VigilSAM3: Reinforcement-Optimized Memory Gating for Ultra-Long Surgical Video Segmentation

Xincheng Yao[1], Yijun Yang[1,4,✉], Zhuojin Song[2], Kangwei Guo[2], Haisu Tao[2], Jian Yang[2], and Lei Zhu[1,3,✉]
leizhu@hkust-gz.edu.cn

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] Southern Medical University, Guangzhou, China
[3] The Hong Kong University of Science and Technology, Hong Kong, China
[4] JD.com, China

Abstract. Laparoscopic hepatectomy demands real-time identification of critical vascular structures, including Glisson's pedicle, hepatic veins, and transient bleeding points, to enable pre-emptive vessel clamping and prevent life-threatening haemorrhage. These targets are frequently obscured and deformable, and in ultra-long procedures spanning thousands of frames they further suffer extreme scale imbalance and intermittent visibility, making memory-based propagation prone to temporal drift where a single corrupted entry cascades forward irreversibly. Memory-based propagation dominates long-range video segmentation, yet existing methods treat all predictions equally when updating the memory. Moreover, efficient intraoperative guidance demands minimal surgeon-model interaction, making autonomous quality control indispensable. We present **VigilSAM3**, built on SAM3's dual-system architecture, which introduces a hierarchical memory firewall that guards memory quality at pixel, class, and frame levels. A causal-Transformer scheduling agent selects among Trust, Correct, and Reset actions, with its policy refined via Group Relative Policy Optimization (GRPO) using class-aware rewards. We further introduce **Hepa-Vein**, the first multi-class large-scale hepatectomy surgical video semantic segmentation dataset. Experiments on Hepa-Vein and DSAD demonstrate state-of-the-art performance, with pronounced gains on small, transient vascular structures. The code and dataset will be released at VigilSAM3.

Keywords: Surgical video segmentation · Reinforcement learning · Segment Anything Model · Hepatectomy

1 Introduction

Automated surgical scene understanding from phase recognition [7] to pixel-wise segmentation, is central to surgical data science [21]. In laparoscopic hepatectomy, the demands are particularly acute: target vascular structures are thin,

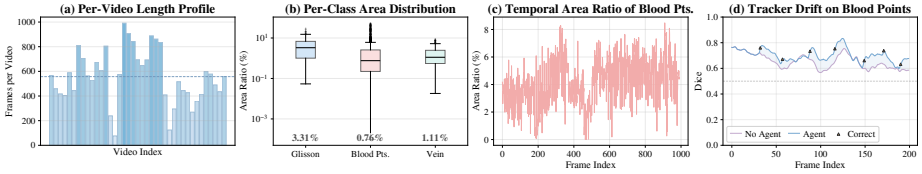

Fig. 1. Hepa-Vein dataset characteristics and motivation. (a) Per-video length profile across the dataset. (b) Foreground area distribution reveals severe scale imbalance. (c) Temporal presence over the longest video shows intermittent appearance of clinically critical structures. (d) A simulated drift illustration: once a low-quality prediction is written into memory, errors can persist without timely correction.

deformable, and frequently obscured, while hepatic venous hemorrhage [14], CO_2 gas embolism [23], and other complications that cause abrupt changes [28] underscore the clinical need for reliable vascular understanding. Existing benchmarks [11,22,3] cover cholecystectomy and colorectal surgery, but not hepatectomy. We introduce **Hepa-Vein**, the first multi-class large-scale hepatectomy surgical video semantic segmentation dataset (Sec. 3.1), annotating three clinically critical foreground classes along the parenchymal transection plane: Glisson's pedicle, hepatic vein, and bleeding point. Real-time segmentation of these structures can guide surgeons to pre-emptively clamp vessels with vascular clips before transection, thereby preventing the majority of major haemorrhagic events during hepatectomy. As Fig. 1 illustrates, Hepa-Vein exhibits three defining challenges: procedures spanning thousands of frames, extreme foreground-scale imbalance (bleeding points $< 0.1\%$ of the frame), and intermittent target visibility. These properties render *temporal drift* the dominant failure mode.

Memory-based video object segmentation has advanced rapidly through space-time correspondences [6], multi-object association [32], hierarchical memory [5], and object-level reading [4]. Promptable foundation models further advance this direction: SAM [13], SAM2 [25], and SAM3 [2] provide strong priors with unified tracking, and surgical adaptations confirm that re-prompting and memory management are critical [17]. Recent SAM2-oriented extensions further study long-horizon memory robustness and domain adaptation: SAM2Long [8] and DAM4SAM [29] focus on error accumulation, distractor interference, and memory selection in long videos, while SurgSAM2 [18] investigates efficient SAM2 adaptation for surgical videos via frame pruning. However, frequent re-prompting demands clinician intervention that is impractical during live surgery, underscoring the need for autonomous correction with minimal human interaction. Nevertheless, all these methods treat the predictions equally when writing to memory; in surgical videos, a single corrupted entry can cascade forward, producing irreversible drift. Yet, no principled mechanism exists to determine *when* predictions are reliable enough to commit and *when* correction should be triggered.

Reinforcement learning (RL) has been explored in surgical contexts, including robotic task automation [10], segmentation refinement [1], and reasoning-driven

surgical understanding [9]. RL offers a natural framework for the above sequential decision problem: in SAM3's dual-system architecture, where a lightweight System 1 tracker (11.7 M params) shares an 815.6 M ViT backbone with a heavier System 2 detector (32.7 M params), each correction incurs significant computational costs and reshapes the shared memory bank that conditions all future predictions, creating compounding, delayed consequences that are difficult to capture with fixed scheduling rules. This memory-coupled scheduling problem is inherently sequential with non-stationary dynamics, making it a compelling target for policy optimization.

We propose **VigilSAM3** ("*Vigil*": continuous vigilance over tracking quality), built on SAM3, in which a causal-Transformer scheduling agent selects Trust, Correct, or Reset at each frame, while a hierarchical memory firewall guards the write path at pixel, class, and frame levels. The policy is initialized by supervised imitation and refined via GRPO [27] with class-aware rewards. Our contributions are: (1) We introduce **Hepa-Vein**, comprising 39 videos and 21,734 annotated frames with three clinically critical foreground classes, to our knowledge, the largest hepatectomy surgical video semantic segmentation dataset. (2) We propose **VigilSAM3**, integrating hierarchical memory gating with RL-optimized correction scheduling to maximally reduce surgeon–model interaction. (3) State-of-the-art results on Hepa-Vein and DSAD validate the synergy between memory gating and reinforcement-driven scheduling, with pronounced gains on small, transient vascular structures.

2 Method

2.1 Overview

VigilSAM3 processes a surgical video in a streaming fashion after initialization at Frame 0 with the System 2 detector (Fig. 2). At each subsequent frame t, the System 1 tracker first produces a draft mask prediction $\mathcal{D}_t$ via memory attention. The scheduling agent π_θ then observes a state composed of three signals: (1) the SAM decoder's quality token $\mathbf{z}_{\text{iou}}$, summarizing predicted mask quality; (2) a pooled memory-conditioned feature $\mathbf{z}_{\text{mem}}$, encoding the current memory state; and (3) per-class quality descriptors $\bar{\mathbf{b}}_t = [\text{mean}; \text{min}; \text{max}]$ aggregated from boundary entropy, mask area, prediction confidence, and IoU statistics. Based on these inputs, the agent outputs an action $a_t \in \{\textsf{Trust}, \textsf{Correct}, \textsf{Reset}\}$ and a continuous frame-level gate $g_t \in (0, 1)$. Under Trust, the tracker prediction passes through the hierarchical memory firewall and is written as regular temporal memory subject to FIFO eviction. Under Correct or Reset, the System 2 detector is invoked to produce an authoritative mask $\mathcal{S}_t$ as a dense prompt to produce the final predict mask $\mathcal{M}_{\sqcup}$, which is promoted to persistent anchored memory never evicted. All memory are further modulated by a pixel-level Memory Writer and per-class quality gates (Sec. 2.2). Training proceeds in two stages: supervised imitation (SFT) followed by GRPO-based policy refinement (Sec. 2.3).

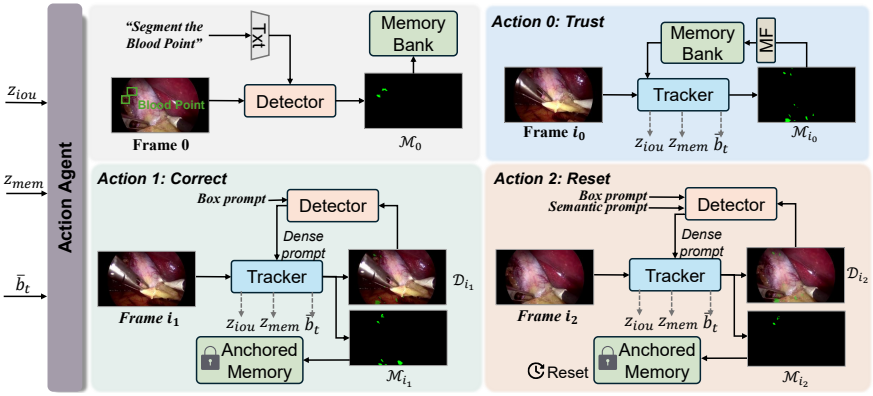

Fig. 2. Inference pipeline of the proposed VigilSAM3 framework. Following initialization at Frame 0, the Action Agent observes the tracking state, including the quality token (z_{iou}), memory features (z_{mem}), and per-class descriptors ($\bar{b}_t$), to dynamically schedule the execution pathway. When Action 0 (Trust) is selected, the lightweight Tracker processes the frame, and its predictions are written to the standard Memory Bank. Conversely, Action 1 (Correct) and Action 2 (Reset) invoke the heavier Detector to provide authoritative spatial corrections, promoting these reliable frames into a persistent Anchored Memory to prevent temporal drift. ***Txt*** represents the SAM3 Text Encoder, and ***MF*** represents the Memory Firewall.

2.2 Hierarchical Memory Quality Gating

On correction frames, we bypass the standard SAM geometric-prompt pathway, which re-predicts rather than adopts the correction, destroying small-target signals via spatial downsampling. Instead, we feed the detector's dense semantic logits $\mathcal{S}_t$ as a mask prompt directly into the tracker decoder before memory encoding, ensuring authoritative rather than merely suggestive corrections.

Memory Firewall. Even with authoritative corrections, unreliable non-corrected frames can still corrupt the memory bank. We introduce a three-level firewall: *(i) Memory Writer:* a lightweight residual CNN refines memory features before writing, suppressing spatially uncertain regions while preserving reliable ones: $\tilde{\mathbf{f}}_t = \mathbf{f}_t + \alpha \cdot \mathcal{W}([\mathbf{f}_t; \sigma(\ell_t)])$, where α is a learnable scalar initialized to zero. *(ii) Per-Class Gate:* an MLP maps an eight-dimensional per-class quality descriptor $\mathbf{g}_{t,c}$, comprising boundary entropy, mask area, prediction confidence, per-class IoU, and object score statistics to a gate $\hat{g}_{t,c} = \sigma(\mathrm{MLP}(\mathbf{g}_{t,c}))$ that independently modulates memory strength per class, allowing, *eg.*, full memory write for a well-tracked organ while suppressing memory for a lost vessel in the same frame. *(iii) Anchor Promotion:* the scheduling agent's action determines whether a frame enters the memory bank as evictable regular memory or as a persistent conditioning anchor, as detailed below.

The **scheduling agent** π_θ is a causal Transformer to observe SAM decoder's quality token, pooled memory-conditioned features, and multi-statistic per-class descriptors ([mean; min; max]), producing action $a_t \in \{\mathsf{Trust}, \mathsf{Correct}, \mathsf{Reset}\}$ and

a continuous frame-level gate g_t. The three levels compose into a single gated memory write for class c:

$$\mathbf{m}_{t,c} = g_t \cdot \hat{g}_{t,c} \cdot \tilde{\mathbf{f}}_{t,c}, \tag{1}$$

where $\tilde{\mathbf{f}}_{t,c}$ is the writer-refined feature (pixel-level), $\hat{g}_{t,c}$ the per-class quality gate (class-level), and g_t the agent-emitted frame-level gate, forming a multiplicative hierarchy that independently modulates memory at spatial, semantic, and temporal granularity. Different actions lead to different memory management regimes: Trust writes the gated tracker prediction as *regular temporal memory*, subject to FIFO eviction, which is suitable for frames where tracking is stable. Correct feeds $\mathcal{S}_t$ as a mask prompt into the tracker and *promotes the frame to a permanent conditioning anchor* that is always attended to and never evicted, providing a persistent high-quality reference. Reset flushes recent temporal memory that may have been contaminated and re-initializes from a fresh System 2 detection as a new anchor. Anchor frames carry a dedicated spatial embedding and fixed temporal position, making them structurally privileged in memory attention, while regular frames compete for a limited FIFO buffer.

2.3 Reinforcement-Driven Long-Horizon Correction Scheduling

The correction decision at frame t does not merely affect t itself: through the memory bank, it determines what every subsequent frame will attend to, creating a *long-horizon credit-assignment problem*. A heuristic threshold (*eg.*, "correct when IoU $< \tau$") cannot capture these temporally compounding effects. A slightly sub-optimal correction in the current frame may prevent catastrophic drift hundreds of frames later, while an unnecessary correction wastes compute and may inject detector noise. We therefore formulate the scheduling problem as a Markov Decision Process and optimize the policy via reinforcement learning.

State and transition. At each frame, the agent observes: $\mathbf{s}_t = [\,\mathbf{z}_{\text{iou}};\ \mathbf{z}_{\text{mem}};\ \bar{\mathbf{b}}_t\,]$, where $\mathbf{z}_{\text{iou}}$ is the SAM decoder's quality token, $\mathbf{z}_{\text{mem}}$ a pooled memory-conditioned visual summary, and $\bar{\mathbf{b}}_t = [\text{mean}; \text{min}; \text{max}]$ aggregated per-class quality descriptors. The agent emits a_t and g_t, after which the memory bank is updated, producing a new state for $t{+}1$. This state–action–memory loop makes the dynamics inherently non-stationary, necessitating trial-and-error learning rather than supervised imitation alone.

Class-aware reward. A naive union-Dice reward is *class-blind*: large foreground classes dominate the score, causing the policy to ignore small but clinically critical targets (*eg.*, blood-point vessels occupying $< 0.1\%$ of the frame). We define softmax inverse-Dice weights

$$w_c = \frac{\exp(-d_{t,c}/\tau)}{\sum_{c'} \exp(-d_{t,c'}/\tau)}, \quad q_t = \textstyle\sum_c w_c\, d_{t,c}, \tag{2}$$

where $d_{t,c}$ is the per-class Dice at frame t. The weights focus credit on underperforming classes. The per-step reward is:

$$r_t = \left(q_t^{\text{final}} - q_t^{\text{draft}}\right) - C(a_t) - \lambda_{\text{rep}} \sum_{k=1}^{\delta} \mathbb{I}[a_{t-k} \in \{\mathsf{C}, \mathsf{R}\}] \cdot \mathbb{I}[a_t \in \{\mathsf{C}, \mathsf{R}\}], \tag{3}$$

where q_t^{draft} and q_t^{final} are the weighted quality scores of the tracker prediction before and after the action, $C(a_t)$ penalizes the computational cost of invoking System 2, and the last term discourages redundant corrections within a δ-frame window. The reward is further augmented by horizon-aware shaping that credits actions whose memory effects propagate positively to future frames.

Group Relative Policy Optimization. For each training clip, K action-sequence trajectories are sampled from π_θ, and GRPO computes group-relative advantages $\hat{A}^{(k)} = (R^{(k)} - \bar{R})/(\sigma_R + \epsilon)$ without a learned critic:

$$\mathcal{L}_{\text{GRPO}} = -\hat{A} \cdot \sum_t \log \pi_\theta(a_t|\mathbf{s}_t) + \lambda_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) - \lambda_H \mathcal{H}(\pi_\theta) + \lambda_{\text{bc}} \mathcal{L}_{\text{CE}}(\pi_\theta, a_t^*). \tag{4}$$

The four terms correspond to the policy-gradient objective, KL anchoring to the SFT reference, an entropy bonus for exploration, and a behavioral-cloning warmstart with oracle actions a_t^*. Eliminating the critic avoids expensive value-function approximation in the high-dimensional, long-horizon surgical setting. Training proceeds in two stages: SFT initializes π_θ with oracle labels, then GRPO refines the policy via environment interaction, enabling strategies such as pre-emptive corrections before drift becomes visible.

3 Experiments

3.1 Dataset

Hepa-Vein Dataset **Hepa-Vein** is a hepatectomy-focused surgical video semantic segmentation dataset from multi-center, comprising 39 videos and 21,734 pixel-wise annotated frames. To our knowledge, it's the largest of dataset for hepatic surgery. Raw footage was collected by two independent surgical teams, each performing laparoscopic hepatectomy procedures lasting over two hours. Clinically informative segments were extracted from the raw recordings, and every retained frame was densely annotated with three foreground classes that are critical along the parenchymal transection plane: *Glisson's pedicle*, *hepatic vein*, and *bleeding point*. To ensure annotation quality, a three-tier clinical review pipeline was adopted: junior annotators produced initial masks, senior surgeons verified and corrected them, and a final expert review resolved remaining ambiguities. We partition the videos into training, validation, and test sets at an approximate 7:1:2 frame ratio, ensuring balanced foreground-class distribution across splits.

DSAD Dataset The Dresden Surgical Anatomy Dataset (DSAD) [3] provides 13,195 laparoscopic images with semantic segmentation from robot-assisted colorectal surgeries, covering eight abdominal organs, the abdominal wall, and two vessel structures. Following prior work, we evaluate on the two vascular classes to assess cross-procedure generalizability of our method beyond hepatectomy.

3.2 Comparison with the State-of-the-Art Methods

Implementation Details and Metrics All experiments run on a single NVIDIA A800 GPU. Every method is trained for 50 epochs (batch size 8, learning rate

Table 1. Quantitative Comparison with Different Methods on Hepa-Vein Datasets. The best values are highlighted in bold. ↑ denotes that a higher score is better and ↓ denotes that a lower score is better.

Method	Venue	Type	Glisson (Dice↑ / Jac↑)	Blood Pt. (Dice↑ / Jac↑)	Vein (Dice↑ / Jac↑)	FPS↑
U-Net [26]	MICCAI'15	image	27.78 / 18.84	26.05 / 19.74	21.50 / 14.51	**152.93**
nnU-Net [12]	Nat.Meth'21	image	27.76 / 18.49	28.32 / 21.97	18.55 / 12.86	60.36
MedSAM [19]	Nat.Comm'24	image	45.32 / 31.90	61.04 / 48.02	77.37 / 65.94	10.73
SAMUS [15]	MICCAI'24	image	42.24 / 31.20	35.86 / 23.92	79.50 / 69.70	29.48
GDKVM [30]	ICCV'25	video	27.72 / 24.99	20.97 / 20.38	31.59 / 26.23	23.33
HRVVS [33]	MICCAI'25	video	56.56 / 45.05	50.62 / 37.07	71.45 / 63.53	40.91
ReSurgSAM2 [17]	MICCAI'25	video	75.49 / 64.66	64.53 / 53.23	78.23 / 67.77	23.53
MedSAM2 [20]	Arxiv'25	video	48.54 / 36.15	14.69 / 9.28	76.24 / 69.34	22.09
MedSAM3 [16]	Arxiv'25	video	78.12 / 68.35	57.79 / 46.29	79.17 / 71.06	26.76
Vivim [31]	TCSVT'25	video	48.81 / 35.23	36.12 / 24.08	77.15 / 65.98	24.23
Ours	– –	video	**80.36** / **70.35**	**66.73** / **54.14**	**82.33** / **75.18**	22.53

Table 2. Quantitative Comparison with Different Methods on DSAD Dataset. The best values are highlighted in bold.

Method	Venue	Type	Mesenteric Artery (Dice↑ / Jac↑ / HD-95↓ / ASSD↓)	Intestinal Veins (Dice↑ / Jac↑ / HD-95↓ / ASSD↓)
nnU-Net [12]	Nat.Meth'21	image	22.65 / 14.91 / 114.78 / 51.28	16.41 / 10.32 / 135.77 / 57.53
IS-Net [24]	ECCV'22	image	51.63 / 40.58 / 77.06 / 34.96	27.09 / 17.69 / 329.85 / 127.75
MedSAM [19]	Nat.Comm'24	image	68.01 / 54.50 / 50.13 / 11.55	59.03 / 43.88 / 55.10 / 13.43
SAMUS [15]	MICCAI'24	image	58.52 / 44.18 / 91.80 / 29.08	58.90 / 44.30 / 30.79 / 8.62
HRVVS [33]	MICCAI'25	video	68.34 / 55.52 / 53.84 / 12.63	42.93 / 30.17 / 108.17 / 36.75
ReSurgSAM2 [17]	MICCAI'25	video	70.89 / 58.83 / 49.99 / 11.06	50.28 / 38.18 / 89.22 / 26.30
MedSAM3 [16]	Arxiv'25	video	73.93 / 61.48 / **45.18** / 12.61	55.12 / 41.26 / 84.75 / 24.75
Vivim [31]	TCSVT'25	video	60.63 / 46.44 / 67.84 / 19.10	46.99 / 32.94 / 110.39 / 31.17
Ours	– –	video	**74.65** / **62.07** / 46.93 / **10.81**	**64.71** / **54.23** / **51.29** / **7.96**

3×10^{-5}); SAM-based methods use their default resolution, others 512×512. We report per-class Dice and Jaccard each class of on both datasets, plus HD-95 and ASSD on DSAD.

Quantitative Comparison Tables 1 and 2 compare VigilSAM3 with representative image- and video-based methods. On Hepa-Vein (Table 1), our method attains the highest Dice and Jaccard across all three foreground classes. The gain is most striking on Blood Point, which is the hardest class due to its extremely small extent and intermittent appearance, where VigilSAM3 reaches a Dice of 66.73, surpassing ReSurgSAM2 by +2.20 and MedSAM3 by +27.41, highlighting the benefit of class-aware GRPO rewards. VigilSAM3 also leads on Glisson (80.36 vs. 78.12) and Vein (82.33 vs. 79.50).

On DSAD (Table 2), VigilSAM3 generalizes to colorectal surgery, obtaining the best Dice, Jaccard, and ASSD on both vascular classes, with particularly pronounced gains on Intestinal Veins (Dice 64.71 vs. 59.03 for MedSAM), confirming cross-procedure transferability of the memory-firewall mechanism.

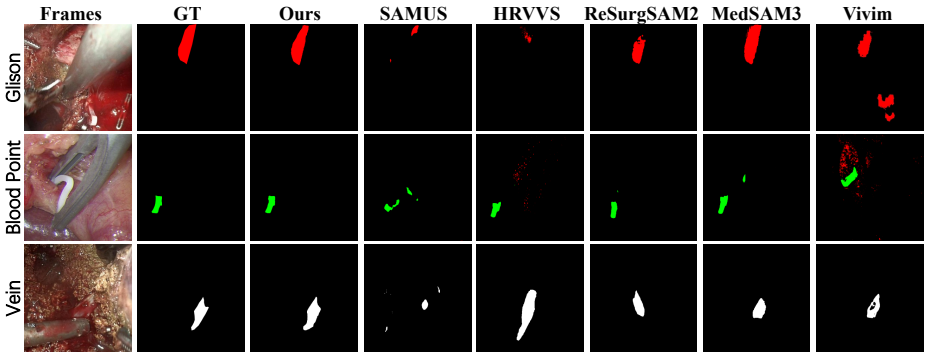

Fig. 3. Qualitative comparison on Hepa-Vein. Each row shows one class results.

Table 3. Ablation study of VigilSAM3 on Hepa-Vein. ***SA*** denotes the scheduling agent. ***MF*** denotes the hierarchical Memory Firewall. ***GRPO*** represents reinforcement alignment of the scheduling policy via Group Relative Policy Optimisation.

Design	SA	MF	GRPO	Dice ↑	Jaccard ↑	HD-95 ↓	ASSD ↓	FPS ↑
M0	–	–	–	66.32	54.79	48.21	15.34	29.41
M1	✓	–	–	67.21	55.61	46.87	15.07	27.12
M2	✓	✓	–	67.95	56.23	46.12	14.88	23.48
M3	✓	–	✓	68.54	56.74	45.36	14.66	22.97
Ours	✓	✓	✓	**69.18**	**57.22**	**44.37**	**14.28**	**22.53**

Qualitative Comparison Figure 3 shows representative predictions on Hepa-Vein. Compared with SAMUS, HRVVS, and other baselines, VigilSAM3 produces more complete masks for Glisson's pedicle, accurately localizes the small and transient Blood Point, and preserves thin distal branches of the hepatic vein, corroborating the quantitative gains from drift-aware memory management.

Ablation Study Table 3 isolates each component's contribution on Hepa-Vein. The scheduling agent (M1) improves Dice by +1.63 over baseline M0, and GRPO alone further lifts it to 68.54 (M3), discovering long-horizon strategies beyond supervised imitation. The Memory Firewall without GRPO (M2) slightly underperforms M1 (−0.74 Dice), as the SFT-only policy triggers too few corrections for Anchor Promotion to take effect; once GRPO calibrates the policy, MF becomes beneficial (Ours vs. M3: +0.64 Dice, −0.98 HD-95), confirming that MF provides the mechanism for structured memory management while GRPO provides the signal for when to deploy it.

4 Conclusion

We presented VigilSAM3, coupling a hierarchical Memory Firewall with GRPO-optimized correction scheduling atop SAM3, and introduced Hepa-Vein, a 39-

video hepatectomy segmentation dataset. State-of-the-art results on both HepaVein and DSAD, together with the observed MF–GRPO synergy, validate drift-aware memory management as a promising direction for reliable long-horizon surgical video understanding. Future work will further explore lightweight GMC, backbone-agnostic validation, and safeguards for rare false anchors.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Project No.82572383), the Guangdong Science and Technology Department (2024ZDZX2004), and the Program of Guangdong Education Department.

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

References

1. Ahmed, F.A., Arsalan, M., Al-Ali, A., Al-Jalham, K., Balakrishnan, S.: Clip-rl: Surgical scene segmentation using contrastive language-vision pretraining & reinforcement learning. In: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS). pp. 863–868. IEEE (2025)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Carstens, M., Rinner, F.M., Bodenstedt, S., Jenke, A.C., Weitz, J., Distler, M., Speidel, S., Kolbinger, F.R.: The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. Scientific Data **10**(1), 3 (2023)
4. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
5. 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)
6. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in neural information processing systems **34**, 11781–11794 (2021)
7. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International conference on medical image computing and computer-assisted intervention. pp. 343–352. Springer (2020)
8. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13614–13624 (2025)
9. Hao, P., Li, S., Wang, H., Kou, Z., Zhang, J., Yang, G., Zhu, L.: Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. arXiv preprint arXiv:2506.19469 (2025)
10. Ho, Y.J., Chiu, Z.Y., Zhi, Y., Yip, M.C.: Surgirl: Toward life-long learning for surgical automation by incremental reinforcement learning. IEEE Robotics and Automation Letters **10**(12), 13145–13152 (2025)

11. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Li, H., Wei, Y.: Hepatic vein injuries during laparoscopic hepatectomy. Surgical Laparoscopy Endoscopy & Percutaneous Techniques **26**(1), e29–e31 (2016)
15. Lin, X., Xiang, Y., Yu, L., Yan, Z.: Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2024)
16. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
17. 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. Springer (2025)
18. 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 preprint arXiv:2408.07931 (2024)
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
20. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
21. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)
22. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 (2023)
23. Otsuka, Y., Katagiri, T., Ishii, J., Maeda, T., Kubota, Y., Tamura, A., Tsuchiya, M., Kaneko, H.: Gas embolism in laparoscopic hepatectomy: what is the optimal pneumoperitoneal pressure for laparoscopic major hepatectomy? Journal of hepato-biliary-pancreatic sciences **20**(2), 137–140 (2013)
24. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly accurate dichotomous image segmentation. In: European Conference on Computer Vision. pp. 38–56. Springer (2022)
25. 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 preprint arXiv:2408.00714 (2024)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

27. Shao, Z., Luo, Y., Lu, C., Ren, Z., Hu, J., Ye, T., Gou, Z., Ma, S., Zhang, X.: Deepseekmath-v2: Towards self-verifiable mathematical reasoning. arXiv preprint arXiv:2511.22570 (2025)
28. Troisi, R.I., Montalti, R., Van Limmen, J.G., Cavaniglia, D., Reyntjens, K., Rogiers, X., De Hemptinne, B.: Risk factors and management of conversions to an open approach in laparoscopic liver resection: analysis of 265 consecutive cases. Hpb **16**(1), 75–82 (2014)
29. Videnovic, J., Lukezic, A., Kristan, M.: A distractor-aware memory for visual object tracking with sam2. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24255–24264 (2025)
30. Wang, R., Sun, Y., Guo, J., Wu, H., Qin, J.: Gdkvm: Echocardiography video segmentation via spatiotemporal key-value memory with gated delta rule. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12191–12200 (2025)
31. Yang, Y., Xing, Z., Yu, L., Fu, H., Huang, C., Zhu, L.: Vivim: a video vision mamba for ultrasound video segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2025)
32. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
33. Yao, X., Yang, Y., Guo, K., Xiao, R., Zhou, H., Tao, H., Yang, J., Zhu, L.: Hrvvs: A high-resolution video vasculature segmentation network via hierarchical autoregressive residual priors. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 266–276. Springer (2025)