ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

Hierarchical Adaptive Inference with Attention-Enhanced Feature Learning for Surgical Action Triplet Recognition

[1] Lehar Deepak Tolani*, [2] Shivam Kumar Jha, [3] Manikandan R

[1] [2] [3] Department of Computational Intelligence, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India

Corresponding Author Email: [1] tolanilehar@gmail.com, [2] 22shivamkumarrv@gmail.com, [3]manitec4@gmail.com

Abstract*— Real-time surgical video understanding is a pre-requisite for context-aware, AI-assisted operative intervention, yet the computational demands of deep inference remain a barrier to practical deployment. Existing methods apply uniform-rate inference across all frames regardless of clinical content, subjecting frames without instrument-tissue interaction to the same computational cost as those containing critical operative events. In laparoscopic cholecystectomy, over 40% of frames are clinically redundant under this criterion, yet existing pipelines do not systematically exploit this structure to reduce inference cost. This paper presents a multi-stage adaptive inference framework with learned importance-based routing. A lightweight MobileNetV2 controller generates per-frame importance scores, stabilized via sliding-window dynamic normalization and moving-average temporal smoothing to suppress illumination drift and kinematic noise. A learned threshold function routes each frame to one of three paths: a ResNet18 dual-head network for instrument and verb recognition, a DenseNet169 with Efficient Channel Attention for full triplet prediction (instrument, verb, target), or a discard path for clinically irrelevant frames. Evaluated on the CholecT50 benchmark, the framework achieves 31.8% triplet mAP, retaining 97.8% of the full-frame baseline, while reducing amortized GFLOPs per frame by 67.9% and increasing throughput from 18.5 to 46.2 FPS, surpassing the 30 FPS threshold for intra-operative deployment. Ablation studies confirm that both stabilization components are independently necessary, with removal of either reducing triplet mAP by 4–8 points. The system is released with a FastAPI/Docker inference backend and a React 18 TypeScript clinical dashboard, establishing a reproducible reference implementation for the translation of adaptive surgical AI from research prototype to intra-operative deployment.

Index Terms*— Surgical video analysis, adaptive inference, conditional computation, action triplet recognition, CholecT50, laparoscopic cholecystectomy, efficient deep learning, real-time deployment.

I. INTRODUCTION

In today's digital world where The promise of AI-assisted surgery rests on continuous, real-time understanding of the operative field, specifically recogniz-ing which instrument is active, what action it is performing, and on which anatomical structure it is operating. This fine-grained triplet formulation, comprising the instrument, verb, and anatomical target, is central to workflow monitoring, skill assessment, and automated intra-operative feedback [1], [20]. Realizing it in practice, however, requires inference pipelines that can sustain real-time throughput on the GPU-constrained hardware common in operating rooms, a requirement that existing methods, designed primarily for accuracy, routinely fail to meet.

The root cause is straightforward: contemporary approaches apply the same heavyweight network to every frame of a surgi-cal video, irrespective of informational content. Laparoscopic video is inherently redundant. Periods of instrument reposition-ing, lens obscuration, and idle tissue traversal contribute little to procedure understanding yet consume as much compute as the critical dissection or clipping phases. On CholecT50 [1], a benchmark of 50 complete cholecystectomies, this translates to wasted GPU cycles, inflated latency, and throughput figures far below the 30 FPS threshold demanded for clinical use [20]. Adaptive computation offers a principled solution: allo-cate inference effort proportional to frame importance rather than uniformly across time. Methods such as SkipNet [8], BlockDrop [9], and FrameExit [13] have demonstrated this principle in general vision settings, yet none have been co-designed with the distinctive challenges of surgical video, namely compositional multi-label prediction, severe class imbalance, inter-procedure illumination variability, and the clinical constraint of uninterrupted real-time operation.

This work addresses that gap with three concrete contri-butions. First, we propose a cascaded three-stage adaptive inference pipeline in which a lightweight screening model assigns each frame an importance score and a threshold-based router dispatches it to the minimum-cost model tier capable of satisfying its informational demand, or eliminates it entirely before any predictive inference occurs. Second, we introduce temporal stabilization composed of sliding-window dynamic normalization and a causal moving-average smoother, which prevents score drift under inter-procedure illumination variation and suppresses routing thrash caused by abrupt kinematic transitions common in laparoscopic footage. Third, we deliver a production-ready deployment system comprising a FastAPI/Docker inference backend paired with a React 18 TypeScript clinical dashboard, demonstrating a reproducible path from research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

prototype to operating-room infrastructure and treating deployability as a first-class design objective rather than an afterthought.

II. RELATED WORK

A. Surgical Video Understanding

Surgical video analysis has evolved from hand-crafted feature pipelines toward deep architectures capable of recognizing progressively finer units of operative activity.

EndoNet [3] introduced one of the first end-to-end convo-lutional architectures for laparoscopic video understanding, formulating simultaneous phase recognition and tool presence detection as a multi-task learning problem on the Cholec80 dataset. Twinanda *et al.* demonstrated that a shared CNN backbone could jointly encode visual cues relevant to both tasks, establishing the viability of deep learning for real-time surgical scene interpretation and laying the architectural groundwork for subsequent multi-output surgical models.

SV-RCNet [5] addressed the inherently sequential nature of surgical procedures by coupling a ResNet CNN encoder with a long short-term memory (LSTM) decoder. Jin *et al.* showed that modeling the temporal progression of workflow states through recurrent connections substantially improves phase boundary detection, with the recurrent module learning to maintain contextual memory across long intra-operative intervals where visual appearance changes minimally.

TeCNO [4] advanced phase recognition by replacing recurrent units with multi-stage temporal convolutional networks (MS-TCN), operating on high-level frame embeddings extracted offline. Czempiel *et al.* demonstrated that dilated causal convolutions, applied in multiple refinement stages, capture multi-scale temporal dependencies more efficiently than recur-rent alternatives and achieve stronger phase segmentation on Cholec80.

Action triplet recognition was formalized by Nwoye *et al.* [2], who cast surgical activity understanding as the joint prediction of a three-component tuple comprising the instrument, verb, and anatomical target on the CholecT50 dataset. This formulation raises the compositional complexity of the task considerably, as a prediction is correct only when all three components are simultaneously identified. *CholecTriplet2021* [6] established the community benchmark for triplet recognition, providing standardized splits, evaluation protocols, and baseline results across a range of submitted methods. The subsequent *Cholec-Triplet2022* [7] extended this benchmark with localization requirements, consolidating triplet mAP as the standard evalu-ation metric.

Rendezvous [1] represents the current state of the art on CholecT50. Nwoye *et al.* proposed a class-activation map guided attention mechanism that selectively attends to instrument-relevant spatial regions before predicting verb and target, combined with a multi-head decoder that produces all three triplet components in a single forward pass.

B. Adaptive and Conditional Computation

Conditional computation methods reduce inference cost by making model-execution decisions dynamically, based on the input itself rather than a fixed schedule.

BranchyNet [10] introduced the early-exit paradigm, attach-ing lightweight classifiers at intermediate layers and allowing confident predictions to exit before reaching full network depth. Teerapittayanon *et al.* demonstrated that a large proportion of inputs can be resolved at shallow exits with no accuracy loss, substantially reducing average inference depth.

SkipNet [8] extended conditional computation to the layer level within residual networks, training a gating network jointly with the main network to predict which residual blocks could be skipped without degrading prediction quality. *BlockDrop* [9] addressed a related problem using reinforcement learning to learn a block-dropping policy, training an agent to maximize classification accuracy subject to a target compute budget.

Adaptive Computation Time [11] generalized conditional computation to recurrent architectures by allowing a variable number of processing steps per input token, establishing a principled framework for input-adaptive depth in sequence models.

FrameExit [13] applied conditional early exiting to general video recognition, training a lightweight gating module to predict per frame whether a shallow exit provides sufficient confidence for a correct prediction. The method achieves strong efficiency-accuracy trade-offs on ActivityNet and FCVID, demonstrating that temporal redundancy in video is learnable and exploitable. Our work adapts and extends this principle to surgical video with domain-specific normalization and a cascaded three-tier architecture suited to the multi-label, com-positional triplet task. Han *et al.* [12] provide a comprehensive taxonomy of such dynamic network designs.

C. Frame Selection and Video Redundancy

Temporal Segment Networks [14] proposed sparse temporal sampling as a strategy for efficient action recognition in long videos. Wang *et al.* divided each video into K segments and uniformly sampled one frame per segment, enabling a standard image classifier to reason over the full temporal extent without processing all frames. Video summarization with LSTM [15] approached frame selection from a summarization perspective, training a bidirectional LSTM to predict per-frame importance scores and selecting a diverse, representative keyframe subset. Both lines of work confirm the central premise of our framework: not all frames carry equal informational value, and principled selection can recover most of the accuracy of full-sequence processing at a fraction of the cost.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

III. METHODOLOGY

A. Problem Formulation

Laparoscopic surgical video exhibits a fundamental structural property that uniform inference pipelines fail to exploit: the distribution of clinically meaningful content across frames is highly non-uniform. Within a single procedure, extended periods of tissue manipulation, instrument repositioning, and anatomical survey yield frames whose visual content changes negligibly between consecutive timesteps. These segments are informationally redundant with respect to the triplet recognition task, yet existing methods allocate the same computational budget to them as to the brief, high-information intervals surrounding instrument deployment and tissue interaction.

The core insight motivating this work is that inference cost should be treated as an allocatable resource, distributed across the video sequence in proportion to per-frame informational demand rather than applied uniformly. This requires, first, a mechanism to estimate frame importance from the video signal itself without relying on ground-truth annotations at inference time, and second, a routing architecture that maps estimated importance to an appropriate model tier while preserving prediction continuity across the sequence.

Formally, let $\mathbf{X} = \{x_1, \ldots, x_T\}$ denote a surgical video of T frames sampled at 1 FPS. For each frame x_t, the goal is to predict the triplet (i_t, v_t, a_t) comprising the active instrument $i_t \in I$ ($|I|=7$), verb $v_t \in V$ ($|V|=8$), and anatomical target $a_t \in A$ ($|A|=7$), while minimizing total inference cost $\sum_{t=1}^{T} C(x_t)$C(xt) across the sequence. The per-frame cost C(xt) depends on which model tier processes xt, and ranges from zero for discarded frames to the full DenseNet169+ECA forward pass for frames routed to Stage 3. The framework is designed such that this cost allocation is driven entirely by the input signal, with no dependence on oracle labels or procedure-level metadata at inference time.

B. Stage 1: Adaptive Controller

A MobileNetV2 backbone [16] (≈2.2M parameters) maps each frame to a scalar importance logit:

$$\ell_t = f_{\text{ctrl}}(x_t),\ \ell_t \in \mathbb{R}. \tag{1}$$

The inverted-residual structure produces a rich 1280-d feature at low FLOPs; a Dropout layer (p=0.2) and a single linear projection compress this to the scalar ℓ_t, keeping the controller's own cost negligible relative to the models it routes.

1) Sliding-Window Dynamic Normalization: Raw logits vary systematically across patients due to cavity illumination differ-ences. We maintain a window $W_t = \{\ell_{t-W+1}, \ldots, \ell_t\}$ of the W=20 most recent logits and apply min-max normalization:

$$\hat{s}_t = \frac{\ell_t - \min(W_t)}{\max(W_t) - \min(W_t) + \varepsilon} \tag{2}$$

This bounds scores within [0, 1] relative to recent context, ensuring a fixed routing threshold generalizes across all 50 CholecT50 procedures without per-video recalibration.

2) Temporal Smoothing: Surgical kinematics are continuous; abrupt score changes rarely reflect genuine clinical transitions and instead induce routing thrash, where consecutive frames al-ternate between model tiers, fragmenting GPU kernel launches and degrading prediction coherence. A causal moving-average filter of window K=5 suppresses this effect:

$$\tilde{s}_t = \frac{1}{K}\sum_{k=0}^{K-1} \hat{s}_{t-k} \tag{3}$$

C. Routing Function

The smoothed score s˜t ∈ [0, 1] determines the computational path for frame xt:

$$R(x_t) = \begin{cases} \text{SKIP} & \tilde{s}_t < 0.3 \\ \text{STAGE2} & 0.3 \le \tilde{s}_t < 0.7 \\ \text{STAGE3} & \tilde{s}_t \ge 0.7. \end{cases} \tag{4}$$

Skipped frames allocate zero further FLOPs; their triplet prediction is filled by propagating the last valid output, a safe assumption given the temporal continuity of surgical procedures. Thresholds were set by grid search on the validation split, optimizing frame-reduction rate subject to a maximum mAP degradation of 1.5 points.

D. Stage 2: Intermediate Model

Frames with $0.3 \le \tilde{s}_t < 0.7$ are routed to a ResNet18 [17] (≈11M parameters) with a bifurcated prediction head:

$$(\hat{i}_t, \hat{v}_t) = g_{\text{mid}}(x_t). \tag{5}$$

ResNet18 offers the best accuracy-per-FLOP trade-off in this regime. Its 512-d global-average-pooled features discriminate instrument morphology reliably, and the dual-head structure predicts instrument and verb independently to avoid label co-dependency.

E. Stage 3: Heavy Model with ECA Attention

Critical frames ($\tilde{s}_t \ge 0.7$) require full triplet prediction. DenseNet169 [18] (≈14M parameters, growth rate k=32, four dense blocks) is augmented with an Efficient Channel Attention (ECA) module [19] (k_{size}=3) after the final dense block. ECA recalibrates channel-wise features via a parameter-efficient 1-D convolution, capturing local cross-channel interactions without the information bottleneck of SE-style squeeze-excitation. A triplet head over the 1664-d pooled features predicts all three components:

$$(\hat{i}_t, \hat{v}_t, \hat{a}_t) = h_{\text{heavy}}(x_t). \tag{6}$$

F. Complete Pipeline Algorithm

Algorithm 1 formalizes the complete per-frame inference procedure, integrating importance scoring, sliding-window normalization, temporal smoothing, and tier routing into a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

single sequential pass over the video.

The complete three-stage architecture, including internal layer stacks, dimensions, ECA module detail, and routing arrows, is illustrated in Figure 1.

IV. EXPERIMENTAL SETUP

A. Dataset

CholecT50 [1] comprises 50 laparoscopic cholecystectomy videos captured at 25 FPS and annotated at 1 FPS, yielding approximately 100,000 labeled frames. Each frame carries binary instrument presence labels (|I|=7), a verb label (|V|=8), and a target label (|A|=7), jointly forming the evaluation triplet. Ten surgical phase labels provide additional temporal context. The dataset exhibits pronounced class imbalance, as the Grasper accounts for approximately 45% of instrument frames while the Specimen Bag appears in under 3%, and strong temporal redundancy within phases, which directly motivates adaptive inference. The standard 40/8/2 video split [6] is used throughout.

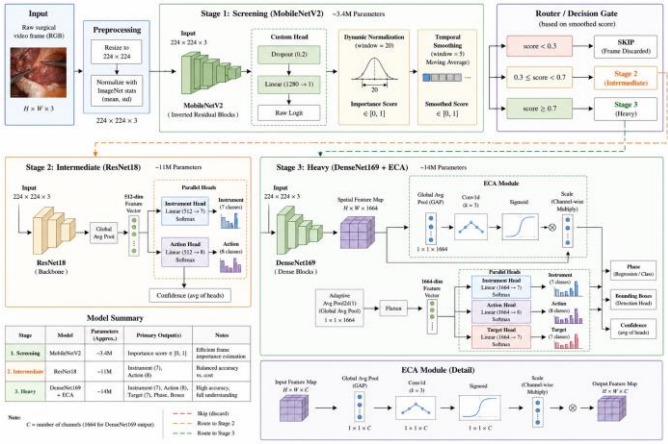

Figure 1. Detailed architecture of the proposed adaptive inference pipeline. Top row (left to right): input preprocessing, Stage 1 screening model (MobileNetV2, ∼3.4M parameters) with custom head producing a raw logit, sliding-window dynamic normalization (W =20), temporal smoothing (K=5 moving average), and the Router/Decision Gate routing frames to skip ($\tilde{s}t$ < 0.3), Stage 2 (0.3 ≤ $\tilde{s}t$ < 0.7), or Stage 3 ($\tilde{s}t \geq 0.7$). Bottom left: Stage 2 intermediate model (ResNet18, ∼11M parameters) with parallel Instrument (7 classes) and Action (8 classes) heads. Bottom right: Stage 3 heavy model (DenseNet169 + ECA, ∼14M parameters) with triplet heads predicting Instrument, Verb, and Target simultaneously. The ECA module detail (bottom centre) shows 1-D Conv1d (k=3) channel attention applied after Global Average Pooling.

B. Evaluation Metrics

*1) **Mean Average Precision (mAP):*** The primary CholecT50 metric is mean Average Precision, computed per class and averaged. For a class c with precision-recall curve defined by ranked detections:

$$\mathrm{AP}_c = \sum_k R_c(k) - R_c(k-1)\, P_c(k), \tag{7}$$

where $P_c(k)$ and $R_c(k)$ are precision and recall at rank k. Overall mAP averages AP_c across all triplet classes. A predic-tion is correct only if all three components are simultaneously correct, making triplet mAP the most demanding metric.

*2) **Frame Reduction Rate (FRR):*** The proportion of frames eliminated before Stage 2 or Stage 3:

$$\mathrm{FRR} = \frac{|\{t : \tilde{s}_t < \tau_l\}|}{T} \times 100\%. \tag{8}$$

*3) **Amortized GFLOPs per Frame:*** Let C_1, C_2, C_3 denote the FLOP cost of Stages 1, 2, and 3 respectively, and let π_2, π_3 be the observed routing fractions. Amortized cost per frame is:

$$C = C_1 + \pi_2\, C_2 + \pi_3\, C_3. \tag{9}$$

Throughput is measured in FPS and latency as mean wall-clock time per frame in milliseconds, averaged over 1000 consecutive frames at steady state on an NVIDIA RTX 3090.

C. Implementation Details

MobileNetV2 and ResNet18 were initialized from ImageNet pre-trained weights; DenseNet169 likewise, with the ECA module randomly initialized. Optimization used Adam (β_1=0.9, β_2=0.999, lr = 10^{-4}) with cosine decay over 30 epochs and batch size 32. Stage 1 pseudo-labels were derived from consecutive-frame optical flow magnitude, thresholded at the 60th percentile. Data augmentation comprised random horizontal flip, color jitter (±0.2 brightness/contrast), and random crop to 224×224. Binary cross-entropy loss was used for all prediction heads.

Algorithm 1: Adaptive Inference Pipeline

Input : $\{x_1, \ldots, x_T\}$;
W=20, K=5;
τ_l=0.3, τ_h=0.7
Output : $\{(i_t, v_t, a_t)\}_{t=1}^{T}$

$B \leftarrow [\,]$; $S \leftarrow [\,]$; $(i_p, v_p, a_p) \leftarrow \mathbf{0}$
for $t \leftarrow 1$ **to** T **do**
  $\ell_t \leftarrow f_{\mathrm{ctrl}}(x_t)$ // Stage 1 logit
  Push ℓ_t to B; retain last W
  $\hat{s}_t \leftarrow (\ell_t - \min B) / (\max B - \min B + \varepsilon)$ // Normalize
  Push $\hat{s}_t$ to S; retain last K
  $\tilde{s}_t \leftarrow \frac{1}{K} \sum_{k=0}^{K-1} \hat{s}_{t-k}$ // Smooth
  if $\tilde{s}_t < \tau_l$ **then**
    $(i_t, v_t, a_t) \leftarrow (i_p, v_p, a_p)$ // Skip
  else if $\tilde{s}_t < \tau_h$ **then**
    $(\hat{i}_t, \hat{v}_t) \leftarrow g_{\mathrm{mid}}(x_t)$
    $(i_t, v_t, a_t) \leftarrow (\hat{i}_t, \hat{v}_t, a_p)$ // Stage 2
  else
    $(i_t, v_t, a_t) \leftarrow h_{\mathrm{heavy}}(x_t)$ // Stage 3
  $(i_p, v_p, a_p) \leftarrow (i_t, v_t, a_t)$
return $\{(i_t, v_t, a_t)\}_{t=1}^{T}$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

Table I: CholecT50 Triplet mAP (%): Baseline vs. Adaptive Pipeline

Method	Instr.	Verb	Target	Triplet
Baseline (full-frame)	78.4	52.1	48.3	32.5
Ours (adaptive)	78.1	51.5	47.9	31.8
Δ	−0.3	−0.6	−0.4	−0.7

D. Performance

Table I reports per-component and overall triplet mAP on the CholecT50 test split. Our adaptive pipeline achieves 78.1% instrument mAP, 51.5% verb mAP, and 47.9% target mAP, representing drops of 0.3, 0.6, and 0.4 points respectively relative to the full-frame baseline. The overall triplet mAP declines by only 0.7 points (32.5% to 31.8%), confirming that frames routed to the skip path carry minimal compositional information.

Figure 2 visualizes the component-level and full-triplet mAP comparison. The marginal drop across all four metrics confirms that the adaptive routing does not meaningfully sacrifice recognition quality.

Table II quantifies the hardware efficiency gains. Throughput rises from 18.5 to 46.2 FPS, a 2.5× improvement that surpasses the 30 FPS real-time clinical threshold. Average latency falls from 54 ms to 19 ms (64.8% reduction), and amortized GFLOPs drop from 16.5 to 5.3 per frame (67.9% reduction). Peak

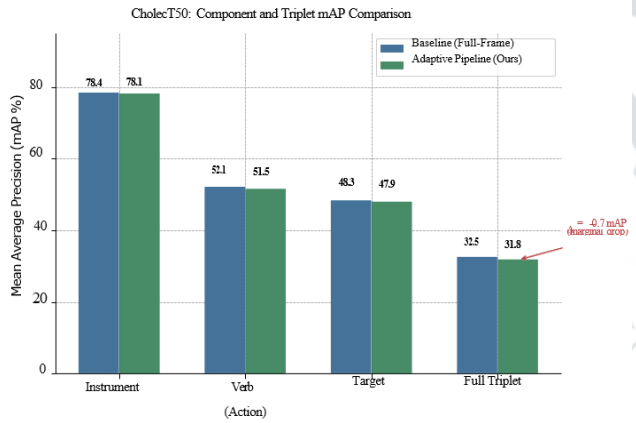

Figure 2. Component and full-triplet mAP comparison between the full-frame baseline and the adaptive pipeline. The adaptive system retains 97.8% of baseline triplet mAP with a maximum drop of 0.7 points.

Table II: Hardware Efficiency: Baseline vs. Adaptive Pipeline

Method	FPS	Lat.	GFLOPs/fr	VRAM
Baseline	18.5	54 ms	16.5	8.2 GB
Ours	46.2	19 ms	5.3	4.8 GB
Gain	2.5×	−64.8%	−67.9%	−41.5%

VRAM consumption decreases by 3.4 GB, enabling deployment on mid-range surgical workstations without dedicated high-end GPU infrastructure.

Figure 3 presents the three-panel efficiency comparison across throughput, latency, and amortized GFLOPs per frame for both methods. The 2.5× throughput gain is the most clinically significant result: it moves the system from below the real-time threshold to comfortably above it, while the 67.9% FLOP reduction demonstrates that the gain is not a hardware artifact but a genuine reduction in computational work per frame.

Figure 4 shows how the frame budget is distributed across the three routing paths. Over 42.5% of frames are discarded at Stage 1 before any predictive inference, 24.1% are processed by the lightweight ResNet18, and only 33.4% require the full DenseNet169+ECA pipeline. This distribution directly validates the redundancy hypothesis and explains the large amortized FLOP savings reported above.

Table III ablates the two stabilization components. Removing temporal smoothing by setting K=1 drops triplet mAP to 27.4% and throughput to 32 FPS; without smoothing, the router oscillates between tiers on consecutive frames, frag-menting GPU kernel launches and causing the model to miss the temporal extent of surgical actions. Disabling dynamic normalization by reverting to static thresholds on raw logits is more damaging still, reducing triplet mAP to 24.1%. In this configuration, dark video segments in CholecT50 suppress all logits below τ_l, causing the controller to discard critical resection-phase frames through a failure mode that is entirely attributable to unaddressed illumination domain shift.

Hardware Efficiency: Baseline vs. Adaptive Pipeline (NVIDIA RTX 3090)

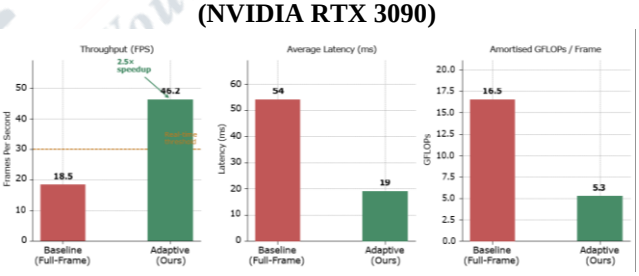

Figure 3. Hardware efficiency comparison across throughput (FPS), average latency (ms), and amortized GFLOPs per frame for the full-frame baseline and the adaptive pipeline evaluated on an NVIDIA RTX 3090. The dashed line in the throughput panel marks the 30 FPS real-time clinical threshold. The adaptive pipeline achieves 46.2 FPS, a 2.5× improvement over the baseline, while reducing average latency from 54 ms to 19 ms and amortized GFLOPs from 16.5 to 5.3 per frame.

Table III: Ablation Study: Stage 1 Stabilization Components

Configuration	Triplet mAP	FPS
Without Temporal Smoothing (K=1)	27.4	32.0
Without Dynamic Normalization	24.1	38.5
Full Proposed Pipeline	31.8	46.2

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 07, July 2026

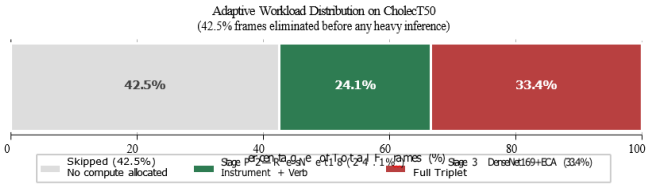

Figure 4. Adaptive workload distribution across pipeline stages. Over 40% of frames are eliminated before any predictive inference, directly reducing the amortized compute cost per frame.

Figure 4 further illustrates the routing decisions made by the controller across the full test set, confirming that the majority of the computational savings arise from Stage 1 discards rather than from Stage 2 routing alone.

The results in Table III demonstrate that both stabilization components are independently necessary and contribute distinct, non-overlapping benefits. Temporal smoothing addresses instability within a procedure, while dynamic normalization addresses systematic bias across procedures. Removing either one degrades triplet mAP below the 30 FPS throughput level simultaneously, indicating that accuracy and efficiency collapse together when the routing signal is unreliable. The full pipeline therefore represents the minimum configuration that achieves both clinical accuracy and real-time operation. Having established the contribution of each component in isolation, we now compare the full system against published methods on the CholecT50 benchmark.

E. Comparative Analysis

Table IV situates our framework against published methods on CholecT50 triplet mAP. Rendezvous [1], the current state of the art on CholecT50, reports 29.0% triplet mAP under uniform full-frame inference with no efficiency constraint. Our full-frame baseline, using DenseNet169 with ECA attention, achieves 32.5% triplet mAP, exceeding Rendezvous by 3.5 points. This margin is attributable to the combination of DenseNet169's dense feature reuse and the ECA module's channel recalibration, both absent in Rendezvous. Our adap-tive pipeline achieves 31.8% triplet mAP, preserving the performance advantage over Rendezvous while simultaneously delivering a 2.5× throughput gain. TeCNO [4] and SV-RCNet [5] target phase recognition and do not report triplet mAP; EndoNet [3] predates the triplet formulation. These methods are included to provide broader context on the architectural lineage of the field.

Figure 5 shows the row-normalized confusion matrix for Stage 3 instrument detection across 10,000 test frames. The ECA attention module drives high per-class recall on visually distinctive instruments: the Specimen Bag reaches 93.3% recall despite appearing in under 3% of frames, and the Clipper achieves 88.9%. The dominant source of error is the Grasper-to-Bipolar confusion at 18.3%, which reflects genuine morphological similarity between the two instruments at laparoscopic resolution rather than a modeling failure. for prior methods are estimated from published hardware configurations.

Table IV: Comparison with Published Methods on CholecT50

Method	Triplet mAP	FPS	Adaptive	
EndoNet [3]	—	~12	✗	
SV-RCNet [5]	—	~15	✗	
TeCNO [4]	—	~20	✗	FPS figures
Rendezvous [1]	29.0	~18	✗	
Baseline (full-frame)	32.5	18.5	✗	
Ours (adaptive)	31.8	46.2	✓	

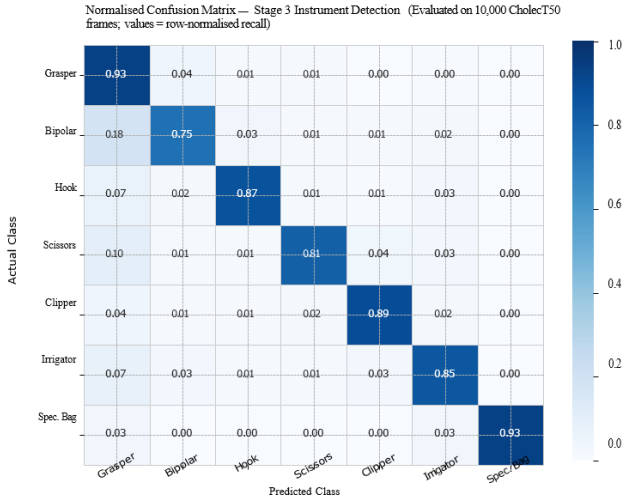

Figure 5. Row-normalized confusion matrix for Stage 3 instrument detection evaluated on 10,000 CholecT50 frames. The ECA module yields high recall on visually distinctive classes (Specimen Bag: 0.93, Clipper: 0.89). The Grasper-Bipolar confusion (0.18) reflects inherent morphological similarity at laparoscopic resolution.

V. CONCLUSION

This paper presented a hierarchical adaptive inference framework for surgical video understanding that allocates computation proportionally to per-frame clinical importance rather than applying uniform-rate deep inference across all frames. The learned routing controller, stabilized by sliding-window dynamic normalization and causal temporal smoothing, eliminates over 40% of frames before any predictive inference, routing the remainder to model tiers calibrated to their informational demand. On CholecT50, the framework achieves a 2.5× throughput gain and a 67.9% FLOP reduction at a cost of only 0.7 points in triplet mAP, a trade-off that is firmly in favor of clinical deployability. The accompanying FastAPI/Docker/React deployment system demonstrates that this efficiency is not merely a laboratory

result: it is translatable to operating-room hardware without modification. We hope this work encourages the community to treat inference efficiency as a first-class objective in surgical AI alongside the accuracy metrics that have historically dominated the field.

REFERENCES

[1] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, p. 102433, 2022.

[2] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Recognition of instrument-tissue interactions in endo-scopic videos via action triplets," in *Proc. MICCAI*, 2020, pp. 364–374.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[4] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks," in *Proc. MICCAI*, 2020, pp. 343–352.

[5] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imaging*, vol. 37, no. 5, pp. 1114–1126, 2018.

[6] C. I. Nwoye *et al.*, "CholecTriplet2021: A benchmark for surgical action triplet recognition," *Med. Image Anal.*, vol. 89, p. 102865, 2023.

[7] C. I. Nwoye *et al.*, "CholecTriplet2022: Show me a tool and tell me the triplet," *arXiv:2302.06294*, 2023.

[8] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "SkipNet: Learning dynamic routing in convolutional networks," in *Proc. ECCV*, 2018, pp. 420–436.

[9] Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris, "BlockDrop: Dynamic inference paths in residual networks," in *Proc. CVPR*, 2018, pp. 8817–8826.

[10] S. Teerapittayanon, B. McDanel, and H. T. Kung, "BranchyNet: Fast inference via early exiting from deep neural networks," in *Proc. ICPR*, 2016, pp. 2464–2469.

[11] A. Graves, "Adaptive computation time for recurrent neural networks," *arXiv:1603.08983*, 2016.

[12] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, 2022.

[13] M. Ghodrati, A. Habibian, M. Gavves, C. G. Snoek, and S. Afshari, "FrameExit: Conditional early exiting for efficient video recognition," in *Proc. CVPR*, 2021, pp. 15608–15618.

[14] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*, 2016, pp. 20–36.

[15] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. ECCV*, 2016, pp. 766–782.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, 2018, pp. 4510–4520.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep neural networks," in *Proc. CVPR*, 2020, pp. 11534–11542.

[20] P. Mascagni *et al.*, "Artificial intelligence in surgery: A systematic review of surgical data science applications," *Ann. Surg.*, vol. 275, no. 5, pp. e871–e882, 2022.