DAVIS17			Ref-YT-VOS			LVOS			SurgDVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAMWISE (Cuttano et al., 2025)	70.9	67.4	74.5	69.2	67.8	70.6	33.6	28.0	39.1	58.5	58.5	58.6
MPG-SAM2 (Rong et al., 2025)	72.4	68.8	76.0	73.9	71.7	76.1	57.8	53.4	62.2	60.4	60.1	60.7
FindTrack (Cho et al., 2025)	74.2	69.9	78.5	70.3	68.6	72.0	43.2	39.4	46.9	38.9	38.4	39.4
LoSh (Yuan et al., 2024)	64.3	61.8	66.8	67.2	65.4	69.0	42.2	34.7	49.7	58.0	58.0	57.9
ReferDINO (Liang et al., 2025b)	68.9	65.1	72.9	69.3	67.0	71.5	56.0	49.5	62.5	56.7	56.8	56.6
DsHmp (He & Ding, 2024)	64.9	61.7	68.1	67.1	65.0	69.1	42.5	37.5	47.5	58.5	57.9	59.0
VD-IT (Zhu et al., 2024)	69.4	66.2	72.6	66.5	64.4	68.5	51.7	46.1	57.2	59.7	59.6	59.8
OnlineRefer (Wu et al., 2023a)	64.7	61.6	67.7	63.5	61.6	65.5	48.0	41.0	55.1	59.8	60.0	59.7
SurgATR (ours)	80.1	76.3	83.9	77.3	75.6	79.1	66.3	62.0	70.7	65.6	65.3	66.0

Table 2: Comparison with state-of-the-art methods on referring video object segmentation. All results are reported as bootstrap means over 1,000 resamples, with videos sampled with replacement.

Mask-guided VOS	DAVIS			YouTube-VOS			LVOS			SurgDVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAM2.1-Large (Ravi et al., 2025)	90.1	86.5	93.7	88.6	85.5	91.8	77.9	73.1	82.7	66.0	65.7	66.3
Surgical SAM2 (Liu et al., 2024)	89.8	86.2	93.4	88.2	85.1	91.3	76.5	71.9	81.1	66.1	65.8	66.4
XMem++ (Bekuzarov et al., 2023)	83.6	80.3	86.9	83.6	81.2	86.0	60.6	57.2	64.0	46.4	46.4	46.3
Cutie (Cheng et al., 2024)	88.8	85.6	91.9	86.5	83.4	89.7	63.5	59.1	67.9	59.8	59.7	59.9
LiVOS (Liu et al., 2025b)	85.1	81.9	88.3	81.3	78.5	84.1	51.2	46.8	55.6	42.6	42.8	42.3
OASIS (Qin et al., 2025)	88.3	85.0	91.6	86.6	83.5	89.8	59.0	53.5	64.6	57.8	57.6	58.0
XMem (Cheng & Schwing, 2022)	86.2	82.9	89.5	85.5	82.3	88.6	52.9	48.1	57.7	51.8	51.6	51.9
MAVOS (Shaker et al., 2025)	86.4	83.2	89.6	86.8	84.1	89.6	64.8	58.7	70.9	58.9	58.2	59.7
AOT (Yang et al., 2021)	83.8	81.1	86.4	84.0	80.7	87.4	60.3	54.6	66.0	49.6	48.7	50.6
SimVOS (Wu et al., 2023b)	88.0	85.0	91.0	84.2	81.1	87.4	60.5	55.8	65.2	58.3	57.6	58.9
SurgATR (ours)	90.2	86.4	94.0	89.6	87.0	92.2	83.7	78.6	88.7	68.2	67.5	68.9

Table 3: Comparison with state-of-the-art methods on mask-guided VOS. All results are reported as bootstrap means over 1,000 resamples, with videos sampled with replacement.

agent is adapted with LoRA (Hu et al., 2022) on the language-model projection layers using rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05, while the visual encoder and multimodal alignment module remain frozen. We train the model with AdamW (Loshchilov & Hutter, 2019) using a learning rate of 1×10^{-4}, weight decay of 0.01, a warm-up ratio of 0.05, and an effective batch size of 16. Training is conducted for two epochs on 8 NVIDIA A40 GPUs.

Datasets and Evaluation Metrics. We evaluate SurgATR on four benchmark families, including DAVIS (Perazzi et al., 2016; Pont-Tuset et al., 2017), YouTube-VOS (Xu et al., 2018), LVOS (Hong et al., 2023), and SurgDVOS, under both mask-guided and referring VOS settings. All methods follow the official dataset splits and evaluation protocols. For evaluation, we report region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$), and their mean ($\mathcal{J}\&\mathcal{F}$) (Perazzi et al., 2016).

5.1 Main Results

Results on SurgDVOS. Tables 2, 3, and 4 report the results on SurgDVOS. Existing methods degrade substantially on this benchmark, indicating the difficulty of preserving instrument identity under dense temporal dynamics. In mask-guided VOS, Surgical SAM2 (Liu et al., 2024) achieves the strongest prior result of 66.1 $\mathcal{J}\&\mathcal{F}$, while SurgATR reaches 68.2, improving by 2.1 points. For referring VOS, SurgATR achieves 65.6 $\mathcal{J}\&\mathcal{F}$, surpassing the strongest prior result of 60.4. MLLM-based methods (Bai et al., 2024; Yuan et al., 2025; Jiang et al., 2026; Qwen Team, 2026) exhibit a larger performance drop on SurgDVOS, further indicating the difficulty of maintaining the intended physical target across fine-grained temporal transitions. These results highlight the importance of explicitly detecting and recovering propagation failures for reliable identity propagation under dense temporal dynamics. Qualitative results in Fig. 6 and Fig. 7 further show that SurgATR can recover the intended instrument after propagation failure and maintain temporally consistent predictions under challenging surgical dynamics such as rapid motion and occlusion.

Referring VOS	Ref-DAVIS17			Ref-YT-VOS			LVOS			SurgDVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VISA (Yan et al., 2024)	69.4	66.3	72.5	61.5	59.8	63.2	–	–	–	–	–	–
VideoLISA (Bai et al., 2024)	68.8	64.9	72.7	63.7	61.7	65.7	40.8	36.3	45.4	23.4	24.1	22.8
GLUS (Lin et al., 2025)	72.9	69.7	76.1	67.3	65.5	69.0	58.1	53.7	62.4	39.5	38.7	40.4
VRS-HQ (Gong et al., 2025)	76.0	72.6	79.4	70.4	68.3	72.5	55.4	50.9	60.0	43.3	42.6	44.0
Qwen3.5 + SAM2	78.3	74.5	82.1	74.7	72.3	77.1	62.6	57.4	67.8	55.8	55.1	56.5
Sa2VA (Yuan et al., 2025)	76.0	71.7	80.3	68.0	66.2	69.8	55.2	49.4	61.1	31.0	31.1	30.8
Refer-Agent (Jiang et al., 2026)	74.6	70.7	78.5	71.3	69.0	73.6	54.1	49.5	58.8	33.1	32.6	33.6
SurgATR (ours)	80.1	76.3	83.9	77.3	75.6	79.1	66.3	62.0	70.7	65.6	65.3	66.0

Table 4: Comparison with state-of-the-art MLLM-based referring VOS methods. All results are reported as bootstrap means over 1,000 resamples, with videos sampled with replacement.

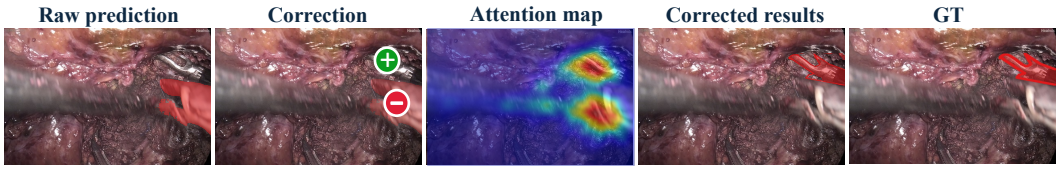


Figure 6: **Visualization of spatial recovery on SurgDVOS.** The recovery agent generates corrective prompts to re-anchor the intended target and recover from erroneous propagation.

Results on Public VOS Benchmarks. We further evaluate SurgATR on DAVIS, YouTube-VOS, and LVOS to assess its generalization beyond surgical videos. As shown in Table 3, SurgATR achieves 90.2, 89.6, and 83.7 $\mathcal{J}\&\mathcal{F}$ on the three mask-guided benchmarks, respectively. Notably, SurgATR improves over the strongest prior methods by 1.0 and 5.8 points on YouTube-VOS and LVOS, respectively, with particularly pronounced gains on the more challenging LVOS benchmark. Similar improvements are observed in referring VOS (Tables 2 and 4), where SurgATR reaches 80.1, 77.3, and 66.3 $\mathcal{J}\&\mathcal{F}$. Compared with Qwen3.5+SAM2, SurgATR further improves by 1.8, 2.6, and 3.7 points across the three benchmarks, demonstrating strong generalizability of SurgATR.

5.2 Ablation Study

We conduct our ablation studies on referring SurgDVOS, whose dense temporal dynamics provide a rigorous testbed for evaluating SurgATR. Further analyses are presented in Appendix B.

Component Ablation. Table 5 evaluates the contribution of the major components in SurgATR. Starting from the baseline at 55.8 $\mathcal{J}\&\mathcal{F}$, agentic recovery (AR) with confidence-based triggering improves performance to 62.0. Replacing confidence-based triggering with trajectory reflection (TR) further raises the score to 64.0 by identifying unreliable propagation from temporal cues. Cycle validation (CV) improves performance to 64.7 by filtering inconsistent recovery hypotheses. These variants directly adopt the recovery branch without Viterbi selection (VS). Finally, VS reaches 65.6 by globally coordinating the baseline and validated recovery trajectories, supporting the contributions of the individual components to reliable identity propagation under dense temporal dynamics.

Agentic Recovery. Table 6 further disentangles the effects of spatial re-anchoring and memory rollback, with TR and VS enabled in all recovery variants. Compared with the no-recovery baseline at 55.8 $\mathcal{J}\&\mathcal{F}$, spatial-only recovery reaches 63.8, whereas memory-only recovery achieves 63.4. Combining spatial re-anchoring and memory rollback improves performance to 65.0, and incorporating CV further increases performance to 65.6, highlighting the benefit of hypothesis filtering.

Trajectory Reflection. To evaluate the effectiveness of trajectory reflection, we compare different triggering strategies. As shown in Fig. 8a, our reflector exhibits the most pronounced response around the failure frame. At the case level, Fig. 8b further shows that our method achieves higher $\mathcal{J}\&\mathcal{F}$ with fewer agentic interventions, demonstrating more efficient failure detection.

Temporal Selection. We study the impact of different temporal selection strategies in Fig. 8c and Fig. 8d. As shown in Fig. 8c, recovery-only, greedy, and reliability-based selection achieve 64.7,

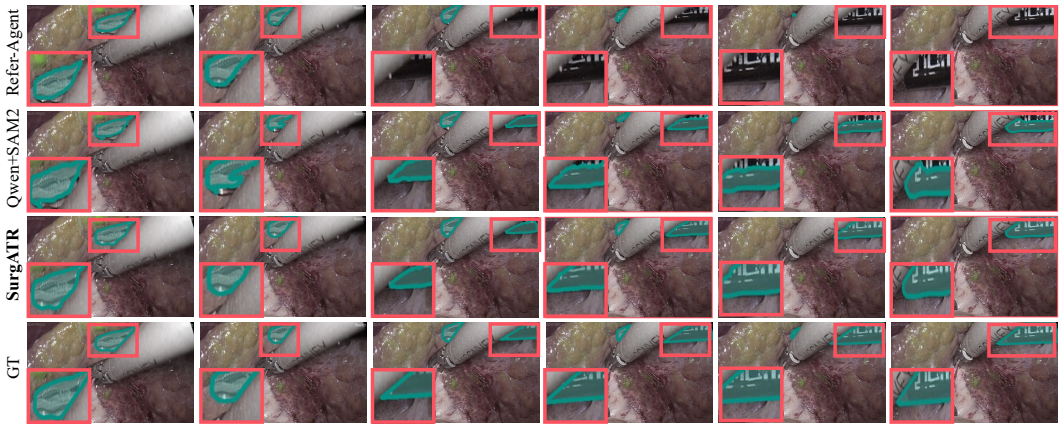


Figure 7: **Qualitative comparison on SurgDVOS.** SurgATR maintains more accurate and temporally consistent tracking, particularly under motion and occlusion.

Table 5: Component ablation on SurgDVOS. For recovery-enabled variants, disabling TR uses confidence-based triggering.

TR	AR	CV	VS	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
✗	✗	✗	✗	55.1	56.5	55.8
✗	✓	✗	✗	61.4	62.6	62.0
✓	✓	✗	✗	63.4	64.5	64.0
✓	✓	✓	✗	64.2	65.1	64.7
✓	✓	✓	✓	65.3	66.0	65.6

Table 6: Ablation of agentic recovery. TR and VS are enabled in all recovery variants. CV is enabled only in the last row.

Variant	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
No Recovery	55.1	56.5	55.8
Spa. Only	63.4	64.2	63.8
Mem. Only	63.0	63.8	63.4
Spa. + Mem.	64.6	65.4	65.0
Spa. + Mem. + CV	65.3	66.0	65.6

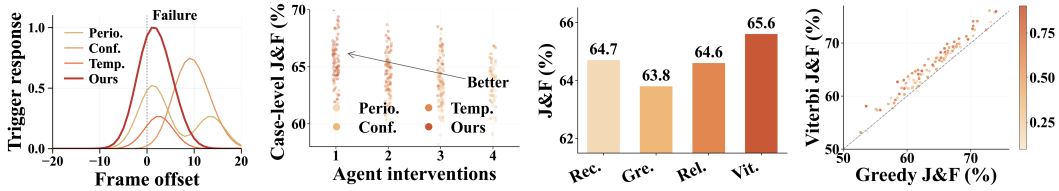


Figure 8: **Trajectory Reflection and Selection Analysis.** (a) Failure-aligned trigger responses of periodic (Perio.), confidence-based (Conf.), temporal-consistency-based (Temp.), and trajectory-reflection-based (Ours) strategies. (b) Case-level segmentation performance versus the number of agent interventions. (c) Comparison of recovery-only (Rec.), greedy (Gre.), reliability-based (Rel.), and Viterbi (Vit.) selection. (d) Case-level comparison of greedy and Viterbi selection.

63.8, and 64.6 $\mathcal{J}\&\mathcal{F}$, respectively. Viterbi selection achieves the best performance of 65.6 by jointly considering reliability and temporal continuity, outperforming recovery-only and greedy selection by 0.9 and 1.8 points, respectively. At the case level, Fig. 8d shows that Viterbi selection outperforms greedy selection, demonstrating the benefit of globally coordinating recoveries.

6 Conclusion

In this work, we study reliable identity propagation under the dense temporal dynamics of high-frame-rate surgical videos. We introduce **SurgDVOS** to better expose propagation failures and **SurgATR** to explicitly detect failures and recover from them through spatiotemporal agentic recovery. By spatially re-anchoring the target and restoring corrupted temporal memory, SurgATR enables more reliable identity recovery in challenging surgical scenes. Experiments across surgical and natural VOS benchmarks demonstrate consistent improvements, showing that reliable spatiotemporal identity propagation provides a foundation for robust surgical scene understanding.

USE OF LARGE LANGUAGE MODELS

We used ChatGPT solely for language polishing and grammatical editing of the authors' manuscript drafts. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This study uses retrospective surgical video data collected from real clinical practice and was approved by the relevant Institutional Review Board (IRB). All data were de-identified before processing and annotation. Clinical metadata are reported only in aggregate form and are not used as model inputs. The proposed method is intended solely for research on surgical video understanding and does not provide clinical diagnoses or treatment recommendations. Dataset release will follow applicable institutional, privacy, and data-use requirements.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction, annotation protocol, task definitions, evaluation settings, and implementation details in the main paper and supplementary material. The complete inference procedure of SurgATR is further documented with algorithmic pseudocode and ablation studies. Code, model configurations, and releasable benchmark resources will be made available to facilitate reproducibility.

REFERENCES

Oluwatosin Alabi, Ko Ko Zayar Toe, Zijian Zhou, Charlie Budd, Nicholas Raison, Miaojing Shi, and Tom Vercauteren. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data*, 12(1):825, 2025.

Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E. Salem, Dominique Lamarque, Christian Daul, Michael A. Riegler, Kim V. Anonsen, Andreas Petlund, Pål Halvorsen, Jens Rittscher, Thomas de Lange, and James E. East. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10:75, 2023.

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 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, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *NeurIPS*, 2024.

Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *IEEE ICCV*, pp. 635–644, 2023.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *IEEE CVPR*, pp. 4985–4995, 2022.

Matthias Carstens, Franziska M. Rinner, Sebastian Bodenstedt, Alexander C. Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R. Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10:3, 2023.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, pp. 640–658, 2022.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *IEEE CVPR*, pp. 3151–3161, 2024.

Suhwan Cho, Seunghoon Lee, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Find first, track next: Decoupling identification and propagation in referring video object segmentation. In *IEEE ICCV Workshops*, pp. 3873–3883, 2025.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing wisdom in SAM2 for text-driven video segmentation. In *IEEE CVPR*, pp. 3395–3405, 2025.

Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE TPAMI*, 47(12):11400–11416, 2025.

Zheng Fang, Xiaoming Qi, Chun-Mei Feng, Jialun Pei, Weixin Si, and Yueming Jin. Spatio-temporal representation decoupling and enhancement for federated instrument segmentation in surgical videos. *IEEE TMI*, 45(5):2256–2268, 2026.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *IEEE CVPR*, pp. 29183–29192, 2025.

Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *IEEE CVPR*, pp. 13332–13341, 2024.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. LVOS: A benchmark for long-term video object segmentation. In *IEEE ICCV*, pp. 13480–13492, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-agent: A collaborative multi-agent system with reasoning and reflection for referring video object segmentation. In *IEEE CVPR*, pp. 39508–39517, 2026.

Woojeong Jin, Jaeho Lee, Heeseong Shin, Seungho Jang, Junhwan Heo, and Seungryong Kim. Agentrvos: Reasoning over object tracks for zero-shot referring video object segmentation. *arXiv preprint arXiv:2603.23489*, 2026.

Tianming Liang, Haichao Jiang, Yuting Yang, Chaolei Tan, Shuai Li, Wei-Shi Zheng, and Jian-Fang Hu. Long-RVOS: A comprehensive benchmark for long-term referring video object segmentation. *arXiv preprint arXiv:2505.12702*, 2025a.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. ReferDINO: Referring video object segmentation with visual grounding foundations. In *IEEE ICCV*, pp. 20009–20019, 2025b.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-local reasoning unified into a single large language model for video segmentation. In *IEEE CVPR*, pp. 8658–8667, 2025.

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.

Haofeng Liu, Mingqi Gao, Xuxiao Luo, Ziyue Wang, Guanyi Qin, Junde Wu, and Yueming Jin. ReSurgSAM2: Referring segment anything in surgical video via credible long-term tracking. In *MICCAI*, pp. 435–445, 2025a.

Qin Liu, Jianfeng Wang, Zhengyuan Yang, Linjie Li, Kevin Lin, Marc Niethammer, and Lijuan Wang. LiVOS: Light video object segmentation with gated linear matching. In *IEEE CVPR*, pp. 8668–8678, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

Sabrina Madad Zadeh, Tom François, Aurélie Comptour, Michel Canis, Nicolas Bourdel, and Adrien Bartoli. SurgAI3.8K: A labeled dataset of gynecologic organs in laparoscopy with application to automatic augmented reality surgical guidance. *Journal of Minimally Invasive Gynecology*, 30(5):397–405, 2023.

Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *IEEE ICCV*, pp. 920–930, 2023.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *IEEE ICCV*, pp. 9226–9235, 2019.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE CVPR*, pp. 724–732, 2016.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, et al. SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2024.

Guanyi Qin, Ziyue Wang, Daiyun Shen, Haofeng Liu, Hantao Zhou, Junde Wu, Runze Hu, and Yueming Jin. Structure matters: Revisiting boundary refinement in video object segmentation. In *IEEE ICCV*, pp. 14431–14442, 2025.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Accessed 29 July 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *ICLR*, 2025.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. MPG-SAM 2: Adapting SAM 2 with mask priors and global context for referring video object segmentation. In *IEEE ICCV*, pp. 23979–23989, 2025.

Tobias Ross, Annika Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020.

Abdelrahman Shaker, Syed Talal Wasim, Martin Danelljan, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Efficient video object segmentation via modulated cross-attention memory. In *IEEE WACV*, pp. 8670–8679, 2025.

Hongqiu Wang, Guang Yang, Shichen Zhang, Jing Qin, Yike Guo, Bo Xu, Yueming Jin, and Lei Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE TMI*, 43(12):4457–4469, 2024.

Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. AutoLaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *MICCAI*, pp. 486–496, 2022.

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *IEEE ICCV*, pp. 2761–2770, 2023a.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *IEEE CVPR*, pp. 4974–4984, 2022.

Qiangqiang Wu, Tianyu Yang, Wei Wu, and Antoni B. Chan. Scalable video object segmentation with simplified framework. In *IEEE ICCV*, pp. 13879–13889, 2023b.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, pp. 585–601, 2018.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *ECCV*, pp. 98–115, 2024.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. LoSh: Long-short text joint prediction network for referring video object segmentation. In *IEEE CVPR*, pp. 14001–14010, 2024.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In *ECCV*, pp. 452–469, 2024.

APPENDIX

We summarize the supplementary material from the following aspects:

SurgDVOS Dataset Analysis. Additional analyses of SurgDVOS, including spatiotemporal dynamics, clinical statistics, instrument diversity, and language annotation visualizations.

Additional Ablation Studies. Further ablation experiments on trajectory reflection, temporal memory recovery, and spatial recovery strategies of SurgATR.

Algorithmic Details. Detailed inference procedure and implementation details of SurgATR.

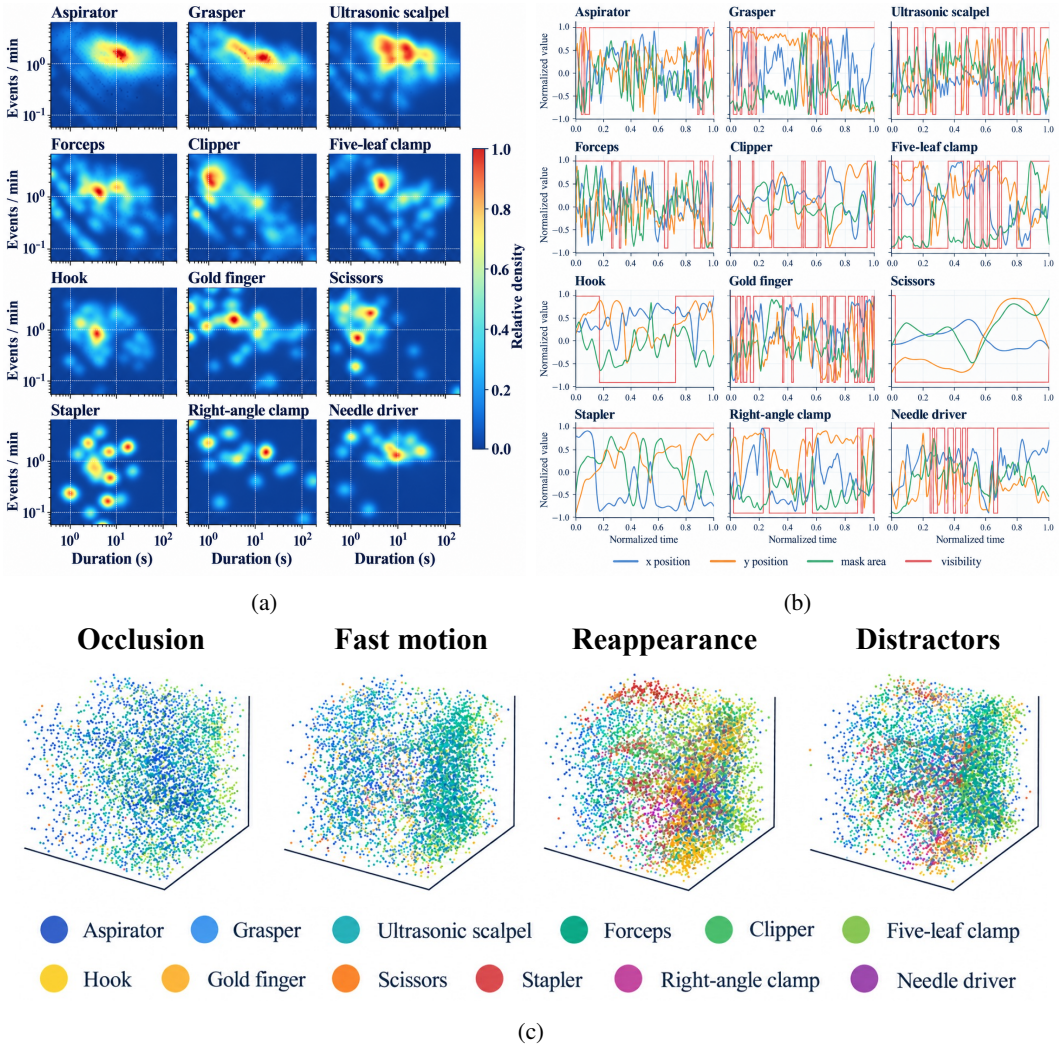


Figure 9: **Spatiotemporal characteristics of SurgDVOS. (a)** Instrument-wise distributions of trajectory duration and temporal event density. **(b)** Temporal evolution of normalized position, mask area, and visibility for representative instrument trajectories. **(c)** Challenge-conditioned spatiotemporal trajectory distributions under occlusion, fast motion, reappearance, and distractor interference, colored by instrument category.

A ADDITIONAL DETAILS OF THE SURGDVOS DATASET

This section provides additional characterization of SurgDVOS, beginning with its spatiotemporal trajectory dynamics, followed by the clinical background of the source cohort, the functional and

Clinical Variable	Value
Patients in Source Cohort	50
Age, mean ± SD (years)	52.48 ± 12.69
Cirrhosis	30 (60.0%)
Anatomical Resection	17 (34.0%)
Wide Resection	20 (40.0%)
Operation Time, median (min)	210.5
Blood Loss, median (mL)	220

Table 7: Clinical characteristics of the patients.

Variable	Category	Value
Tumor Location	Right Lobe	41 (82.0%)
	Left / Caudate Lobe	9 (18.0%)
Type	HCC	33 (66.0%)
Tumor Diameter	< 5 cm	35 (70.0%)
	5–10 cm	15 (30.0%)
Tumor Number	Single	41 (82.0%)
	Multiple	9 (18.0%)

Table 8: Tumor characteristics of the patients.

visual diversity of surgical instruments, and representative language-based annotations. Together, these analyses provide a more comprehensive view of the temporal, spatial, clinical, and semantic diversity represented in SurgDVOS.

Spatiotemporal Analysis. We further characterize the trajectory diversity of SurgDVOS from instrument-, temporal-, and challenge-specific perspectives, as shown in Fig. 9. The instrument-wise distributions in Fig. 9a reveal substantial variation in both trajectory duration and temporal event density, with several categories exhibiting multi-modal distributions spanning short, highly dynamic trajectories and longer sequences with sustained temporal activity. Fig. 9b further illustrates the temporal evolution of normalized spatial position, mask area, and visibility, revealing diverse motion patterns, scale variation, and intermittent visibility across instrument categories. Finally, Fig. 9c groups spatiotemporal trajectories by major challenge type. Occlusion, fast motion, reappearance, and distractor interference occur across multiple instrument categories, indicating that these challenges are not confined to specific tools or spatial configurations, highlighting the heterogeneous and densely evolving spatiotemporal dynamics captured by SurgDVOS.

Clinical Background. SurgDVOS is constructed from laparoscopic liver resection videos collected in real clinical practice. The source cohort comprises 50 patients, covering diverse anatomical conditions and operative settings. Table 7 summarizes the clinical characteristics of the source cohort. The cohort comprises 50 patients, with a mean age of 52.48±12.69 years. Cirrhosis is present in 60.0% of patients, while anatomical and wide resections account for 34.0% and 40.0% of cases, respectively. The median operation time is 210.5 minutes, with a median blood loss of 220 mL, providing additional context on the operative characteristics of the cohort. Table 8 further summarizes the anatomical and pathological characteristics of the tumors. Most tumors are located in the right lobe (82.0%), while 18.0% are located in the left or caudate lobe. Hepatocellular carcinoma (HCC) accounts for 66.0% of cases. In terms of tumor size, 70.0% of tumors are smaller than 5 cm, while 30.0% measure between 5 and 10 cm. Most patients have a single tumor (82.0%), whereas 18.0% have multiple tumors. Together, these characteristics reflect variation in tumor location, size, and multiplicity in SurgDVOS.

Instrument Categories and Visual Characteristics. SurgDVOS covers 12 surgical instrument categories spanning diverse surgical functions, morphologies, and motion patterns. Table 9 and Fig. 10 summarize their major functional and visual characteristics. Several categories share similar elongated structures, articulated tips, or closely related appearances, while individual instruments may undergo substantial changes in scale, orientation, articulation, and visibility due to tissue interaction and occlusion. These characteristics introduce considerable visual ambiguity and pose substantial challenges for maintaining physical target identity over time.

More Data Visualizations. In Fig. 11 and Fig. 12, we provide additional visualizations of the referring annotations in SurgDVOS.

B Additional Ablation Studies

We further analyze three key design choices in SurgATR on the referring VOS setting of SurgDVOS, including the temporal cues for trajectory reflection, the temporal memory rollback strategy, and the spatial recovery action space. All experiments follow the same implementation and evaluation protocol as in the main paper.

Table 9: Functions and visual characteristics of the 12 surgical instrument categories in SurgDVOS.

Instrument	**Function**	**Morphological or Motion Characteristics**
Aspirator	Removes blood and other fluids from the operative field.	Long hollow shaft with a narrow tip; frequently enters and exits the field and may be obscured by fluid.
Grasper	Grasps, retracts, and manipulates tissue.	Paired articulated jaws with frequent opening, closing, and rotation; often co-occurs with similar-looking graspers.
Ultrasonic scalpel (US)	Cuts and coagulates tissue using ultrasonic energy.	Elongated shaft with an asymmetric active jaw; commonly moves along tissue boundaries during dissection.
Forceps	Performs precise grasping, holding, or blunt dissection.	Slender opposing jaws with a small working tip; visually similar to graspers at low resolution.
Clipper	Applies ligating clips to vessels or ducts.	Paired jaws carrying a clip; typically approaches a target structure, closes briefly, and withdraws.
Five-leaf clamp (FLC)	Retracts organs or tissue to expose the operative field.	Expandable fan-shaped head with substantial shape changes between folded and deployed states.
Hook	Hooks, separates, or retracts tissue.	Thin shaft with a small curved tip; rapid tip motion and weak contrast can make its boundary difficult to identify.
Gold finger	Supports blunt dissection and organ retraction.	Articulating curved tip with substantial appearance changes during bending and rotation.
Scissors	Cuts tissue or sutures.	Two crossing blades with repeated opening and closing; the blades may appear narrow or merged when closed.
Stapler	Places staples while sealing or dividing tissue.	Large elongated jaw assembly that may occupy a substantial image region and is frequently truncated by the image boundary.
Right-angle clamp (RAC)	Performs blunt dissection and passes ligatures around structures.	Right-angled distal jaws whose rotation produces pronounced changes in projected shape and tip orientation.
Needle driver	Holds and controls a surgical needle during suturing.	Short robust jaws with a slender shaft and rapid fine-scale motion during needle manipulation.

Table 10: Ablation of temporal cues used in trajectory reflection.

Area	Overlap	Motion	Presence	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
✓				62.7	63.5	63.1
	✓			63.4	64.2	63.8
		✓		63.0	64.0	63.5
			✓	62.9	63.9	63.4
✓	✓	✓		64.6	65.4	65.0
✓	✓	✓	✓	65.3	66.0	65.6

Table 11: Ablation of temporal memory rollback strategies.

Rollback Strategy	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
No rollback	63.0	63.8	63.4
Previous frame $(t-1)$	63.5	64.2	63.9
Fixed $t-5$	64.2	64.9	64.6
Fixed $t-10$	63.8	64.5	64.2
Latest reliable anchor	65.3	66.0	65.6

Trajectory Reflection Cues. Table 10 evaluates the temporal cues used for failure reflection. Among individual cues, temporal overlap performs best, reaching 63.8 $\mathcal{J}\&\mathcal{F}$, while motion and target presence achieve 63.5 and 63.4, respectively, and area variation reaches 63.1. Combining area, overlap, and motion further improves performance to 65.0, and incorporating target presence achieves the full performance of 65.6. These results show that the four cues provide complementary evidence for identifying unreliable propagation under dense temporal dynamics.

Temporal Memory Recovery. Table 11 compares different memory rollback strategies. Rolling back to the previous frame only improves $\mathcal{J}\&\mathcal{F}$ from 63.4 to 63.9, as the nearby memory state may already contain propagation errors. Fixed rollback offsets achieve 64.6 at $t-5$ and 64.2 at $t-10$, indicating that a predefined temporal distance cannot consistently accommodate failures of different durations. In contrast, restoring the latest reliable anchor reaches 65.6, supporting adaptive failure-aware memory recovery.

Recovery Action Space. Table 12 further studies the spatial recovery actions used by the recovery agent. Positive-point prompting reaches 63.2 $\mathcal{J}\&\mathcal{F}$, while adding negative points improves the score to 63.8 and box re-initialization obtains 63.6. A fixed mixed policy further reaches 64.7, but remains below the 65.6 achieved by adaptive agent selection. This demonstrates the benefit of selecting corrective actions according to the ob-

Table 12: Ablation of the spatial recovery action.

Recovery Strategy	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Positive point only	62.8	63.6	63.2
Positive + negative points	63.4	64.2	63.8
Box re-initialization	63.2	64.0	63.6
Fixed mixed policy	64.3	65.1	64.7
Adaptive agent selection	65.3	66.0	65.6

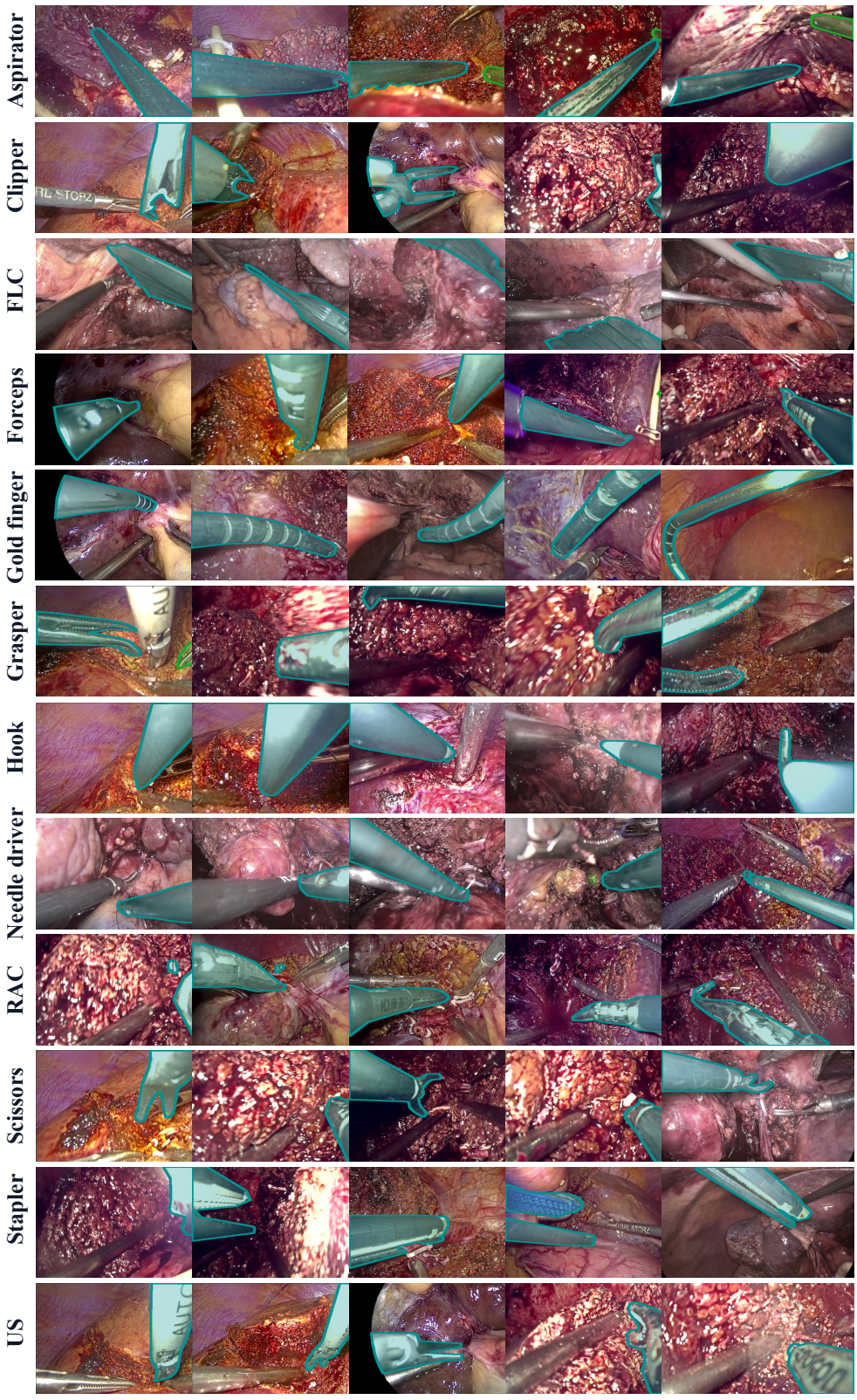
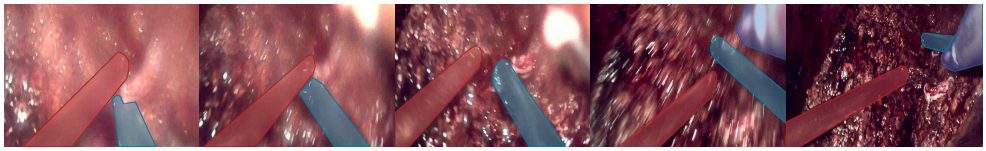
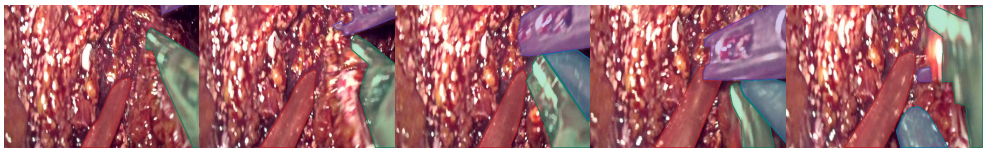


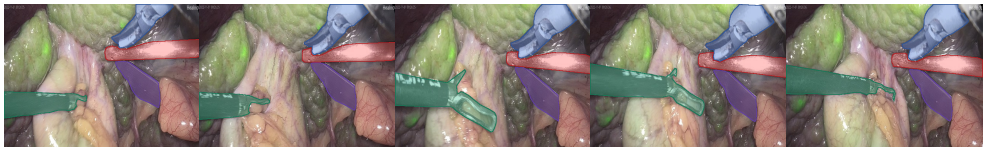
Figure 10: Visualization of annotated instruments in SurgDVOS.



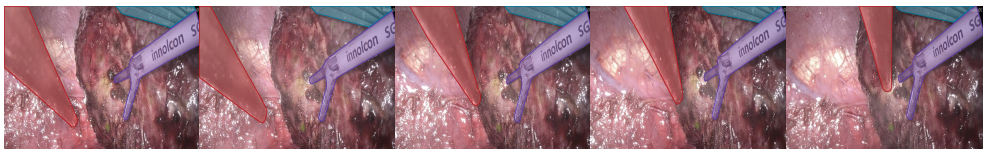
the aspirator clearing the blood-obscured lower-left operative field
the aspirator partly visible at the lower-right edge while the operative field is obscured by blood and mist
the grasper retracting tissue from the upper edge away from the central working area
the ultrasonic scalpel treating the blood-covered tissue surface in the upper-right field



the aspirator clearing the lower-central field beneath the ultrasonic scalpel
the second aspirator clearing the middle-right field beside the ultrasonic scalpel and opposite the first aspirator
the grasper positioned in the upper-right above the ultrasonic scalpel and the first aspirator
the ultrasonic scalpel treating tissue in the middle-right field between the grasper and the first aspirator



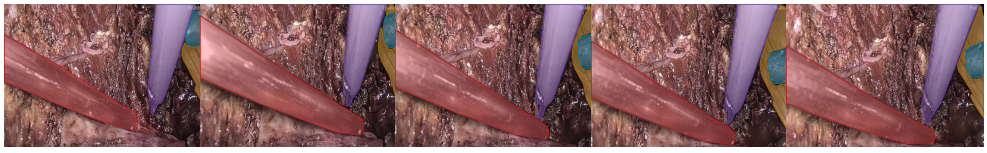
the aspirator touching the right edge in the upper-right of the image
the five-leaf clamp in the middle-right portion of the image
the grasper touching the left edge in the middle-left of the image
the ultrasonic scalpel touching the top edge in the upper-right of the image



the aspirator located along the top-left boundary in the upper-left of the image
the five-leaf clamp located along the top boundary in the upper-right of the image
the ultrasonic scalpel positioned at the upper-right of the image against the right boundary



the grasper visible in the lower-left of the image beside the bottom edge
the left grasper touching the top-left edge in the upper-left of the image
the ultrasonic scalpel positioned at the middle-left of the image against the left boundary

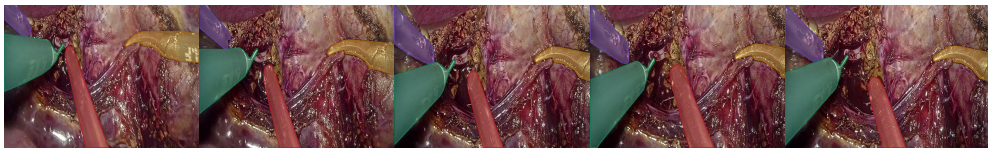


the central aspirator clearing the middle-central field to the left of the ultrasonic scalpel
the upper-right aspirator clearing the field above the ultrasonic scalpel and beside the five-leaf clamp
the five-leaf clamp retracting tissue in the upper-right field below the second aspirator and above the ultrasonic scalpel
the ultrasonic scalpel treating tissue in the middle-right field between the first aspirator and the upper-right second aspirator–five-leaf clamp pair

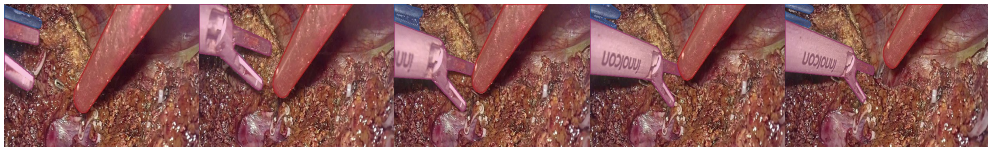
Figure 11: Representative referring annotations in SurgDVOS.



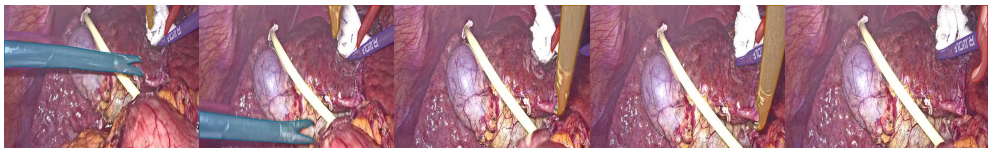
the aspirator clearing the lower-left field beneath the grasper and to the left of the second clipper
the central clipper applying a clip above the aspirator and to the right of the grasper
the grasper retracting tissue in the upper-central field above the aspirator and to the left of the second clipper



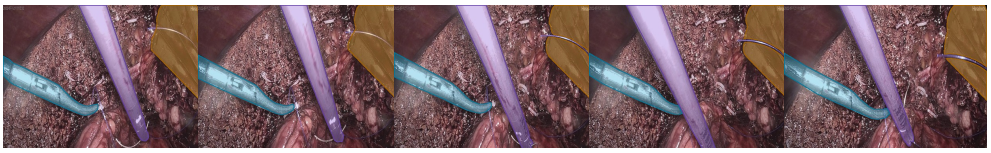
the aspirator clearing the central field between the left-sided grasper and the hook and the right-sided forceps
the forceps stabilizing tissue in the middle-right field to the right of the aspirator, the grasper, and the hook
the grasper retracting tissue in the middle-left field above the hook and to the left of the aspirator and the forceps
the hook dissecting or coagulating tissue in the middle-left field below the grasper



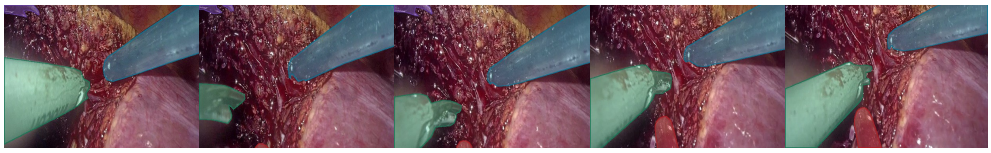
the upper aspirator located along the top boundary in the upper-center of the image
the grasper touching the top-left edge in the upper-left of the image
the ultrasonic scalpel touching the left edge in the upper-left of the image



the aspirator visible in the upper-right of the image beside the top-right edge
the left forceps located along the left boundary in the upper-left of the image
the upper forceps located along the top boundary in the upper-right of the image
the grasper positioned at the upper-right of the image against the top-right boundary



the aspirator clearing the lower-right field below the five-leaf clamp and to the right of the needle driver
the forceps stabilizing tissue in the middle-left field to the left of the needle driver and below the five-leaf clamp
the five-leaf clamp retracting tissue in the upper-right field above the aspirator
the needle driver handling the needle in the middle-central field between the forceps and the right-sided aspirator–five-leaf clamp pair



the lower aspirator clearing the lower-left field below the second aspirator and the ultrasonic scalpel
the upper aspirator clearing the upper-central field above the first aspirator and to the right of the grasper
the grasper retracting tissue at the upper-left edge above the first aspirator and to the left of the second aspirator
the ultrasonic scalpel treating tissue in the middle-left field above the first aspirator and below the grasper

Figure 12: Additional referring annotation examples in SurgDVOS.

served failure pattern rather than applying a fixed recovery strategy.

C ALGORITHMIC DETAILS

Algorithm 1 summarizes the complete inference procedure of SurgATR. The tracker first initializes the target at frame t_0 and performs baseline propagation, while trajectory reflection monitors recent target evolution. A lower reflection score indicates less reliable propagation, and recovery is triggered when the score falls below the threshold τ. The recovery agent then determines whether spatial re-anchoring, temporal memory rollback, or both are required.

Coordinated Recovery. Here, M_{t-1}^k denotes the input memory used by recovery branch k to process frame t. When temporal memory rollback is selected, SurgATR restores this memory to the most recent reliable anchor state by setting $M_{t-1}^k \leftarrow M_{t_a}^0$. Memory entries accumulated after t_a are discarded, and the intervening frames are not replayed. Otherwise, the branch retains the current baseline memory by setting $M_{t-1}^k \leftarrow M_t^0$. The tracker then processes the current frame t, using a corrective prompt when spatial re-anchoring is selected. When both interventions are selected, the corrective prompt is applied with the restored memory, preventing the discarded memory entries from influencing the current prediction.

Independent Recovery Branches. Each recovery hypothesis maintains an independent tracker state, leaving the baseline propagation unchanged. Cycle validation traces the recovered target backward toward the reliable anchor and retains the hypothesis only when the reverse trajectory remains reliable, *i.e.*, when its reflection score is no lower than τ. The state at frame t is preserved during reverse validation so that validated branches can subsequently propagate forward from frame $t+1$. The baseline and validated recovery trajectories are jointly resolved through global Viterbi selection.

Algorithm 1 SurgATR: Agentic Trajectory Recovery and Global Viterbi Selection

Require: Video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, task input q, tracker $\mathcal{T}$, recovery agent π_ϕ
Ensure: Final segmentation trajectory $\{\hat{Y}_t\}_{t=t_0}^{T}$

Initialize target prompt p_{t_0} from q
Initialize baseline state $(\hat{Y}^0_{t_0}, M^0_{t_0}, c^0_{t_0}, o^0_{t_0})$ from p_{t_0}
$\mathcal{K} \leftarrow \{0\}$ ▷ baseline and validated recovery branches
for $t = t_0 + 1, \ldots, T$ **do**
    Propagate baseline trajectory:
    $(\hat{Y}^0_t, M^0_t, c^0_t, o^0_t) \leftarrow \mathcal{T}(I_t, M^0_{t-1})$
    Compute trajectory cues:
    $h_t = \{h_t^{\text{area}}, h_t^{\text{overlap}}, h_t^{\text{motion}}, h_t^{\text{presence}}\}$
    if $\mathcal{R}(h_{t-w:t}) < \tau$ **then**
        Retrieve the most recent reliable anchor t_a
        Query the recovery agent for corrective interventions
        Initialize independent recovery branch k
        if temporal memory rollback is applied **then**
            $M^k_{t-1} \leftarrow \text{Copy}(M^0_{t_a})$ ▷ restore anchor memory; no replay
        else
            $M^k_{t-1} \leftarrow \text{Copy}(M^0_t)$ ▷ retain current baseline memory
        end if
        if spatial re-anchoring is applied **then**
            Generate corrective prompt p^k_t
            $(\hat{Y}^k_t, M^k_t) \leftarrow \mathcal{T}(I_t, M^k_{t-1}; p^k_t)$
        else
            $(\hat{Y}^k_t, M^k_t) \leftarrow \mathcal{T}(I_t, M^k_{t-1})$
        end if
        Trace backward to t_a using a separate validation state ▷ preserve M^k_t
        Compute reverse trajectory cues $h^{k,\text{rev}}_{t:t_a}$
        if $\mathcal{R}(h^{k,\text{rev}}_{t:t_a}) \geq \tau$ **then**
            $\mathcal{K} \leftarrow \mathcal{K} \cup \{k\}$ ▷ pass cycle validation
            for $u = t + 1, \ldots, T$ **do**
                $(\hat{Y}^k_u, M^k_u) \leftarrow \mathcal{T}(I_u, M^k_{u-1})$
            end for
        else
            Reject recovery branch k
        end if
    end if
end for
for $t = t_0, \ldots, T$ **do**
    for each available branch $k \in \mathcal{K}$ **do**
        Compute branch score $U_t(k)$
    end for
end for
Run global Viterbi selection over the baseline and validated branches
Backtrack the optimal branch sequence $s^*_{t_0:T}$
for $t = t_0, \ldots, T$ **do**
    $\hat{Y}_t \leftarrow \hat{Y}_t^{\,s^*_t}$
end for
return $\{\hat{Y}_t\}_{t=t_0}^{T}$