Supplementary Information

Is Segment Anything Model 2 All You Need for Surgery Video Segmentation? A Systematic Evaluation

Cheng Yuan[1,†], Jian Jiang[1,†], Kunyi Yang[1,†], Lv Wu[1], Rui Wang[1], Zi Meng[1], Haonan Ping[1], Ziyu Xu[1], Yifan Zhou[1], Wanli Song[1], Hesheng Wang[2], Yueming Jin[3], Qi Dou[4], Yutong Ban[1,*]

[†]These authors contributed equally to this work.

[1]UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[3]Department of Biomedical Engineering and Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

* Corresponding author: Yutong Ban (yban@sjtu.edu.cn)

Supplementary Note 1

This supplement provides additional details to support the main text, including SAM2's technical implementation, dataflow specifications, and comprehensive segmentation results across all evaluated datasets.

1 Dataflow and Technical Details of SAM2

The Segment Anything Model (SAM) [1] revolutionized image segmentation through its prompt-based zero-shot approach, eliminating the need for task-specific training. Building on this foundation, SAM2 extends these capabilities to video processing while maintaining image segmentation performance. Both architectures share core components (image encoder, prompt encoder, and mask decoder), but SAM2 introduces critical video-oriented modifications.

1.1 Architectural Modifications

The SAM2 image encoder employs a Hiera encoder [2] pretrained with Masked Autoencoder (MAE) [3], producing multi-scale feature embeddings through three key modifications: First, relative positional biases are removed from all encoder layers to improve computational efficiency. Second, a window-based interpolation scheme replaces absolute positional encoding for global positional embeddings. Third, hierarchical feature extraction is implemented with progressive downsampling rates of [4×, 8×, and 16×] to preserve spatial details.

The novel memory attention module enables temporal processing through two primary mechanisms. A stack of transformer blocks first performs self-attention on current frame features, then engages in cross-attention with a memory bank that utilizes 2D spatial RoPE [4] for positional encoding. This memory bank dynamically integrates information from previous frames through a dedicated memory encoder that fuses mask predictions with image embeddings.

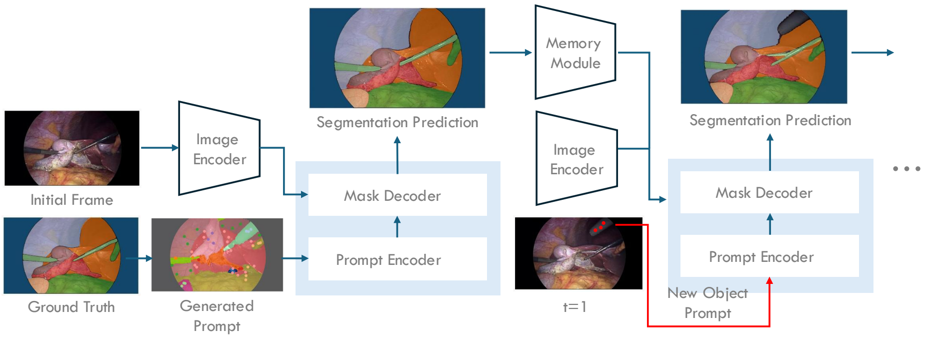


Figure 1: SAM2 surgical video processing pipeline. Three-stage architecture showing (1) prompt generation from surgical tool detectors, (2) initial segmentation using hybrid point-box prompts, and (3) temporal propagation through memory-guided mask refinement.

1.2 Surgical Implementation Pipeline

As shown in Fig. 1, our surgical adaptation implements three key enhancements. The first enhancement involves instrument-aware prompting that combines bounding boxes from real-time tool detectors with anatomical landmarks identified in the surgical field. Secondly, memory reinitialization protocols automatically reset the memory bank during critical phase transitions such as instrument changes or surgical step transitions. Finally, an automatic occlusion handling system leverages the mask decoder's occlusion prediction head to resolve complex tool-tissue interactions through iterative mask refinement.

2 Comprehensive Segmentation Results

This part presents extended qualitative evaluations across seven surgical benchmarks, demonstrating SAM2's capabilities in diverse clinical scenarios. Figures 2–**??** showcase performance on EndoVis2018, Cholecseg8k, and EndoNerf. The visualizations highlight three critical capabilities: (1) precise multi-instance segmentation in crowded surgical fields, (2) temporal consistency during complex tool-tissue interactions, and (3) robust handling of surgical domain challenges including blood occlusion, smoke artifacts, and specular reflections. For conciseness, we focus on representative cases that exemplify SAM2's advantages over image-only foundation models, with quantitative analysis provided in the main text.

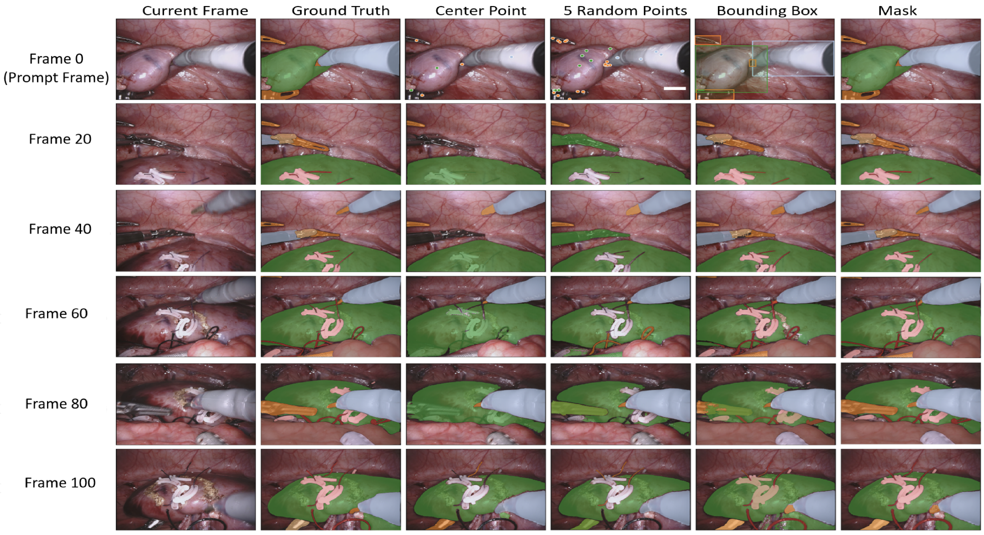


Figure 2: Visualization of segmentation results on the EndoVis2018 dataset.

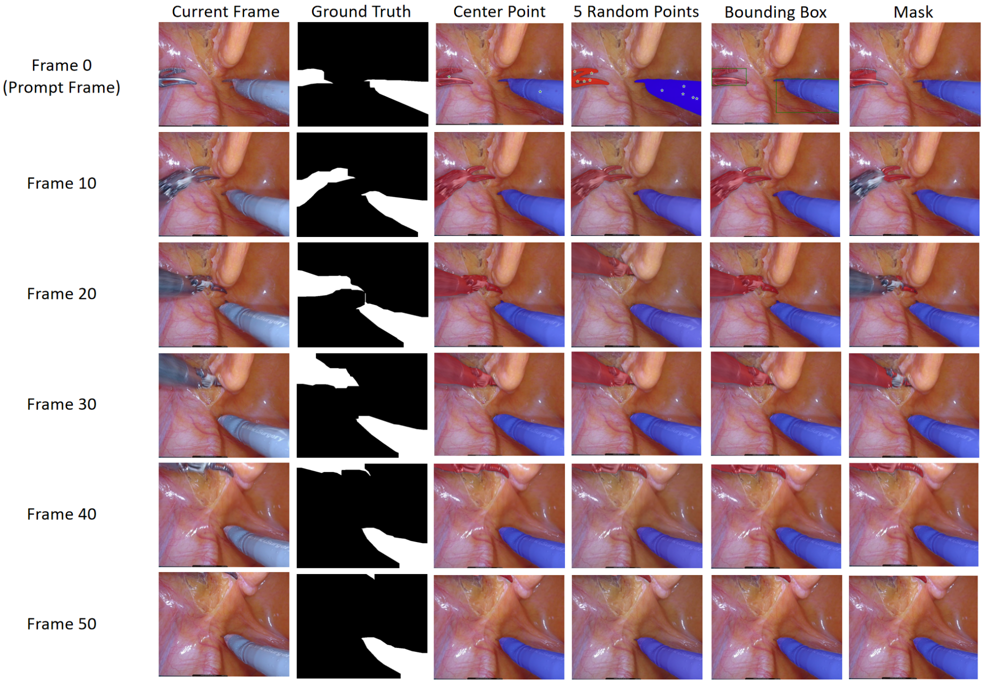


Figure 3: Visualization of segmentation results on the EndoNerf dataset.

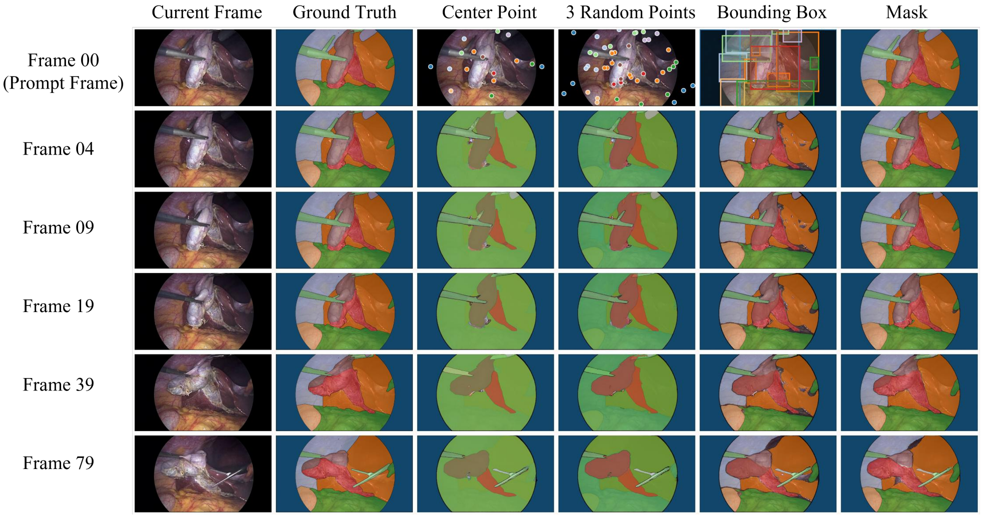


Figure 4: Visualization of segmentation results on the Cholecseg8k dataset.

Supplementary Note 2 This supplement presents ablation studies evaluating the impact of prompt quality, point quantity, and prompt perturbations on SAM2's performance, along with an analysis of failure cases in various surgical datasets.

Table 1: Quantitative evaluation on Cholecseg8k dataset with different prompts.

Method	Reinitialization	Mean IoU	Mean Dice	Tissues IoU	Instruments IoU	Tissues Dice	Instruments Dice
Mask2Former[5]	-	69.10%	-	-	-	-	-
HRNet32[5]	-	61.10%	-	-	-	-	-
HRNet32 + SP-TCN[5]	-	**65.37%**	-	-	-	-	-
Swin base[5]	-	68.42%	-	-	-	-	-
Swin base + SP-TCN[5]	-	**69.38%**	-	-	-	-	-
SAM2-Points-Random 1 (Standard)	-	63.43%	71.53%	55.47%	69.86%	70.18%	76.40%
SAM2-Points-Random 1	60 frames	65.09%	73.32%	57.28%	72.22%	71.34%	79.29%
SAM2-Points-Random 1	30 frames	64.72%	73.06%	56.83%	70.81%	71.32%	77.70%
SAM2-Points-Random 1	10 frames	65.90%	74.28%	58.02%	75.13%	72.08%	82.23%
SAM2-Points-Random 1	5 frames	65.77%	74.18%	57.64%	75.72%	71.75%	82.83%
SAM2-Points-Random 3 (Standard)	-	71.24%	79.17%	63.76%	77.64%	77.14%	84.06%
SAM2-Points-Random 3	60 frames	71.64%	79.50%	64.03%	79.29%	77.37%	85.76%
SAM2-Points-Random 3	30 frames	72.58%	80.39%	65.46%	78.99%	78.43%	85.59%
SAM2-Points-Random 3	10 frames	**73.48%**	**81.30%**	66.25%	81.63%	79.09%	88.23%
SAM2-Points-Random 3	5 frames	73.34%	81.23%	66.15%	82.01%	77.14%	84.06%
SAM2-Points-Center (Standard)	-	61.98%	70.08%	60.72%	76.00%	67.20%	82.97%
SAM2-Points-Center	60 frames	62.08%	70.20%	60.73%	77.03%	67.15%	83.93%
SAM2-Points-Center	30 frames	62.38%	70.40%	60.97%	78.12%	67.26%	84.85%
SAM2-Points-Center	10 frames	62.68%	70.80%	61.21%	79.60%	67.50%	86.29%
SAM2-Points-Center	5 frames	62.65%	70.84%	61.38%	80.21%	67.44%	86.98%
SAM2-Points-Center (1 negative point)	-	62.62%	70.75%	61.93%	75.69%	68.16%	82.81%
SAM2-Points-Center (3 negative points)	-	61.54%	69.88%	61.64%	73.52%	67.57%	80.20%
SAM2-Bbox (Standard)	-	83.34%	89.00%	82.15%	79.98%	89.13%	85.85%
SAM2-Bbox	60 frames	84.31%	89.89%	83.07%	81.66%	89.92%	87.58%
SAM2-Bbox	30 frames	84.93%	90.49%	83.71%	83.12%	90.35%	89.04%
SAM2-Bbox	10 frames	86.11%	91.57%	84.88%	84.83%	91.34%	90.74%
SAM2-Bbox	5 frames	**86.50%**	**91.90%**	85.32%	85.57%	91.63%	91.49%
SAM2-Mask (Standard)	-	88.95%	92.73%	88.40%	81.62%	93.49%	86.83%
SAM2-Mask	60 frames	90.14%	93.69%	89.64%	84.04%	94.27%	89.05%
SAM2-Mask	30 frames	91.14%	94.48%	90.59%	86.24%	94.86%	90.95%
SAM2-Mask	10 frames	92.85%	95.77%	92.40%	89.19%	95.98%	93.35%
SAM2-Mask	5 frames	**93.93%**	**96.48%**	93.55%	91.04%	96.60%	91.49%

3 Ablation Study

Introducing noise (scale, shift, etc.) into the initial prompts shows that SAM2's performance can be significantly influenced by the quality of the prompts. As expected, these noises significantly impact its performance compared to using accurate instructions. As shown in Table 2, adding noise to the prompt will have a significant influence on the accuracy of the segmentation performance. For bounding box prompting, the IoU and Dice decrease from 83.34% to 45.44% and from 89.00% to 51.28%. For mask prompting, the IoU and Dice decrease from 88.95% to 48.36% and 92.73% to 55.23%. These results indicated the importance of providing SAM2 with clear and accurate guidance to achieve optimal segmentation results.

The number of points provided in point prompts significantly impacts SAM2's performance. Contrary to initial expectations, simply increasing the number of points does not lead to continuous improvement. We experimented with two point sampling strategies: random sampling and grid sampling, both derived from ground-truth masks. With random sampling, performance improved slightly when increasing the number of points in the single-digit range. However, performance began to decline after 10 points and seriously decreased beyond 30 points (Fig. 5). For example, using five random points resulted in an mIoU of 26.20% and an mDice of 32.28%. Increasing the number of points to 60 significantly degraded performance, resulting in an mIoU of 5.20% and an mDice of 8.71%. The grid sampling showed similar trends. Performance increased slightly as the grid spacing decreased, but then began to decline (Fig. 5). With a grid spacing of 32 pixels, the mIoU was 20.54% and the mDice was 26.11%. Reducing the spacing to 4 pixels resulted in a significant performance drop, with an mIoU of 5.77% and an mDice of 9.52%. These results suggest that an optimal number of points exists for point prompts, and exceeding this threshold can be detrimental to SAM2's performance.

Table 2: Cholecseg8k - Ablation Study on Noise Perturbation

Method	**Noised**	**Mean**		**IoU**		**Dice**	
		IoU	Dice	Tissues	Instruments	Tissues	Instruments
SAM2-Bbox (Standard)	-	**83.34**	**89.00**	82.15	79.98	89.13	85.85
SAM2-Bbox	0.1 scale	80.01	86.29	79.69	79.00	86.48	85.16
SAM2-Bbox	0.3 scale	70.24	77.52	70.60	69.46	78.70	75.81
SAM2-Bbox	0.5 scale	61.69	69.28	60.18	62.16	68.56	68.30
SAM2-Bbox	0.1 shift	67.18	74.88	66.84	62.95	75.10	69.84
SAM2-Bbox	0.3 shift	47.86	54.41	47.97	43.81	54.76	47.88
SAM2-Bbox	0.5 shift	45.44	51.28	44.98	37.47	51.15	40.94
SAM2-Bbox	0.1 shift+scale	65.82	73.33	65.48	63.54	73.42	70.38
SAM2-Bbox	0.3 shift+scale	47.34	53.96	45.92	39.74	52.97	43.86
SAM2-Bbox	0.5 shift+scale	46.59	52.87	44.48	41.33	51.05	45.30
SAM2-Mask (Standard)	-	**88.95**	**92.73**	88.40	81.62	93.49	86.83
SAM2-Mask	0.1 noise	68.37	76.30	67.44	55.22	75.84	62.49
SAM2-Mask	0.3 noise	55.01	62.84	53.43	53.98	61.53	59.63
SAM2-Mask	0.5 noise	48.36	55.23	48.55	41.47	55.45	46.48

4 Failure Case Analysis

The SAM2 model, when applied to various surgical datasets, encountered several challenges that impacted its performance in surgical tool tracking and segmentation tasks. These challenges primarily fell into two categories: (1) failure cases in datasets with ground truth annotations, and (2) issues arising from AutoMask generation and its application as pseudo labels in datasets lacking ground truth. In the first category, key issues included difficulties with prompting and initialization, instrument identification and confusion, and maintaining temporal consistency. The second category revealed challenges in generating accurate initial segmentations and propagating them to subsequent frames. This analysis explores these failure cases in detail, providing insights into areas for improvement and future research directions.

In datasets with ground-truth annotations, the SAM2 model encountered several significant challenges, primarily related to prompting and initialization. The quality and placement of initial prompts proved crucial in guiding accurate segmentation. This was particularly evident in various scenarios encountered during the model's application. Multipart objects presented unique challenges in prompting, especially in the Cholecseg8k dataset. When dealing with large multi-part objects in single-point prompting, incorrect prompt placements often occur due to overlapping object centers (Fig. 8-(a)). This highlighted the complexity of accurately identifying and segmenting objects with multiple components. The bounding box prompt faced difficulties with complex and irregularly shaped objects. In the EndoVis2018 dataset, for example, closely connected objects with overlapping boxes posed significant challenges in bounding box prompt experiments (Fig. 8-(b) and (c)). This underscored the limitations of bounding box approaches when dealing with intricate object geometries and spatial relationships. Incomplete object appearance in prompt frames, such as when only a tooltip was visible, significantly affected segmentation in subsequent frames (Fig. 8-(d)). This scenario highlighted the model's sensitivity to partial object visibility during the prompting phase, emphasizing the need for robust handling of incomplete visual information.

The model also struggled with instrument identification and confusion, particularly when dealing with similar instruments in dynamic scenarios. In the EndoVis17 dataset, the model occasionally confused new instruments appearing in locations similar to previously prompted ones. This issue was particularly evident when an instrument was removed from the frame and

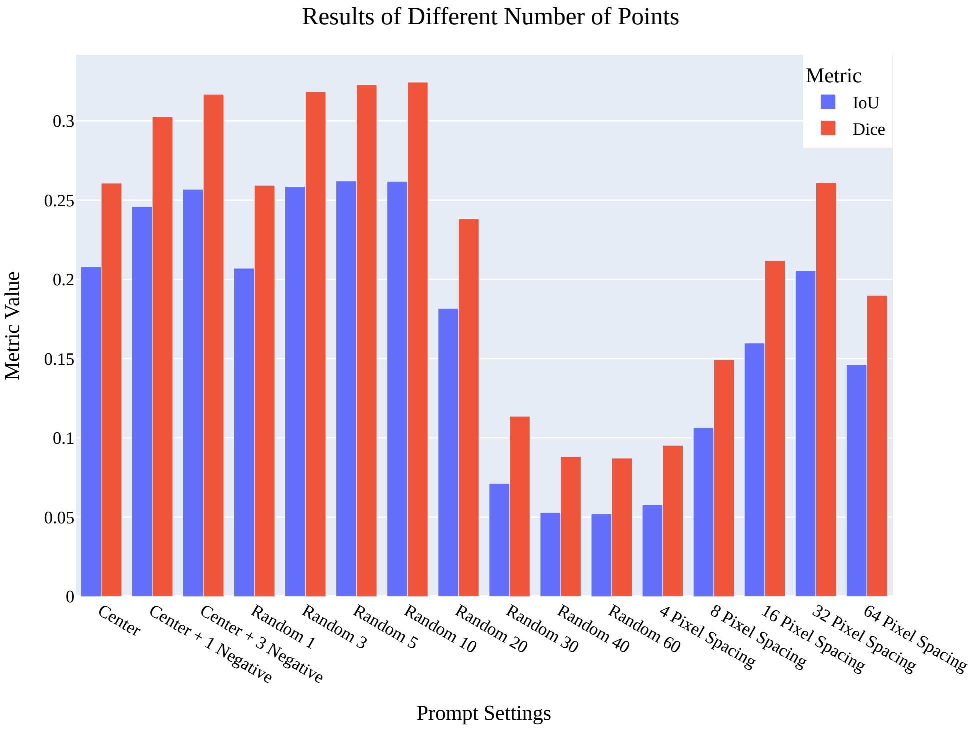


Figure 5: Ablation Study on Increasing the Number of Points on Endoscapes2023 Dataset.

then reintroduced, indicating the need for better long-term memory or context understanding. Instrument blurriness further contributed to identification problems, making it challenging for SAM2 to accurately distinguish between different instruments.

For datasets lacking ground-truth annotations, such as SurgToolLoc and Cholec80, an AutoMask generation approach was used for the first frame as a pseudo label. This introduced a second category of challenges related to the accuracy of the initial segmentation and its propagation to subsequent frames. In cases where the initial frame was clear and contained surgical tools, both point and mask prompts performed well in generating initial segmentations. However, when the initial frame was blurry or lacked surgical tools, a later frame had to be selected for initialization, potentially introducing inconsistencies in the tracking process.

Segmentation boundary issues manifested in various forms in the SurgToolLoc dataset when applying AutoMask-generated pseudo labels. The oversegmentation of tissues resulted in large regions being divided into multiple irregular small areas, indicating issues with spatial continuity (Fig. 9-(a)). Additionally, erroneous merging occurred as shown in Fig. 9-(b), where surgical instruments were sometimes segmented together with each other, blurring the boundaries between distinct elements. Incorrect boundary delineation was another notable issue, with single instruments sometimes incorrectly divided or unidentified after exiting and re-entering the frame, as illustrated in Fig. 9-(c)). The Cholec80 dataset revealed additional challenges in the AutoMask approach, as the SAM2 model struggled to effectively segment different parts of internal organs. This stemmed from its inability to distinguish different

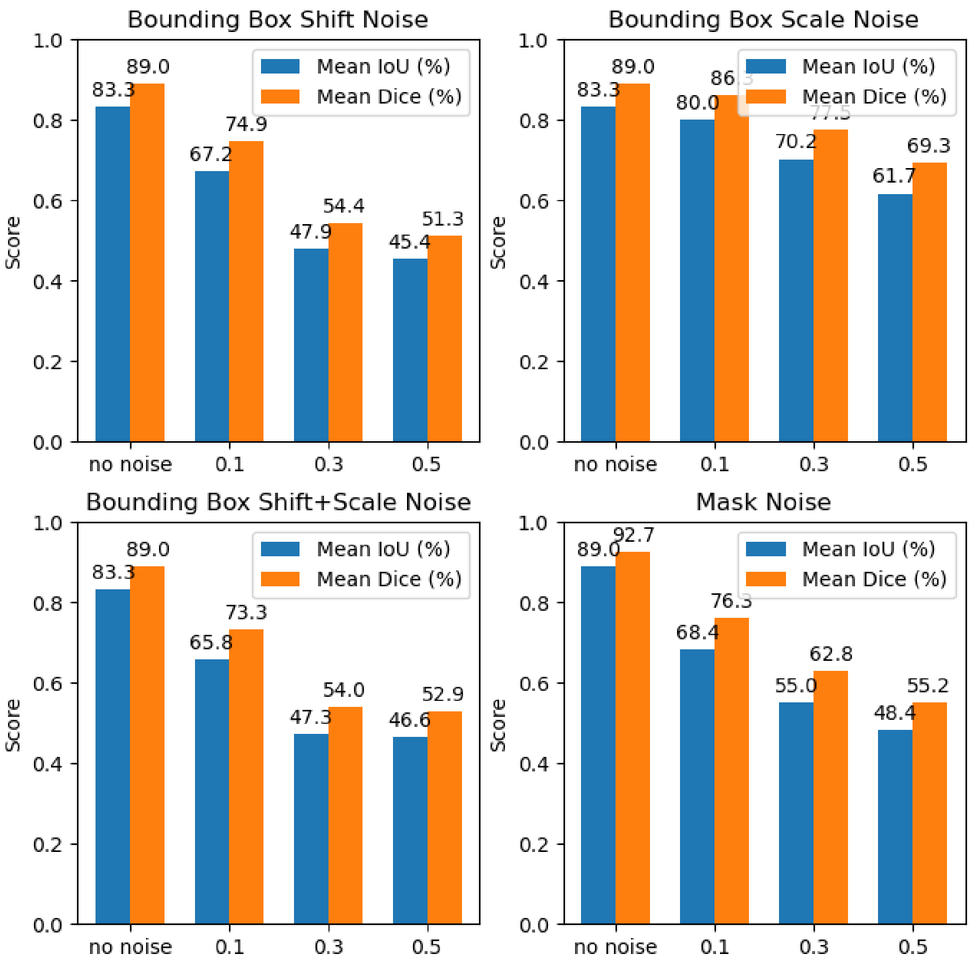
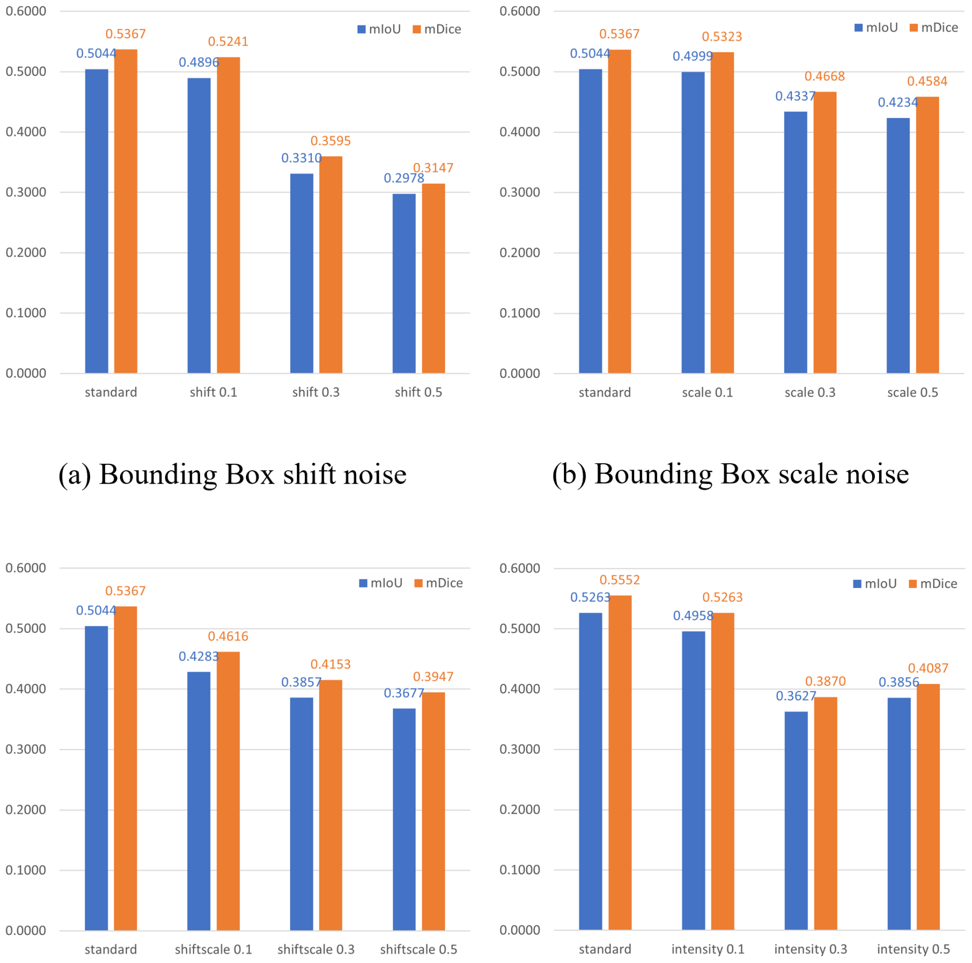


Figure 6: Ablation Study on Noise Perturbation on Cholecseg8k Dataset.

parts of internal organs in the initial frame used for mask generation, leading to inconsistent segmentations in subsequent frames (Fig. 9-(d)).

To address such challenges and improve the performance of SAM2 model in surgical applications, future research should focus on several key areas. For datasets with ground truth, enhancing the model's ability to handle complex prompting scenarios, improving instrument identification for similar tools, and developing better strategies for maintaining temporal consistency are crucial. For datasets requiring AutoMask generation, refining the initial segmentation process, improving the model's ability to propagate segmentations accurately across frames, and developing more robust techniques for handling varying image qualities and content is essential. By addressing these challenges, the utility of the SAM2 model in real-world surgical applications can be significantly advanced, potentially leading to more accurate and reliable tool tracking and segmentation in various surgical scenarios.



(a) Bounding Box shift noise

(b) Bounding Box scale noise

(c) Bounding Box shift+scale noise

(d) Mask noise

Figure 7: Ablation Study on Noise Perturbation on Endovis17 Dataset.

References

[1] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[2] Chaitanya Ryali et al. "Hiera: A hierarchical vision transformer without the bells-and-whistles". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29441–29454.

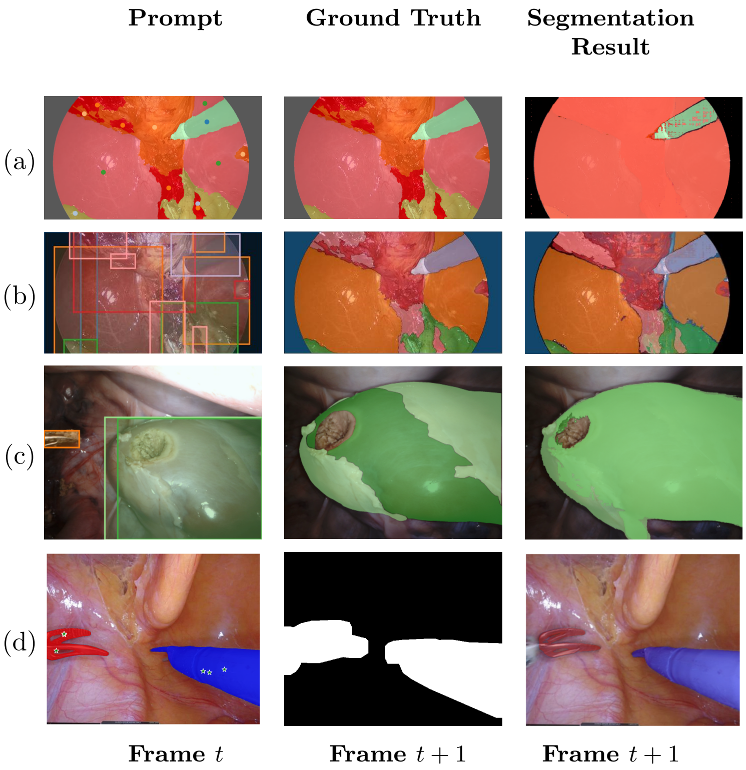


Figure 8: Failure cases analysis in datasets with ground truth annotations. Common failure cases due to (a) Single-point prompting for multi-part objects (Cholecseg8k). (b) Bounding box prompting for complex objects (Cholecseg8k). (c) overlapping bounding box prompts (EndoVis2018). (d) Incomplete object appearance prompt (EndoNerf).

[3] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.

[4] Byeongho Heo et al. "Rotary position embedding for vision transformer". In: *arXiv preprint arXiv:2403.13298* (2024).

[5] Maria Grammatikopoulou et al. "A spatio-temporal network for video semantic segmentation in surgical videos". In: *International Journal of Computer Assisted Radiology and Surgery* 19.2 (2024), pp. 375–382.

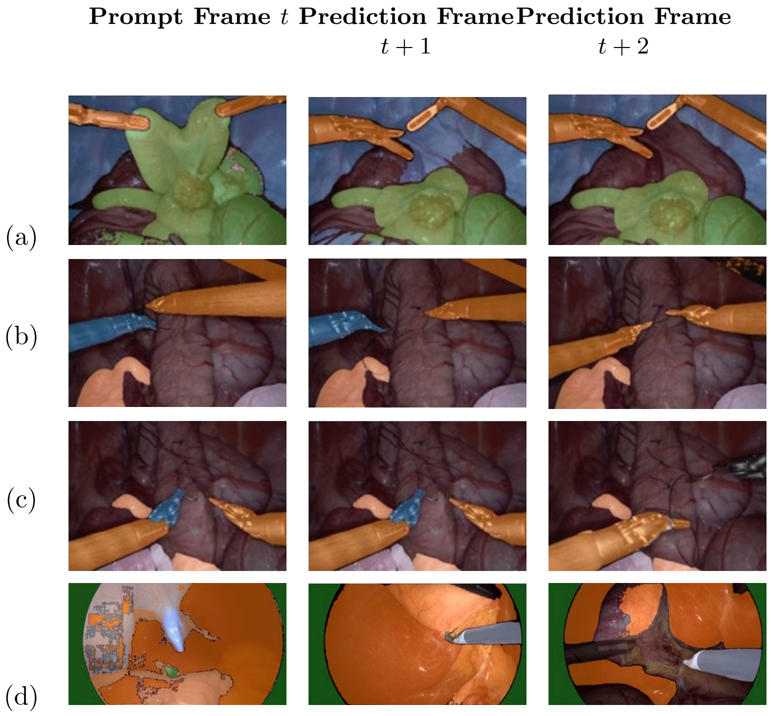


Figure 9: Failure Cases Analysis for Unlabeled Surgical Videos. Common failure case due to (a) Over-segmentation of tissues using Auto-generated pseudo labels. (SurgToolLoc) (b) Erroneous merging of surgical instruments (SurgToolLoc). (c) Incorrect boundary delineation of AutoMask. (SurgToolLoc) (d) Difficulty in distinguishing different tissues (Cholec80)