SurgLQA: Scalable Long-Horizon Surgical Video Question Answering

Diandian Guo, Xikai Yang, Ruiyang Li, Jialun Pei ✉, Pheng-Ann Heng

The Chinese University of Hong Kong
peijialun@gmail.com

Abstract. Surgical Video Question Answering (VideoQA) provides a promising paradigm for dynamic intraoperative interpretation, enabling real-time decision support and context-aware retrieval in clinical environments. Nevertheless, existing approaches are predominantly restricted to images or short clips, limiting their ability to model long-range procedural dynamics and causal dependencies across extended surgical workflows. To address this challenge, we propose SurgLQA, a unified long-horizon VideoQA framework for scalable surgical reasoning. This framework incorporates Faithful Temporal Consolidation (FTC), which leverages intrinsic temporal cues to construct compact long-range representations while preserving fine-grained temporal fidelity. Further, we develop Temporally-Grounded Multi-Policy Scaling (TMS), an adaptive test-time inference paradigm that strategically adjusts policy-level reasoning capacity within temporally grounded contexts. To facilitate systematic evaluation, we restructured a long-duration colonoscopy VideoQA benchmark, Colon-LQA, and conducted extensive experiments on Colon-LQA and REAL-Colon-VQA. Experimental results demonstrate that our approach achieves consistent performance gains in long-range reasoning with temporally grounded inference. Code link: SurgLQA.

Keywords: Surgical scene understanding · Video question answering · Temporal modeling.

1 Introduction

With modern surgical procedures becoming increasingly complex and prolonged, clinical environments demand a new paradigm capable of dynamically understanding and reasoning across the entire surgical process [22], rather than relying solely on stage-specific analysis of predefined objectives [3,12]. Cognitive decision-making in clinical settings frequently relies on cross-stage contextual information and a holistic understanding of the evolution of surgical procedures [15,14]. In this context, surgical Video Question Answering (VideoQA) has emerged as a promising paradigm by integrating natural language with surgical video interactions, thereby facilitating flexible workflow reasoning within a unified inference framework [18]. This technology not only provides real-time intraoperative decision assistance [11,8] but also supports postoperative review, surgical quality assessment, and data-driven clinical research [17,23,16,24].

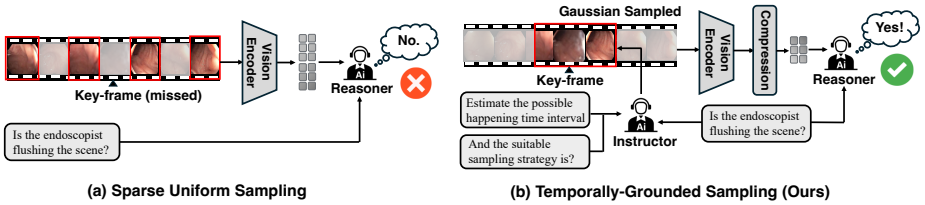

Fig. 1. (a) Existing VideoQA methods directly encode uniformly sampled frames, which may overlook subtle temporal evidence in keyframes and incur increased computational overhead; (b) SurgLQA adopt a temporally grounded sampling policy and compression mechanism to construct focused video representations.

However, scalable surgical VideoQA remains fundamentally constrained by long-horizon temporal reasoning. Unlike general video understanding tasks, surgical procedures often span tens of minutes to hours, where clinically critical events occur sparsely and depend on long-range causal relationships across different operation stages. Consequently, effective surgical VideoQA needs to meet the following requirements: 1) precise temporal modeling across extended durations, 2) accurate localization of sparse yet critical surgical events, and 3) scalable collaborative reasoning under limited computational resources.

Existing surgical VQA research faces two intertwined limitations in both benchmarks and modeling paradigms [9,5,7]. On the data side, most existing benchmarks emphasize image-level or short clip–level settings [25,5], prioritizing static visual cues such as instrument type or spatial layout. Corresponding question–answer pairs are typically anchored to narrowly defined frames [9,19] or brief video segments [5], diverging from real clinical workflows where reasoning involves cross-episode evidence aggregation over extended temporal horizons. On the modeling side, most surgical vision–language models rely on fixed uniform or sparse sampling strategies [1,5] despite the inherently heterogeneous dynamics of surgical workflows. Consequently, such strategies may overlook sparsely distributed yet clinically critical events (see Fig. 1), undermine question-conditioned temporal reasoning, and compromise scalability for long videos. Collectively, these limitations leave long surgical VideoQA largely unexplored.

In light of these challenges, we propose SurgLQA, a unified surgical VideoQA framework for scalable and temporally grounded procedural reasoning. SurgLQA models long-duration surgical workflows through *Faithful Temporal Consolidation* (FTC), which leverages intrinsic temporal cues to compress frame-level representations into compact yet temporally faithful embeddings. Building upon this representation, we introduce *Temporally-Grounded Multi-Policy Scaling* (TMS), an adaptive inference mechanism that dynamically modulates reasoning granularity across temporally localized contexts. As illustrated in Fig. 1, TMS embraces temporally grounded adaptive reasoning instead of sparse uniform sampling, enabling dynamic concentrate on clinically relevant segments to achieve scalable and reliable long-horizon surgical VideoQA. Additionally,

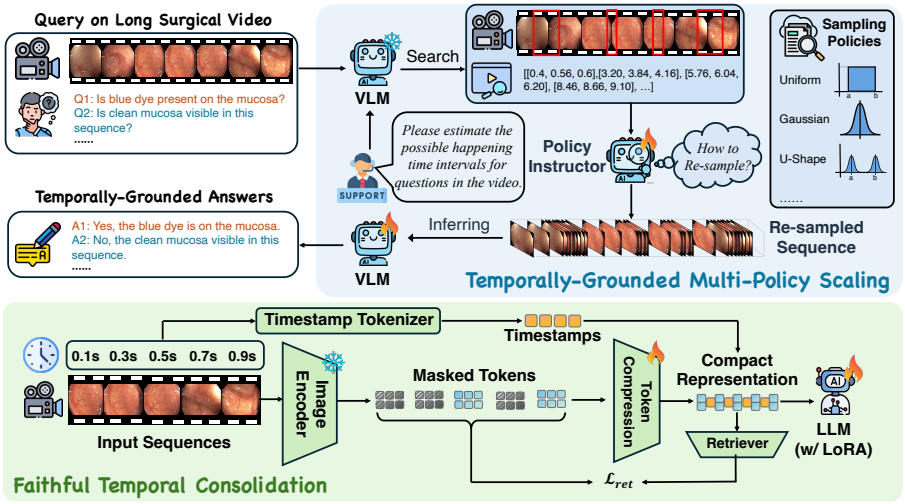

Fig. 2. Overview of the proposed framework for long surgical VideoQA.

we restructure a long-duration colonoscopy VideoQA benchmark, Colon-LQA, by concatenating multiple temporally ordered real video segments with corresponding question–answer pairs to construct extended sequences [6]. Extensive experiments on Colon-LQA and REAL-Colon-VQA [5] demonstrate that SurgLQA improves long-range reasoning through event-level temporal grounding.

2 Methodology

2.1 Framework Overview

Fig. 2 illustrates the proposed framework for long surgical VideoQA. We first employ *Faithful Temporal Consolidation* (FTC) to compress long video streams into compact, temporally structured representations for efficient long-horizon modeling. Next, we apply *Temporally-Grounded Multi-Policy Scaling* (TMS) to perform query-conditioned temporal grounding. Specifically, TMS identifies candidate intervals relevant to the query and selects an appropriate sampling policy for each grounded window, generating a query-adaptive re-sampled sequence with refined temporal resolution. The re-sampled sequence is then directly fed into the vision–language model for the final answer generation.

2.2 Faithful Temporal Consolidation

Typically long surgical videos generate massive visual tokens, rendering exhaustive encoding computationally prohibitive. To this end, we propose *Faithful Temporal Consolidation* (FTC), a learnable spatiotemporal token compression module that reduces redundancy while preserving temporal fidelity. Let an input

video be denoted as $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$, $\mathbf{I}_t \in \mathbb{R}^{C\times H\times W}$. The video is partitioned into non-overlapping 3D blocks of size $p_t \times p_s \times p_s$, yielding

$$\mathbf{V}' \in \mathbb{R}^{N\times C\times p_t\times p_s\times p_s}, \quad N = \frac{T}{p_t}\frac{H}{p_s}\frac{W}{p_s}. \tag{1}$$

Then, we apply a stack of temporal-aware 3D convolution layers with SiLU activation to the sequence $\mathbf{V}'$, encouraging the model to learn compact and informative spatiotemporal representations $\{\mathbf{Z}_l\}$:

$$\mathbf{Z}_l = \sigma(\mathrm{Conv3D}_l(\mathbf{Z}_{l-1})), \quad l = 1, \ldots, L. \tag{2}$$

where $\sigma(\cdot)$ denotes the SiLU activation function. To preserve explicit temporal cues for long-horizon reasoning, we adopt an interleaved timestamp-token design, constructing a text prompt $\tau_i =$ ‘‘`timestamp:` t_i `seconds`’’ and tokenize it as T_i. The sequence is formed by interleaving timestamp and visual tokens:

$$S = [T_1; V_1; T_2; V_2; \cdots; T_{N'}; V_{N'}], \quad T_i = \phi_{\mathrm{tokenizer}}(\tau_i), \tag{3}$$

where V_i denotes the consolidated tokens of the i-th temporal unit. This explicit temporal marking can retrieve temporal boundaries in long videos by reading inserted timestamp tokens. Besides, we bring an auxiliary retrieving loss $\mathcal{L}_{\mathrm{ret}}$ to preserve pre-compression fidelity. Giving pre-compression tokens $\mathbf{Z}_1$ and consolidated representations $\mathbf{Z}_{\mathrm{FTC}}$, we randomly sample an index set $\mathcal{M}\subset\{1, \ldots, N\}$. To prevent trivial copying from temporally adjacent tokens, random masking is applied to the compressed representation, followed by a lightweight retriever $g(\cdot)$ to predict masked tokens. The retrieving objective is formulated as

$$\mathcal{L}_{\mathrm{ret}} = \sum_{i\in\mathcal{M}} \left\| \left[g(\mathrm{Mask}(\mathbf{Z}_{\mathrm{FTC}}))\right]_i - \mathbf{z}_i^{(1)} \right\|_2^2, \tag{4}$$

where $\mathbf{z}_i^{(1)}$ denotes the i-th pre-compression token for retrieving. By reconstructing masked intermediate features, FTC encourages the compressed representation to preserve temporal fidelity under token compression.

2.3 Temporally-Grounded Multi-Policy Scaling

Current video reasoning study relies on fixed uniform sampling during inference, overlooking the highly heterogeneous temporal dynamics. Certain visual attributes, such as illumination conditions and imaging modality, remain relatively stable and are distributed uniformly over time. In contrast, critical procedural cues, *e.g.*, transient occlusions, sudden scope movements, or the appearance of specific instruments, often occur sparsely and may only persist for brief moments. As illustrated in Fig. 3, different question types exhibit distinct preferences over temporal sampling patterns, indicating that a single static sampling policy insufficiently accommodate diverse temporal characteristics. Motivated by this observation, we propose Temporally-Grounded Multi-Policy Scaling (TMS) to dynamically select sampling policies conditioned on queries.

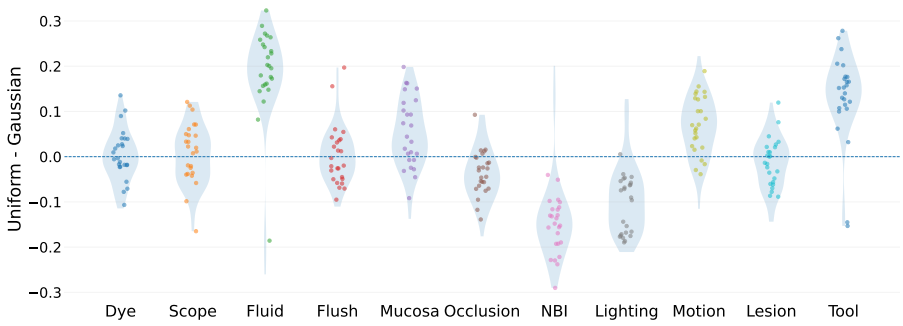

Fig. 3. Instructor reveals distinct sampling preferences across question types. Temporally fine-grained events (*e.g.*, fluid, motion) favor Gaussian sampling, while global static questions (*e.g.*, lighting) benefit more from uniform sampling.

TMS first identifies temporally relevant windows and leverages a lightweight policy instructor to determine the most suitable sampling distribution for each window. Given a video $V=\{f_t\}_{t=1}^{T}$ and a query q, TMS construct a compact yet evidence-dense resampled sequence V^* that preserves task-critical temporal cues. We augment the query with an additional instruction: *"Please estimate the possible happening time intervals for questions in the video."* The model outputs a set of predicted intervals, from which we extract their centers $\{\mu_m\}$ as temporal anchors to construct query-relevant temporal windows $\mathcal{B} = \cup_m B(\mu_m, w)$:

$$B(\mu_m, w) = \{t \in \{1, \ldots, T\} : |t - \mu_m| \leq w\}. \tag{5}$$

Within each window B_m, we maintain a candidate policy set $\mathcal{P} = \{p_1, \ldots, p_K\}$ representing different temporal sampling patterns (*i.e.*, Gaussian, Uniform, Dense, U-shape). Rather than manually selecting a fixed policy, a policy instructor generates appropriate proposals conditioned on the window context and the query:

$$\hat{p} = \text{Instructor}(B_m, q). \tag{6}$$

Given B_m, the selected policy $\hat{p}$ shapes the resampling distribution to produce a compact sequence V^*, which is then fed into the vision–language model for answer generation. The resampled sequence is then fed into the vision–language model for the final answer generation. The overall training objective consists of:

$$\mathcal{L} = \mathcal{L}_{\text{QA}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}} + \lambda_{\text{policy}}\mathcal{L}_{\text{policy}}. \tag{7}$$

Here, $\mathcal{L}_{\text{QA}}$ denotes the next-token prediction loss computed on the final answer generated from V^*. The instructor is trained via cross-entropy over the K candidate strategies to optimize $\mathcal{L}_{\text{policy}}$. The coefficients λ_{ret} and λ_{policy} balance the contributions of retrieval alignment and policy supervision, respectively.

Table 1. Performance comparison on the Colon-LQA benchmark. † denotes zero-shot evaluation without task-specific fine-tuning.

Model	In-template				Out-of-template			
	BLE-4	ROU-L	MET	K-ACC	BLE-4	ROU-L	MET	K-ACC
Qwen3VL† [2]	6.26	31.18	29.73	41.42	2.11	29.84	26.27	38.35
InternVL3† [27]	4.37	28.05	34.62	33.18	2.64	27.57	32.88	40.12
SurgicalGPT [18]	13.26	44.13	48.84	47.58	9.74	39.27	42.05	38.46
PitVQA [9]	24.73	51.78	55.06	44.17	18.67	45.18	47.73	36.28
Qwen3VL [2]	63.87	72.94	74.24	63.46	30.47	60.14	61.57	53.18
MedGemma [20]	24.68	48.83	50.74	41.86	16.97	40.26	42.16	34.37
InternVL3 [27]	57.28	73.36	72.18	60.25	27.53	56.47	57.73	49.63
VideoLLaMA3 [26]	60.17	70.26	75.73	61.64	29.27	58.64	59.97	51.06
SurgViVQA [5]	52.46	66.74	68.17	54.63	25.86	52.47	53.66	46.94
SurgLQA(Ours)	**74.37**	**85.03**	**83.00**	**66.67**	**38.01**	**69.50**	**65.85**	**65.73**

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate SurgLQA on the restructured long-horizon benchmark Colon-LQA and REAL-Colon-VQA [5]. Specifically, we stitch consecutive colonoscopy clips to form long, continuous videos and aggregate their associated question–answer pairs into extended VideoQA samples. Each long-video instance spans approximately 34 seconds, posing challenges for long-range retrieval and reasoning. Following previous studies [5], we evaluate our model using BLEU-4, ROUGE-L, METEOR, and Keyword Accuracy (K-ACC), and report both in-template and out-of-template results. All ablation studies are conducted on Colon-LQA to evaluate each component's contribution to long-horizon VideoQA.

3.2 Implementation Details

All experiments are conducted on two NVIDIA RTX 4090 GPUs. The model is trained for three epochs using the AdamW [13] optimizer with a weight decay of 0.01 and a base learning rate of 5×10^{-4}. The batch size is 2 with gradient accumulation over 8 steps. The visual encoder is initialized from SigLIP2-Large 300M [21], and the language backbone adopts the pre-trained Qwen3-2B [1] model. For policy selection, we use a lightweight instructor built upon BERT-base-uncased [4]. We employ LoRA [10] with rank 8 and scaling factor 16 for parameter-efficient adaptation of the language backbone while keeping its original weights frozen. Other task-specific modules are optimized during training.

3.3 Comparison with State-of-the-art Methods

Table 1 and Table 2 compare our method with state-of-the-art approaches on the Colon-LQA long-video benchmark and REAL-Colon-VQA [5]. Due to the poor zero-shot performance observed, we fine-tune all compared methods under

Table 2. Performance comparison on the REAL-Colon-VQA [5] benchmark. † denotes zero-shot evaluation without task-specific fine-tuning.

Model	In-template				Out-of-template			
	BLE-4	ROU-L	MET	K-ACC	BLE-4	ROU-L	MET	K-ACC
Qwen3VL† [2]	7.31	40.34	44.87	45.40	6.42	36.42	41.19	47.86
InternVL3† [27]	7.13	43.90	54.31	68.67	3.87	32.10	44.34	53.73
SurgicalGPT [18]	14.93	47.85	52.36	33.33	12.35	42.87	50.49	46.67
PitVQA [9]	64.55	78.48	79.99	54.13	23.63	50.03	53.22	42.93
Qwen3VL [2]	82.40	88.75	89.10	72.35	35.82	64.50	63.41	66.28
MedGemma [20]	69.85	80.92	82.40	60.78	27.16	52.80	50.43	48.62
InternVL3 [27]	72.30	83.15	85.22	65.91	29.44	55.36	53.90	52.84
VideoLLaMA3 [26]	76.47	85.60	84.75	68.34	37.08	59.82	58.55	66.97
SurgViVQA [5]	71.98	82.85	84.11	63.20	31.19	53.62	54.89	47.73
SurgLQA(Ours)	**87.69**	**92.93**	**91.85**	**79.60**	**44.89**	**72.98**	**69.59**	**78.93**

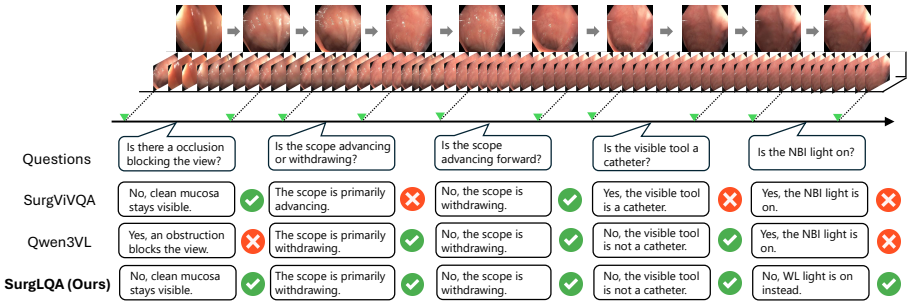

Fig. 4. Qualitative comparison with representative models on Colon-LQA.

a unified task-specific training protocol (same splits, resolution, and metrics) following their default implementations to ensure fair and meaningful evaluation.

On the Colon-LQA long-video bench, our model achieves superior performance across all four metrics, highlighting its long-range understanding capabilities. Under out-of-template settings, where question formulations differ from those seen during training, our method maintains clear advantages, suggesting deeper semantic reasoning rather than reliance on template-specific cues. Fig. 4 further demonstrates improved temporal coherence and contextual consistency over extended video sequences. On REAL-Colon-VQA, while the benefit of long-horizon modeling is naturally reduced, our method still achieves competitive performance, indicating strong generalization to standard short-video benchmarks.

3.4 Ablation Study

Contributions of Key Components. As shown in the left of Fig. 5, introducing FTC consistently improves all metrics, indicating more effective preservation of critical temporal cues with reduced redundancy. Adding Temporal Grounding (TG) further enhances performance, particularly in keyword accuracy, reflecting the benefit of aligning temporal sampling with query-relevant intervals. With

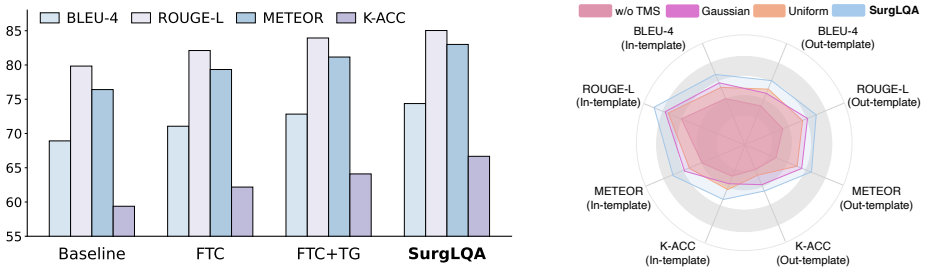

Fig. 5. Left: Ablations of key components in SurgLQA. Right: Ablations of TMS.

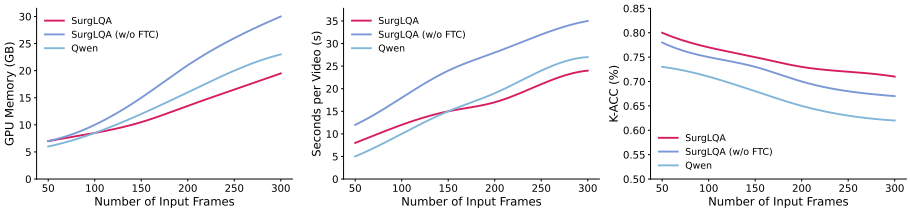

Fig. 6. Scalability analysis with increasing input frames.

Policy Instructor (PI), the model achieves better overall results, demonstrating the advantage of adaptive test-time policy selection over fixed strategies.

Ablations for TMS. The right part of Fig. 5 compares different inference strategies within TMS. As shown in the figure, removing TMS yields the weakest performance, while incorporating temporal grounding with a single policy improves all metrics, demonstrating the benefit of focusing on temporally relevant intervals. The better results are achieved with full policy selection, indicating that adaptive multi-policy scaling further enhances long-range reasoning.

Scalability Analysis. As illustrated in Fig. 6, removing FTC substantially increases memory and runtime while reducing K-ACC as the frame budget grows. Direct frame-level scaling thus incurs significant computational overhead without effectively preserving temporal evidence. In contrast, SurgLQA maintains lower resource consumption and consistently higher K-ACC, confirming the effectiveness of temporally faithful compression for efficient long-horizon reasoning.

4 Discussion

Clinical Relevance and Deployment. The practical utility of surgical VideoQA frameworks heavily depends on their clinical relevance and computational feasibility in real-time intraoperative environments. Combined with the runtime efficiency analysis illustrated in Fig. 6, SurgLQA achieves low-latency inference and exceptional computational scalability under extended temporal horizons, strongly supporting its potential for seamless integration into practical clinical workflows and real-time intraoperative decision assistance.

Beyond Counterparts. Unlike generic long-video methods that miss sparse, fine-grained surgical events via uniform sampling or token compression, SurgLQA targets long-horizon sparse evidence retrieval. FTC preserves temporal fidelity via a dedicated retrieval loss, while TMS replaces rigid grids with query-conditioned grounding and multi-policy re-sampling to capture critical details.

5 Conclusion

This paper introduces SurgLQA, a unified long-horizon framework for surgical video question answering. Faithful Temporal Consolidation models extended surgical procedures by constructing compact yet temporally coherent representations that preserve fine-grained procedural cues. Temporally-Grounded Multi-Policy Scaling further regulates reasoning granularity across temporally salient contexts, enabling grounded modeling of temporal dynamics within long surgical workflows. Experimental results on the restructured long-duration benchmark Colon-LQA and REAL-Colon-VQA demonstrate consistent improvements in long-range reasoning enabled by event-level temporal grounding.

Acknowledgment. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N, and by the Hong Kong Innovation and Technology Fund, under Project GHP/252/23SZ.

Disclosure of Interests. The authors have no competing interests.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report. In: arXiv preprint arXiv: 2511.21631 (2025)
3. Cao, J., Yip, H.C., Chen, Y., et al.: Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. Nature Communications **14**, 6676 (2023)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) NAACL. pp. 4171–4186 (2019)
5. Drago, M.O., Carlini, L., Balyemez, P.C., Pierantozzi, D., Lena, C., Hassan, C., Stoyanov, D., Momi, E.D., Bano, S., Hoque, M.I.: Surgvivqa: Temporally-grounded video question answering for surgical scene understanding. In: arXiv preprint arxiv: 2511.03325 (2025)
6. Durante, Z., Singh, S., Khatua, A., Agarwal, S., Tan, R., Lee, Y.J., Gao, J., Adeli, E., Fei-Fei, L.: Videoweave: A data-centric approach for efficient video understanding. In: arXiv preprint arXiv: 2601.06309 (2026)
7. Gautam, S., Storås, A.M., Midoglu, C., Hicks, S.A., Thambawita, V., Halvorsen, P., Riegler, M.A.: Kvasir-vqa: A text-image pair gi tract dataset. In: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications. pp. 3–12 (2024)

8. Guo, D., Si, W., Li, Z., Pei, J., Heng, P.A.: Surgical workflow recognition and blocking effectiveness detection in laparoscopic liver resection with pringle maneuver. AAAI **39**(3), 3220–3228 (2025)
9. He, R., Khan, D.Z., Mazomenos, E.B., Marcus, H.J., Stoyanov, D., Clarkson, M.J., Islam, M.: Pitvqa++: Vector matrix-low-rank adaptation for open-ended visual question answering in pituitary surgery. arXiv preprint arXiv:2502.14149 (2025)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
11. 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)
12. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: Skit: a fast key information video transformer for online surgical phase recognition. In: IEEE ICCV. pp. 21074–21084 (2023)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
14. Pei, J., Guo, D., Zhang, J., Lin, M., Jin, Y., Heng, P.A.: S^2former-or: Single-stage bi-modal transformer for scene graph generation in or. IEEE Transactions on Medical Imaging **44**(1), 361–372 (2025)
15. Pei, J., Zhang, J., Qin, G., Wang, K., Jin, Y., Heng, P.A.: Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. IEEE transactions on Medical Imaging (2025)
16. Pei, J., Zhou, Z., Guo, D., Li, Z., Qin, J., Du, B., Heng, P.A.: Synergistic bleeding region and point detection in laparoscopic surgical videos. In: IEEE CVPR. pp. 1–10 (2026)
17. Qiu, P., Wu, C., Liu, S., et al.: Quantifying the reasoning abilities of llms on clinical cases. Nature Communications **16**, 9799 (2025). https://doi.org/10.1038/s41467-025-64769-1, https://doi.org/10.1038/s41467-025-64769-1
18. Seenivasan, L., Islam, M., Kannan, G., Ren, H.: Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In: MICCAI. pp. 281–290 (2023)
19. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: MICCAI. pp. 33–43. Springer (2022)
20. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
21. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025)
22. Varghese, C., Harrison, E.M., O'Grady, G., Topol, E.J.: Artificial intelligence in surgery. Nature Medicine pp. 1–12 (2024)
23. Wu, J., Holm, F., Chen, C., Wang, A., Hu, Y., Ye, X., et al.: Unisurg: A video-native foundation model for universal understanding of surgical videos (2026)
24. Yang, S., Zhou, F., Mayer, L., et al.: Large-scale self-supervised video foundation model for intelligent surgery. npj Digital Medicine (2026)
25. Yuan, K., Kattel, M., Lavanchy, J.L., Navab, N., Srivastav, V., Padoy, N.: Advancing surgical vqa with scene graph knowledge. International journal of computer assisted radiology and surgery **19**(7), 1409–1417 (2024)
26. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)

27. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)