Robust Colonoscopy Video Polyp Segmentation via Quality-Aware Memory and Prototype Alignment

Sijing Xiong[1], Yunfei Yin✉[1], Argho Dey[1], Zheng Yuan[1], Zumnan Timbi[1], Swachha Ray[1], and Hongyu Liu[1]

College of Computer Science, Chongqing University, Chongqing, China
yinyunfei@cqu.edu.cn, {202414021016t,arghodey}@stu.cqu.edu.cn

Abstract. In colonoscopy video polyp segmentation, degradations such as motion blur, specular reflections, and occlusions often render single-frame features incomplete and unreliable. However, most existing methods rely heavily on current-frame features to construct queries for cross-frame retrieval and alignment, leading to performance drops under such degradations. To address this, we propose a query-enhanced cascaded memory framework integrating Quality Awareness (QA) and Prototype Alignment (PA). Specifically, (1) to construct stable query representations for robust retrieval, we propose a quality-aware sensory memory mechanism that adaptively gates memory updates based on observation reliability and state consistency; (2) to mitigate appearance drift, we propose a prototype alignment mechanism that maintains appearance prototypes as long-term anchors. By leveraging stable queries to retrieve relevant prototypes, our framework enables reliable cross-frame association. On the most challenging Unseen-Hard subset of SUN-SEG, our method improves the best baseline by 2.9% in Dice and 4.2% in Sensitivity (Sen), demonstrating superior robustness under degraded conditions.

Keywords: Polyp segmentation · Video segmentation · Quality-aware memory · Prototype alignment · Degradation robustness.

1 Introduction

Accurate polyp segmentation in colonoscopy is essential for preventing colorectal cancer progression. Deep learning has substantially advanced polyp segmentation on static images [9, 26, 28]. Extending to videos requires effective temporal modeling to leverage cross-frame information for enhanced segmentation accuracy. As illustrated in Fig. 1, clinical videos frequently exhibit imaging degradations and appearance variations due to camera motion and intestinal peristalsis, which distort frame-level cues and hinder reliable temporal aggregation.

Existing video methods mainly follow three temporal modeling paradigms. Alignment-and-fusion approaches [10, 25] propagate details from reference frames

✉ Corresponding author: Yunfei Yin (yinyunfei@cqu.edu.cn).

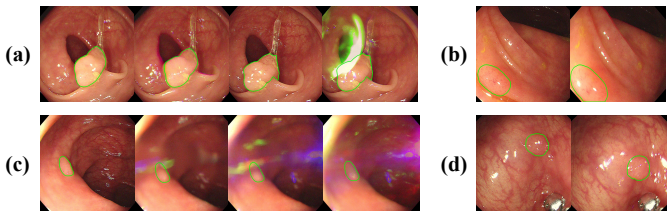

Fig. 1. Common challenges in colonoscopy videos. (a) Motion blur and specular reflections. (b) Small polyp with low contrast against the background. (c) Illumination/color shifts and water-induced occlusions. (d) Large inter-frame polyp displacement.

to compensate for local degradations. Memory read–write mechanisms [4, 12, 16, 18, 21] retrieve relevant historical information from stored frames to enhance current representations. Spatiotemporal attention models [3, 5, 13] aggregate long-range spatiotemporal context via attention. However, most of these approaches rely heavily on current-frame features to construct queries for cross-frame matching or retrieval. When the current observation is degraded, it results in incomplete or unreliable features and consequently unreliable queries, leading to erroneous correspondences and noise propagation through temporal fusion. Moreover, appearance drift across frames further challenges reliable cross-frame association.

To enable reliable query construction under degradations, we propose a query-enhanced cascaded memory framework. We build stable queries from quality-filtered short-term states and use them to retrieve long-term appearance anchors. Specifically, we recursively maintain a compact Sensory Memory [2] storing quality-filtered foreground/background states. Memory updates are gated by observation reliability and state consistency to prevent degraded frames from corrupting the stored representations. We fuse Sensory Memory with current-frame features to construct the query, which then retrieves long-term appearance prototypes to mitigate drift. By cascading quality-aware query construction with prototype-based retrieval, this framework reduces reliance on single-frame observations and enables stable cross-frame association. The main contributions are summarized as follows:

1. We propose a query-enhanced cascaded memory framework with Quality Awareness (QA) and Prototype Alignment (PA) modules to reduce reliance on single-frame observations and enable stable cross-frame association.
2. We introduce a quality-aware sensory memory mechanism that adaptively gates updates based on observation reliability and state consistency, constructing robust queries under degradations.
3. We propose a prototype alignment mechanism that maintains appearance prototypes as long-term anchors, mitigating appearance drift across frames.

Our method achieves leading or comparable performance on SUN-SEG. In particular, on the most challenging Unseen-Hard subset, it improves the best baseline by 2.9% in Dice and 4.2% in Sensitivity (Sen).

2 Method

Given a colonoscopy video $\{\mathbf{I}_t\}_{t=1}^{T}$ with $\mathbf{I}_t \in \mathbb{R}^{3\times H\times W}$, our goal is to predict a polyp probability mask $\hat{\mathbf{M}}_t \in [0,1]^{H\times W}$ for each frame, with robustness to common clinical degradations. We employ a shared encoder ϕ to extract per-frame features $\mathbf{X}_t = \phi(\mathbf{I}_t) \in \mathbb{R}^{C\times H\times W}$, along with reference frame feature $\mathbf{X}_t^r = \phi(\mathbf{I}_t^r)$ for temporal association, where $\mathbf{I}_t^r = \mathbf{I}_{t-1}$.

As illustrated in Fig. 2(a), each frame is processed by a query-enhanced cascaded memory: (1) the QA stage maintains a short-term Sensory Memory via quality-gated updates and constructs a stable query representation from the filtered state; (2) the PA stage employs this query to retrieve long-term appearance anchors from a prototype bank, followed by reference frame alignment and decoding to yield the final mask $\hat{\mathbf{M}}_t$.

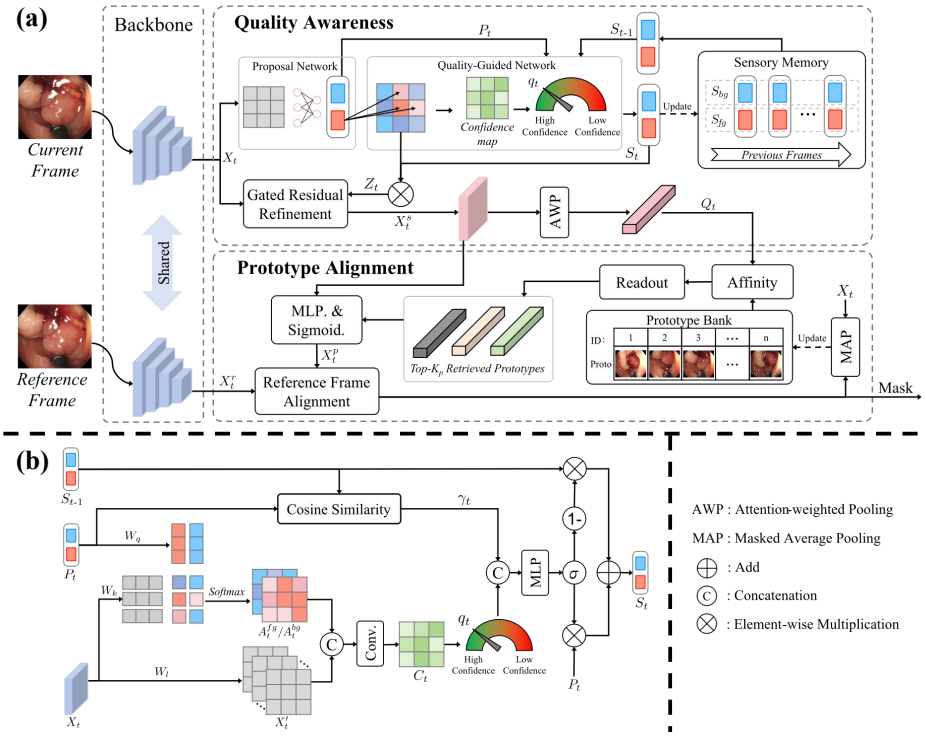

Fig. 2. (a) Overview of the proposed framework with Quality Awareness (QA) and Prototype Alignment (PA) modules. (b) Detail of the Quality-Guided Network (QGN) that predicts gate g_t from reliability q_t and consistency γ_t.

2.1 Quality Awareness

To reduce reliance on single-frame observations, we maintain a recurrent Sensory Memory $\mathbf{S}_t \in \mathbb{R}^{K_s \times D}$ with $K_s = 2$, storing compact foreground (fg) and background (bg) slot representations. We adopt slot-based representations [15, 19], which provide fixed cardinality and preserve slot identity across frames, enabling selective gating of state updates. Specifically, a Proposal Network (PN) produces proposal slots $\mathbf{P}_t$ from $\mathbf{S}_{t-1}$ and $\mathbf{X}_t$, and a Quality-Guided Network (QGN) predicts a write gate g_t based on observation reliability and state consistency. The updated memory state is then integrated with $\mathbf{X}_t$ to construct a stable query $\mathbf{Q}_t$ for subsequent prototype retrieval.

Proposal Network. Given $\mathbf{X}_t$, we generate slot proposals via Slot Attention [19] with recurrent initialization to preserve slot identity. We set $\mathbf{P}_t^{(0)} = \mathbf{S}_{t-1}$ for $t > 1$, and $\mathbf{P}_1^{(0)} = \mathbf{S}_0$ where $\mathbf{S}_0$ is produced by a learnable slot generator [15]. At each iteration $l = 1, \ldots, T_p$, we first aggregate features by $\tilde{\mathbf{P}}_t^{(l)} = \mathrm{SlotAttn}(\mathbf{X}_t, \mathbf{P}_t^{(l-1)})$ and then update slots with a GRU: $\mathbf{P}_t^{(l)} = \mathrm{GRU}(\mathbf{P}_t^{(l-1)}, \tilde{\mathbf{P}}_t^{(l)})$. After T_p iterations, we obtain $\mathbf{P}_t = \mathbf{P}_t^{(T_p)} \in \mathbb{R}^{K_s \times D}$.

Quality-Guided Network. To prevent error accumulation under degraded or abruptly shifting observations, the QGN predicts a write gate g_t from observation reliability q_t and state consistency γ_t (Fig. 2(b)), motivated by [16, 18].

Observation Reliability. Observation reliability quantifies the confidence of the current frame in supporting a fg/bg slot decomposition. Given $\mathbf{X}_t$ and proposal slots $\mathbf{P}_t$, we apply Slot Attention [19] to obtain fg/bg assignment maps $\mathbf{A}_t^{\mathrm{fg}}, \mathbf{A}_t^{\mathrm{bg}} \in \mathbb{R}^{H \times W}$. We concatenate $\mathbf{X}_t$ with both maps and apply a per-pixel MLP to predict a confidence map $\mathbf{C}_t \in [0,1]^{H \times W}$. The frame-level reliability $q_t \in (0,1)$ is obtained by global average pooling over $\mathbf{C}_t$.

State Consistency. Observation reliability q_t captures within-frame separability, but reliable frames may still yield proposals $\mathbf{P}_t$ that deviate from the historical state $\mathbf{S}_{t-1}$ (*e.g.*, due to fast motion or occlusion), which can disrupt state evolution. We therefore measure state consistency as the mean cosine similarity between corresponding slots: $\gamma_t^{(k)} = \mathrm{Cos}\big(\mathbf{s}_{t-1}^{(k)}, \mathbf{p}_t^{(k)}\big) \in [-1, 1]$, and define $\gamma_t = \frac{1}{K_s} \sum_{k=1}^{K_s} \gamma_t^{(k)}$, where $\mathbf{s}_{t-1}^{(k)}$ and $\mathbf{p}_t^{(k)}$ denote the k-th historical and proposed slots, respectively. Together, q_t and γ_t serve as dual signals to gate memory updates, suppressing contributions from degraded or abruptly shifted frames.

Gated State Update and Query Construction. We predict a write gate $g_t \in (0,1)$ from $[q_t, \gamma_t]$ using a lightweight MLP with sigmoid, and update the historical slot state by

$$\mathbf{S}_t = g_t \mathbf{P}_t + (1 - g_t)\mathbf{S}_{t-1} \in \mathbb{R}^{K_s \times D} \tag{1}$$

The gate is driven by slot coherence rather than segmentation correctness. We then project $\mathbf{S}_t$ to a slot-guided feature map $\mathbf{Z}_t \in \mathbb{R}^{C \times H \times W}$ via the fg/bg

assignment maps $\mathbf{A}_t^{\mathrm{fg}}, \mathbf{A}_t^{\mathrm{bg}}$, concatenate it with $\mathbf{X}_t$, and predict a residual $\Delta\mathbf{X}_t$ and a gating map $\mathbf{m}_t \in (0,1)^{C\times H\times W}$ to obtain $\mathbf{X}_t^s = \mathbf{X}_t + \mathbf{m}_t \odot \Delta\mathbf{X}_t$. Finally, we predict a normalized spatial weight map $\mathbf{W}_a \in \mathbb{R}^{H\times W}$ from $\mathbf{X}_t^s$ and form the query by attention-weighted pooling:

$$\mathbf{Q}_t = \sum_{h,w} \mathbf{W}_a(h,w)\, \mathbf{X}_t^s(h,w) \in \mathbb{R}^C \tag{2}$$

2.2 Prototype Alignment

To provide long-term appearance anchors, we maintain a prototype bank $\mathcal{M} = \{\mathbf{G}_k\}_{k=1}^{N_p}$, where each $\mathbf{G}_k \in \mathbb{R}^C$ summarizes the polyp appearance of a single frame via masked average pooling, serving as a region-focused appearance anchor. The quality-aware query $\mathbf{Q}_t$ from the QA stage is then used to compute retrieval weights over $\mathcal{M}$ and aggregate the most relevant prototypes. Compared with full-frame memories used in VOS [4, 21], this compact representation is more efficient for small polyp regions and reduces background-induced texture distractions under appearance drift.

Prototype Retrieval and Aggregation. Given $\mathbf{Q}_t$, we compute bilinear retrieval weights over the prototype bank:

$$\alpha_t^k = \frac{\exp\big(\mathbf{Q}_t^\top \mathbf{W}_p \mathbf{G}_k/\tau\big)}{\sum_{j=1}^{N_p} \exp\big(\mathbf{Q}_t^\top \mathbf{W}_p \mathbf{G}_j/\tau\big)}, \quad k = 1,\dots,N_p \tag{3}$$

where $\mathbf{W}_p$ is a learnable projection matrix and τ is a temperature parameter. We retain the top-K_p prototypes with the largest α_t^k and aggregate them as $\mathbf{G}_t = \sum_{k\in\mathcal{I}_t} \alpha_t^k\, \mathbf{G}_k$, where $\mathcal{I}_t$ indexes the selected prototypes. The aggregated prototype $\mathbf{G}_t$ is then used for channel-wise feature recalibration.

Channel-Modulated Alignment. Given the aggregated prototype $\mathbf{G}_t$, we apply channel attention [11] to recalibrate the QA-enhanced feature $\mathbf{X}_t^s$. We compute channel-wise weights $\mathbf{s}_t = \sigma\big(\mathbf{W}_2\, \delta(\mathbf{W}_1\, \mathbf{G}_t)\big) \in (0,1)^C$, where $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU and sigmoid, and obtain the prototype-aligned feature via channel-wise scaling: $\mathbf{X}_t^p = \mathbf{s}_t \odot \mathbf{X}_t^s$, where $\odot$ denotes channel-wise multiplication. The resulting $\mathbf{X}_t^p$ is then passed to reference frame alignment and decoding.

Adaptive Prototype Update. To track intra-case appearance drift, we update the prototype bank using the predicted mask $\hat{\mathbf{M}}_t$. We extract a candidate prototype by masked average pooling:

$$\tilde{\mathbf{G}}_t = \frac{\sum_{h,w} \hat{\mathbf{M}}_t(h,w)\, \mathbf{X}_t(h,w)}{\sum_{h,w} \hat{\mathbf{M}}_t(h,w) + \varepsilon} \in \mathbb{R}^C \tag{4}$$

where ε avoids division by zero. Every L frames, we apply an LFU-based (Least Frequently Used) [4] replacement, substituting the least-used prototype in $\mathcal{M}$ with $\tilde{\mathbf{G}}_t$.

Table 1. Quantitative comparison on SUN-SEG. Best in bold.

	Method	SUN-SEG-Easy						SUN-SEG-Hard					
		S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice	S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice
Seen	2/3D [22]	89.5	90.9	81.9	85.3	80.8	85.6	84.9	86.8	75.3	80.5	72.6	80.9
	PNS-Net [13]	90.6	91.0	83.6	86.0	82.7	86.1	87.0	89.2	78.7	82.2	77.4	82.3
	PNS+ [14]	91.7	92.4	84.8	87.8	83.7	88.8	88.7	92.9	80.6	84.9	78.0	85.5
	FLA-Net [17]	90.6	92.2	84.6	86.7	85.1	87.5	85.9	89.2	77.8	81.0	78.5	80.9
	SLT-Net [5]	92.7	96.1	89.4	91.4	88.8	90.6	89.4	94.3	84.7	87.4	85.1	86.6
	SALI [12]	94.4	**97.2**	91.0	92.5	92.4	92.7	91.6	95.2	86.5	88.9	89.8	89.1
	Ours	**94.9**	97.0	**91.6**	**93.0**	**94.4**	**93.2**	**92.4**	**95.9**	**87.8**	**90.1**	**91.7**	**90.0**
Unseen	2/3D [22]	78.6	77.7	65.2	70.8	60.3	72.2	78.6	77.5	63.4	68.8	60.7	70.6
	PNS-Net [13]	76.7	74.4	61.6	66.4	57.4	67.6	76.7	75.5	60.9	65.6	57.9	67.5
	PNS+ [14]	80.6	79.8	67.6	73.0	63.0	75.6	79.7	79.3	65.3	70.9	62.3	73.7
	FLA-Net [17]	72.2	69.7	54.7	59.7	50.6	63.6	72.1	70.1	54.0	59.2	52.2	62.8
	SLT-Net [5]	84.8	89.3	77.5	81.7	74.7	79.2	84.4	90.4	75.7	79.5	76.0	78.1
	MAST [3]	84.5	89.8	77.0	81.9	75.5	78.4	86.1	91.4	77.7	81.6	81.1	80.3
	SALI [12]	87.0	92.0	79.4	83.1	81.1	82.5	87.4	92.0	79.0	82.2	83.0	82.2
	Ours	**87.3**	**92.4**	**80.1**	**84.0**	**83.4**	**83.3**	**88.9**	**93.2**	**81.8**	**84.3**	**87.2**	**85.1**

Reference Frame Alignment and Mask Prediction. While the prototype-based modulation in PA provides semantically stable channel-level calibration, it may not preserve fine-grained spatial details. To recover such details, we align the reference feature $\mathbf{X}_t^r$ with the prototype-aligned feature $\mathbf{X}_t^p$ via deformable convolution [6]. Specifically, we predict spatial offsets $\Delta_t = f_{\text{offset}}([\mathbf{X}_t^p, \mathbf{X}_t^r])$ and obtain the aligned reference feature $\hat{\mathbf{X}}_t^r = \text{DCN}(\mathbf{X}_t^r, \Delta_t)$. The aligned feature $\hat{\mathbf{X}}_t^r$ is then fused with $\mathbf{X}_t^p$ via non-local attention [24] and passed to the decoder to produce the final mask $\hat{\mathbf{M}}_t$.

2.3 Loss Function

For a training clip of length T_c, we compute the per-frame objective $\mathcal{L}_{\text{total}}^{(t)} = \mathcal{L}_{\text{seg}}^{(t)} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}^{(t)}$ and average over time: $\mathcal{L} = \frac{1}{T_c}\sum_{t=1}^{T_c}\mathcal{L}_{\text{total}}^{(t)}$. Each $\mathcal{L}_{\text{total}}^{(t)}$ comprises a segmentation loss $\mathcal{L}_{\text{seg}}^{(t)}$ and a confidence regularizer $\mathcal{L}_{\text{conf}}^{(t)}$.

Main Segmentation Loss. We supervise $\hat{\mathbf{M}}_t$ with the F3Net loss [27] as $\mathcal{L}_{\text{seg}}^{(t)} = \mathcal{L}_{\text{BCE}}(\hat{\mathbf{M}}_t, \mathbf{Y}_t) + \mathcal{L}_{\text{IoU}}(\hat{\mathbf{M}}_t, \mathbf{Y}_t)$, where $\mathbf{Y}_t$ is the ground-truth mask.

Confidence Correctness Loss. The QA predicts a confidence map $\mathbf{C}_t$ indicating the reliability of slot assignments. Using the foreground assignment map $\mathbf{A}_t^{\text{fg}}$ and ground truth $\mathbf{Y}_t$, we define a pixel-wise assignment-correctness target:

$$\mathbf{R}_t = \mathbf{Y}_t \odot \mathbf{A}_t^{\text{fg}} + (1 - \mathbf{Y}_t) \odot (1 - \mathbf{A}_t^{\text{fg}}) \in [0,1]^{H \times W} \tag{5}$$

where $\odot$ denotes element-wise multiplication. We regress $\mathbf{C}_t$ to $\mathbf{R}_t$ with an ℓ_1 loss to obtain $\mathcal{L}_{\text{conf}}^{(t)}$, with $\lambda_{\text{conf}} = 0.1$.

3 Experiments

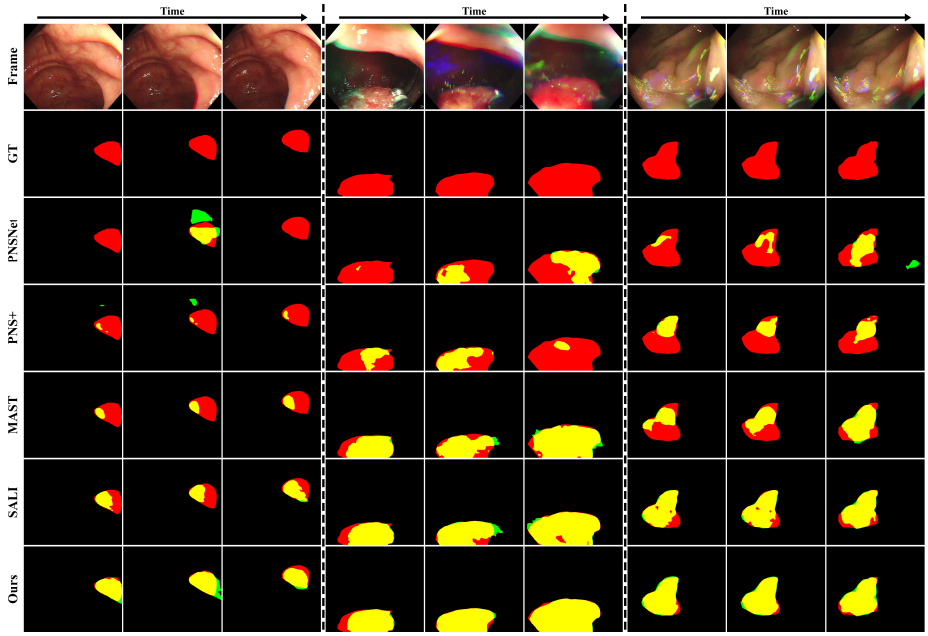

Fig. 3. Qualitative comparisons on challenging sequences. Each group shows the raw frame, ground truth, and predictions from different methods.

3.1 Experimental Setup

Datasets and Metrics. We use SUN-SEG [14] with 112 training videos (19,544 frames) and four test subsets: Seen-Easy (33 videos, 4,719 frames), Seen-Hard (17 videos, 3,882 frames), Unseen-Easy (86 videos, 12,351 frames), and Unseen-Hard (37 videos, 8,640 frames). Seen/Unseen indicates whether the case appears in training, and Hard contains heavier degradations. Following [14], we report case-level results using S_α [7], E_ϕ^{mn} [8], F_β^w [20], F_β^{mn} [1], Sen, and Dice.

Implementation Details. We implement our framework in PyTorch with a pretrained PVTv2 backbone [23], resizing frames to 352×352. Training uses Adam (lr 10^{-4}, cosine annealing) for 50 epochs with batch size 4 on 5-frame clips. For QA, we set $T_p = 3$. For PA, we set $N_p = 20$ to balance accuracy and efficiency, retrieve the top $K_p = 3$ prototypes, and update every $L = 5$ frames.

3.2 Comparison with State-of-the-Art Methods

Quantitative Comparisons. Table 1 reports results on SUN-SEG against representative video polyp segmentation methods, including CNN-based baselines

Table 2. Ablation study of the QA module and the PA module.

QA	PA	SUN-SEG-Seen-Hard						SUN-SEG-Unseen-Hard					
		S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice	S_α	E_ϕ^{mn}	F_β^w	F_β^{mn}	Sen	Dice
		89.3	93.3	83.6	86.6	85.0	86.4	83.5	87.5	74.5	78.9	73.9	78.3
✓		89.6	94.1	86.0	87.3	86.6	87.9	85.8	90.1	77.2	80.9	78.1	80.6
	✓	90.3	94.5	85.7	87.6	87.7	87.8	86.9	91.0	78.6	81.4	80.9	81.2
✓	✓	**92.4**	**95.9**	**87.8**	**90.1**	**91.7**	**90.0**	**88.9**	**93.2**	**81.8**	**84.3**	**87.2**	**85.1**

and temporal propagation models (2/3D, PNS-Net, PNS+, FLA-Net), recent temporal modeling methods (SLT-Net, MAST), and the strong recent method SALI. Overall, our method achieves the best Dice and Sen across all subsets. Notably, the gains are most pronounced on the Unseen splits: we outperform SALI by 2.3%/0.8% (Sen/Dice) on Unseen-Easy, and by 4.2%/2.9% on the challenging Unseen-Hard split, demonstrating superior robustness under severe degradations and stronger generalization.

Qualitative Comparisons. Fig. 3 shows three challenging sequences. In the first group, baselines yield false positives or boundary leakage on a small polyp with large inter-frame displacement. In the second, severe water flushing induces temporal fluctuation and boundary distortion, while ours remains complete and stable. In the third, specular reflections deform contours in baselines, whereas ours preserves boundaries. These observations agree with the gains on Hard subsets. Overlay colors denote FN (red), FP (green), and TP (yellow).

3.3 Ablation Studies

To assess each component's contribution, we conduct ablation studies on the Hard subsets, with results in Table 2. The baseline uses the backbone encoder, reference-frame alignment, and decoder without QA or PA. On Unseen-Hard, adding QA alone raises Dice from 78.3% to 80.6%; adding PA alone improves it to 81.2%; combining both further boosts Dice to 85.1%, with similar trends on Seen-Hard. These results confirm their complementary benefits under degraded conditions and improved generalization.

4 Conclusion

We propose a query-enhanced cascaded memory framework for robust colonoscopy video polyp segmentation, integrating QA and PA. The QA module stabilizes temporal association by suppressing unreliable observations, while the PA module mitigates appearance drift via long-term prototypes. On SUN-SEG, our method improves the strongest baseline by 4.2% Sen and 2.9% Dice on the most challenging Unseen-Hard subset. Future work includes improving sensitivity to small polyps, as well as validating its utility in broader clinical scenarios.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (grant number 62262045), the Fundamental Research Funds for the Central Universities (grant number 2023CDJYGRH-YB11), and the Open Funding of SUGON Industrial Control and Security Center (grant number CUIT-SICSC-2025-03).

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

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1597–1604. IEEE (2009)
2. Atkinson, R., Shiffrin, R.: Human memory: A proposed system and its control processes. In: Psychology of Learning and Motivation. pp. 89–195. Academic Press (1968)
3. Chen, G., Yang, J., Pu, X., Ji, G.P., Xiong, H., Pan, Y., Cui, H., Xia, Y.: Mixture-attention siamese transformer for video polyp segmentation. Artificial Intelligence in Medicine **170**, 103278 (2025)
4. Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European conference on computer vision. pp. 640–658. Springer (2022)
5. Cheng, X., Xiong, H., Fan, D.P., Zhong, Y., Harandi, M., Drummond, T., Ge, Z.: Implicit motion handling for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13864–13873 (2022)
6. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 764–773. IEEE (2017)
7. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557. IEEE (2017)
8. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 698–704. AAAI Press, Stockholm, Sweden (2018)
9. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
10. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video CNNs through representation warping. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4463–4472. IEEE (2017)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7132–7141. IEEE (2018)
12. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: SALI: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024. pp. 531–541. Springer Nature Switzerland, Cham (2024)

13. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 142–152. Springer (2021)
14. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
15. Lee, M., Cho, S., Lee, D., Park, C., Lee, J., Lee, S.: Guided slot attention for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 3807–3816. IEEE (2024)
16. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
17. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 497–507. Springer Nature Switzerland, Cham (2023)
18. Liu, Y., Yu, R., Yin, F., Zhao, X., Zhao, W., Xia, W., Yang, Y.: Learning quality-aware dynamic memory for video object segmentation. In: European Conference on Computer Vision – ECCV 2022. pp. 468–486. Springer, Cham (2022)
19. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Advances in neural information processing systems **33**, 11525–11538 (2020)
20. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2014)
21. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 9225–9234. IEEE (2019)
22. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020. pp. 295–305. Springer International Publishing, Cham (2020)
23. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: PVT v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
24. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7794–7803. IEEE (2018)
25. Wang, X., Chan, K.C., Yu, K., Dong, C., Loy, C.C.: EDVR: Video restoration with enhanced deformable convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1954–1963 (2019)
26. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 699–708. Springer (2021)
27. Wei, J., Wang, S., Huang, Q.: F3Net: Fusion, feedback and focus for salient object detection. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 12321–12328. AAAI Press (2020)

28. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021. pp. 120–130. Springer, Cham (2021)