ColonSAM: 3D Polyp Semantic Segmentation with Deep Reinforcement Learning

Syn Nguyen and Yu Tian

July 2025

Abstract

Colorectal cancer (CRC) is a leading cause of cancer-related mortality, with survival rates exceeding 95% when detected early but dropping below 35% in later stages. Accurate and temporally consistent polyp segmentation in colonoscopy videos is essential for early detection, yet existing models often rely on manual annotation and lack temporal robustness. In this work, we propose a temporal extension of the Plug-and-Play Point Prompt Optimizer (PPO) to enable zero-shot video segmentation using SAM2. Our method leverages a deep reinforcement learning agent to optimize point prompts within a dual-space graph, combining feature similarity and spatial distance. To address the absence of visible polyps in early frames, we introduce a confidence-based frame selection strategy that identifies the best reference frame. Optimized masks from SAM2 are then propagated across frames to ensure consistency. To the best of our knowledge, this is the first automatic prompting approach for SAM2 applied to 3D or video-based medical imaging. Experiments on the SUNSEG dataset demonstrate that our method produces accurate early-frame segmentation and maintains mask quality over short sequences, highlighting its potential for annotation-free medical video analysis.

1 Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related death and the third most commonly diagnosed malignancy worldwide. It is estimated to cause millions of new cases and deaths annually. Critically, the survival rate exceeds 95% when detected at an early stage, but drops below 35% in later stages of the disease [1]. This stark contrast highlights the importance of early and accurate detection, particularly of polyps, which are often precursors to CRC.

Segmenting anatomical structures in colonoscopy videos is thus a vital task in computer-aided diagnosis. High-quality, temporally consistent segmentation of polyps can assist clinicians during real-time procedures and reduce the likelihood of missed detections, 22% 28% missing rate [4]. However, manual annotation of video frames is time-consuming, and most segmentation models struggle with domain shifts, low-light conditions, or motion artifacts common in endoscopy data.

Recent advances in foundation models such as the Segment Anything Model (SAM) have enabled zero-shot segmentation guided by point prompts [5]. However, SAM's performance is highly sensitive to prompt quality. To address this, PPO (Plug-and-Play Point Prompt Optimizer) was introduced by Liu et al. (2025) [3], leveraging deep reinforcement learning to optimize point prompt distribution based on both visual and spatial features. While PPO improves segmentation performance in static images, it does not address temporal coherence, which is crucial for video tasks like colonoscopy.

In this work, we propose a temporal extension of PPO for video-based polyp segmentation on the SUNSEG [2] colonoscopy dataset. Our approach addresses two key challenges: (1) early frames in

colonoscopy videos often lack visible polyps, and (2) there is a need to maintain consistent prompts and masks across frames. Instead of relying on the first frame, we compute prompt confidence scores across all frames during inference and select the most confident frame as the reference. A DRL agent then optimizes prompts, which are passed to SAM2 for segmentation. The resulting mask is propagated to other frames to achieve temporally coherent, zero-shot segmentation.

To the best of our knowledge, this is the first automatic prompting approach for SAM2 applied to 3D or video-based medical imaging tasks. This highlights the potential of prompt-based zero-shot segmentation in clinical video workflows, such as colonoscopy.

2

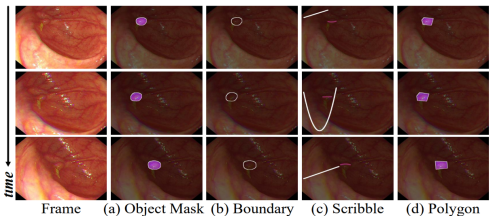


Figure 1: SUNSEG: Colonoscopy Dataset

2 Related Work

2.1 Segment Anything Models and SAM2

The Segment Anything Model (SAM) [Kirillov et al., 2023] introduced a flexible and general-purpose vision model capable of performing zero-shot segmentation using various prompt types (points, boxes, text). SAM2 [Ravi et al., 2024] extended this model to support segmentation across both images and videos, offering improved generalization by leveraging the large-scale SA-1B and SA-V datasets. SAM2 provides high-quality mask generation using point-based prompts without additional fine-tuning, making it suitable for zero-shot video segmentation. However, its segmentation quality remains highly sensitive to prompt distribution, motivating the need for intelligent prompt optimization strategies.

2.2 Prompt Optimization and PPO

To address SAM's prompt sensitivity, Liu et al. (2025) introduced PPO (Plug-and-Play Point Prompt Optimizer), a reinforcement learning framework that automatically refines the spatial distribution of point prompts. PPO formulates prompt optimization as a graph-based decision process, using a Q-network to update prompts based on self-supervised reward functions derived from feature and physical distance matrices. While PPO improves SAM's performance on static images, it does not generalize to temporal data, limiting its use in video segmentation tasks such as colonoscopy.

2.3 SAM Variants and Yolo-SAM

Several works have explored integrating SAM with other vision backbones or adapting it to specific domains. YoloSAM 2 [Mansoori et al., 2024], for example, combines YOLOv8's object detection capabilities with SAM's segmentation head to support real-time, bounding-box-guided instance segmentation. This fusion give SAM the ability to segment semantically. However, the process requires finetuning Yolo on medical dataset, which is resource intensive.

3 Method

We build on the PPO framework proposed by Liu et al. (2025), which utilized deep reinforcement learning to optimize point prompts generated by DINOv2. With the prompts distribution, we then used SAM2 to generate segmentation masks. The approach constructs a dual-space heterogeneous graph which is a combination of two homogeneous graphs representing features, and physical space. A Q-network is trained to iteratively refine the prompt layout using three actions: add, remove, or restore nodes (prompt points).

While FM-PPO performs well on natural image segmentation tasks such as FSS-1000 (achiev-

ing 91.9% mDSC), performance on medical datasets remains suboptimal. Moreover, the original PPO framework is limited to static image segmentation and does not account for temporal coherence across video frames.

3.1 Our Contribution Temporal PPO

We propose a temporal extension of PPO that propagates and optimizes point prompts across frames in a video sequence using SAM2. This enables the model to maintain consistency over time and adaptively refine prompts based on both spatial and temporal information.

Additionally, since the first frame of colonoscopy videos would not have the polyps, the agent will generate the initial prompts for all frames during inference alongside with confidence score. The frame with the highest confidence score will be used as the reference frame and propagate to other frames using SAM2. The model architecture is detailed in figure 2

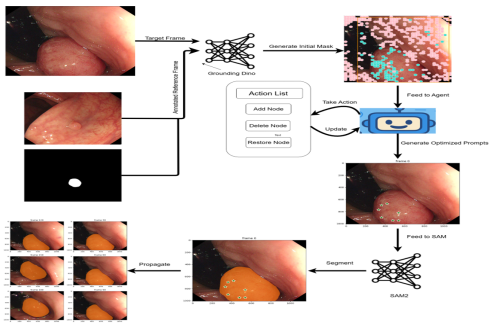


Figure 2: ColonSAM2 Architecture by Adapting Deep Reinforcement Learning to SAM2.

3.2 Training PPO Agent

For each video-based segmentation task, we use the first frame (along with its annotated mask) as the reference image. From this reference, we generate initial point prompts based on feature similarity using DINOv2. These prompts serve as the starting state for the PPO agent.

The agent operates within a dual-space heterogeneous graph environment and executes one of three actions at each step:

- **Add**: Introduce a new prompt point.
- **Remove**: Delete an existing prompt point.
- **Restore**: Re-activate a previously removed prompt point.

The Q-network is updated based on a reward function that encourages:

- High similarity between positive prompt points (homogeneous foreground),
- High dissimilarity between positive and negative prompt points (foreground vs. background),
- Sufficient spatial separation between prompts to avoid redundancy or over-segmentation.

As in Liu et al. (2025), the reward is computed using changes in feature and spatial distances between successive states. The agent iteratively refines the prompt distribution until convergence, after which the optimized prompts are passed to SAM2 to obtain the final segmentation mask.

3.3 Zero-shot SAM2 inference

Our approach operates in a zero-shot fashion without fine-tuning SAM2. Since SAM2 is pretrained on large-scale datasets (e.g., SA-V and SA-1B) [citation], it can generalize well to unseen visual patterns. Using the optimized prompts produced by the PPO agent, SAM2 generates mask for the selected high confident frame. This mask is then propagated across video frames to generate consistent and accurate segmentations throughout the sequence.

4 Results

Figure 3 illustrates the effectiveness of our proposed Temporal PPO framework on a sample sequence from the SUNSEG colonoscopy dataset. In

this example, the agent successfully generated positive point prompts concentrated in the foreground region, accurately identifying the location of the polyp.

Using these optimized prompts, SAM2 produced a segmentation mask that closely matches the ground truth (GT) for the reference frame. Furthermore, the segmentation is successfully propagated across subsequent frames, preserving boundary quality and shape consistency throughout the video.

This demonstrates that:

- The PPO agent effectively distinguishes between foreground and background regions via point prompt refinement.
- SAM2 can perform zero-shot segmentation with high precision when provided with well-optimized prompts.
- The temporal propagation mechanism maintains segmentation quality across frames, even without additional supervision.

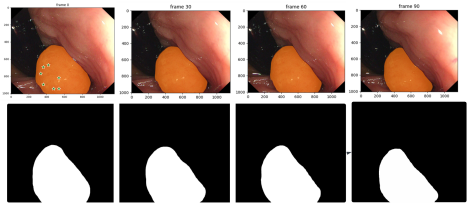


Figure 3: Case 2_8 Test Easy Seen.

4.1 Failure Case

Figure 4 presents a case where the PPO agent successfully optimized positive prompts in the foreground region, resulting in an accurate segmentation of the reference frame by SAM2. The optimized mask was then propagated through the first half of the video with minimal degradation, maintaining strong alignment with the target polyp across frames.

However, in the second half of the video, the segmentation mask gradually degrades and eventually disappears. This failure is likely due to:

- Significant changes in camera angle or lighting
- Motion blur or occlusions
- A lack of adaptive prompt updating during propagation

This result highlights a current limitation of our approach: while temporal PPO performs well in short, stable sequences, its performance may drop in longer or more dynamic videos where additional prompt re-optimization may be necessary.

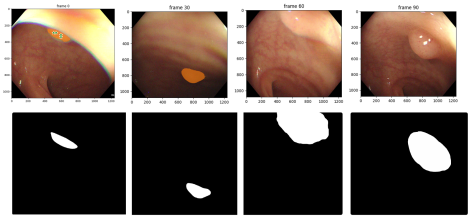


Figure 4: Case 99 Test Easy Unseen.

Third result segmentation was completely off.

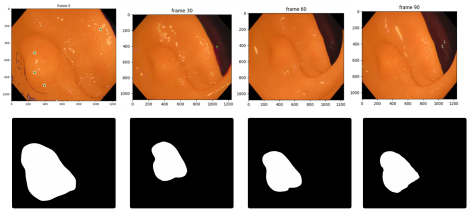


Figure 5: Case 90_1 Test Easy Seen.

5 Conclusion

In this work, we proposed a temporal extension of the PPO (Point Prompt Optimizer) framework to enable zero-shot, temporally coherent polyp segmentation in colonoscopy videos using SAM2. Our method leverages DINOv2 features and a reinforcement learning agent to generate and refine point prompts for the reference frame, then propagates the resulting mask across the video sequence. Through experiments on the SUNSEG dataset, we demonstrated that our approach can effectively guide SAM2 to produce accu-

rate segmentation in early frames and maintain reasonable consistency across short video segments.

While the results are promising, several challenges remain. In longer or more dynamic sequences, we observed that mask quality degrades due to factors such as camera motion, lighting changes, or occlusions — especially in the absence of adaptive prompt updates during propagation. Additionally, the reliance on a single reference frame limits robustness in cases where polyp appearance varies significantly across time.

5.1 Future Work

To address these limitations, future work may explore:

- Online prompt refinement, where prompts are dynamically re-optimized at key frames.
- Multi-frame reference selection or confidence-aware updating to improve stability.
- Incorporation of optical flow or temporal attention to better model motion and guide mask propagation.
- Extending evaluation to real-time or clinical settings with live endoscopy data.

Overall, this study shows that combining prompt optimization with strong vision foundation models like SAM2 opens new opportunities for effective, annotation-free medical video segmentation — particularly in early disease detection scenarios like colorectal cancer screening.

References

[1] J. Bernal, J. Sánchez, and F. Vilarino. "Towards Automatic Polyp Detection with a Polyp Appearance Model". In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182. DOI: 10.1016/j.patcog.2012.03.002. URL: https://doi.org/10.1016/j.patcog.2012.03.002.

[2] Alexander Kirillov et al. "Segment Anything". In: *arXiv preprint arXiv:2304.02643* (2023). URL: https://arxiv.org/abs/2304.02643.

[3] Xueyu Liu et al. "Title of the Paper (replace with actual title)". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025, pp. 4332–4342.

[4] J. Garcia-Borreguero Puyal et al. "Endoscopic Polyp Segmentation Using a Hybrid 2D/3D CNN". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Lima, Peru: Springer, 2020, pp. 295–305. DOI: 10.1007/978-3-030-59725-2_29. URL: https://doi.org/10.1007/978-3-030-59725-2_29.

[5] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: *arXiv preprint arXiv:2408.00714* (2024). URL: https://arxiv.org/abs/2408.00714.