BEYOND AUROC: EVALUATING TEMPORAL STABILITY, FALSE-POSITIVE LOAD, AND UNCERTAINTY CALIBRATION IN CAPSULE ENDOSCOPY VIDEO AI

Anonymous Authors

Anonymous Institution

ABSTRACT

Frame-level performance metrics (e.g., AUROC, accuracy) can substantially overestimate real-world reliability for medical video AI because they ignore temporal consistency and reviewer burden. Using the Kvasir Capsule dataset, we report a complementary evaluation suite capturing: (i) temporal stability on lesion segments, (ii) false-positive workload on normal segments, and (iii) uncertainty calibration for risk-aware deployment. Although the classifier achieves near-ceiling frame-level discrimination (AUROC $= 0.999$, AUPRC $= 0.998$, F1 $= 0.978$), our temporal analysis reveals clinically relevant failure modes: severe prediction flicker on some lesion segments (worst TJI $= 322.6$ flips/1000 frames; worst TDP $= 0.19$) and concentrated false-positive burden on specific normal segments (worst VRBI $= 65.6$). Uncertainty analysis indicates well-calibrated confidence (ECE $= 0.007$) and meaningful uncertainty-error coupling (UEC $= 0.42$), supporting human-in-the-loop safety use. Our results show that near-ceiling frame-level performance can coexist with temporally localized instability and workload concentration that remain hidden under conventional evaluation metrics.

1. INTRODUCTION

Artificial intelligence (AI) systems for medical imaging are increasingly evaluated and reported using frame-level discrimination metrics such as AUROC, accuracy, and F1 scores. While these metrics are appropriate for static image classification, unfortunately they are insufficient for AI-enabled medical video devices, where predictions unfold over time and clinical interpretation depends on temporal coherence, inference stability, and workload impact.

Medical video AI systems, including capsule endoscopy and colonoscopy detection devices, operate on continuous image streams rather than isolated frames. In these settings, lesions persist across multiple frames, and clinical usability depends not only on whether a lesion is detected, but how consistently it is detected over time, how early it is flagged, and how frequently false alarms disrupt review. A model that achieves near-perfect frame-level AUROC may still exhibit temporally unstable behavior (prediction flicker) or concentrated false-positive bursts that increase reviewer fatigue. These failure modes are largely invisible to conventional performance metrics. In capsule endoscopy in particular, videos may span tens of thousands of frames, with abnormalities occupying only a small fraction of the stream. As a result, high accuracy can be achieved even when false alarms meaningfully increase cognitive load or when lesion detections are intermittent. Furthermore, uncertainty calibration plays a critical role in human-in-the-loop deployment, where clinicians must interpret confidence scores under challenging imaging conditions such as motion blur, debris, and low illumination.

In this work, we evaluate a capsule endoscopy classifier using a video-aware assessment framework that extends beyond frame-level discrimination. Specifically, we quantify (i) temporal detection stability on lesion segments, (ii) false-positive burden on normal segments as a proxy for review workload, and (iii) uncertainty calibration and uncertainty–error coupling. Using the Kvasir Capsule dataset, we demonstrate that despite near-ceiling frame-level performance (AUROC = 0.999), substantial temporal instability and workload concentration can be observed in specific segments.

Our results highlight the need for temporal reliability, burden-aware, and uncertainty-informed evaluation frameworks for medical video AI systems, aligning with emerging regulatory science priorities in AI-enabled device assessment.

2. RELATED WORK

AI systems for medical imaging are commonly evaluated using frame-level discrimination metrics such as AUROC, sensitivity, specificity, and F1-score. While these metrics are appropriate for static image classification, they do not fully characterize performance in medical video applications where predictions evolve temporally. In colonoscopy computer-aided detection (CADe), prior work has shown that false-positive alerts and their temporal persistence can substantially influence clinical usability and workflow efficiency [1]. Inconsistent definitions of alert duration and false-positive events have also been identified as barriers to fair benchmarking across CADe systems [2]. These findings highlight the need for evaluation protocols that account for temporal prediction behavior in video-based AI systems.

Deep learning approaches for capsule endoscopy have demonstrated strong diagnostic performance, often reporting

near-ceiling AUROC values for lesion detection [3]. Reviews of AI in gastrointestinal endoscopy confirm significant improvements in sensitivity and specificity across multiple lesion types [4]. However, most capsule endoscopy studies evaluate performance at the frame level and do not explicitly analyze temporal detection persistence, prediction instability, or false-positive burden across video segments. In parallel, uncertainty quantification has emerged as an important component of trustworthy medical AI, with calibration metrics such as Expected Calibration Error (ECE) and Brier score widely used to assess probabilistic reliability [5, 6].

Motivated by these gaps, we propose a video-aware evaluation framework for capsule endoscopy that jointly analyzes temporal detection stability, false-positive review burden, and uncertainty calibration. Our results show that near-ceiling frame-level performance can coexist with temporally localized instability and workload concentration that remain hidden under conventional evaluation metrics.

3. METHODOLOGY

We propose a video-aware evaluation framework for capsule endoscopy AI that extends conventional frame-level discrimination metrics to characterize temporal detection stability, false-positive review burden, and uncertainty reliability. The framework is model-agnostic and applicable to any classifier producing frame-wise abnormality probabilities on a temporally ordered video stream.

Let a capsule endoscopy video be represented as a sequence of frames

$$\mathcal{V} = \{I_t\}_{t=1}^{T},$$

with ground-truth labels $y_t \in \{0, 1\}$ indicating normal (0) or lesion-visible (1) frames. A classifier outputs a probability

$$p_t = P(y_t = 1 \mid I_t),$$

which is converted to a binary decision using threshold τ

$$\hat{y}_t = I(p_t \geq \tau). \tag{1}$$

Rather than evaluating frames independently, we group temporally adjacent frames with identical ground-truth labels into maximal segments

$$S_k = \{t_k^{(s)}, \ldots, t_k^{(e)}\},$$

such that $y_t = c_k$ for all $t \in S_k$, where $c_k \in \{0, 1\}$. Segment-based evaluation enables reliability analysis over temporally coherent clinical events rather than isolated frames.

Frame-level baseline. As a reference, we compute conventional discrimination metrics including AUROC, AUPRC, accuracy, precision, recall, specificity, and F1-score. These metrics quantify global class separability but do not capture temporal prediction behavior.

Temporal reliability. For lesion segments ($c_k = 1$), we measure prediction stability using three metrics. Prediction instability is quantified by the Temporal Jitter Index (TJI), defined as the normalized number of prediction transitions

$$\mathrm{TJI}(S_k) = \frac{1000}{|S_k| - 1} \sum_{t=t_k^{(s)}+1}^{t_k^{(e)}} |\hat{y}_t - \hat{y}_{t-1}|. \tag{2}$$

Sustained lesion coverage is measured by Temporal Detection Persistence (TDP), defined as the longest continuous run of positive detections normalized by segment length

$$\mathrm{TDP}(S_k) = \frac{L_k^{\max}}{|S_k|}, \tag{3}$$

where $L_k^{\max}$ is the longest contiguous subsequence with $\hat{y}_t = 1$. Early detection capability is captured by First Detection Latency (FDL), defined as the delay between lesion onset and the first positive prediction within the segment.

False-positive burden. For normal segments ($c_k = 0$), we quantify reviewer burden caused by false alerts. The false-positive density is

$$\mathrm{FPR}_{1000}(S_k) = \frac{1000}{|S_k|} \sum_{t \in S_k} \hat{y}_t. \tag{4}$$

Let $B_{k,j}$ denote contiguous false-positive bursts. Their average duration defines the false-positive persistence. We combine alert frequency and duration into a Video Review Burden Index

$$\mathrm{VRBI}(S_k) = \mathrm{FPR}_{1000}(S_k) \times \mathrm{FPD}(S_k), \tag{5}$$

which approximates localized reviewer workload caused by repeated false alerts.

Uncertainty reliability. To evaluate probabilistic reliability, we compute Expected Calibration Error (ECE), Brier score, and negative log-likelihood. In addition, we assess whether model uncertainty increases appropriately under failure by measuring the correlation between predictive entropy and frame-level error, referred to as uncertainty–error coupling. Together, frame-level discrimination, temporal stability, false-positive burden, and uncertainty calibration provide complementary views of video AI reliability. This multi-dimensional evaluation reveals deployment-relevant failure modes that remain hidden when assessment is restricted to conventional frame-level metrics.

4. EXPERIMENTS

4.1. Dataset and Evaluation Protocol

We evaluated the proposed framework using the publicly available *Kvasir Capsule* dataset. Frames were categorized as abnormal or normal and temporally ordered using frame indices embedded in filenames. Following the segmentation

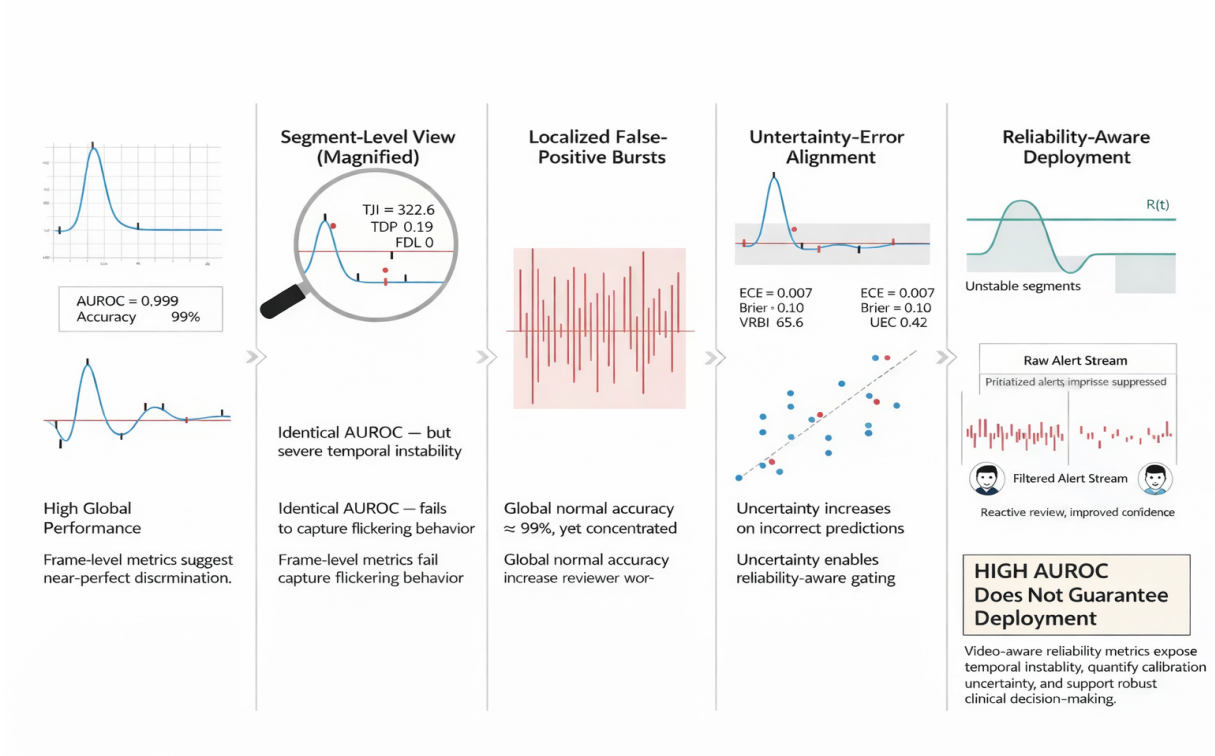

Fig. 1. Limitations of frame-level evaluation for medical video AI. (A) Near-ceiling AUROC and accuracy fail to capture temporal behavior. (B) Lesion segments with similar frame-level performance exhibit different stability (TJI). (C) Localized false-positive bursts despite high specificity (VRBI). (D) Predictive uncertainty correlates with errors. (E) Unified framework integrating stability, burden, and uncertainty.

protocol in Section 3, frames were aggregated into maximal contiguous ground-truth segments representing lesion-visible and normal intervals. To ensure stable temporal estimates, segments shorter than 30 frames were excluded. After filtering, the evaluation set contained 12 lesion segments and 35 normal segments used for temporal reliability and false-positive burden analysis. Unless otherwise stated, results are reported using mean values together with tail percentiles and worst-case statistics to capture variability across segments.

4.2. Frame-Level Baseline

We first report conventional frame-level discrimination metrics to establish a reference baseline. At the calibrated threshold ($\tau \approx 0.758$), the classifier achieves near-ceiling performance (Table 1). These results confirm strong global separability between abnormal and normal frames. However, frame-level metrics do not characterize temporal prediction behavior across video segments.

Table 1. Frame-level discrimination performance.

Metric	Value
AUROC	0.999
AUPRC	0.998
Accuracy	0.988
Balanced Accuracy	0.983
Precision	0.984
Recall	0.973
Specificity	0.994
F1-score	0.978
TP / FP / TN / FN	1255 / 21 / 3413 / 35

4.3. Temporal Reliability and False-Positive Burden

We next evaluate temporal detection stability on lesion segments and false-positive burden on normal segments using the metrics defined in Section 3. Although lesion onset is typically detected with minimal delay (low FDL), stability varies substantially across segments. While many segments exhibit consistent detection, tail statistics reveal severe prediction flicker in some cases, exceeding 300 transitions per 1000 frames. Correspondingly, worst-case detection persistence drops to 0.19, indicating intermittent lesion highlighting despite high frame-level sensitivity. Despite high global specificity, false positives are unevenly distributed over time. Certain normal segments exhibit concentrated bursts of alerts, reaching up to 65 false positives per 1000 frames. Such localized spikes

Table 2. Temporal reliability metrics on lesion segments (length $\geq$ 30 frames).

Metric	Mean	Percentile	Worst Case
TJI (per 1000 frames)	39.7	p95 = 293.1	322.6
TDP	0.912	p10 = 0.571	0.194
FDL (frames)	0.053	p95 = 0	2

Table 3. False-positive burden on normal segments (length $\geq$ 30 frames).

Metric	Mean	Percentile	Worst Case
FP per 1000 frames	4.45	p95 = 18.18	65.57
FP Duration (frames)	0.387	p95 = 1.0	1.0
VRBI	7.20	p95 = 24.13	65.57

Table 4. Uncertainty and calibration metrics.

Metric	Value
Expected Calibration Error (ECE)	0.007
Brier Score	0.010
Negative Log Likelihood	0.036
Uncertainty–Error Correlation (UEC)	0.42
Uncertainty-weighted Detection Score	0.992

remain invisible to aggregate specificity but may increase reviewer workload during video inspection. We further evaluated probabilistic reliability using calibration and uncertainty metrics in Table 4. The low calibration error indicates reliable probability estimates, while the positive uncertainty–error correlation suggests that predictive uncertainty inc

4.4. Case Study: Hidden Failure Modes

To illustrate limitations of frame-level evaluation, we compare two lesion segments with similar AUROC contributions but different temporal behavior (Fig. 2). A stable *Ileocecal valve* segment (48 frames) exhibits TJI = 0 and TDP = 1.00, indicating uninterrupted detection. In contrast, a *Reduced mucosal view* segment (31 frames) shows severe instability with TJI = 322.6 and TDP = 0.19 despite identical early detection (FDL = 0). These examples highlight that models with near-identical frame-level metrics may exhibit markedly different temporal stability and false-positive burden characteristics.

5. CONCLUSION AND FUTURE WORK

We presented a video-aware reliability evaluation framework for capsule endoscopy AI that extends conventional frame-level discrimination metrics by incorporating temporal detection stability, false-positive review burden, and uncertainty calibration. While the evaluated model achieves near-ceiling AUROC and accuracy, our analysis reveals that such aggregate metrics can mask clinically relevant behaviors, including

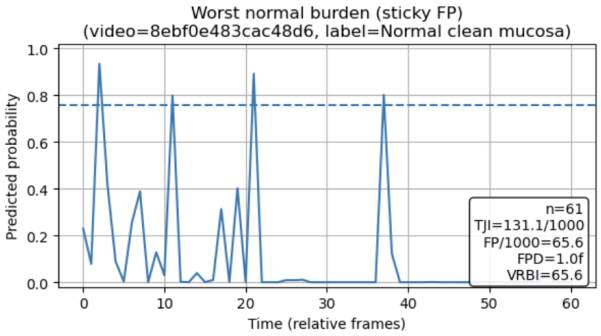

Fig. 2. Example of temporal prediction instability in a lesion segment.

temporal prediction instability and localized bursts of false-positive alerts across video segments. By explicitly quantifying these reliability dimensions, the proposed framework provides a more deployment-relevant assessment of medical video AI systems and highlights limitations that remain hidden under standard evaluation protocols. Future work will extend this analysis to larger capsule endoscopy datasets and diverse model architectures, and explore reliability-aware training strategies that directly optimize temporal stability and reduce alert burden, ultimately supporting more robust and clinically trustworthy AI-assisted capsule review.

6. REFERENCES

[1] Cesare Hassan et al., "False-positive alerts in computer-aided detection systems for colonoscopy: clinical implications," *Endoscopy International Open*, vol. 9, no. 9, pp. E1310–E1318, 2021.

[2] Erik C. Brand et al., "Standardization of false-positive alert definitions in colonoscopy computer-aided detection systems," *Gastrointestinal Endoscopy*, vol. 94, no. 1, pp. 123–131, 2021.

[3] Zhaoshen Ding et al., "Deep learning for detecting small-bowel diseases in capsule endoscopy images," *Gastroenterology*, vol. 157, no. 4, pp. 1044–1054, 2019.

[4] Dimitris K. Iakovidis et al., "Deep learning in gastrointestinal endoscopy: a systematic review," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 110–122, 2020.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.

[6] Jakob Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, pp. 1513–1589, 2023.