SURGSTREAM: A WORKFLOW-ANCHORED PROACTIVE STREAMING FRAMEWORK AND BENCHMARK FOR REAL-TIME SURGICAL VIDEO UNDERSTANDING

Ruipeng Ma[1,2], *Qing Xu*[3], *Jie Liang*[2,✉], *Zhen Chen*[1,✉]

[1]Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University
[2]Eastern Institute of Technology, Ningbo [3]University of Nottingham, UK

ABSTRACT

Surgical video understanding is essential for intraoperative assistance such as workflow monitoring, instrument preparation and procedural documentation. Delivering such assistance in real time requires a model to jointly perceive the ongoing scene, remember past surgical events, and respond before clinically relevant moments pass. Existing approaches provide limited support for jointly retaining procedural evidence and deciding when to respond in streaming surgical video. In this work, we propose SurgStream, a workflow-anchored proactive streaming framework for real-time surgical video understanding. We first introduce Workflow-Anchored Memory (WAM) to record structured instrument–action–target observations and fold them into phase-level summaries, providing compact and queryable surgical memory. Deadline-aware Transition Learning (DTL) is then devised to supervise response states within clinical windows, enabling timely responses before procedural deadlines. We further introduce SurgStream-Bench, comprising 11.7K streaming interactions across various tasks covering real-time perception, backward memory queries, forward proactive QA, proactive alerts and caption memory. Experiments on SurgStream-Bench show gains over state-of-the-art methods, especially in memory and forward proactive QA. The source code will be available.

***Index Terms*—** Surgical video understanding, streaming video LLM, workflow memory, proactive response

1. INTRODUCTION

Real-time surgical video understanding plays a crucial role in intraoperative assistance [1, 2], such as phase-aware instrument preparation, procedural documentation and progress tracking. Such assistance involves understanding the current scene, recalling past events and delivering responses within relevant procedural windows. Figure 1 illustrates these interaction requirements and the temporal modes considered in this work. Surgical videos are long, visually homogeneous and densely structured by the workflow [3]. For time-sensitive assistance, delayed responses may lose their practical value, making real-time surgical interaction challenging.

Early surgical video-language models primarily focused on visual question answering on images or complete videos [4, 5, 6, 7]. Subsequently, general and medical vision-language models [8, 9, 10] achieved significant progress in visual understanding, and real-time surgical commentary [11] narrated the current frames on the fly. More recently, streaming video large language models (LLMs) [12, 13, 14, 15, 16] have enabled causal perception and proactive response on continuous video streams, and OneStreamer [17] further unifies perception, memory and proactive response in a single model.

Despite this progress, existing methods are not explicitly designed around surgical workflows. Free-text memory does not explicitly preserve phase boundaries or structured event counts, making retrospective queries difficult to support reliably. Response supervision at a single annotated time [17] also leaves the interval of acceptable responses implicit. These limitations motivate workflow-structured memory, supervision within valid response windows, and evaluation of both answer content and timing.

To address these limitations, we propose SurgStream, a workflow-anchored proactive streaming framework that integrates compact surgical memory with response supervision within valid procedural windows. For memory construction, we introduce Workflow-Anchored Memory (WAM) to record stable instrument–action–target changes and consolidate each completed phase into a compact *Summary*, preserving evidence for retrospective queries. We further devise Deadline-aware Transition Learning (DTL) to resample response targets within label-derived clinical windows in each epoch, encouraging early responses once evidence is available and before procedural deadlines. Moreover, we construct SurgStream-Bench, spanning diverse tasks across five clinical scenarios and five temporal modes. Extensive experiments demonstrate that SurgStream outperforms competing methods, with particularly strong gains in memory and forward proactive QA.

2. METHODOLOGY

As shown in Fig. 2, SurgStream performs real-time surgical video interaction within a single streaming video LLM. Given a video stream $\{x_t\}_{t=1}^{T}$ arriving at 1 fps, the model reads at each step t the causal context $C_t = [M_t; V_t; D_t]$, consisting of the memory M_t, the recent visual window $V_t = \{x_{t-W+1}, \ldots, x_t\}$ and the dialogue D_t, and emits a control state $s_t \in \{$*Silence*, *Standby*, *Response*, *Observe*, *Summary*$\}$ followed by optional text. On this basis, WAM and DTL anchor what to remember and when to respond to the surgical workflow without new network modules, and SurgStream-Bench evaluates real-time surgical video interaction.

2.1. Workflow-Anchored Memory

To provide compact and queryable surgical memory, we propose WAM, which anchors both the write points and the content of memory to the surgical workflow instead of letting the model decide the boundaries of free-text captions. For training-target construction, let u_t denote the structured surgical state at time t, *i.e.*, the set of instrument–action–target triplets or present instruments. When u_t changes and the new state remains stable for κ seconds, the model writes an *Observe* record $o_\tau = \langle t, \Delta u_t \rangle$ at $\tau = t + \kappa - 1$, where Δu_t denotes the elements newly appearing at time t relative to the preceding state. The record lists these elements as "[mm:ss] instrument | verb | target". Let O_k collect the *Observe* records o_τ whose event times t fall within phase k, $[b_k, b_{k+1})$. After the phase

✉ Corresponding author.

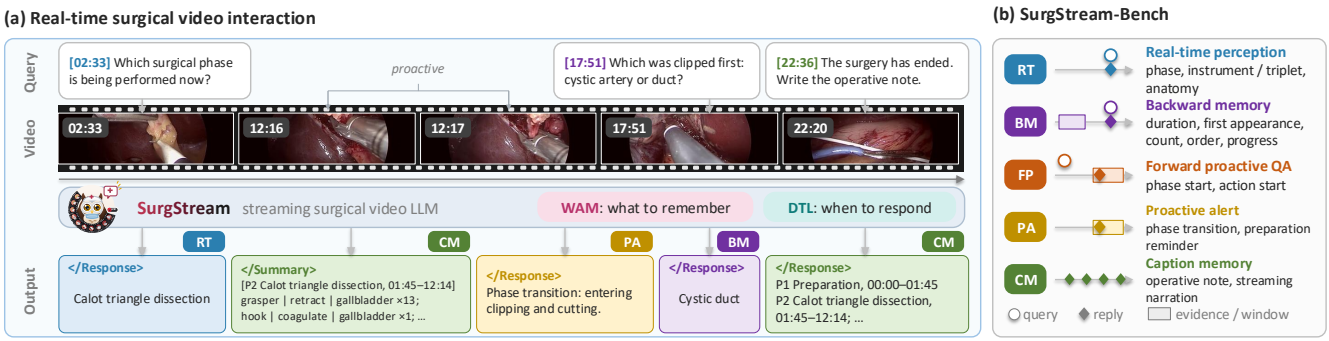


Fig. 1. Illustration of SurgStream and SurgStream-Bench. (a) An example of streaming surgical interaction, including current-scene questions, retrospective queries, proactive alerts and memory writing. (b) The five temporal modes evaluated by SurgStream-Bench.

boundary b_{k+1} is confirmed, the model writes a *Summary* record $\sigma_k = \langle p_k, [b_k, b_{k+1}), \{(c, n_k(c))\}\rangle$, where p_k is the phase label and $n_k(c)$ counts the onsets of element c in O_k (*e.g.*, clip×3). We replace O_k with σ_k and retain the other memory records R_t:

$$\begin{aligned} R_t &= \{\, m \in M_t \mid m \notin O_k \,\}, \\ M_{t+1} &= R_t \cup \{\sigma_k\}. \end{aligned} \tag{1}$$

WAM retains phase summaries for completed phases and detailed observations for the current phase, reducing the accumulation of historical records. Phase intervals and event counts support retrospective queries, while retained *Observe* records preserve precise event times. Expert annotations define these training targets; at inference, the model generates memory records solely from causal video context. The memory loss L_{mem} is token-averaged cross-entropy conditioned on the state and causal context.

2.2. Deadline-aware Transition Learning

To enable timely responses within valid clinical windows, we propose DTL, which extends generic sparse supervision [17] beyond a single annotated response time. Each response target j is assigned a clinical window $[a_j, d_j]$, where a_j is the earliest admissible response time and d_j is the deadline, both derived from annotations and task-specific timing rules (*e.g.*, a preparation reminder is due at the next phase onset). To favour early responses, we resample the target time within the valid window at each epoch using inverse transform sampling. Let $U_j \sim \text{Uniform}(0,1)$ and $\Delta_j = d_j - a_j + 1$. The sampled delay is:

$$\tau_j = -\eta \log\left[1 - U_j\left(1 - e^{-\Delta_j/\eta}\right)\right]. \tag{2}$$

The response time is then:

$$r_j = \min\{d_j,\ a_j + \lfloor \tau_j \rfloor\}. \tag{3}$$

Here, timestamps are discretized at one-second intervals, and $\eta > 0$ controls the decay of the sampling distribution. For $a_j = d_j$, we set $r_j = a_j$. We assign $s^*_{r_j}$ =*Response*, s^*_t =*Standby* for $t \in [a_j - \epsilon, r_j)$, and *Silence* elsewhere. Following [17], the supervised set Ω retains output anchors (including the resampled r_j) and uses quota-based sampling over state-transition groups to select other positions. The state loss is:

$$L_{\text{DTL}} = -\frac{1}{|\Omega|}\sum_{t\in\Omega} \log \pi_\theta(s^*_t \mid C_t). \tag{4}$$

Keeping Ω sparse limits the contribution of abundant *Standby* positions to the state loss. Redrawing r_j in each epoch with earlier times favoured by Eq. (2) encourages the model to respond soon after the earliest admissible time while respecting the clinical deadline.

SurgStream is trained jointly with the cross-entropy of the reply text, L_{mem} and L_{DTL}. At inference, the states are emitted in the order *Summary*→*Observe*→*Response* until *Silence* or *Standby*. Inference uses a fixed visual window and retained memory.

2.3. SurgStream-Bench

We construct SurgStream-Bench by converting expert annotations from Cholec80 [3] and CholecT50 [24] into streaming interactions. Phase, instrument and action-triplet labels are aligned along each procedure. Shared recordings are identified using the official overlap mapping [25], represented once with merged labels, and assigned to procedure-level splits to prevent leakage.

We organize the interactions into five temporal modes: real-time perception (RT), backward memory (BM), forward proactive QA (FP), proactive alert (PA) and caption memory (CM). Across these modes, tasks cover five clinical scenarios: workflow (S1), instrument and action (S2), anatomy (S3), documentation (S4) and narration (S5). Current-scene questions use labels at the query time. Past labels support queries on durations, first appearances, counts and order. Phase and triplet labels yield summaries and operative notes.

Proactive targets pair an expected answer with an earliest admissible response time and a deadline. Phase-transition alerts follow annotated transitions, while preparation reminders precede the next phase. This construction evaluates both whether the answer is correct and whether it arrives within the task-specific response window.

RT and BM use accuracy or task-specific tolerances. FP timing accuracy (T-Acc) checks the first reply after the query for correct content within $[a_j, d_j]$. PA timing F1 (T-F1) uses one-to-one temporal matching with the same requirements. CM uses structured F1 (S-F1) over phase intervals and triplet events. Exact vocabulary matching avoids an LLM judge. Avg_{M} and Avg_{S} average mode and scenario scores, each computed as the mean of its task scores.

3. EXPERIMENTS

3.1. Dataset

SurgStream-Bench uses Cholec80 (80 videos, 7 phases and 7 tools) and CholecT50 (50 videos with phase labels and 100 triplet classes), sampled at 1 fps. After resolving shared recordings, the training, validation and test splits contain 53/11/21 procedures spanning 32.5/6.1/15.5 h, respectively. The benchmark contains 11,667 interaction instances (7,315/1,320/3,032 across the splits), including 1,893 FP/PA instances with annotation-derived response windows for evaluating answer content and timing.

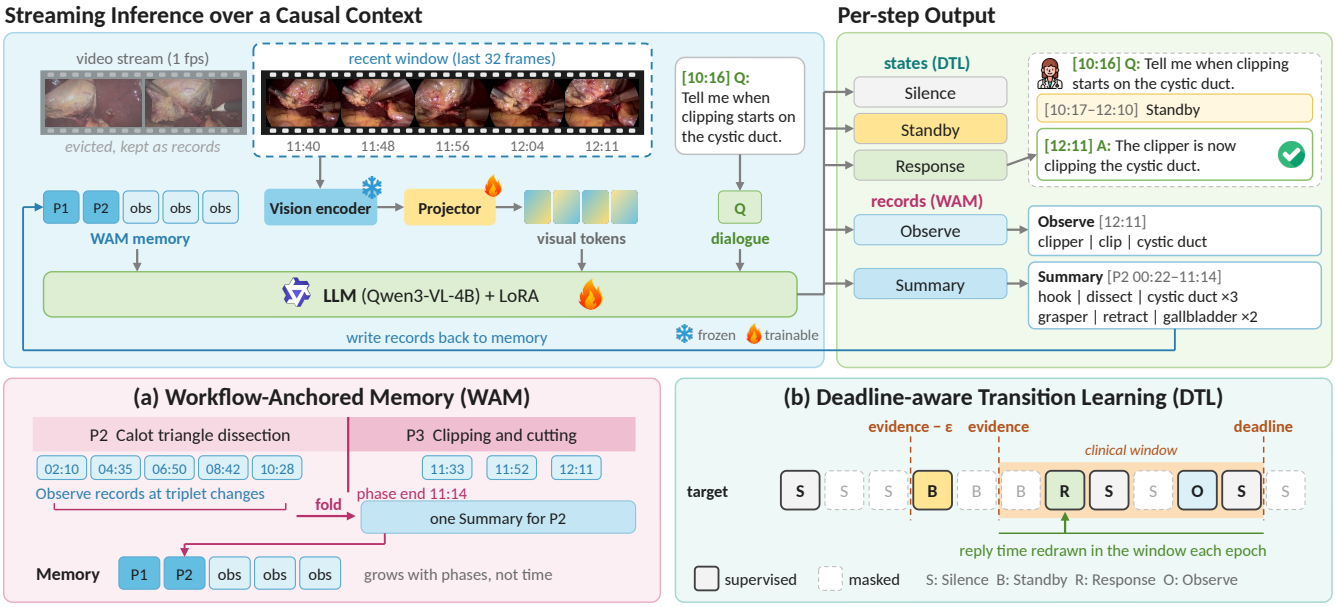


Fig. 2. Overview of SurgStream. At each step, the LLM reads WAM records, the recent visual window and pending dialogue, and emits a control state with optional text. (a) WAM writes *Observe* records at triplet changes and folds them into a *Summary* at each phase boundary. (b) DTL redraws the response time within the clinical window in each epoch and supervises states on a sparse mask.

Table 1. Comparison on SurgStream-Bench (%). †: fine-tuned. Proactive: the model autonomously decides when to respond. Periodic: the model is queried every 5 s. **Bold** marks the better result in each pair.

Methods	Size	Response mode	RT Acc↑	BM Score↑	FP T-Acc↑	PA T-F1↑	CM S-F1↑	Avg_M ↑	S1	S2	S3	S4	S5	Avg_S ↑
Qwen3-VL [8]	4B	Periodic	41.73	29.93	0.43	0.30	4.26	15.33	12.11	30.13	46.95	0.68	8.52	19.68
Qwen3.5 [18]	4B	Periodic	42.70	27.77	2.69	1.08	5.98	16.04	13.06	28.88	43.46	5.38	7.93	19.74
Gemma 4 [19]	E4B	Periodic	42.89	19.88	0.86	1.60	6.62	14.37	14.17	13.06	52.39	4.64	9.08	18.67
MedGemma [20]	4B	Periodic	34.54	11.14	0.43	0.73	3.48	10.06	8.84	11.04	36.66	0.89	5.17	12.52
SurgVidLM [21]	7B	Periodic	36.90	17.93	0.00	0.00	0.62	11.09	12.23	10.22	49.16	0.00	1.24	14.57
Hulu-Med [10]	4B	Periodic	42.77	29.70	0.00	0.78	5.95	15.84	13.18	30.15	42.23	6.01	7.28	19.77
MedVLM [22]	7B	Periodic	55.35	27.80	0.54	1.84	2.43	17.59	13.75	33.36	53.13	0.68	4.85	21.16
EndoChat [23]	13B	Periodic	39.62	27.92	0.00	0.21	0.86	13.72	5.29	29.57	55.37	2.33	0.48	18.61
Qwen3-VL† [8]	4B	Periodic	82.04	52.33	21.84	**24.37**	20.68	40.25	35.04	62.50	71.87	22.95	**26.73**	43.82
SurgStream	4B	Periodic	**86.26**	**58.62**	**27.43**	20.71	**22.83**	**43.17**	**35.44**	**71.19**	**77.93**	**25.15**	25.68	**47.08**
OneStreamer† [17]	4B	Proactive	83.27	59.10	10.87	14.78	15.16	36.64	32.99	62.72	74.81	18.49	24.47	42.70
SurgStream	4B	Proactive	**84.07**	**64.13**	**23.13**	**16.64**	**25.22**	**42.64**	**36.84**	**67.70**	**78.06**	**31.45**	**28.79**	**48.57**

Table 2. Ablation study under proactive inference.

WAM	DTL	RT↑	BM↑	FP↑	PA↑	CM↑	Avg_M ↑	Avg_S ↑
		83.27	59.10	10.87	14.78	15.16	36.64	42.70
✓		82.74	62.46	14.53	17.58	22.84	40.03	46.74
	✓	83.48	61.97	20.23	14.37	22.34	40.48	45.96
✓	✓	84.07	64.13	23.13	16.64	25.22	42.64	48.57

Phase-transition alerts use $[b_k, b_k + 10\,\text{s}]$, and preparation reminders use $[b_k - 60\,\text{s}, b_k]$. Duration, first-appearance and count answers use tolerances of 20%, 10 s and ±1, respectively. Progress questions use S-F1. CM matches phase boundaries and triplet events within 30 s and requires narration within 10 s after each phase.

3.2. Implementation Details

We fine-tune OneStreamer-4B [17], built on Qwen3-VL-4B [8], using LoRA (rank 16) on NVIDIA RTX 6000 Ada GPUs. Training comprises 1 epoch of memory writing and 2 epochs of joint interaction training. We use AdamW with an initial learning rate of 10^{-4}, cosine decay and a global batch size of 32. We set $W = 32$, $\kappa = 3\,\text{s}$, $\epsilon = 5\,\text{s}$ and $\eta = 10\,\text{s}$, and use greedy decoding. Each FP/PA instance and *Summary* write point receives two additional training windows. We drop memory records, jitter timestamps and replace elements.

OneStreamer† matches our training data, window sampling and epochs. Offline VLMs use causal frame sampling, periodic querying and model-specific inputs. All methods share task definitions, temporal matching and scoring rules. We use Qwen3-VL-4B-Instruct [8], Qwen3.5-4B [18], Gemma 4-E4B-it [19], MedGemma 1.5-4B-it [20], Hulu-Med-4B [10], EndoChat-13B [23] and SurgVidLM-7B [21]. MedVLM uses the uAI-NEXUS-MedVLM-1.0a-7B-RL checkpoint from MedGRPO [22].

3.3. Comparison with State-of-the-Art Methods

Table 1 reports zero-shot and fine-tuned comparisons. Zero-shot VLMs score below 3% on FP and PA.

Periodic evaluation. Under 5 s polling, SurgStream achieves Avg_M/Avg_S of 43.17%/47.08%, exceeding Qwen3-VL† by 2.92/3.26 percentage points. It leads on 10 of 12 metrics, with FP improving from 21.84% to 27.43%.

Proactive evaluation. With autonomous response timing,

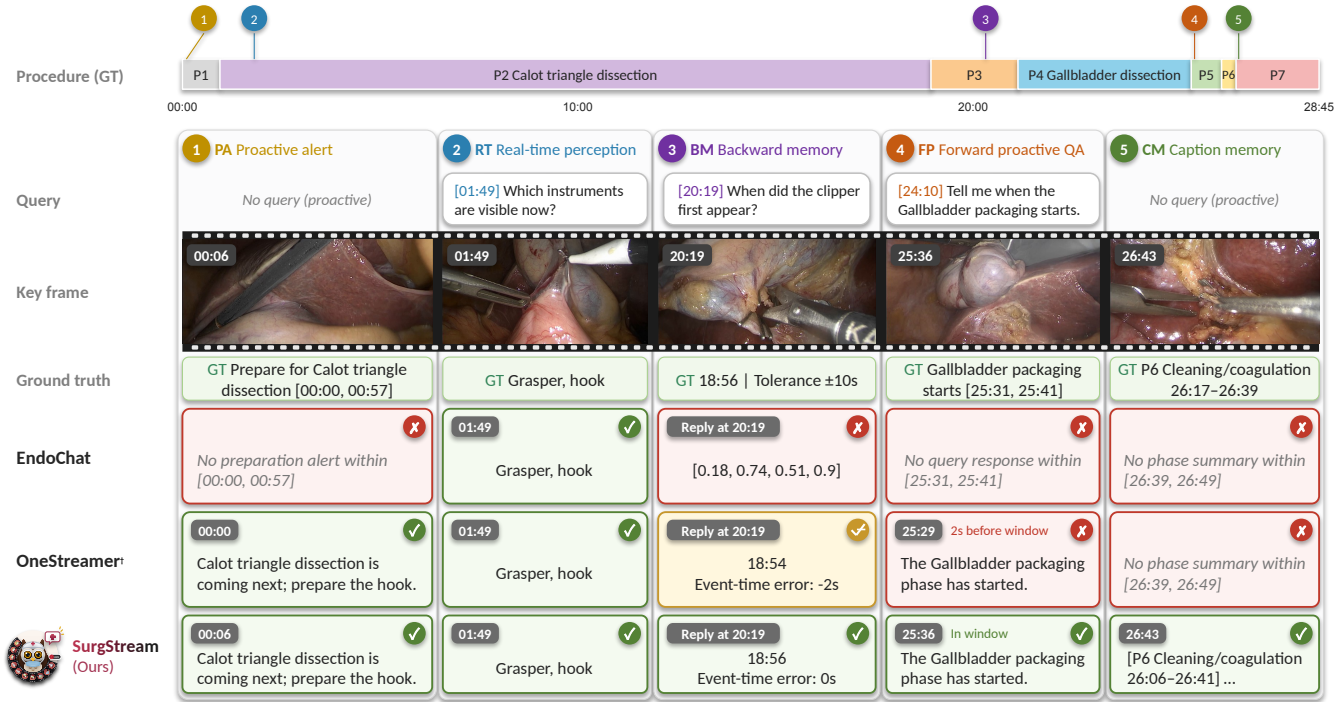


Fig. 3. Qualitative comparison of EndoChat, OneStreamer† and SurgStream on one test procedure across five temporal modes. Responses are excerpted. BM compares event-time estimates (±10 s tolerance), and the CM checkmark denotes a matched phase record.

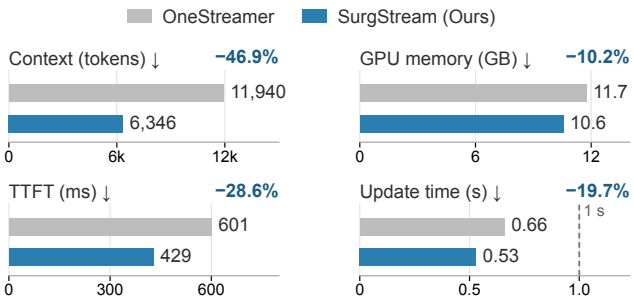


Fig. 4. Efficiency comparison of SurgStream and OneStreamer on two full-length test procedures at 1 fps. Update time is averaged over video steps. GPU memory is reported in GB.

SurgStream achieves Avg_M/Avg_S of 42.64%/48.57%, exceeding OneStreamer† by 6.00/5.87 points and improving all 12 metrics. The largest mode gains are on FP (+12.26) and CM (+10.06); documentation (S4) gains most across scenarios (+12.96). Figure 3 shows accurate recall and timely responses, unlike OneStreamer†'s premature phase-start reply. Overall, SurgStream performs strongly under both periodic polling and proactive response, exceeding the corresponding fine-tuned baseline in both aggregate scores.

3.4. Efficiency Analysis

In a separate profiling experiment, we compare SurgStream with OneStreamer on two full-length test procedures at 1 fps (Fig. 4). SurgStream reduces peak context, median question-call TTFT and peak GPU memory by 46.9%, 28.6% and 10.2%, respectively. WAM summarizes completed-phase observations, reducing historical text while retaining the recent visual window. Mean update time is 0.53 s excluding frame I/O, below the 1 s input interval. These results demonstrate the inference efficiency of SurgStream when processing full-length surgical videos as continuous streams.

3.5. Ablation Study

We ablate WAM and DTL against OneStreamer† (Table 1), which uses free-caption memory and sparse state supervision (PSTL) [17], with the same data and training schedule. As shown in Table 2, introducing WAM improves BM by 3.36 and CM by 7.68 percentage points, DTL improves FP by 9.36 and CM by 7.18 points, and combining both achieves the best Avg_M of 42.64% and Avg_S of 48.57%. WAM brings the larger gain on BM and PA, whereas DTL brings the larger gain on FP. The combination performs best on all modes except PA, where WAM alone is higher (17.58% *vs.* 16.64%). Together, WAM improves retrospective recall and DTL improves timely response performance, yielding the best aggregate performance when combined. Proactive interaction remains challenging, as it requires both correct content and valid response timing.

4. CONCLUSION

We introduced SurgStream for streaming surgical video understanding. Workflow-Anchored Memory (WAM) consolidates timestamped observations into phase summaries, providing compact, queryable surgical memory. Deadline-aware Transition Learning (DTL) resamples target response times within valid clinical windows to guide timely proactive interaction. Experiments on SurgStream-Bench show gains in memory and forward proactive QA with lower context and inference costs. Proactive timing remains challenging, and evaluation is currently limited to laparoscopic cholecystectomy. Future work will improve response timing and evaluate generalization across a broader range of surgical procedures.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by Cholec80 [3] and CholecT50 [24]. Ethical approval was not required as confirmed by the licenses attached with the open access data.

6. ACKNOWLEDGEMENTS

This work is supported by PolyU Start-up Fund (P0060371). We thank The Hong Kong Polytechnic University and Eastern Institute of Technology, Ningbo for providing the computing resources.

7. REFERENCES

[1] Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al., "Surgical data science for next-generation interventions," *Nat. Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.

[2] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al., "Surgical data science – from concepts toward clinical translation," *Med. Image Anal.*, vol. 76, pp. 102306, 2022.

[3] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[4] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K. Krishna, and Hongliang Ren, "Surgical-VQA: Visual question answering in surgical scenes using transformer," in *MICCAI*. 2022, pp. 33–43, Springer.

[5] Kun Yuan, Manasi Kattel, Joël L. Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy, "Advancing surgical VQA with scene graph knowledge," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 7, pp. 1409–1417, 2024.

[6] Runlong He, Mengya Xu, Adrito Das, Danyal Z. Khan, Sophia Bano, Hani J. Marcus, Danail Stoyanov, Matthew J. Clarkson, and Mobarakol Islam, "PitVQA: Image-grounded text embedding LLM for visual question answering in pituitary surgery," in *MICCAI*. 2024, pp. 488–498, Springer.

[7] Jiajie Li, Garrett Skinner, Gene Yang, Brian R. Quaranto, Steven D. Schwaitzberg, Peter C. W. Kim, and Jinjun Xiong, "LLaVA-Surg: Towards multimodal surgical assistant via structured surgical video learning," *arXiv preprint arXiv:2408.07981*, 2024.

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al., "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[9] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al., "InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[10] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al., "Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding," *arXiv preprint arXiv:2510.08668*, 2025.

[11] Jingyi He, Yue Zhou, Long Bai, Kun Yuan, Nassir Navab, and Yuan Bi, "SurgOnAir: Hierarchy-aware real-time surgical video commentary," in *MICCAI*. 2026, pp. 573–582, Springer.

[12] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou, "VideoLLM-online: Online video large language model for streaming video," in *CVPR*, 2024, pp. 18407–18418.

[13] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao, "VideoLLM knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format," in *Findings of EMNLP*, 2025, pp. 6338–6359.

[14] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang, "Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction," in *CVPR*, 2025, pp. 24045–24055.

[15] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou, "LiveCC: Learning video LLM with streaming speech transcription at scale," in *CVPR*, 2025, pp. 29083–29095.

[16] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han, "StreamingVLM: Real-time understanding for infinite video streams," in *ICLR*, 2026, pp. 61463–61475.

[17] Xiangyu Zeng, Yuandong Yang, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Qingyi Si, Dingyu Yao, Changlian Ma, Haoran Chen, Xinyu Chen, et al., "OneStreamer: Unifying perception, memory, and proactive response in streaming video interaction," *arXiv preprint arXiv:2610.01762*, 2026.

[18] Qwen Team, "Qwen3.5: Towards native multimodal agents," `https://qwen.ai/blog?id=qwen3.5`, Feb. 2026.

[19] Gemma Team, "Gemma 4 technical report," *arXiv preprint arXiv:2607.02770*, 2026.

[20] Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al., "MedGemma 1.5 technical report," *arXiv preprint arXiv:2604.05081*, 2026.

[21] Guankun Wang, Junyi Wang, Wenjin Mo, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, Nicolas Padoy, et al., "SurgVidLM: Towards multi-grained video understanding with large language model in robot-assisted surgery," in *ICRA*, 2026, pp. 13177–13184.

[22] Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, and Ziyan Wu, "Med-GRPO: Multi-task reinforcement learning for heterogeneous medical video understanding," in *CVPR*, 2026, pp. 2788–2798.

[23] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al., "EndoChat: Grounded multimodal large language model for endoscopic surgery," *Med. Image Anal.*, vol. 107, pp. 103789, 2026.

[24] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, pp. 102433, 2022.

[25] CAMMA, "CAMMA dataset overlaps: Video overlaps across Cholec80, CholecT50 and Endoscapes," `https://github.com/CAMMA-public/camma_dataset_overlaps`, 2025.