BRAIN: Bi-directional Motion Reasoning with Dynamic Memory Pruning for Surgical Video Segmentation

Chuanzhen Xu, Shumeng Li, Jian Zhang*, and Yinghuan Shi

State Key Laboratory of Novel Software Technology, Nanjing University, China
{chuanzhenxu, lism}@smail.nju.edu.cn, {zhang.jian, syh}@nju.edu.cn

Abstract. Accurate instrument tracking and segmentation is essential for robot-assisted surgery, where failures pose direct threats to patient safety. Current methods focus on dealing with the large intra-class variations of different instruments, neglecting the critical process of instrument's abrupt disappearance and reappearance, which can significantly mislead advanced models' predictions. We identify two critical challenges that hinder addressing of these issues: instrument misidentification due to abrupt positional shifts and redundant memory caused by long-term instrument absence. To address these two challenges, we present **B**idirectional motion **R**easoning with dyn**A**mic memory prun**I**ng for surgical video segmentatio**N** (BRAIN). The Bidirectional Motion Reasoning module first detects abrupt positional shifts from forward and backward trajectories and corrects errors via a mask reselection mechanism. Then, a Dynamic Memory Pruning module maintains a compact yet diverse memory bank by filtering out redundant frames captured during instrument absence. Extensive experiments demonstrate that our framework significantly outperforms existing state-of-the-art methods by 5.27% and 7.28% on EndoVis18 and EndoVis17 benchmarks, respectively. Our code is available at https://github.com/ChuanzhenXu/BRAIN.

Keywords: Surgical Instrument Segmentation · Abrupt Positional Shift · Long-term Instrument Absence · Dynamic Memory Pruning.

1 Introduction

Surgical instrument segmentation is a fundamental task for computer-assisted intervention systems [1,2,3,4,5], and forms a core component of downstream perception in robot-assisted surgery [6,7]. And robotic surgery has the potential to fundamentally change interventional medicine [8,9,10,11]. Unlike natural videos, surgical scenes exhibit highly unstable temporal dynamics: instruments frequently disappear due to occlusion, smoke, blood, or leaving the field of view, referred to as *instrument absence*, and may later reappear at completely different spatial locations, a phenomenon we term *abrupt positional shift*, both of which

* Corresponding author.

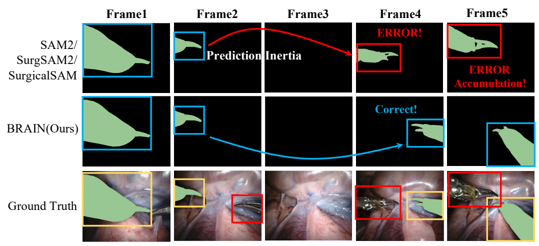

Fig. 1. Example of an abrupt positional shift. We compare the tracking results of advanced methods over five consecutive frames. When the shift occurs, all competing models incorrectly recognize the Suction Instrument as the Bipolar Forceps. This error arises from prediction inertia, where the models keep following the last known position and mistake visually similar instruments.

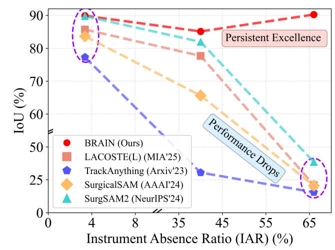

Fig. 2. Impact of IAR on segmentation performance. Methods perform similarly at low IAR, but competitors degrade as IAR increases, while BRAIN maintains robust performance across the range. Notably under high absence (IAR $> 65\%$), baselines collapse.

significantly challenge accurate instrument tracking and segmentation, as shown in Fig. 1. Although current video segmentation methods [12,13,14] achieve impressive performance when instruments move smoothly, they still cannot handle these two issues well because of two critical challenges we observed.

First, **motion inertia**: after an abrupt positional shift, models tend to keep tracking the last seen location and may latch onto visually similar but incorrect instruments (Fig. 1, top row). While such behavior may result in only accuracy degradation in generic tracking tasks, the same error in a surgical setting, where a robot depends on this tracking for guidance, could lead to unintended instrument manipulation, posing a direct and unacceptable risk to patient safety. Therefore, handling abrupt positional shifts is not just about model accuracy but a prerequisite for deploying autonomous surgical systems.

Second, **limited long-term reasoning** caused by prolonged instrument absence. To precisely characterize this issue, we define the Instrument Absence Ratio (IAR) as the proportion of frames in which the target instrument is absent within its active interval (from t_{start} to t_{end}):

$$\text{IAR} = \frac{\sum_{t=t_{\text{start}}}^{t_{\text{end}}} \mathbb{1}_{\text{absent}}(t)}{t_{\text{end}} - t_{\text{start}} + 1}, \tag{1}$$

where $\mathbb{1}_{\text{absent}}(t) = 1$ indicates the instrument is absent at frame t. As illustrated in Fig. 2, when the IAR is low, baseline models achieve over 80% segmentation accuracy. However, when the IAR exceeds 60%, their performance drops below 40%. Instrument absence is a very common issue in the real surgical videos, for instance, the absence ratio of prograsp forceps (sequence 5) can reach up to 66.9% in the EndoVis18 dataset. This finding reveals a fundamental limitation: current models lack the long-term reasoning needed to re-identify instruments after a long period of instrument absence, leading to tracking failure when instruments reappear and error accumulation when misidentified.

Consequently, motion inertia and limited long-term reasoning are two inherent failure modes of current methods, hindering the accurate tracking and segmentation of surgical instruments. Augmented with memory mechanism, recent foundation models [12,13,14,33,34,35] have shown impressive performance on general video segmentation tasks. However, it still cannot handle these two issues well because temporally correlated memory frames can amplify motion inertia and the redundant memory caused by instrument absence reduces the effectiveness of memory guidance for long-term reasoning. Motivated by these challenges, we propose **B**idirectional motion **R**easoning with dyn**A**mic memory prun**I**ng for surgical video segmentatio**N** (BRAIN) framework. BRAIN consists of two modules: a bidirectional motion reasoning (BMR) module and a dynamic memory pruning (DM) module. To mitigate instrument misidentification under abrupt positional shifts, often caused by motion inertia, the BMR module first detects abrupt positional shifts according to the motion consistency between forward and backward trajectories and corrects errors via a mask reselection mechanism. Additionally, since the redundant memory caused by instrument absence in the memory bank results in the limited long-term reasoning, the DM module is introduced to filter out redundant frames, forming a compact yet diverse memory set. Unlike SAM2's fixed-length memory updates, our adaptive pruning strategy retains the most informative frames across temporal contexts, which is beneficial for challenging surgical scenes involving long occlusions and reappearances. Moreover, different from recent surgical foundation models [33,34,35] that mainly improve generic video representation, memory quality, or referring-based initialization, BRAIN explicitly detects and corrects abrupt reappearance errors through bidirectional motion consistency and removes absence-induced redundant memories. In summary, our main contributions are as follows:

1. We formulate the critical challenges of abrupt positional shifts and prolonged instrument absence as a direct threat to the safety of robot-assisted surgery.
2. We propose a BMR module to address motion inertia by detecting and correcting errors of abrupt positional shifts through bidirectional trajectories.
3. We introduce a DM module that adaptively filters out redundant memory frames caused by instrument absence to enhance the long-term reasoning.

2 Methodology

We first describe the BMR module to detect and correct instrument misidentification due to abrupt positional shifts, then introduce the DM module to prune redundant memories for long-term reasoning, as shown in Fig. 3.

2.1 Bidirectional Motion Reasoning Module

To detect errors caused by motion inertia, the BMR module leverages forward and backward trajectories to detect the motion inconsistency between the start and end frames. Then the mask is reselected based on the motion consistency.

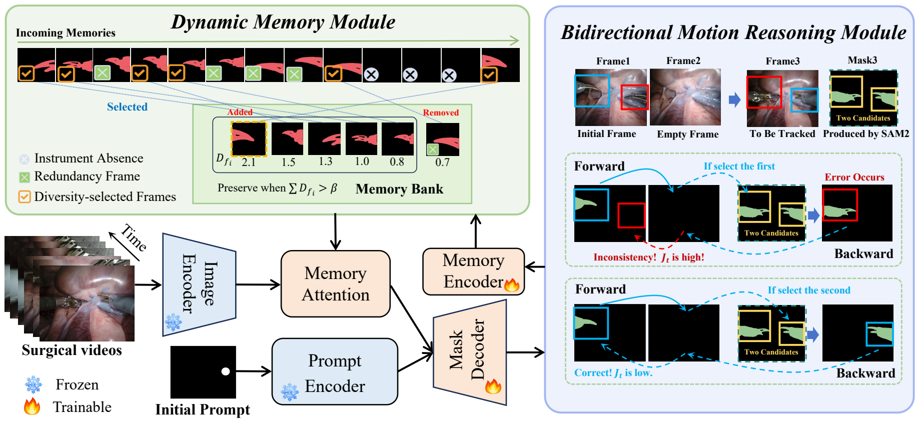

Fig. 3. Overview of our BRAIN architecture.

Forward Motion Tracking: We adopt the Kalman filter [15,16,17,18] to track the trajectory of surgical instruments and predict their positions in the next frame. It is a simple and effective motion tracker that can maintain smooth and physically consistent trajectories by optimally fusing prediction and observation. It alternates between *prediction* and *update* steps to accurately estimate current position of the instrument. We define the motion state vector at time t as $\mathbf{x}_t = (p_x, p_y, p_a, p_h, v_x, v_y, v_a, v_h)^\top \in \mathbb{R}^8$, where p and v denote the center coordinates and velocity of the instrument respectively. The observation $\mathbf{z}_t = (p_x, p_y, p_a, p_h)^\top \in \mathbb{R}^4$ contains the center coordinates and aspect ratio of current SAM2 mask proposal, which provides noisy measurements of the instrument's position and aspect ratio. The prediction step uses the previous state estimate and the motion model to predict the current state. The update step is used to correct the state estimate based on the actual observation:

$$\hat{\mathbf{x}}_{t|t-1} = F_t \hat{\mathbf{x}}_{t-1|t-1}, \tag{2}$$

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^\top + Q_t, \tag{3}$$

where $F_t \in \mathbb{R}^{8\times 8}$ is the state transition matrix, $Q_t \in \mathbb{R}^{8\times 8}$ is the process noise covariance matrix and $P_{t-1|t-1} \in \mathbb{R}^{8\times 8}$ is the state covariance matrix of the previous step in the Kalman filter.

For the i-th SAM2 candidate mask $M_{t,i}$ at time t, we compute the forward motion score $s^{t,i}_{\text{motion}}$ by calculating the Intersection over Union (IoU) between the mask bounding box $\mathcal{B}(M_{t,i})$ and the estimated bounding box $\mathcal{B}(\hat{\mathbf{x}}_{t|t-1})$ derived from the tracker's predicted coordinates: $s^{t,i}_{\text{motion}} = \text{IoU}\big(\mathcal{B}(\hat{\mathbf{x}}_{t|t-1}), \mathcal{B}(M_{t,i})\big)$. Then we combine the motion score with the IoU score predicted by SAM2 for each mask $s^{t,i}_{\text{SAM2}}$ to get the final motion score: $s^{t,i}_{\text{final}} = \gamma * s^{t,i}_{\text{motion}} + (1-\gamma) * s^{t,i}_{\text{SAM2}}$. For mask $M_{t,i}$ with $s^{t,i}_{\text{final}}$, if $s^{t,i}_{\text{final}} > k_{\text{tracker}}$, we employ it to update the Kalman filter's state estimate. The update step is based on the actual observation $\mathbf{z}_t =$

$(p_x, p_y, p_a, p_h)^\top$ derived from the mask $M_{t,i}$:

$$G_t = P_{t|t-1} H_t^\top \left(H_t P_{t|t-1} H_t^\top + R_t \right)^{-1}, \tag{4}$$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + G_t \left(\mathbf{z}_t - H_t \hat{\mathbf{x}}_{t|t-1} \right), \tag{5}$$

$$P_{t|t} = \left(I - G_t H_t \right) P_{t|t-1}, \tag{6}$$

where $G_t \in \mathbb{R}^{8\times4}$, $R_t \in \mathbb{R}^{4\times4}$, and $I \in \mathbb{R}^{8\times8}$ denote the Kalman gain, observation-noise covariance, and identity matrices, and $H_t \in \mathbb{R}^{4\times8}$ is the observation matrix that extracts the position components (p_x, p_y, p_a, p_h) from the state vector.

Rauch-Tung-Striebel Backward Smoothing: Forward tracking alone cannot recover from misidentification: an incorrect prediction may persist and propagate errors. Therefore, instead of only considering forward trajectory tracking, we also employ the backward trajectory tracking to further correct the errors. Given a temporal window $[t-L, \ldots, t]$, we employ the current prediction to backtrack the trajectory to test whether it is consistent with the historical trajectory. Given the forward Kalman estimates $\hat{\mathbf{x}}_{k|k}$, $\hat{\mathbf{x}}_{k|k-1}$, $P_{k|k}$ and $P_{k|k-1}$ for $k \in [t-L, \ldots, t]$, we perform Rauch-Tung-Striebel [19] (RTS) smoothing within the local window. The smoothed estimate at the end of the window is initialized as $\hat{\mathbf{x}}^s_{t|t} = \hat{\mathbf{x}}_{t|t}$, $P^s_{t|t} = P_{t|t}$. Then for $k = t-1, \ldots, t-L$, we compute the RTS gain $C_k = P_{k|k} \, F_{k+1}^\top \, P_{k+1|k}^{-1}$, and update the smoothed state and covariance:

$$\hat{\mathbf{x}}^s_{k|t} = \hat{\mathbf{x}}_{k|k} + C_k \left(\hat{\mathbf{x}}^s_{k+1|t} - \hat{\mathbf{x}}_{k+1|k} \right), \tag{7}$$

$$P^s_{k|t} = P_{k|k} + C_k \left(P^s_{k+1|t} - P_{k+1|k} \right) C_k^\top. \tag{8}$$

By iteratively performing the above steps, we obtain the smoothed sequence $\{\hat{\mathbf{x}}^s_{k|t}\}_{k=t-L}^t$ and $\{P^s_{k|t}\}_{k=t-L}^t$. To detect abrupt positional shifts, we measure the deviation of the forward prediction from the RTS-smoothed trajectory. We define the residual $\mathbf{r}_t = \hat{\mathbf{x}}_{t|t-1} - F_t \, \hat{\mathbf{x}}^s_{t-1|t}$, and compute the motion consistency score: $J_t = \mathbf{r}_t^\top \left(F_t \, P^s_{t-1|t} F_t^\top + Q_t \right)^{-1} \mathbf{r}_t$. Intuitively, J_t is the Mahalanobis distance between the current motion prediction and the trajectory manifold inferred from fixed-lag smoothing. A small J_t indicates smooth, continuous motion, whereas a large J_t reveals a possible teleport or abrupt re-entry after disappearance.

Motion-Guided Mask Reselection: With the motion score $s^{t,i}_{\text{motion}}$ and the motion consistency score J_t, we can compute the final score $S_{t,i}$ for each candidate mask $M_{t,i}$ for mask selection. If $J_t \geq \tau$, which indicates inconsistent motion (e.g., an abrupt positional shift or reappearance), we will use the appearance score to select the mask $S_{t,i} = s^{t,i}_{\text{app}}$. The appearance score $s^{t,i}_{\text{app}}$ is the cosine similarity between the embedding vector extracted from $M_{t,i}$ and that of the most confident mask in the memory bank.

If the motion consistency score $J_t < \tau$, which means the motion is consistent, we will use both the motion score and the appearance score to select the mask:

$$S_{t,i} = \lambda_{\text{motion}} s^{t,i}_{\text{motion}} + (1 - \lambda_{\text{motion}}) s^{t,i}_{\text{app}}, \tag{9}$$

where the fusion weights $\lambda_{\text{motion}} = e^{-n}$ is dynamically adjusted according to the instrument's absence duration, and n is the number of instrument absence frames since the last confident observation.

With the final score $S_{t,i}$, we can select the final mask as $M_t^\star = M_{t,\arg\max_i S_{t,i}}$.

2.2 Dynamic Memory Pruning Module

The memory bank performs a critical role in SAM2, which stores the initial frame and the six most recent frames. While this provides some temporal context for the video processing, it becomes less informative when instruments are occluded by anatomical structures or temporarily disappear. To address this, we propose a DM module that reduces redundancy in the memory bank.

Specifically, after the mask embedding of a target frame is added to the memory bank, we compute pairwise dissimilarities among all embeddings in the memory bank $B = \{f_0, f_1, f_2, \ldots, f_b\}$ with the size of $b+1$ including the initial frame. For each embedding $f_i \in B$, a dissimilarity score is calculated by aggregating its dissimilarities with all other embeddings $f_j \in B$ $(j \neq i)$:

$$D_{f_i} = \frac{1}{b-1} \sum_{f_j \in B, i \neq j}^{b-1} 1 - \text{sim}(f_i, f_j), \tag{10}$$

where D_{f_i} denotes the dissimilarity of embeddings f_i. We introduce a redundancy-aware selection mechanism that dynamically prunes embeddings with low dissimilarity in the memory bank. We sort dissimilarity scores in descending order, select the l highest to form $B' = \{f'_1, \ldots, f'_l\}$, and compute the sum of their dissimilarities: $\sum_{f'_i \in B'} \left(D_{f'_i}\right) > \beta$, where β denotes the redundancy threshold used to determine whether embeddings should be pruned. If the sum of dissimilarity score of the subset B' is greater than β, we will prune the subset $B \setminus B'$ from the memory bank, and $\setminus$ denotes the set difference. To ensure memory bank reliability, we enforce a minimum retention policy that always preserves the reference frame (the first frame where the target instrument appears) and keeps at most six additional memory frames (i.e., $b \leq 6$), regardless of redundancy. This ensures stable segmentation guidance even after prolonged target absence.

3 Experiments

Datasets: EndoVis17 [20] contains 8 training and 2 validation sequences (225 frames each), and EndoVis18 [21] contains 11 training and 4 validation sequences (149 frames each). Following Shvets [22], we preprocess both datasets for consistency with prior work. Adopting SurgSAM2's annotation protocol [23], we augment category-level labels with instance-level annotations for instance-wise evaluation and realistic generalization assessment.

Evaluation and Implementation Details: We follow the standard video object segmentation protocol and exclude the first and last frames of each sequence. We report Challenge IoU (CIoU), IoU, Dice, Boundary F1 (F), and J&F, where

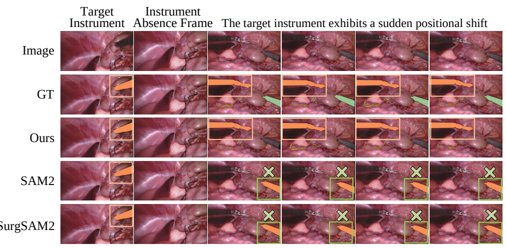

Fig. 5. Visualization of predicted segmentation masks on EndoVis18 dataset.

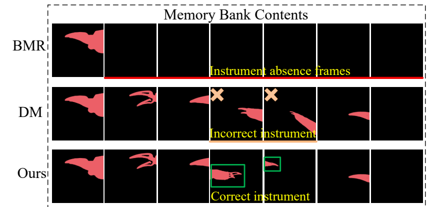

Fig. 6. Memory bank visualization under different settings.

J&F is the average of IoU (J) and F. Experiments run on an NVIDIA RTX A6000 (48GB) with bfloat16. We use a ViT-Small backbone initialized from SAM2 pretrained weights and train at 512×512. Following SAM2's alternating image/video schedule, we fine-tune only the mask decoder and memory encoder, and extend the SAM2 baseline with multi-mask output and IoU prediction. For video training, we use batch size 12, sequences up to 8 frames, and up to 3 objects per sequence; for image training, batch size 32 for 32 epochs. Learning rates are 2×10^{-4} (mask decoder) and 2×10^{-5} (memory encoder). At inference, we process the first frame at 1024×1024 and subsequent frames at 512×512. BRAIN is insensitive to γ, k_{tracker}, L; τ and β are interpretable. We set $L, \tau, \gamma, k_{\text{tracker}}$ to $15, 0.5, 0.3, 0.8$. The redundancy threshold β is set to 6 (Fig. 4).

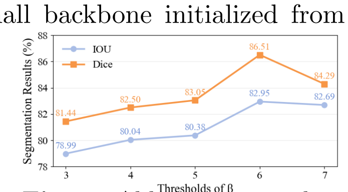

Fig. 4. Ablation on redundancy threshold β.

3.1 Quantitative Results

We quantitatively evaluate BRAIN on EndoVis17 and EndoVis18. As shown in Table 1, BRAIN significantly outperforms SAM2 and SurgSAM2 under identical settings. Fig. 5 shows that SAM2 and SurgSAM2 often mis-segment instruments, mainly due to instrument absence and abrupt positional shift. This abrupt discontinuity causes the model to associate a nearby instrument around the original position with the target, and the error then propagates to later frames. In contrast, BRAIN, with dynamic motion awareness and memory-guided prompts, correctly re-identifies the target after positional shifts, yielding more consistent tracking and improved segmentation under challenging dynamics.

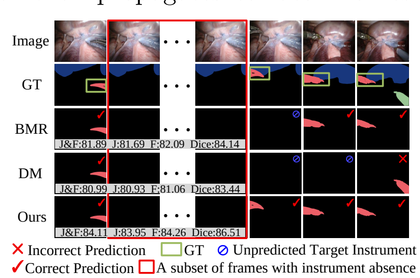

Fig. 7. Ablation results on EndoVis18.

3.2 Ablation Studies

We perform ablations on EndoVis18 to assess BMR and DM. Removing DM reverts to SAM2's fixed memory (six recent

Table 1. Comparative Results (%) on EndoVis18 and EndoVis17 Dataset.

Methods	EndoVis18									EndoVis17								
	One Point			Five Points			Full Mask			One Point			Five Points			Full Mask		
	J	F	Dice	J	F	Dice	J	F	Dice	J	F	Dice	J	F	Dice	J	F	Dice
SAM2	71.50	73.00	74.20	76.20	76.30	80.00	78.40	78.60	81.70	81.10	83.80	85.10	82.00	85.90	86.90	85.90	89.10	90.20
SurgSAM2	72.60	73.80	76.70	80.90	80.70	84.90	81.90	82.00	85.30	82.70	84.20	87.30	86.90	89.10	91.40	88.20	90.60	92.30
BRAIN	81.25	81.64	83.80	82.17	82.50	84.74	83.95	84.26	86.51	87.75	89.17	89.60	90.75	93.70	94.13	90.83	93.71	94.18

Table 2. Comparison results (%) on Dataset EndoVis18.

Category	Method	CIoU	IoU	Instrument Classes IoU						
				BF	PF	LND	SI	CA	MCS	UP
Task-specific Model	ISINet	73.00	70.94	73.83	48.61	30.98	37.68	0.00	88.16	2.16
	TraSeTR	76.20	-	76.30	53.30	46.50	40.60	13.90	86.30	17.50
	S3Net	76.20	74.02	77.22	50.87	19.83	50.59	0.00	92.12	7.44
	MATIS (Full)	84.26	79.12	83.52	41.90	66.18	70.57	0.00	92.96	23.13
	QPD	77.77	78.43	82.80	60.94	19.96	49.70	0.00	**93.93**	0.00
	LACOSTE (L)	86.78	85.68	85.71	77.68	67.97	76.39	45.29	93.27	22.27
SAM-based Model	MT-RCNN + SAM	78.49	78.49	79.83	74.86	43.12	62.88	16.74	91.62	23.45
	Mask2Former + SAM	78.72	78.72	85.95	82.31	44.08	0.00	49.80	92.17	13.18
	TrackAnything (5 Points)	65.72	60.88	72.90	31.07	64.73	10.24	12.28	61.05	17.93
	PerSAM	49.21	49.21	51.26	34.40	46.75	16.45	15.07	52.28	25.62
	PerSAM (Fine-Tune)	52.21	52.21	57.19	36.13	53.86	14.34	25.94	54.66	18.57
	SurgicalSAM	80.33	80.33	83.66	65.63	58.75	54.48	39.78	88.56	21.23
	BRAIN (Ours)	**92.45**	**85.81**	**89.95**	**85.12**	**77.11**	**77.00**	**65.91**	92.76	**61.25**

frames) while keeping BMR. While this improves over the baseline, BMR alone fails after long absences with distant reappearance because memory is dominated by empty frames (Fig. 6, first row), leading to errors in Fig. 7. Conversely, keeping DM but removing BMR improves overall performance but is sensitive to abrupt reappearances: without motion cues, incorrect masks may be stored (Fig. 6, second row), and recovery typically needs several correct post-reappearance frames (Fig. 7). These results show BMR and DM are complementary: BMR improves mask quality and reduces drift, enabling DM to maintain informative memory for robust long-term tracking under occlusion and motion discontinuity.

3.3 Comparison with Other Methods

To validate BRAIN, we compare it with task-specific methods [24,25,26,27,28] and SAM-based approaches [12,23,29,30,31,32] across multiple metrics using results reported in the original papers. Such comparisons may be imperfect because some methods are prompt-free at inference and many prior works report only category-level masks [23]. Following SurgSAM2 [23], we report instance-level evaluation by segmenting same-class instruments separately. As shown in Table 2, BRAIN achieves SOTA CIoU on EndoVis18 and leads across categories (Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Suction Instrument (SI), Clip Applier (CA), Monopolar Curved Scissors (MCS), and Ultrasound Probe (UP)), improving CA and UP by 20.62% (from 45.29% to 65.91%) and 35.63% (from 25.62% to 61.25%), respectively. These gains mainly stem from instrument absence and abrupt positional shifts: BRAIN's motion-aware tracking reduces errors and propagation, surpassing strong task-specific models (e.g., LACOSTE) and SAM-based baselines (e.g., SurgicalSAM [36]). Under identical settings, it consistently outperforms SAM2 and SurgSAM2 on

J (IoU), F, and Dice (Table 1), indicating improved temporal consistency. Besides, BRAIN (46.75 FPS, 0.02 s/frame, 3.23 GB) adds negligible overhead over SAM2 (52.18 FPS, 0.02 s/frame, 3.14 GB), since extra modules use simple matrix computation, RTS uses a short temporal window (L=15), and DM prunes redundant frames to further speed up the process.

4 Conclusion

We present BRAIN for surgical video segmentation, explicitly targeting instrument absence and abrupt positional shifts. BRAIN combines bidirectional motion reasoning to detect and correct motion inconsistencies with dynamic memory pruning module to prune redundant frames and maintain a compact, informative memory bank. Experiments on two public benchmarks show consistent gains over prior methods, highlighting the value of motion-guided memory management for reliable and safer surgical video understanding.

Acknowledgments. This work was supported by NSFC Project (62506162, 62506005, 62192783), Jiangsu Science and Technology Project (BK20251241), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), and the "111 Center" (No. B26023).

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

References

1. Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.A., Stoyanov, D.: Exploring intra- and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging 41(11), 2991–3002 (2022)
2. Loftus, T.J., Tighe, P.J., Filiberto, A.C., Efron, P.A., Brakenridge, S.C., Mohr, A.M., Rashidi, P., Upchurch, G.R., Bihorac, A.: Artificial intelligence and surgical decision-making. JAMA Surgery 155(2), 148–158 (2020)
3. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science: from concepts toward clinical translation. Medical Image Analysis 76, 102306 (2022)
4. Curtis, N.J., Foster, J.D., Miskovic, D., Brown, C.B.S., Hewett, P.J., Abbott, S., Hanna, G.B., Stevenson, A.R.L., Francis, N.K.: Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surgery 155(7), 590–598 (2020)
5. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9522–9531 (2021)
6. Sheng, M., Fan, J., Liu, D., Kikinis, R., Cai, W.: AMNCutter: affinity-attention-guided multi-view normalized cutter for unsupervised surgical instrument segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4533–4544 (2025)

7. Ni, Z.L., Bian, G.B., Wang, G.A., Zhou, X.H., Hou, Z.G., Chen, H.B., Xie, X.L.: Pyramid attention aggregation network for semantic segmentation of surgical instruments. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(07), pp. 11782–11790 (2020)
8. Ciuti, G., Webster III, R.J., Kwok, K.W., Menciassi, A.: Robotic surgery. Nature Reviews Bioengineering (2025)
9. Hao, P., Wang, H., Yang, G., Zhu, L.: Enhancing visual reasoning with LLM-powered knowledge graphs for visual question localized-answering in robotic surgery. IEEE Journal of Biomedical and Health Informatics (2025)
10. Liang, Z., Chiu, Z.Y., Richter, F., Yip, M.C.: Differentiable rendering-based pose estimation for surgical robotic instruments. arXiv (2025)
11. Gorpas, D., Phipps, J., Bec, J., Ma, D., Dochow, S., Yankelevich, D., Sorger, J., Popp, J., Bewley, A., Gandour-Edwards, R., et al.: Autofluorescence lifetime augmented reality as a means for real-time robotic surgery guidance in human patients. Scientific Reports 9(1), 1187 (2019)
12. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: segment anything in images and videos. arXiv (2024)
13. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical SAM 2: segment medical images as video via segment anything model 2. arXiv (2024)
14. Kikuchi, S., Kouno, A., Matsuzaki, H.: Memory-enhanced temporal learning: leveraging SAM2's memory modules for consistent video segmentation. In: MICCAI Workshop (MSB EMERGE) (2025)
15. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering (1960)
16. Yin, J., Wu, F., Su, H., Huang, P., Yu, Q.: Improvement of SAM2 algorithm based on Kalman filtering for long-term video object segmentation. Sensors 25(13), 4199 (2025)
17. Yang, C., Huang, H., Chai, W., Jiang, Z., Hwang, J.: SAMURAI: adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv (2024)
18. Yang, Q., Yao, Y., Cui, M., Bo, L.: MoSAM: motion-guided segment anything model with spatial-temporal memory selection. arXiv (2025)
19. Rauch, H.E., Tung, F., Striebel, C.T.: Maximum likelihood estimates of linear dynamic systems. AIAA Journal 3(8), 1445–1450 (1965)
20. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv (2019)
21. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv (2020)
22. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: Proceedings of the International Conference on Machine Learning and Applications, pp. 624–628 (2018)
23. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical SAM 2: real-time segment anything in surgical video by efficient frame pruning. arXiv (2024)
24. González, C., Bravo-Sánchez, L., Arbelaez, P.: ISINet: an instance-based approach for surgical instrument segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 595–605 (2020)

25. Baby, B., Thapar, D., Chasmai, M., Banerjee, T., Dargan, K., Suri, A., Banerjee, S., Arora, C.: From forks to forceps: a new framework for instance segmentation of surgical instruments. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6191–6201 (2023)
26. Zhao, Z., Jin, Y., Heng, P.A.: TraseTr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 11186–11193 (2022)
27. Dhanakshirur, R.R., Shastry, K.N.A., Borgavi, K., Suri, A., Kalra, P.K., Arora, C.: Learnable query initialization for surgical instrument instance segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 728–738 (2023)
28. Wang, Q., Zhao, S., Xu, Z., Zhou, S.K.: LACOSTE: exploiting stereo and temporal contexts for surgical instrument segmentation. Medical Image Analysis 99, 103387 (2025)
29. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5188–5197 (2019)
30. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: segment anything meets videos. arXiv (2023)
31. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv (2023)
32. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1290–1299 (2022)
33. Tian, Q., Liao, H., Huang, X., Yang, B., Lei, D., Ourselin, S., Liu, H.: EndoMamba: An Efficient Foundation Model for Endoscopic Videos via Hierarchical Pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 224–234 (2025)
34. Yin, M., Wang, F., Ye, X., Meng Y., Fu, Z.: Memory-Augmented SAM2 for Training-Free Surgical Video Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 328–337 (2025)
35. Liu, H., Gao, M., Luo, X., Wang, Z., Qin, G., Wu, J., Jin, Y.: ReSurgSAM2: Referring Segment Anything in Surgical Video via Credible Long-term Tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 435–445 (2025)
36. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: SurgicalSAM: efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38(7), pp. 6890–6898 (2024)