AdaBIMBA: Adaptive Token Compression for Long Endoscopy Video Understanding

Ka-Wai Yung*, Pooya Mobadersany*,†, Chaitanya Parmar, Krishna Chaitanya, Fabio Gunderson, Lindsey Surace, Louis R. Ghanem, Tommaso Mansi, Gabriela Oana Cula, Kristopher Standish, and Pablo F. Damasceno

Johnson & Johnson

*These authors contributed equally to this work and share first authorship.

†Corresponding author: Pooya Mobadersany (pmobader@its.jnj.com).

Abstract. Vision–language models (VLMs) show strong potential for automating endoscopy-video interpretation, including detailed description, disease-severity assessment, and temporal event localization. However, most surgery and endoscopy focused VLMs remain trained primarily on static images, despite reliable assessment requiring long-range temporal reasoning over continuous procedures. This modality mismatch limits their applicability in real clinical workflows. Existing surgical video VLM datasets rely on noisy, public sources with limited clinical reliability or contain only short clips that fail to capture full procedural context. To address these gaps, we curated a long-video dataset of endoscopy recordings from a clinical trial of patients with Inflammatory Bowel Disease (IBD). We additionally introduce AdaBIMBA, a novel adaptive token compression module comprising (i) a Token Scorer that predicts frame importance and (ii) a Mamba-based compressor that adjusts per-frame compression ratios accordingly. AdaBIMBA enables processing substantially more frames under the same memory budget while preserving clinically relevant content, achieving state-of-the-art performance across all question categories on our IBD dataset.

Keywords: Vision-Language Models · Long-form Video Understanding · Endoscopic Video Analysis · Adaptive Token Compression.

1 Introduction

Vision–language models (VLMs) have advanced rapidly, driven by large-scale instruction datasets [25,17,7,16,2]. Building on this, medical VLMs have emerged for domain-specific tasks [20,38,15,11,37], with recent work extending to surgical settings that demand modeling complex intraoperative workflows [13,33,27,18,11,40]. However, most medical VLMs are trained on static images and therefore lack the temporal reasoning needed for continuous, long procedural videos. Existing surgical-video VLM datasets either source from public platforms (e.g., YouTube), introducing noisy labels and limited clinical validation, or rely on journal supplementary clips often under one minute [34,14][1]. These constraints limit clinical

[1] These datasets were not publicly available at the time of writing.

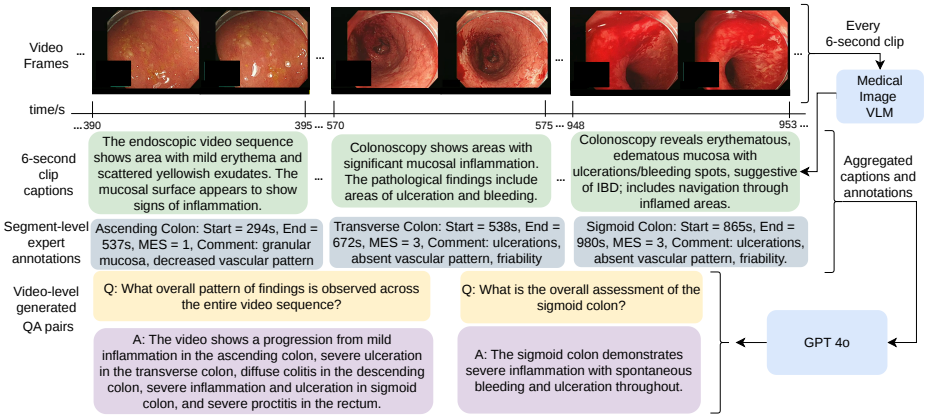

Fig. 1. Creation of training and testing QA pairs. Expert annotations provide start and end timestamps for each anatomic segment and assign MES with explanations. A medical image VLM generates captions for every 6-second clip. All captions and annotations are aggregated and fed to GPT-4o to produce video-level QA pairs.

reliability and prevent models from learning full procedural context. To bridge this gap, we curated a long endoscopy video dataset from an inflammatory bowel disease (IBD) clinical trial. Each recording captures the entire endoscopy examination and is paired with high-quality expert annotations detailing bowel segment–level severity and findings, enabling models to learn from the complete procedural workflow rather than isolated events [19,4].

While long videos provide richer temporal context, they introduce substantial computational challenges during training. Memory and compute demands scale quadratically with sequence length in transformers [32], making end-to-end training of high-resolution, long-duration videos impractical. A common workaround is uniform frame subsampling [6,30], which reduces input length but inevitably discards informative content. This is particularly problematic in endoscopy, where key clinical events (e.g., bleeding, ulceration) are brief and occupy only a small fraction of the procedure. Uniform sampling risks missing these events entirely, degrading both representation learning and downstream performance.

Prior work has explored strategies to mitigate the computational burden of long-video modeling; however, important limitations remain. LongVU [28] propose heuristic saliency based on cosine similarity to compress similar tokens across successive frames, but lacks task adaptivity. Dynamic-VLM [35] applies global compression rates conditioned on total video length, which does not account for uneven clinical relevance over time. AKS [31] selects keyframes based on prompt relevance but discards unselected frames entirely, risking the loss of subtle but critical transitions.

To address these challenges, we propose AdaBIMBA, a novel module that adaptively compresses frame tokens based on their importance to both the visual input and the question. AdaBIMBA employs a two-stage training strategy: In

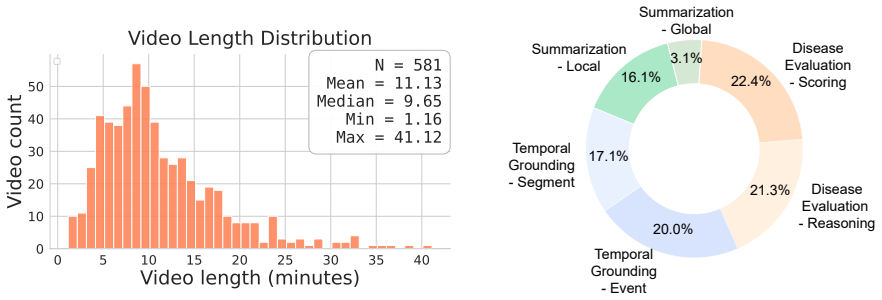

Fig. 2. Distribution of the dataset. (Left) Histogram of video lengths. (Right) Proportion of samples across QA categories generated by GPT-4o for training and testing.

stage 1, a Token Scorer is trained to predict frame importance using input-token gradients as supervision. In stage 2, a Mamba [9]-based compressor applies adaptive, per-frame compression ratios derived from these scores. This design enables processing of substantially more frames under the same memory budget while preserving clinically relevant content essential for long endoscopy reasoning.

We evaluate AdaBIMBA on our IBD dataset across six question types—Local Summarization, Global Summarization, Disease Evaluation Scoring, Disease Evaluation Reasoning, Segment-Level Temporal Grounding, and Event-Level Temporal Grounding—and demonstrate state-of-the-art performance with substantially higher input-frame budgets. Our contributions are threefold: **(1)** we introduce AdaBIMBA, a frame-token compressor that adaptively reduces tokens based on question-aware visual importance, without relying on attention weights and thus remaining compatible with FlashAttention [8]; **(2)** through extensive experiments, we show that AdaBIMBA enables training on significantly longer video sequences, yielding consistent gains across all question categories; **(3)** to the best of our knowledge, we present the first application of VLMs to full-length endoscopy procedures, extending VLMs toward realistic clinical workflows.

2 Dataset

We collected 581 full-length endoscopy recordings for assessing Ulcerative Colitis (UC), a form of IBD. Each video spans the entire procedure from insertion to withdrawal. Expert endoscopists (3-8+ years experience with 58.7% inter-rater agreement) annotated the start and end boundaries of all six anatomic segments—ileum; ascending (A-), transverse (T-), descending (D-), and sigmoid (S-) colon; and rectum—and assigned a segment-specific Mayo Endoscopic Subscore (MES, 0–3) indicating inflammation severity, with higher scores reflecting more active disease. As MES captures the worst appearance within a segment, experts also identified the most representative frame and documented key findings (e.g., bleeding, erosions) informing the score.

Table 1. Example of generated QA pairs and their formats for training and testing.

Category	Type	Question	Answer	Train	Test
Summarization	Local	How would you summarize the rectum's findings?	...exhibits progressing inflammation with absent vascular pattern and friability.	Sentence	Multiple Choice
	Global	What is the overall summary of the disease presentation across all examined regions?	...shows a presentation from normal mucosa in the ascending colon... to severe inflammation in the rectum.	Sentence	Multiple Choice
Disease Evaluation	Scoring	What is the MES for the rectum?	The Mayo Endoscopic Subscore is 3.	Sentence	Sentence
	Reasoning	What endoscopic findings justify the MES?	As seen with deep ulceration at frame 298.	Sentence	Multiple Choice
Temporal Grounding	Segment	Where is the rectum segment?	The rectum segment is seen at frame 125–158.	Sentence	Sentence
	Event	Which frames capture the transition of the sigmoid colon from normal mucosa to mild inflammation?	The sigmoid colon transitions from normal mucosa at frame 98–103 to mild inflammation at frame 103–125.	Sentence	Multiple Choice

While expert annotations provide high-quality data for segment-level temporal grounding and MES evaluation, the dataset lacks descriptive narration of procedural progression. Inspired by [29], we used HuatuoGPT-Vision [5]—a medical image VLM trained on a large scale medical data including endoscopy imagery—to caption every 6-second clip. These captions, combined with expert annotations, are used to prompt GPT-4o [21] to generate multiple question–answer (QA) pairs supporting procedural summarization and event-level temporal grounding, as shown in Fig. 1. To ensure factual correctness, GPT-4o is instructed to prioritize expert annotations when captions conflict. Fig. 2 shows distributions of video lengths and question types, and Table 1 provides examples of QA pairs. The dataset is split at patient level and contains 10,153 QA pairs from 441 training videos and 3,545 from 140 testing videos. During training, all QA pairs are provided in sentence form; during testing, Segment and Scoring tasks remain sentence-based, while the remaining tasks adopt the multiple-choice (MC) format used in [27,36,18]. To reduce answer bias, MC options are filtered by querying Qwen3-VL [2] 2B without video input; pairs repeatedly guessed correctly were removed.

3 Methods

We use Qwen3-VL 8B as our base VLM. As with other LLaVA-style [17] models, long-video training is constrained by rapidly increasing GPU memory usage as token length grows. BIMBA [12] mitigates this by inserting query tokens into the video token sequence and processing them with a Mamba selective-scan layer to capture long-range dependencies and key features. Only the processed query tokens are retained, reducing sequence length and memory consumption.

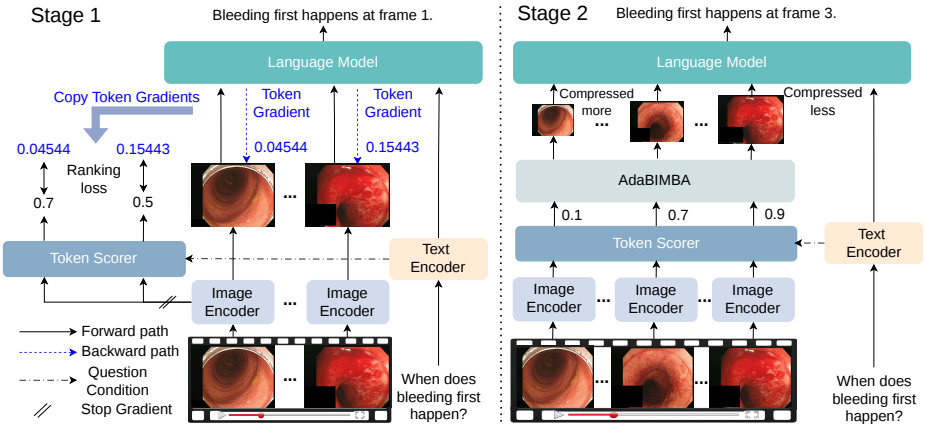

Fig. 3. AdaBIMBA staged training. All video frames are used up to a maximum frame count; longer videos are subsampled to fit this limit. All components except the text encoder are trainable. **(Stage 1)** The Token Scorer predicts frame-token importance, supervised using input-token gradients (blue dotted lines) with a ranking loss. **(Stage 2)** Predicted scores guide AdaBIMBA's adaptive compression, allocating more query tokens to high-importance frames and fewer to low-importance ones, increasing maximum frame capacity compared to baseline, while preserving key information. See Fig. 4 for scoring and compression details. For illustration, we assume one token per frame.

A key distinction between natural and endoscopy videos lies in the distribution of meaningful events. In natural videos, actions often span extended periods, whereas clinically relevant events in endoscopy—such as bleeding or ulceration—are brief and occupy only a small portion of the procedure. BIMBA uniformly inserts query tokens across the sequence, implicitly assuming equal importance throughout. This premise is suboptimal for endoscopy, where most frames have limited diagnostic value and rare but critical events warrant higher attention.

We therefore propose AdaBIMBA, a learnable adaptive-compression module that adjusts compression ratio according to frame-token importance (Fig. 3). AdaBIMBA is trained in two stages: an initial warm-up stage to stabilize the Token Scorer, followed by training with adaptively compressed inputs. In stage 1 (Fig. 3, left), token compression is disabled, and we maximize the number of frames that fit in GPU memory. The encoded features are processed by the Token Scorer—a lightweight Mamba with adaptive normalization [23] from DiT [22] for question-conditioned scoring (Fig. 4, left). Direct supervision using attention scores is infeasible under kernel-based attention (e.g., FlashAttention), as the full attention matrix is not materialized. Instead, motivated by prior work showing that input gradients provide an effective indicator for feature importance [1], we train the Scorer with a ranking loss based on the Spearman correlation between

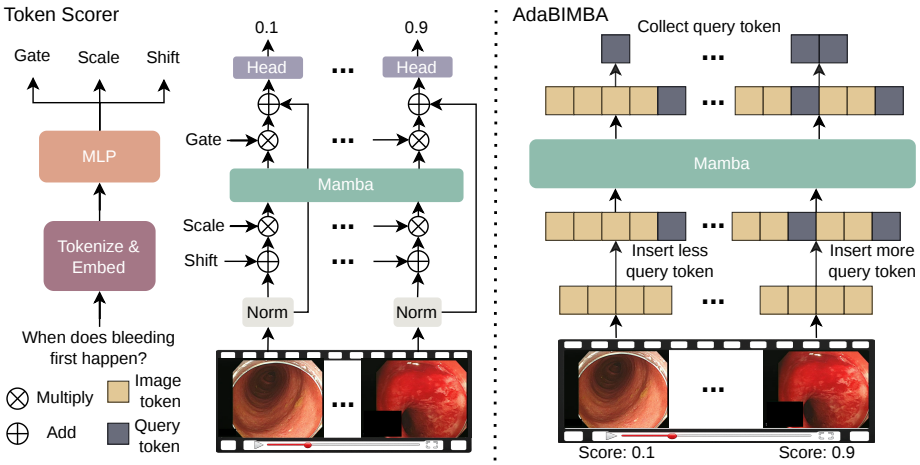

Fig. 4. (Left) Token Scorer: A Mamba estimates the importance across frames. The question embedding is projected into three parameters and injected via adaptive normalization, enabling question-conditioned importance scoring. **(Right)** AdaBIMBA compression: Frame scores determine the query token allocation per frame. Original frame tokens and the inserted query tokens are concatenated and processed with a Mamba selective-scan layer, after which only the processed query tokens are retained.

predicted scores p_i and input gradients y_i, where $r(\cdot)$ denotes the rank operator:

$$\mathcal{L}_{\text{Spearman}} = 1 - \frac{\sum_{i=1}^{n} \left(r(p_i) - \overline{r(\mathbf{p})}\right)\left(r(y_i) - \overline{r(\mathbf{y})}\right)}{\sqrt{\sum_{i=1}^{n} \left(r(p_i) - \overline{r(\mathbf{p})}\right)^2} \sqrt{\sum_{i=1}^{n} \left(r(y_i) - \overline{r(\mathbf{y})}\right)^2}} \tag{1}$$

In Stage 2 (Fig. 3, right), training proceeds with three main changes: 1) AdaBIMBA is activated, enabling adaptive token compression guided by the Scorer's importance scores; 2) The number of input frames is increased to the GPU limit after compression; 3) The Token Scorer is no longer supervised with the ranking loss. Fig. 4, right, shows the compression mechanism, which extends BIMBA with a key enhancement: variable per-frame compression rates. Frames are assigned to compression bands using a first-fill strategy—high-importance frames receive more query tokens, low-importance frames fewer—allocating token capacity to critical frames while supporting more total frames under the same memory budget. As band allocation is non-differentiable, we adopt a Straight-Through Estimator (STE) [3] with a softmax relaxation to enable end-to-end training.

4 Results and Discussion

Experiments were conducted on two NVIDIA A100 (80GB) GPUs with mixed precision training, using a batch size of 1 and gradient accumulation over 16

Table 2. Performance by question category across five inference runs. Results are in the format mean±std. Acc denotes accuracy; best results are in blue and underscored.

	Train	Summarization		Disease Evaluation		Temporal Grounding		
Metrics		Acc↑	Acc↑	F1↑	Acc↑	Acc↑	IoU↑	
Methods	Max Frame	Local	Global	Scoring	Reasoning	Event	Segment	Average
Qwen3-VL[2][2]	/	55.3±0.9	61.6±1.4	33.2±0.5	26.8±0.7	29.9±3.4	3.7±0.1	35.1±0.6
SurgPub-Video[14][3]	512	49.2±0.8	48.5±2.9	36.6±1.8	31.7±0.8	27.3±1.3	31.5±0.6	34.5±0.9
Video-XL2[24]	512	50.1±1.7	49.2±4.1	36.7±1.4	32.9±0.6	31.7±2.1	23.4±0.3	37.3±1.3
SurgVidLM[34][3]	128	57.1±1.2	60.0±4.8	40.8±1.1	33.8±1.0	24.3±1.0	31.5±0.3	41.2±1.2
Qwen3-VL	128	64.8±1.5	64.9±2.2	50.8±1.5	37.9±0.9	23.8±1.8	44.4±0.5	47.8±0.8
BIMBA[12]	512	62.8±0.9	66.1±1.8	53.3±1.2	40.1±0.5	35.6±1.6	46.6±0.4	50.8±0.3
AdaBIMBA	512	66.6±1.4	67.4±1.9	55.4±0.5	40.5±0.5	37.3±2.0	47.8±0.7	52.5±0.6
AdaBIMBA	1024	70.6±1.1	70.1±0.6	56.5±1.2	43.3±1.1	38.3±0.5	49.3±0.4	54.7±0.4

steps. We apply LoRA [10], training Stage 1 for 2 epochs and Stage 2 for 3 epochs, with a learning rate of $5e^{-5}$. We report the mean and standard deviation over five inference runs with different seeds to account for stochastic variation in the decoding process. Table 2 reports results by question types, using F1 for Scoring task, Intersection over Union (IoU) for Segment, and accuracy for all other tasks. Methods relying on temporal compression, such as Video-XL2 and SurgPub-Video, reduce token count via aggressive temporal token reduction and therefore consistently underperform. SurgVidLM uses a two-stage pipeline that first generates a summary and then an answer; under heavy subsampling, errors accumulate across stages and reduce accuracy. In contrast, BIMBA allows spatial-only compression, enabling more frames to be processed and achieving higher accuracy. At the equivalent frame count to BIMBA, AdaBIMBA outperforms BIMBA across all categories and improves average performance (Average) from 50.8 to 52.5. AdaBIMBA's adaptive compression allows higher token compression rates to further maximizing frame count to 1024, delivering best performance across all evaluation categories (paired t-test $p<0.02$ in Event, $p<0.005$ in other categories). While AdaBIMBA scores lower in MES=3, this drop is not statistically significant ($p=0.335$) with overlapping CIs. A plausible explanation is that mild–moderate disease has sparse/localized findings benefiting more from adaptive allocation, while severe disease often involves diffuse pathology already captured in shorter windows.

Table 3 provides a breakdown of Scoring performance (measured by F1), as well as additional Segment metrics using Recall@1 (R1) at IoU of 0.3, 0.5, and 0.7, following recent studies [26,39]. AdaBIMBA shows consistent improvement across all settings except MES 3. Table 4 summarizes the contribution of each AdaBIMBA component. Removing question conditioning produces consistent performance drops, especially for Local and Reasoning, indicating that

[2] No fine-tuning onto the IBD dataset.

[3] Original code was not available at the time of writing; reported results are from our reimplementation of the cited methods using Qwen3-VL for comparability.

Table 3. Class breakdown by MES and additional temporal grounding metrics.

	Disease Evaluation – Scoring				Temporal Grounding – Segment		
Methods	0	1	2	3	$R1@0.3\uparrow$	$R1@0.5\uparrow$	$R1@0.7\uparrow$
Qwen3-VL	59.8±2.0	35.8±3.8	40.4±2.1	52.1±1.7	59.6±1.0	44.7±0.7	29.4±0.8
BIMBA	59.6±0.5	34.1±1.5	39.3±3.8	53.5±2.5	62.1±0.6	45.8±0.5	31.6±0.5
AdaBIMBA	64.6±0.8	39.0±1.7	44.1±3.4	49.6±4.5	65.7±1.3	50.5±0.9	34.3±0.5

Table 4. Ablation results for the design components of AdaBIMBA.

	Summarization		Disease Evaluation		Temporal Grounding	
Methods	Local	Global	Scoring	Reasoning	Event	Segment
AdaBIMBA	70.6±1.1	70.1±0.6	56.5±1.2	43.3±1.1	38.3±0.5	49.3±0.4
✗ Question Condition	68.5±0.6	68.5±1.4	55.4±1.0	40.3±1.2	37.2±1.6	48.5±0.9
✗ Token Scorer	65.0±1.4	69.0±1.9	54.0±1.7	42.4±0.6	36.9±1.6	46.0±2.3
✗ Stage 1	67.5±1.3	69.9±1.8	52.2±0.6	42.2±0.7	37.3±1.4	47.6±0.8

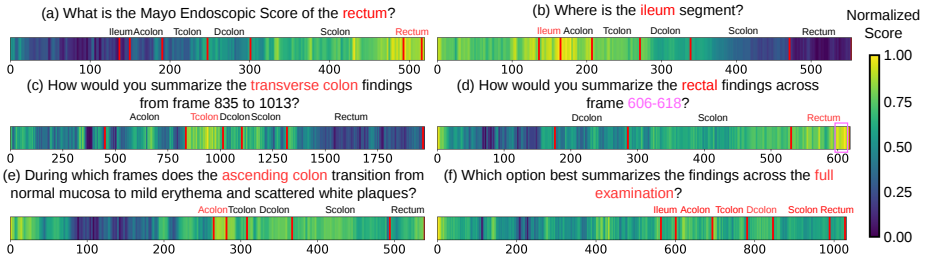

Fig. 5. Examples of predicted score distribution from the Token Scorer conditioned on the question. Red lines indicate segment boundaries.

Summarization – Local

Question: How would you summarize the findings in the descending colon?
Ground Truth: "The descending colon displays mild inflammation characterized by decreased vascularity and mild erythema."
Qwen3-VL: "The descending colon shows mild erythema and scattered erosions, consistent with mild inflammation."
BIMBA: "The descending colon demonstrates normal mucosa with no evidence of inflammation, ulceration, or bleeding."
AdaBIMBA: "The descending colon exhibits mild erythema with patchy loss of vascular patterns, indicative of early inflammatory changes."

Temporal Grounding – Event

Question: What frames show the transition of the descending colon to mild inflammation?
Ground Truth: "The descending colon transitions to mild inflammation between frame 224 - 244."
Qwen3-VL: "The descending colon transitions to mild inflammation between frame 256 and 292, with partial vascular loss and erythema noted." *(IoU=0)*
BIMBA: "The descending colon transitions to mild inflammation between frame 259 and 307, with visible erythema and vascular changes." *(IoU=0)*
AdaBIMBA: "The descending colon transitions to mild inflammation with erythema between frame 210 and 255." *(IoU=44.4)*

Fig. 6. Examples of responses from AdaBIMBA and baseline methods. Representative question-relevant frames are shown in chronological order for visualization only; all models receive the full video as input.

conditioning frame importance on the question improves relevance. Replacing the Token Scorer with random scores severely degrades Local and Segment, confirming that learned importance estimates are critical for allocating compression budgets effectively. Eliminating Stage 1 warm-up, which stabilizes the scorer before compression is enabled, yields the largest decline in Scoring accuracy,

showing that STE-based training is too noisy without the initial uncompressed optimization.

Fig. 5 visualizes examples of Scorer outputs: panels (a, b) show Scoring and Segment questions for the rectum and ileum; (c, d) show Local questions, with (c) summarizing the transverse colon and (d) describing a specific portion of the rectum; (e) shows an Event question on the ascending colon. In all cases, score peaks correspond to clinically relevant regions. Panel (f) presents a Global question, which requires information spanning the entire video and therefore produces a more uniform score distribution. Fig. 6 shows examples of responses from Qwen3-VL, BIMBA, and AdaBIMBA. AdaBIMBA produces more comprehensive and clinically aligned summaries (left) and more accurate temporal grounding (right). These examples illustrate how adaptive token allocation enables AdaBIMBA to capture subtle lesions and localize events more reliably than the baselines.

5 Conclusion

To our knowledge, this work represents the first application of vision–language models to long-form, clinically acquired endoscopy procedures. To support this, we curated a dataset of full endoscopy recordings with expert annotations and introduced AdaBIMBA, a learnable adaptive token-compression module combining a Token Scorer for frame-importance estimation with an importance-based adaptive compressor. AdaBIMBA can process substantially more frames under fixed memory constraints while preserving clinically relevant detail, showing state-of-the-art performance across all question categories. This capability may support physicians in reviewing endoscopy procedures by summarizing procedural events and providing structured assessments of disease severity and anatomic segments, thereby facilitating efficient review and consistent, data-driven interpretation. Although promising, current evaluation is limited to a single clinical cohort; future work will extend validation to additional independent clinical trial cohorts to assess generalizability, robustness, and potential clinical utility.

Disclosure of Interests. Ka-Wai Yung contributed to this work during an internship at Johnson & Johnson. All other authors were employees of Johnson & Johnson at the time this research was conducted and may own Johnson & Johnson stock or stock options.

References

1. Baehrens, D., Schroeter, T., Harmeling, S., et al.: How to explain individual classification decisions. J. Mach. Learn. Res. **11**, 1803–1831 (2010)
2. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report. ArXiv **abs/2511.21631** (2025)
3. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. ArXiv **abs/1308.3432** (2013)
4. Chaitanya, K., Damasceno, P.F., Fadnavis, S., et al.: Arges: Spatio-temporal transformer for ulcerative colitis severity assessment in endoscopy videos. In: MICCAI Workshop. pp. 201–211 (2024)
5. Chen, J., Gui, C., Ouyang, R., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. ArXiv **abs/2406.19280** (2024)
6. Chen, L., Wei, X., Li, J., et al.: Sharegpt4video: Improving video understanding and generation with better captions. NeurIPS (2024)
7. Dai, W., Li, J., Li, D., et al.: Instructblip: Towards general-purpose vision-language models with instruction tuning. NeurIPS (2023)
8. Dao, T., Fu, D., Ermon, S., et al.: Flashattention: Fast and memory-efficient exact attention with io-awareness. NeurIPS (2022)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. ArXiv **abs/2312.00752** (2024)
10. Hu, E.J., Shen, Y., Wallis, P., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
11. Huang, J., He, R., Khan, D.Z., et al.: Surgicalvlm-agent: Towards an interactive ai co-pilot for pituitary surgery. ArXiv **abs/2503.09474** (2025)
12. Islam, M.M., Nagarajan, T., Wang, H., et al.: Bimba: Selective-scan compression for long-range video question answering. In: CVPR. pp. 29096–29107 (2025)
13. Jin, J., Jeong, C.W.: Surgical-llava: Toward surgical scenario understanding via large language and vision models. ArXiv **abs/2410.09750** (2024)
14. Li, Y., Yang, X., Xu, D., et al.: Surgpub-video: A comprehensive surgical video dataset for enhanced surgical intelligence in vision-language model. ArXiv **abs/2508.10054** (2025)
15. Lin, T., Zhang, W., Li, S., et al.: HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In: ICML (2025)
16. Liu, H., Li, C., Li, Y., et al.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024)
17. Liu, H., Li, C., Wu, Q., et al.: Visual instruction tuning. NeurIPS (2023)
18. Liu, S., Zheng, B., Chen, W., et al.: Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. In: NeurIPS (2023)
19. Mobadersany, P., Parmar, C., Damasceno, P.F., et al.: Harnessing temporal information for precise frame-level predictions in endoscopy videos. In: MICCAI. pp. 295–305 (2024)
20. Nath, V., Li, W., Yang, D., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: CVPR. pp. 14788–14798 (2025)
21. OpenAI: Gpt-4o (May 2024), `https://openai.com/index/hello-gpt-4o/`
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: CVPR. pp. 4195–4205 (2023)

23. Perez, E., Strub, F., De Vries, H., et al.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
24. Qin, M., Liu, X., Liang, Z., et al.: Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. ArXiv **abs/2506.19225** (2025)
25. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
26. Ren, S., Yao, L., Li, S., et al.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR. pp. 14313–14323 (2024)
27. Schmidgall, S., Cho, J., Zakka, C., et al.: Gp-vls: A general-purpose vision language model for surgery. ArXiv **abs/2407.19305** (2024)
28. Shen, X., Xiong, Y., Zhao, C., et al.: LongVU: Spatiotemporal adaptive compression for long video-language understanding. In: ICML (2025)
29. Shu, Y., Liu, Z., Zhang, P., et al.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: CVPR. pp. 26160–26169 (2025)
30. Song, E., Chai, W., Wang, G., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR. pp. 18221–18232 (2024)
31. Tang, X., Qiu, J., Xie, L., et al.: Adaptive keyframe sampling for long video understanding. In: CVPR. pp. 29118–29128 (2025)
32. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. NeurIPS (2017)
33. Wang, G., Bai, L., Nah, W.J., et al.: Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. ArXiv **abs/2405.10948** (2024)
34. Wang, G., Wang, J., Mo, W., et al.: Surgvidlm: Towards multi-grained surgical video understanding with large language model. ArXiv **abs/2506.17873** (2025)
35. Wang, H., Nie, Y., Ye, Y., et al.: Dynamic-vlm: Simple dynamic visual token compression for videollm. In: CVPR. pp. 20812–20823 (2025)
36. Wang, W., He, Z., Hong, W., et al.: Lvbench: An extreme long video understanding benchmark. In: CVPR. pp. 22958–22967 (2025)
37. Wu, C., Chen, X., Wu, Z., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: CVPR. pp. 12966–12977 (2025)
38. Yang, L., Xu, S., Sellergren, A., et al.: Advancing multimodal medical capabilities of gemini. ArXiv **2405.03162** (2024)
39. Zeng, X., Li, K., Wang, C., et al.: Timesuite: Improving MLLMs for long video understanding via grounded tuning. In: ICLR (2025)
40. Zeng, Z., Zhuo, Z., Jia, X., et al.: Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. ArXiv **abs/2506.02555** (2025)