From Perception to Anticipation: Forecasting Vessel–Instrument Interactions in Endoscopic Surgery under Unreliable Observations

Yueyao Chen[1†], Kai-Ni Wang[1†], Hon-Chi Yip[2], and Qi Dou[1*]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
yychen@cse.cuhk.edu.hk, kainiwang@cuhk.edu.hk, qidou@cuhk.edu.hk
[2] Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China
hcyip@surgery.cuhk.edu.hk

[†] Equal contribution. [*] Corresponding author.

Abstract. Vessel safety in endoscopic surgery, particularly in Endoscopic Submucosal Dissection (ESD), is currently supported by segmentation-based visual guidance, which provides spatial awareness but cannot anticipate dangerous vessel–instrument interactions before they occur. To address this limitation, we formulate vessel–instrument interaction anticipation as a new task that predicts the upcoming interaction state from a short observation window of frames and predicted segmentation masks. This task presents a fundamental tension between observation reliability and temporal sufficiency, as both visual and segmentation-derived inputs are inherently noisy, yet the brief observation window offers no margin to absorb such noise, with different interaction categories further exhibiting heterogeneous temporal evidence patterns. To address these coupled challenges, we propose a hierarchical framework that explicitly decouples reliability estimation from temporal reasoning. A cross-modal latent recovery module enforces consistency between visual and geometric observations through an information bottleneck, learning a shared representation that prevents segmentation noise from propagating into temporal reasoning. A dual asynchronous memory module separates transient dynamics from stable interaction prototypes through differentiated admission and retrieval strategies, enabling robust temporal reasoning under unreliable observations. We construct the first multi-case annotated ESD dataset for this task and demonstrate that the proposed framework achieves 70.80% Macro F1 at 97.82 FPS, confirming both strong anticipation performance and real-time clinical applicability.

Keywords: Surgical Safety · Interaction Anticipation · Temporal Modeling

1 Introduction

Avoiding vessel injury is critical for safe endoscopic surgery [1–3], as unintended instrument–vessel contact can trigger bleeding that complicates the procedure

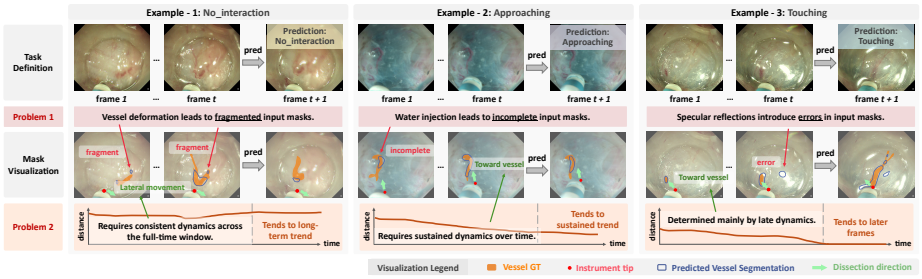

Fig. 1. Task definition and challenges of vessel–instrument interaction anticipation. Problem 1: Unstable vessel segmentation introduces observation noise. Problem 2: Interaction states depend on different temporal evidence distributions.

and compromises patient safety. This risk is particularly acute in Endoscopic Submucosal Dissection (ESD) [4–6], where the instrument operates within a confined submucosal field with dense vascular exposure [7–9]. Current computer-assisted approaches address vessel safety through segmentation, delineating vessel regions in individual frames [10, 11]. However, segmentation describes the current scene but cannot capture how the instrument–vessel relationship is evolving [12]. An instrument rapidly approaching a vessel produces the same segmentation output as one held stationary at equal distance [13], offering no advance warning of impending contact [14]. This motivates a shift from static vessel perception to **proactive interaction anticipation** to alert the operator before risk materializes.

Realizing this goal requires predicting the upcoming interaction state from a short observation window of frames and their segmentation masks. Since manual annotation is unavailable during actual procedures, the masks must be generated by a segmentation model and inevitably contain boundary jitter, partial disappearance, and false detections [15], corrupting both visual features and the geometric measurements derived from them (Fig. 1). In longer sequences, such noise could be absorbed through temporal redundancy, but a brief observation window offers limited margin for error, creating a tension between observation reliability and temporal sufficiency. Moreover, not all interaction states depend on the same part of the observation [16] (Fig. 1). *Touching* depends heavily on the final frames where the instrument and vessel converge, while *approaching* requires a sustained spatial trend across the full sequence. A single corrupted frame therefore carries category-dependent consequences for prediction.

These challenges demand models capable of selective and reliable temporal reasoning. Existing temporal models in surgical video analysis have achieved strong results in phase recognition and action anticipation [17–19], but rely on an implicit assumption that observations faithfully represent the underlying surgical scene. When this assumption breaks down, the model must first evaluate whether each frame constitutes genuine interaction evidence or segmentation noise before extracting temporal patterns. Such architectures conflate these two functions

into a single learned process, so that noise is encoded indistinguishably from signal and propagated through temporal aggregation. Addressing this problem requires explicitly decoupling reliability estimation from temporal modeling, and adapting temporal reasoning to the evidence structure specific to each interaction category.

Following this principle, we propose a framework that explicitly separates reliability estimation from temporal modeling. Specifically, a cross-modal latent recovery module learns a shared interaction representation from which both visual features and geometric descriptors can be reconstructed. By enforcing consistency between two independently noisy observation channels, this module recovers interaction structure that neither modality can provide alone, preventing observation noise from propagating into temporal reasoning. Building on these stabilized representations, a dual asynchronous memory module maintains a transient memory that captures recent dynamics sensitive to late-stage decisive interaction cues, alongside a prototypal memory that selectively accumulates only reliable observations to model stable interaction patterns over time. Together, differentiated retrieval strategies and cross-modal alignment supervision enable robust temporal reasoning under noisy observations.

Our contributions are as follows. (1) We introduce vessel–instrument interaction anticipation as a new task for surgical safety in ESD, and construct the first multi-case annotated dataset for this problem. (2) We propose a cross-modal latent recovery mechanism that enforces structural consistency between visual and geometric observations, learning a shared representation resilient to observation noise in either modality. (3) We design a dual asynchronous memory module with reliability-aware training that decouples transient dynamics from stable interaction prototypes, enabling effective anticipation under heterogeneous temporal evidence. (4) Experiments on our ESD dataset validate the effectiveness of the proposed framework, achieving 70.80% Macro F1 and 97.82 FPS, demonstrating strong anticipation performance and real-time capability. The code is available at: https://github.com/med-air/EndoInteractionAnticipation.

2 Method

2.1 Overview

This work presents a hierarchical framework that transforms unreliable multimodal observations into stable interaction predictions through representation recovery and reliability-aware temporal reasoning. The goal is to anticipate the vessel–instrument interaction state one second into the future from a short observation window of an ESD video clip, where 16 frames are uniformly sampled from the first three seconds, each paired with a predicted segmentation mask. The model learns from these observations to predict the interaction category at the target timestep, where no observation is available.

Let $I = \{(x_i, s_i)\}_{i=1}^{T}$ denote the observation window of T frames, where x_i is the i-th RGB frame and s_i is the corresponding segmentation mask. The

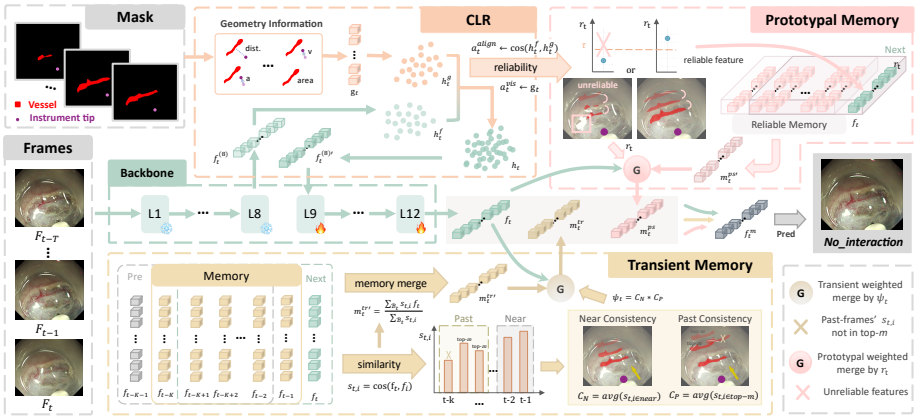

Fig. 2. Overview of the proposed framework for vessel–instrument interaction anticipation under unreliable observations. Cross-modal latent recovery **(CLR)** stabilizes representations, while dual memories combine **transient** consistency and reliability-filtered **prototypal** evidence for future interaction prediction.

model produces a representation f^m used for predicting the interaction category $y_{T+1} \in \{no\_interaction,\ approaching,\ touching\}$ at the next timestep:

$$f^m = F_{\text{memory}}\big(F_{\text{latent}}(I)\big), \tag{1}$$

where $F_{\text{latent}}(\cdot)$ maps visual and geometric observations into a shared latent space via cross-modal consistency (Sec. 2.2), and $F_{\text{memory}}(\cdot)$ integrates temporal evidence through dual asynchronous memory (Sec. 2.3).

2.2 Cross-Modal Latent Recovery

To address inconsistencies between visual and geometric observations under unreliable segmentation, we introduce a cross-modal latent recovery module that aligns the two modalities within the intermediate layers of the ViT-B [20] encoder. At timestep t, a visual representation $f_t^{(8)}$ is extracted from intermediate transformer features, and a geometric descriptor g_t is computed from the predicted segmentation mask s_t. The geometric descriptor encodes the instrument–vessel spatial relationship through measurements of spatial proximity, relative motion, and structural visibility. The recovered latent state is projected back to the token space to refine the representation for subsequent transformer blocks.

Rather than directly fusing these observations, we introduce a latent interaction state $h_t \in \mathbb{R}^m$ with that serves as an information bottleneck to retain modality-consistent interaction structure while suppressing modality-specific noise. The visual and geometric features are first projected into modality-specific latent representations and then fused to produce the shared latent state h_t. The latent

state is required to reconstruct both observations through learnable mappings $\tilde{f}(\cdot)$ and $\hat{g}(\cdot)$, enforced by

$$\mathcal{L}_{\mathrm{CLR}} = \left\| f_t^{(8)} - \tilde{f}(h_t) \right\|_2^2 + \lambda \left\| g_t - \hat{g}(h_t) \right\|_2^2 . \tag{2}$$

Because h_t must explain both visual and geometric observations, only modality-consistent interaction structure is preserved. The recovered latent interaction state h_t is projected back to the token space to refine the representation before temporal reasoning.

2.3 Dual Asynchronous Memory

This module performs temporal reasoning using two memories with complementary roles: a Transient buffer for short-term consistency cues and a Prototypal buffer for reliability-filtered long-term evidence. Their different admission and readout rules enable joint modeling of decisive interaction dynamics and stable long-term interaction structure over time.

Transient Memory. A FIFO buffer $\mathcal{B}_t$ of size K stores the most recent backbone features without quality filtering. At each timestep, cosine similarities between f_t and the buffered entries are computed and partitioned into near-frame and past-frame groups to assess local continuity and short-term pattern consistency respectively. These two consistency measures are integrated to produce a confidence $\psi_t \in (0, 1)$, which modulates the softmax-weighted aggregation of buffered features to yield the transient readout m_t^{tr}.

Reliability Estimation. At each timestep t, two complementary reliability signals are available: an appearance reliability a_t^{vis} derived from geometric temporal statistics, and an alignment reliability a_t^{align} measuring visual–geometric consistency in the latent space. These are fused through a learned gate $\gamma_t = \sigma(w_\gamma^\top h_t + b_\gamma)$ into a unified reliability score:

$$r_t = \gamma_t \, a_t^{\mathrm{vis}} + (1 - \gamma_t) \, a_t^{\mathrm{align}}, \tag{3}$$

which governs prototypal memory operations.

Prototypal Memory. A separate FIFO buffer $\mathcal{P}_t$ of capacity K' admits only frames whose reliability exceeds a threshold ($r_t > \tau$), storing both the feature f_t and its reliability score r_t. Unreliable frames leave the buffer unchanged, preventing noise from contaminating long-term evidence. The prototypal representation is obtained by combining the current feature with a reliability-weighted aggregation of stored features, where r_t controls their relative contributions:

$$m_t^{\mathrm{ps}} = r_t f_t + (1 - r_t) \frac{\sum_{i \in \mathcal{P}_t} r_i \, f_i}{\sum_{i \in \mathcal{P}_t} r_i + \epsilon}, \tag{4}$$

The two memory features are first adaptively balanced and then integrated with the current observation through a residual compensation:

$$f_t^m = f_t + z\big[\lambda_t \, m_t^{\mathrm{tr}} + (1 - \lambda_t) \, m_t^{\mathrm{ps}}\big], \tag{5}$$

where $\lambda_t = \sigma(W_\lambda f_t)$ adaptively balances short-term dynamics and long-term prototypes, and z is a learnable scaling factor.

Table 1. Overall and class-wise performance comparison.

Group	Model	F1	Pre	Rec	Acc	$\mathbf{F1}_{\mathrm{No}}$	$\mathbf{F1}_{\mathrm{App}}$	$\mathbf{F1}_{\mathrm{Tou}}$
Visual	ResNet+LSTM	41.11	51.41	44.40	48.89	63.41	43.24	16.67
	ResNet+TCN	45.37	57.78	49.37	46.67	48.48	47.62	40.00
	ViT-B	43.30	62.65	42.42	51.11	68.00	28.57	33.33
Geometry	Geo+LSTM	53.02	59.14	58.48	51.11	43.75	48.65	66.67
Concat	ResNet+Geo	55.53	55.15	57.95	55.56	60.00	37.04	69.57
	ViT-B+Geo	55.54	55.60	56.96	57.78	68.29	46.15	52.17
Ours	**Full Model**	**70.80**	**71.34**	**71.52**	**71.11**	**74.42**	**64.29**	**73.68**

2.4 Consistency-Constrained Optimization

The model is optimized using a combination of classification loss and cross-modal consistency loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_c \, \mathcal{L}_{\mathrm{CLR}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss, $\mathcal{L}_{\mathrm{CLR}}$ enforces cross-modal consistency, and λ_c is a weighting coefficient. The consistency objective improves robustness to unreliable observations.

3 Experiments and Results

Dataset. We construct the first vessel–instrument interaction anticipation dataset from 30 real ESD surgical cases, yielding 3330 four-second clips with clip-level interaction labels spanning *no_interaction*, *approaching*, and *touching*. The first three seconds serve as the observation window and the interaction state at the unobserved fourth second as the ground-truth prediction target, forming a 1 s anticipation setting. Videos are captured at 50 fps with 16 frames uniformly sampled per clip, each paired with a SAM2-generated vessel segmentation mask. The dataset is split at the case level into 22 training and 8 testing cases, yielding 2430 training clips and 900 testing clips. Each clip is independently annotated by two trained annotators, with disagreements resolved by consensus.

Implementation. ViT-B pretrained via self-supervised learning on 100 unlabeled ESD videos serves as the visual backbone. For each frame, a 10-D geometric descriptor $g_t \in \mathbb{R}^{10}$ is computed from the predicted vessel mask, which achieves 84.64 Dice and 73.37 IoU on an independently annotated test set. The active instrument tip center is manually annotated to isolate vessel-mask uncertainty from tip-localization errors. The descriptor encodes normalized tip-to-vessel distance with its temporal differences and moving average, normalized tip coordinates, vessel area ratio, visibility indicators, and tip-inside-vessel status. The model is optimized with AdamW ($\mathrm{lr} = 2.5 \times 10^{-4}$). The first eight transformer layers are frozen, the remaining four are fine-tuned after a 10-epoch warm-up. All experiments use a single NVIDIA A100 GPU with batch size 16.

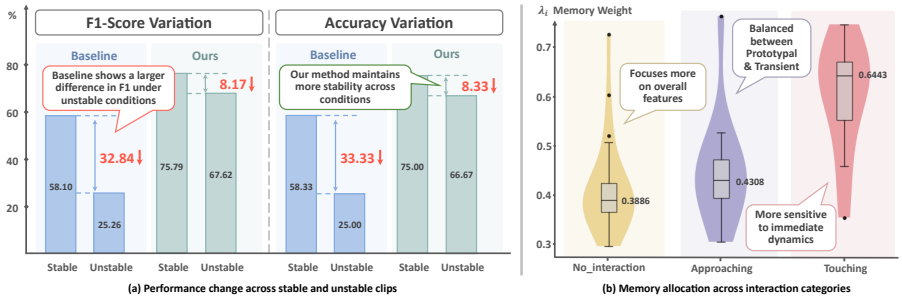

Fig. 3. Robustness and memory allocation analysis. (a) Comparison of performance change between stable and unstable clips, showing smaller degradation of the proposed method. (b) Comparison of memory allocation across interaction categories, showing a shift from prototypal to transient dominance.

Performance is assessed using Macro F1, Precision, Recall, Accuracy, and per-class F1.

Comparative Experiments. Since no prior benchmark exists for this task, we construct a systematic baseline suite spanning visual-only, geometry-only, and concatenation paradigms. As shown in Table 1, the proposed method achieves 70.80 Macro F1, outperforming the strongest baseline by over 15 points. Visual-only models [21, 22, 20] plateau below 46 F1, reflecting susceptibility to segmentation perturbations and limited spatial reasoning. Geometry-only modeling [23] captures spatial discriminability but yields uneven per-class performance under noisy segmentation. Naive concatenation provides moderate gains yet struggles with the ambiguous *approaching* state. These results confirm that reliable anticipation under observation uncertainty requires explicit cross-modal consistency modeling rather than simply combining stronger feature representations.

Ablation Study. Three components are progressively added to the ViT-B+Geo baseline (Table 2). CLR yields the largest single improvement, primarily by stabilizing *touching* recognition, confirming that cross-modal consistency is the dominant factor under observation noise. Adding Dual Asynchronous Memory brings further gains on both *touching* and *approaching*, reflecting that interaction states with different temporal dynamics benefit from heterogeneous memory modeling. Jointly optimizing the classification objective with the CLR loss achieves the final improvement to 70.80 F1, with the largest gain on *approaching*, indicating that enforcing cross-modal recoverability complements the classifier for ambiguous interaction states. Each component targets a distinct failure mode, and progressive gains are interpretable rather than incidental.

Robustness under Observation Instability. Stable and unstable subsets are constructed by selecting the most stable and most unstable clips from each interaction category based on visual and segmentation-derived instability indicators (Fig. 3 (a)). The baseline suffers a 32.84 F1 drop from stable to unstable clips, while the proposed method degrades by only 8.17. The performance gap

Table 2. Ablation study of the proposed framework CLR: Cross-modal Latent Recovery; Dual: Dual Asynchronous Memory; $\mathcal{L}_{\text{CLR}}$: Loss from CLR.

Method	CLR	Dual	$\mathcal{L}_{\text{CLR}}$	**F1**	**Pre**	**Rec**	**Acc**	**F1**$_{\text{No}}$	**F1**$_{\text{App}}$	**F1**$_{\text{Tou}}$
Baseline	–	–	–	55.54	55.60	56.96	57.78	68.29	46.15	52.17
+CLR	✓	–	–	65.13	65.97	64.64	66.67	75.56	46.15	73.68
+Dual	✓	✓	–	69.06	72.80	70.51	66.67	68.42	54.55	84.21
Full Model	✓	✓	✓	**70.80**	**71.34**	**71.52**	**71.11**	**74.42**	**64.29**	**73.68**

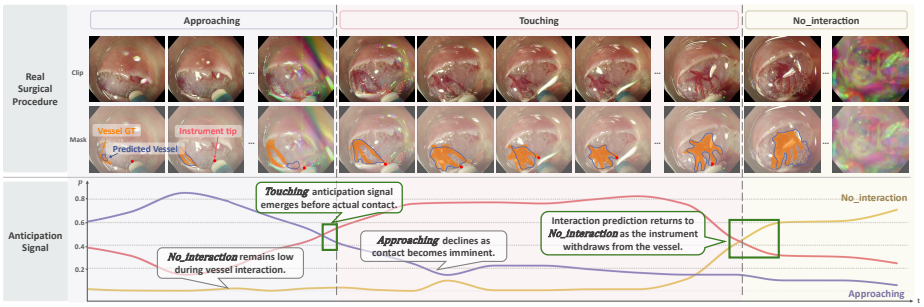

Fig. 4. Anticipation signals predicted on a real ESD clinical video. The touching signal rises before actual instrument–vessel contact, demonstrating proactive warning capability in continuous surgical workflow.

on unstable clips (66.67 vs. 25.00 Acc) confirms that the proposed mechanisms are most effective precisely when observations are most corrupted.

Memory Allocation across Interaction Categories. Fig. 3 (b) confirms that the model adaptively differentiates observation quality based on interaction semantics. *Touching* yields the highest memory weight (0.6443), reflecting greater reliance on transient observations during physical contact, while *no_interaction* receives the lowest (0.3886), relying more on prototypal features due to fewer distinctive temporal cues.

Anticipation on Continuous Surgical Video. Fig. 4 visualizes predicted interaction signals over a continuous real ESD surgical video spanning all three interaction states. During the *approaching* phase, the *touching* probability rises as the instrument converges toward the vessel. Crucially, the *touching* signal peaks approximately 0.5–1 second before actual contact, demonstrating genuine anticipatory capability rather than frame-level reaction. As the instrument withdraws, the prediction transitions smoothly back to *no_interaction*, reflecting coherent temporal reasoning across the full interaction cycle. This continuous and timely anticipation behavior suggests practical potential for real-time intraoperative warning during surgical procedures.

4 Conclusion

This work introduces vessel–instrument interaction anticipation as a new task for proactive surgical safety, and proposes a framework that recovers reliable interaction representations from jointly noisy observations and performs temporally adaptive reasoning through dual asynchronous memory with quality-gated admission. Experiments on our ESD dataset validate both the anticipation accuracy and temporal stability of the approach. The current evaluation is conducted on a single dataset, and validating the approach across broader clinical settings remains an important direction for future work. We hope this work offers a useful step toward anticipatory safety systems in surgical video analysis, where timely prediction of dangerous instrument–tissue interactions can directly contribute to safer procedures and better patient outcomes.

5 Acknowledgements

This work described in this paper was fully supported by a grant from the ANR/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and the French National Research Agency (Project No. A-CUHK402/23), and by the Innovation and Technology Fund of the Hong Kong Special Administrative Region, China (Project No. ITS/311/23FP).

Disclosure of Interests

The authors have no competing interests to declare.

References

1. Yang, H., Chen, C., Chen, Y., Scheppach, S., Yip, H. C., Dou, Q.: Uncertainty estimation for safety-critical scene segmentation via fine-grained reward maximization. *Adv. Neural Inf. Process. Syst.* **36**, 36238–36249 (2023).
2. Kamitani, Y., Nonaka, K., Misumi, Y., Isomoto, H.: Safe and efficient procedures and training system for endoscopic submucosal dissection. *Journal of Clinical Medicine* **12**(11), 3692 (2023).
3. Visrodia, K., Dobashi, A., Bazerbachi, F., Poneros, J., Sethi, A.: Endoscopic submucosal dissection facilitating techniques among non-experts: A systematic literature review. *Digestive Diseases and Sciences* **68**(6), 2561–2584 (2023).
4. Lau, K. C., Yam, Y., Chiu, P. W. Y.: An advanced endoscopic surgery robotic platform for removal of early-stage gastrointestinal cancer using endoscopic submucosal dissection. *HKIE Trans.* **28**, 186–198 (2021).
5. Rashid, M. U., Alomari, M., Afraz, S., Erim, T.: EMR and ESD: Indications, techniques and results. *Surgical Oncology* **43**, 101742 (2022).
6. Scheppach, M. W., Yip, H. C., Chen, Y., Yang, H., Cao, J., Chua, T., Dou, Q., Meng, H. M. L., Yam, Y., Chiu, P. W.: Feasibility of real-time artificial intelligence-assisted anatomical structure recognition during endoscopic submucosal dissection. *Endosc. Int. Open* **13**, e1–e8 (2025).

7. Steinbrück, I., Faiss, S., Dumoulin, F. L., Oyama, T., Pohl, J., von Hahn, T., Schmidt, A., Allgaier, H.-P.: Learning curve of endoscopic submucosal dissection (ESD) with prevalence-based indication in unsupervised Western settings: A retrospective multicenter analysis. *Surgical Endoscopy* **37**(4), 2574–2586 (2023).
8. Takao, M., Bilgic, E., Kaneva, P., Waschke, K., Endo, S., Nakano, Y., Kawara, F., Tanaka, S., Ishida, T., Morita, Y., et al.: Development and validation of an endoscopic submucosal dissection video assessment tool. *Surgical Endoscopy* **35**, 2671–2678 (2021).
9. Mitsuyoshi, Y., Ide, D., Ohya, T. R., Ishihoka, M., Yasue, C., Chino, A., Igarashi, M., Nakashima, A., Saito, S., Fujisaki, J., et al.: Training program using a traction device improves trainees' learning curve of colorectal endoscopic submucosal dissection. *Surgical Endoscopy*, 1–8 (2021).
10. Draganov, P. V., Aihara, H., Karasik, M. S., Ngamruengphong, S., Aadam, A. A., Othman, M. O., Sharma, N., Grimm, I. S., Rostom, A., Elmunzer, B. J., et al.: Endoscopic submucosal dissection in North America: A large prospective multicenter study. *Gastroenterology* **160**(7), 2317–2327 (2021).
11. Pattarajierapan, S., Saito, Y., Takamaru, H., Toyoshima, N., Wisedopas, N., Wanpiyarat, N., Lerttanatum, N., Khomvilai, S.: Learning curve of colorectal endoscopic submucosal dissection of an endoscopist experienced hands-on training in Japan. *Journal of Gastroenterology and Hepatology* (2024).
12. Li, J., Jin, Y., Chen, Y., Yip, H.-C., Scheppach, M., Chiu, P. W.-Y., Yam, Y., Meng, H. M.-L., Dou, Q.: Imitation learning from expert video data for dissection trajectory prediction in endoscopic surgical procedure. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 494–504 (2023).
13. Wang, H., Long, Y., Chen, Y., Yip, H.-C., Scheppach, M., Chiu, P. W.-Y., Yam, Y., Meng, H. M.-L., Dou, Q.: Learning dissection trajectories from expert surgical videos via imitation learning with equivariant diffusion. *Med. Image Anal.* **103**, 103599 (2025).
14. Nwoye, C. I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Recognition of instrument–tissue interactions in endoscopic videos via action triplets. In: *Med. Image Comput. Comput.-Assist. Interv.*, pp. 364–374 (2020).
15. Yu, Z., Du, C., Liang, H., Zheng, X., Ma, Z., Wu, M., Ao, M., Lao, Q.: Endoscopic artifact inpainting for improved endoscopic image segmentation. In: *Med. Image Comput. Comput.-Assist. Interv.*, pp. 191–201 (2025).
16. Zhang, B., Goel, B., Sarhan, M. H., Kejriwal Goel, V., Abukhalil, R., Kalesan, B., Stottler, N., Petculescu, S.: Surgical workflow recognition with temporal convolution and transformer for action segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **18**(4), 785–794 (2023).
17. Chen, Y., Wang, K.-N., Tayupo, D., Huaulmé, A., Nyangoh Timoh, K., Jannin, P., Dou, Q.: DSTED: decoupling temporal stabilization and discriminative enhancement for surgical workflow recognition. *International Journal of Computer Assisted Radiology and Surgery*, 1–12 (2026).
18. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.-A.: Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging* **40**(7), 1911–1923 (2021).
19. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L. C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: LoViT: Long video transformer for surgical phase recognition. *Medical Image Analysis* **99**, 103366 (2025).
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations* (2021).

21. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., Heng, P.-A.: SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. In: *Med. Image Comput. Comput.-Assist. Interv.*, pp. 182–190 (2017).
22. Cao, J., Yip, H.-C., Chen, Y., Scheppach, M., Luo, X., et al.: Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. *Nat. Commun.* **14**(1), 6676 (2023).
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997).