Data Engineering and Applications

Research Paper

FNO-LSTM: Context-Aware Feature Fusion for Edge Camera Event Summarization with Reliability Gating

Horváth János[1,*] and Szilárd Miklós[1]

[1] Faculty of Technical Informatics, University of Pannonia, Veszprém, 8200, Hungary
*Corresponding author: horvath.j@mik.uni-pannon.hu

Abstract. Edge cameras can find many local incidents, but sending full streams or disconnected alarm clips over constrained links overloads these links and gives operators redundant evidence. This paper proposes a context-aware event summarization model that combines Fourier Neural Operators and Long Short-Term Memory networks. The operator learns the cross-window temporal dynamics in a compressed spectral space, and the recurrent branch maintains the order and duration of events. Scene state, object interaction, motion intensity, illumination and device load are fused by a reliability-gated context module prior to diversity-aware key-clip selection. A four-camera edge testbed was evaluated using 18,420 annotated events from traffic, campus, warehouse and perimeter scenes. At a 12-clip budget, the proposed model reached 0.842 F1, 0.781 mean average precision and 0.736 temporal intersection over union, and outperformed the best temporal convolution baseline by 4.7, 4.3 and 5.1 percentage points, respectively. Reduced the transmitted video volume by 91.6%, maintained 27.8 frames per second at a 15 W edge module, and kept the 95th-percentile summary delay at 428 ms. Ablation experiments show that spectral modelling and reliability-gated context fusion are complementary; they perform well under light-variation and partial-occlusion conditions. Based on the above results, operator-recurrent fusion can generate compact, ordered and operationally feasible camera summaries without continuous cloud inference.

Keywords: *Fourier Neural Operator, Context-Aware Fusion, Temporal Key-Clip Selection, Embedded Video Analytics*

Received on 14 February 2023, Accepted on 26 May 2023, Published on 07 June 2023

Introduction

Edge cameras are no longer simple collection devices; they are now local sensing nodes that can detect vehicles, people, equipment status and safety accidents. The backhaul requirement is relatively small; however, a problem has arisen: many cameras output large numbers of overlapping detection data in a short time. Traditional clip retention retains the evidence but requires the operator to view several angles; alert-only transmission lacks the time sequence needed to determine the cause. Therefore, high-efficiency video understanding at the edge requires a summary that indicates what happened, when it changed, and which contextual factors influenced the confidence level. Compact neural video models have improved embedded perception [1]. Temporal action localization also provides some assistance in delineating event boundaries [2]. Research on video summarization emphasizes representativeness and diversity [3]. Research on edge intelligence indicates that latency, energy consumption, and communication must be optimized simultaneously, rather than being handled independently during deployment [4].

Most existing summarizers rank frames or short segments based on appearance and motion features, then add a diversity constraint. This formulation is suitable for edited consumer videos, but continuous edge streams have long idle periods, abrupt illumination changes, camera vibrations, repeated object tracks, and hardware-dependent frame loss. A high motion score may be due to rain or a moving shadow and not an actionable event. Recurrent networks maintain order but lose long-range correlations, and temporal convolutions have a fixed effective field. Transformer models can handle long-range dependencies, but they typically require a large

https://doi.org/10.64972/dea.2023.v2i2.3597d:90-104

amount of memory and introduce latency, making them difficult to use in low-power modules [5]. Neural operators are alternatives that learn transformations of functions and can express global temporal structure through spectral modes [6]. Their application in event summarization is not widely known, and their use should change feature fusion. Multi-modal gating has been employed to suppress noisy inputs [7], resource-aware inference has improved the stability of edge services [8], but these ideas have not been combined in a single operator-recurrent summary pipeline.

In this paper, we develop a context-aware FNO-LSTM model for event summarization on edge cameras. The Fourier Neural Operator branch learns long-term variations over a fixed-length window, and a long short-term memory branch retains local chronology and event continuity. A reliability gate consists of visual, motion, interaction, environmental, and device state features, while a budget selector tends to favor coverage, boundary fidelity, and semantic diversity. Add a causal buffer to the model and implement quantized inference and a load-aware fallback schedule. Investigate whether spectral and recurrent representations are complementary, if context gating is still effective in the presence of corruption, and whether the resulting accuracy meets the requirements for real-time edge deployment. Evaluation has four scene families, five baselines, controlled ablations, perturbation tests, and measurements of throughput, power, memory, transmitted volume, and end-to-end delay.

Related Work

Edge Video Summarization and Event Selection

Video summarization methods are divided into three categories: event-driven abstraction, keyframe selection, and key shot selection. Unsupervised learning is conducted through methods such as adversarial target optimization and diversity; supervised learning uses labeled data. Recently, some researchers have improved temporal consistency by combining convolutional features with recurrent scoring [9]. Attention mechanisms can connect distant parts but may select visually prominent but operationally redundant locations [10]. Event cameras in fixed locations also present a problem: the background is relatively regular, and most of the value lies in rare transitions. Therefore, a suitable summary should keep the start time, peak interaction time and end time of the event rather than selecting a representative frame.

Edge deployment has changed the evaluation objectives. Cloud-oriented summaries often overlap with manual selection, but embedded systems also need to meet latency budgets, memory constraints, thermal limits, and communication goals. Collaborative Edge Systems can distribute analytics across camera and gateway resources [11], but due to unreliable links, each camera must retain a locally meaningful fallback. Based on the above observations, an architecture with a large temporal field and a bounded online state has been proposed.

Temporal Modeling and Context-Aware Fusion

Long Short-Term Memory networks are still suitable for streaming event models due to their compact, causal hidden state. Temporal convolutions are suitable for stable parallel computation, and transformers have flexible global attention. Each choice has a certain deployment trade-off. Neural operators learn mappings on discretized functions, and it has been proven that truncated Fourier modes can effectively represent non-local dynamics [12]. Fourier Neural Operator performs multiple spectral convolutions and pointwise mixing to spread information across the entire window without using a deep stack of small kernels [13]. Operator studies have focused on physical systems and not on compressed surveillance events; issues of causality, window overlap and edge quantization have not been addressed.

Context-aware fusion adds auxiliary variables only when they improve the current decision. Scene labels, illumination estimates, object density, track consistency and device health all have different noise patterns. Fixed concatenation can cause the corrupted context to dominate the representation. Reliability Gating and Uncertainty Calibration provide adaptive weighting mechanisms [14]. Resource-aware scheduling can reduce the resolution or skip spectral updates during a transient overload [15]. The remaining defect is a unified model that considers contextual reliability, long-term temporal dynamics, local event sequences, and summary budget optimization as interrelated components of the edge camera pipeline.

Context-Aware Event Representation

Problem Formulation and Processing Pipeline

A camera stream is divided into overlapping windows of $T = 64$ feature steps, with each step representing eight decoded frames. A light-weight detector and tracker output appearance embeddings, class probabilities, bounding-box geometry, track velocity and interaction cues. The environmental context includes standardized lighting, background motion entropy, weather or indoor state indicators, and scene identifiers. Device context records decoder-queue occupancy, recent frame-drop rate, temperature and free accelerator memory. The goal is to learn a sequence of event-importance scores and boundary probabilities that select no more than B clips.

Figure 1 is the processing flow of causal acquisition and context estimation, dual-branch temporal inference, clip selection, and local packaging. The generated summary includes timestamps and event tags, allowing the edge gateway to query the summary content without reopening the entire source stream.

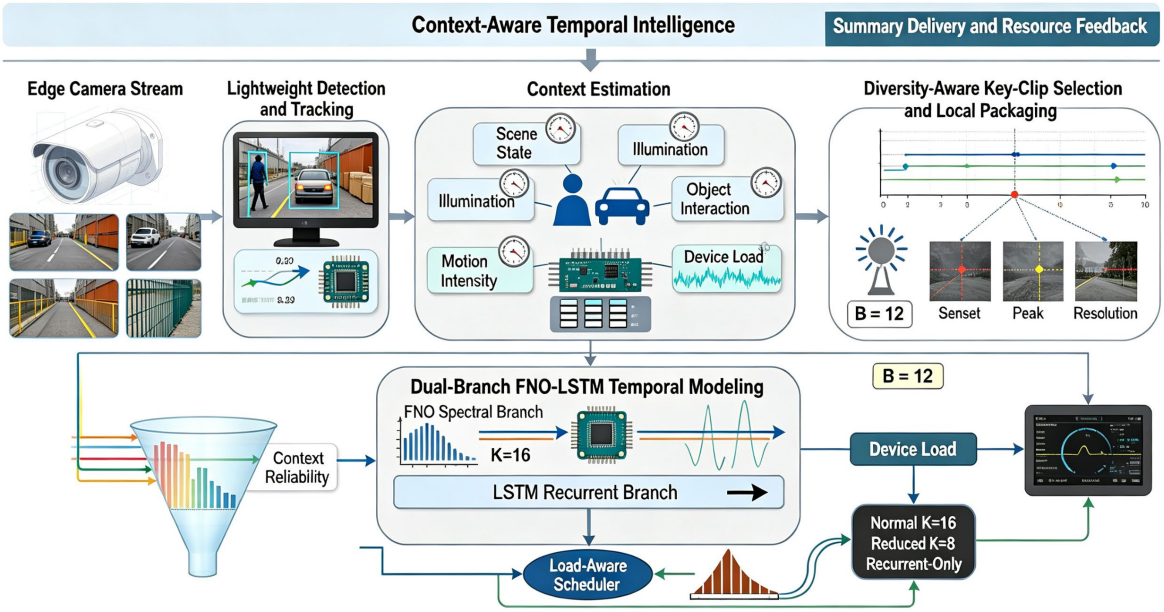

Figure 1. End-to-end workflow of the context-aware edge-camera event summarization system.

Reliability-Gated Context Fusion

Let x_t denote the visual and interaction feature at time step t, c_t the associated context vector, and r_t the reliability descriptor. Separate projections first map the heterogeneous measurements into a common d-dimensional feature space:

$$z_t = W_x x_t + b_x, h_t = W_c c_t + b_c \quad \text{Eq.(1)}$$

The fusion gate is derived from the projected visual content, contextual representation and reliability indicators. An element-wise gate is used because an illumination disturbance may invalidate appearance channels without reducing the reliability of track velocity or object-interaction features:

$$g_t = \sigma\left(W_g \begin{bmatrix} z_t \\ h_t \\ r_t \end{bmatrix} + b_g\right) \quad \text{Eq.(2)}$$

The fused representation contains a residual path for the primary visual feature. Consequently, auxiliary context cannot completely erase strong event evidence produced by the detector:

$$f_t = z_t + g_t \odot h_t \quad \text{Eq.(3)}$$

Context features are aligned with the middle timestamp of the corresponding visual frame. The old values are retained for no longer than 2 seconds, and a new record of the high-speed section is computed incrementally. Missing environmental measurements are indicated by both a zero-filled value and an explicit availability bit. This Design will not confuse a valid zero-illumination measurement with an unavailable input.

https://doi.org/10.64972/dea.2023.v2i2.3597d:90-104

Scene Identity is presented as a learned embedding. Continuous contextual variables are robustly normalized using the median and interquartile range of the training set. Camera identifiers are not provided to the fusion module to prevent the gate from learning camera-specific features. A small proportion of the training batches randomly swap scene embeddings among samples from the same domain. Augmentation helps the model learn the operating environment instead of fixed camera identities.

Reliability labels are obtained by synthetically masking context channels and then comparing the short-term context prediction with the corresponding observation. A large prediction residual, stale sensor timestamp, or high missing-value ratio reduces the reliability of the corresponding feature. When the reliability target of the context is low, the fusion output will be normalized toward the primary visual projection. Therefore, well-organized contextual information can support event decision-making, while still using the visual evidence provided by the detector as the default information path.

A device-load token is added only after semantic feature fusion. Hardware stress will therefore alter the inference schedule but not change the meaning of the event representation. The split is required because queue congestion or accelerator temperature needs to affect how the model operates, but it should not change the visual features of an event directly.

Summary Objective and Causal Selection

At each time step, the network outputs an importance value, start-boundary probability, end-boundary probability, event-class probability and semantic embedding. The utility of a candidate segment s is the average importance, boundary confidence and classification evidence.

$$u(s) = \alpha\ mean_{\,t\in s}(p_t) + \beta\sqrt{p_{start}(s)p_{end}(s)} + \gamma p_{class}(s) \qquad \text{Eq.(4)}$$

The selector retains clips under a total duration budget while penalizing temporal overlap and semantic duplication. Given candidate segments $\mathcal{S}$, the selected subset is defined as

$$\mathcal{S}^* = \underset{\mathcal{A}\subseteq\mathcal{S}}{arg\,max}\left[\sum_{s_i\in\mathcal{A}} u(s_i) - \eta \sum_{\substack{s_i,s_j\in\mathcal{A}\\ i<j}} IoU(s_i,s_j) - \mu \sum_{\substack{s_i,s_j\in\mathcal{A}\\ i<j}} cos(e_i,e_j)\right] \qquad \text{Eq.(5)}$$

subject to

$$\sum_{s_i\in\mathcal{A}} |s_i| \le B \qquad \text{Eq.(6)}$$

Here, e_i is the semantic embedding of segment s_i, IoU measures temporal overlap, and B is the maximum permitted summary duration. The first penalty discourages repeated temporal coverage, while the second suppresses clips that convey nearly identical semantic evidence.

Model training jointly optimizes importance regression, boundary estimation, event classification, and summary diversity:

$$\mathcal{L} = \mathcal{L}_{imp} + 0.6\mathcal{L}_{bnd} + 0.8\mathcal{L}_{cls} + 0.12\mathcal{L}_{div} \qquad \text{Eq.(7)}$$

Validation Experiments are used to find the coefficients. Regression Importance and event classification have the highest weights because they determine whether a relevant event is included in the candidate set. Boundary supervision enhances temporal accuracy, and the diversity regulariser avoids using most of the summary budget for repeatedly covering the same part of an event.

Only when the predicted end boundary is older than the 16-step look-ahead interval during online inference is a clip finalized. A small delay keeps the original cause and can distinguish between a sudden increase in a phenomenon and a slow decrease. Neighboring Selections with a separation of less than 0.8 s are merged. For a long event, the selector may save the onset, peak and closure clips separately instead of selecting only a middle section.

The budget selection is achieved through a two-stage process. First, non-maximum suppression removes regions with a temporal overlap greater than 0.75 that belong to the same predicted class. A lazy greedy search then assesses the marginal utility of each remaining segment after applying diversity and overlap penalties. Therefore, its number of comparisons is less than that of full subset optimization.

https://doi.org/10.64972/dea.2023.v2i2.3597d:90-104

The selector is for a safety-critical event with a calibrated confidence exceeding the validation threshold and has one slot. The reservation will not exceed the total-time budget. When multiple important events need to use the same reserved slot, confidence, time coverage and track continuity determine which one is selected.

Select the timestamps, map them back to source-frame indices, and then copy the video clips. Mapping avoids the accumulation of timestamp drift caused by feature subsampling. The emitted metadata records the utility components, aggregate context reliability, active model mode and calibration version, etc., so an operator can determine why a specific clip was selected.

FNO-LSTM Architecture and Edge Deployment

Spectral-Recurrent Encoder

Figure 2 presents the dual-branch temporal encoder. The spectral branch applies a one-dimensional real Fourier transform over the current feature window, multiplies the lowest $K = 16$ modes by learned complex weights, and returns the representation to the temporal domain. In parallel, the recurrent branch processes the same fused features using two unidirectional long short-term memory layers.

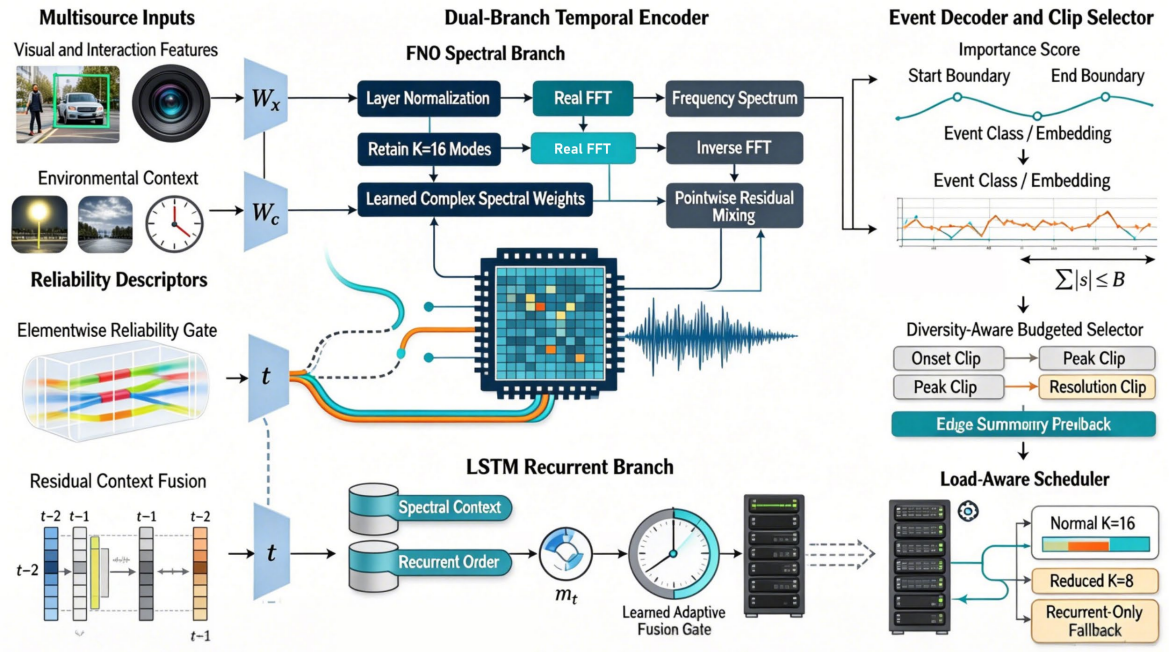

Figure 2. Internal architecture of the reliability-gated FNO-LSTM encoder and multi-head event decoder.

The outputs of the two branches are normalized, aligned through a learned fusion gate, and then passed to the event prediction head. Overlapping windows retain the final recurrent state from the previous window but recompute their Fourier Transforms. Thus, the frequency coefficients for a previous window position will not be discontinuous after the window moves.

For spectral layer l, the Fourier Neural Operator update is expressed as

$$v_{l+1} = \phi(\mathcal{F}^{-1}[R_l \odot \mathcal{F}(v_l)] + W_l v_l + b_l) \quad \text{Eq.(8)}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the forward and inverse Fourier transforms, respectively. R_l contains the learned complex weights for the retained modes, W_l is a pointwise linear transformation, and ϕ is a nonlinear activation function.

The intermediate spectral output before nonlinear activation is

$$y_l = \mathcal{F}^{-1}[R_l \odot \mathcal{F}(v_l)] + W_l v_l \quad \text{Eq.(9)}$$

The pointwise linear path retains high-frequency local information that may not be included in the truncated Fourier modes. The spectral term spreads information over the entire time window, and the pointwise term keeps step-specific details.

Rather than expanding every input, forget, output, and cell-update gate separately, the recurrent branch is represented by the standard single-line LSTM state update:

$$(q_t, c_t) = LSTM(f_t, q_{t-1}, c_{t-1}) \quad \text{Eq.(10)}$$

Here, f_t is the context-fused input feature, q_{t-1} is the preceding hidden state, c_{t-1} is the preceding cell state, and q_t and c_t are their updated values. This compact formulation retains the complete recurrent relationship while preventing an unnecessarily long multiline equation.

The spectral and recurrent representations are combined using a feature-dependent coefficient rather than a fixed summation:

$$a_t = \sigma\left(W_a \begin{bmatrix} y_t \\ q_t \end{bmatrix} + b_a\right) \quad \text{Eq.(11)}$$

The fused temporal representation is subsequently calculated as

$$m_t = a_t \odot y_t + (1 - a_t) \odot q_t \quad \text{Eq.(12)}$$

The spectral representation is expected to receive greater weight during gradual scene evolution, periodic motion, and long events containing temporally separated phases. The recurrent representation is more influential during short interactions and near event boundaries. Layer normalization is applied before adaptive fusion to maintain comparable activation scales and reduce quantization imbalance between the two branches.

Training, Quantization, and Edge Scheduling

The number of training cycles is 60, and AdamW along with a cosine learning rate schedule is used. Class-balanced event sampling avoids having too many pedestrian and vehicle passages in rare safety events. Training augmentation includes random temporal crops, brightness drift, track dropout between 5% and 30%, shuffled environmental flags, simulated decoder congestion, and missing context channels.

Windows in the same event are assigned to a single data split to avoid temporal leakage. A Gaussian kernel is used to smooth the importance target around the annotated key moment, and a three-step tolerance is employed for boundary targets. Mixed-precision training is used in the training.

Deployment uses per-channel int8 weights and int8 activations. Fourier multiplication, LSTM cell states, normalization parameters, and final logits remain in fp16. For a floating-point value x, affine quantization is defined as

$$q = clip\left(round\left(\frac{x}{S}\right) + z, q_{min}, q_{max}\right) \quad \text{Eq.(13)}$$

The corresponding reconstructed value is

$$\hat{x} = s(q - z) \quad \text{Eq.(14)}$$

where s is the quantization scale and z is the zero point.

Calibration uses 2,000 representative windows and minimizes the deviation between floating-point and quantized summary logits:

$$s^* = \underset{s}{arg\,min} \sum_{x \in \mathcal{D}_{cal}} \left\| \ell_{fp}(x) - \ell_{ints}(x; s) \right\|_2^2 \quad \text{Eq.(15)}$$

The calibration set includes all scene types, illumination quartiles and device-load states. Fourier activations have separate positive and negative ranges because the reconstructed temporal values are approximately symmetric; sigmoid-gate activations use an unsigned range. The recurrent cell state remains in fp16 to avoid saturation over an extended idle period.

Five epochs of quantization-aware fine-tuning have recovered most of the initial accuracy loss. Each deployment package includes the quantization scales and a checksum bound to the camera software version. A failed checksum or unsupported accelerator kernel will cause a controlled switch to a signed recurrent fallback package. This rule prevents optimization mismatches from quietly changing event scores after updates in the air.

A load controller observes decoder-queue occupancy q_t^{load}, normalized module temperature u_t, and recent framedrop rate d_t. Its combined load score is

$$\rho_t = \lambda_q q_t^{load} + \lambda_u u_t + \lambda_d d_t \quad \text{Eq.(16)}$$

The active execution mode is selected according to

$$M_t = \begin{cases} Normal, & \rho_t < 0.55 \\ Reduced, & 0.55 \le \rho_t < 0.78, \\ Recurrent\text{-}only, & \rho_t \ge 0.78 \end{cases} \quad \text{Eq.(17)}$$

Normal operation retains $K = 16$ Fourier modes. Reduced-mode execution uses $K = 8$ modes and updates the spectral branch every second window. The recurrent-only fallback is entered only after sustained overload and retains the reliability gate, recurrent branch, prediction heads, and clip selector.

Hysteresis avoids fast switching of the execution mode. The controller will not return to the normal working state for ten consecutive low-load windows. The policy will avoid fluctuating power consumption and prevent a sudden spike in delayed spectral calculations following a reduction in temporary workload.

Complexity and Implementation Details

For a window containing T temporal steps, feature dimension d, and K retained Fourier modes, the dominant computational cost consists of the fast Fourier transform, complex spectral mixing, pointwise projection, and recurrent processing:

$$\mathcal{C}(T, d, K) = \mathcal{O}(Td\log T + Kd^2 + Td^2) \quad \text{Eq.(18)}$$

The first term represents the forward and inverse Fourier transforms. The second term corresponds to complex multiplication over the retained modes, while the third term covers pointwise and recurrent feature transformations. Because $K \ll T$ in the deployed configuration, the spectral branch provides a global temporal field without requiring dense pairwise attention.

The implemented model contains 5.8 million parameters and occupies 22.7 MB in its int8 deployment package, including detector-side projection layers. Ring buffers hold 80 feature steps. Zero-copy tensors transfer detector outputs to the summarization module, reducing memory movement between inference stages.

Camera saves the source video in an encrypted circular buffer. Copy the selected event clips by reference until packaging is done to avoid redundant video decoding and storage. If the summary transmission fails, use a transaction record to resume the processing from where it left off without duplicating already transferred clips.

Semantic Inference and Storage Reliability are handled independently. A valid model prediction does not mean that a clip has been durably packaged, and a storage retry does not make the neural model score the same event again. The two are decoupled, with the aggregator and the detector front-end/storage system both retaining the original time information and selected targets.

Experimental Evaluation and Analysis

Experimental Setup and Overall Accuracy

The 18,420 annotated events for the four fixed-camera domains were divided into urban intersections, campus walkways, warehouse aisles and fenced perimeters. Events include entry, exit, stopping, queue formation, near-collision, restricted-zone intrusion, unattended object, loading activity, fall-like motion and camera obstruction. A total of 312 hours of recordings were taken at 1920x1080 resolution and 25-30 frames per second. Split into 60% training, 20% validation and 20% test sets, divided by day and camera. The same detector features were used to retrain the five baselines: bidirectional LSTM, temporal convolution, transformer-lite, importance-only diversity selection and a recurrent context-concatenation model. Wu et al. [16] demonstrated that graph-based temporal relations can preserve complementary information across summarized video segments. Yang et al. [17] emphasized the influence of candidate anchoring on temporal boundary localization. Temporal localization performance was evaluated using segment-level detection and boundary-quality criteria [18]. Edge measurements in this study used a fixed warm-up interval and a consistent power-sampling protocol.

https://doi.org/10.64972/dea.2023.v2i2.3597d:90-104

Two rounds of annotation were carried out. The first pass labelled event classes, onsets, peaks, resolutions and unusable intervals; the second pass selected up to three representative moments and resolved discrepancies. The median onset disagreement before adjudication was 0.38s, and Cohen's kappa for class agreement was 0.86. Events with camera restarts or more than 40% frame loss are excluded, but regular packet loss, motion blur, and partial occlusion are retained. The important goal is to standardize the average scores of the annotators, rather than a binary choice. This Design provided a smooth training signal and maintained a strict segment-level test. Macro F1 treated every class equally, mean average precision assessed ranked event confidence, and temporal intersection over union measured boundary quality. A predicted segment matched one reference only when class agreed and overlap exceeded 0.5. The summary coverage score counted reference moments contained by at least one selected clip; redundancy was the mean pairwise overlap of semantic embeddings above 0.8 cosine similarity.

Figure 3(a) compares F1, mean average precision, and temporal intersection over union. The proposed model obtained 0.842 F1, 0.781 mean average precision, and 0.736 temporal intersection over union. Transformer-lite was the strongest comparator at 0.804, 0.746, and 0.700, followed closely by temporal convolution at 0.795, 0.738, and 0.685. The 4.7-point F1 gain over temporal convolution was largest for long events with several phases. Figure 3(b) reports domain F1 values of 0.858 for traffic, 0.846 for campus, 0.832 for warehouse, and 0.817 for perimeter video. The perimeter result was lower because rain, insects, and low illumination produced uncertain short tracks, yet reliability gating limited the decline to 4.1 points relative to traffic. Event-balanced evaluation avoids allowing frequent pedestrian passages to conceal performance on rare safety events [19].

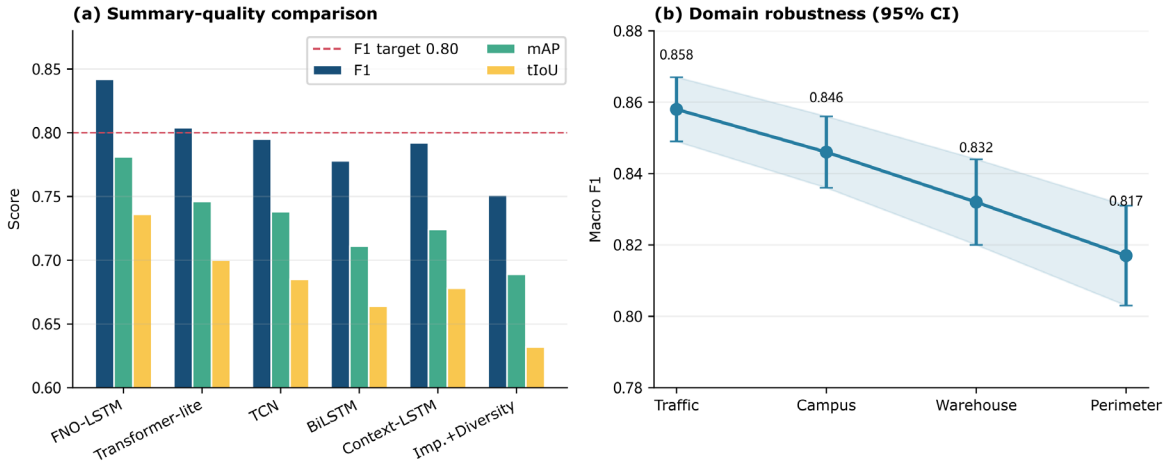

Figure 3. Overall summarization quality: (a) metric comparison across six methods; (b) domain-wise F1 with 95% bootstrap intervals.

The restricted-zone intrusion at the class level was 0.903 F1 because its spatial rule and entry transition were visually stable. Near-collision reached 0.794, with most misses occurring when one vehicle was almost stationary at the window boundary. Unattended-object F1 was 0.811, and false positives occurred when a worker temporarily placed a parcel during normal loading. The boundary head of the model reduced the average excess clip duration to 0.93s for importance-only selection from 2.14s. Shorter clips did not reduce coverage; rather, the reference-moment coverage increased from 0.842 to 0.881 because the duration budget could be allocated to more separate stages. At a 12-clip budget, an average summary had about 9.6 clips and was about 38.7 seconds long per hour of the source video. Human reviewers finished incident triage 34.2% faster with these summaries than with detector-triggered clips, although this auxiliary study only involved eight reviewers and was not used for training targets.

Ablation, Robustness, and Representation Behavior

Figure 4(a) removes one component at a time. Excluding the Fourier operator reduced F1 from 0.842 to 0.816; excluding the LSTM reduced it to 0.821; replacing reliability gating with concatenation produced 0.824; removing the diversity term produced 0.829. The paired reductions show that global spectral context and local recurrent order are complementary rather than interchangeable. In Figure 4(b), illumination corruption at severity 0.6 lowered F1 by 5.2 points with gating and 10.1 points with concatenation. Under 25% track dropout, the

corresponding declines were 4.4 and 8.3 points. Calibration practices described in [20] motivated reporting both mean accuracy and expected calibration error.

Figure 4(c) plots accuracy against retained Fourier modes. Eight modes achieved 0.834 F1 at 24.9 ms per window, while 16 modes reached 0.842 at 29.6 ms. Increasing to 24 modes changed F1 by only 0.2 points and raised latency to 37.8 ms, supporting the selected operating point. Figure 4(d) is the reliability-gate values: appearance context averaged 0.81 in normal light but dropped to 0.42 under strong exposure shifts; track context stayed above 0.70 until dropout exceeded 20%; device-load context increased from 0.18 to 0.76 after queue occupancy reached 70%. Therefore, the gate learned condition-specific redistribution rather than a single global attenuation.

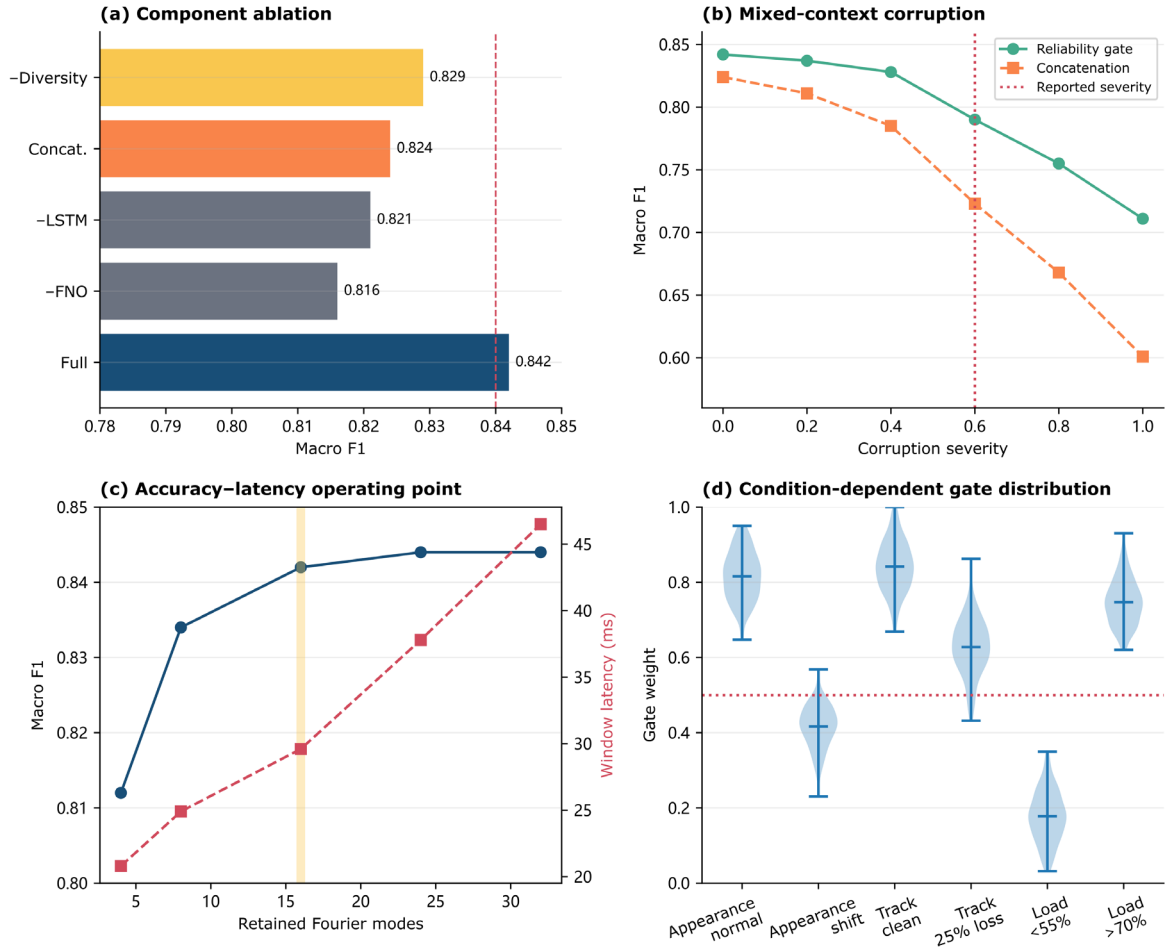

Figure 4. Component and robustness analysis: (a) ablation effects; (b) corruption-response curves; (c) Fourier-mode accuracy–latency trade-off; (d) context-gate distributions.

Extra Perturbations Separated Semantic Robustness from Simple Signal Quality. Camera vibration caused a 3-pixel median displacement and reduced the proposed F1 by 1.9 points; the importance-only baseline dropped by 5.7 points because the background motion increased nearly all segment scores. Artificially add rain streaks to reduce the mean average precision from 0.781 to 0.751. A stale scene-state flag had no effect when the timestamp age was available to the reliability network, but removing the age descriptor resulted in a 2.6-point drop. When all the context channels were valid, replacing the learned gate with direct concatenation narrowed the difference to 0.7 points; thus, the main advantage of the gate was fault isolation, not additional model capacity. A capacity-matched concatenation network had 5.9 million parameters and was still outperforming by 1.6 points on the uncorrupted set and 5.4 points under mixed corruption. The above controls prevent attributing the improvement solely to an expanded projection layer.

The single heat map in Figure 5(a) displays normalized confusion among ten event classes. Most errors occurred between stopping and queue formation (7.3%) and between loading activity and unattended object (5.8%), both

https://doi.org/10.64972/dea.2023.v2i2.3597d:90-104

of which share stationary objects before their later outcome becomes visible. Figure 5(b) examines event duration. F1 rose from 0.781 for events shorter than 2 s to 0.866 for events between 8 and 20 s, then remained at 0.851 beyond 20 s. The operator branch improved the longest group by 7.8 points over temporal convolution. Long-range temporal modeling is valuable when the event identity depends on an early cause and a later consequence [21].

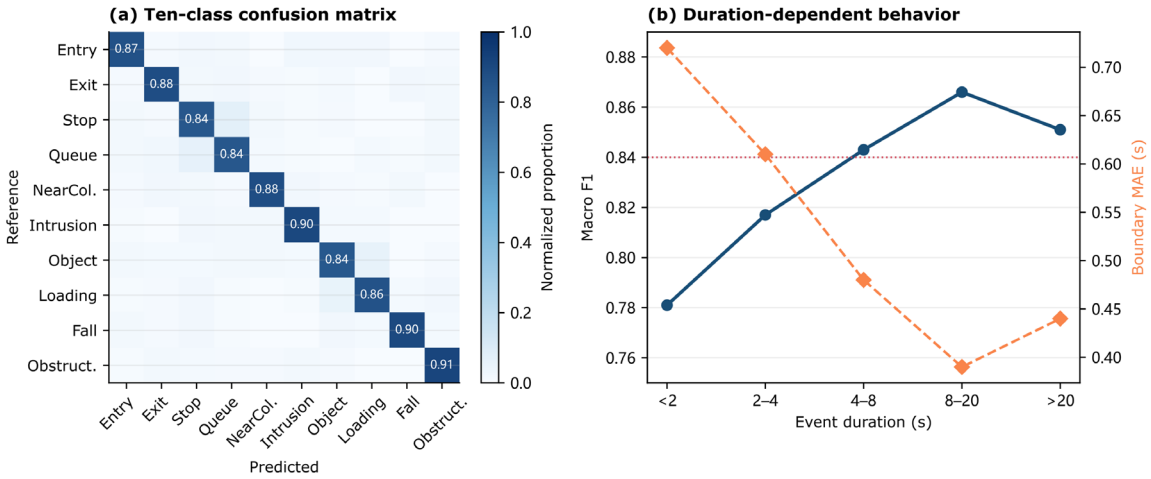

Figure 5. Event-level behavior: (a) normalized ten-class confusion heat map; (b) accuracy and boundary error across duration bins.

Temporal Inspection shows two repeated failure modes. First, an event that started in the last four steps of a window was sometimes given a delayed start probability because only limited pre-event context was available. Window overlap recovered most of the cases in the next inference pass, but 2.1% of short incidents were emitted one step late. Second, visually similar object handoffs can be divided into two events if the occlusion lasts longer than the tracker memory. Extended the tracker's memory from 1.2s to 2.0s, improved the handoff F1 score by 1.4 points, but increased identity switches in dense traffic. The selected Environment supports stable cross-domain operation. Spectral activation analysis showed that the concentrated energy modes were 2-6 for queue formation and 7-12 for short interactions. Removing the highest retained modes mainly damaged falls and near-collisions, and removing the lowest modes damaged stopping and unattended-object events. The above distribution shows that a combined spectral and recurrent encoder is more robust than either one alone.

Edge Efficiency, Summary Economy, and Statistical Analysis

Deployment measurements used an edge module set to 15 W and an ARM-only gateway for the comparison. Figure 6 shows an informative Pareto scatter plot, where the size of the markers represents memory and the color represents power. The proposed int8 model has achieved a frame rate of 27.8, used 1.42 GB of peak memory, drawn 13.6 W, and achieved an F1 score of 0.842. Transformer-lite reached a frame rate of 18.9 at 2.31 GB and 14.8 W. Temporal convolution was faster at 31.6 fps, but it had an F1 score of 0.795. The proposed point is near the accuracy-throughput trade-off frontier and still exceeds the 25-frames-per-second real-time requirement. Quantization reduced model storage by 73.4 per cent and lowered the energy per processed minute from 11.8 Wh to 8.2 Wh with a 0.6-point F1 loss. Hardware-aware comparison is needed because the number of parameters is not directly related to memory traffic or spectral-kernel efficiency [22].

A component timing trace assigned 8.7ms to feature projection and context estimation, 6.4ms to the Fourier branch, 9.1ms to the LSTM, 2.8ms to prediction heads, and 2.6ms to selection and packaging for a normal window. The total of 29.6ms excludes detector time because all baselines share the same detector output, but the full pipeline is still real-time at 27.8 fps. Peak allocation occurs during the inverse transform, reduced by reusing the pre-allocated complex workspace. On the ARM-only gateway, fp32 inference reached 6.7 frames per second, and int8 recurrent fallback reached 12.4 frames per second. This way is suitable for the gateway review and not for direct full-rate processing. Thermal tests used a closed box and a programmable heater. The scheduler reduced the spectral frequency before clock throttling started, and the maximum observed module temperature was 78.4°C instead of 83.7°C under always-normal inference.

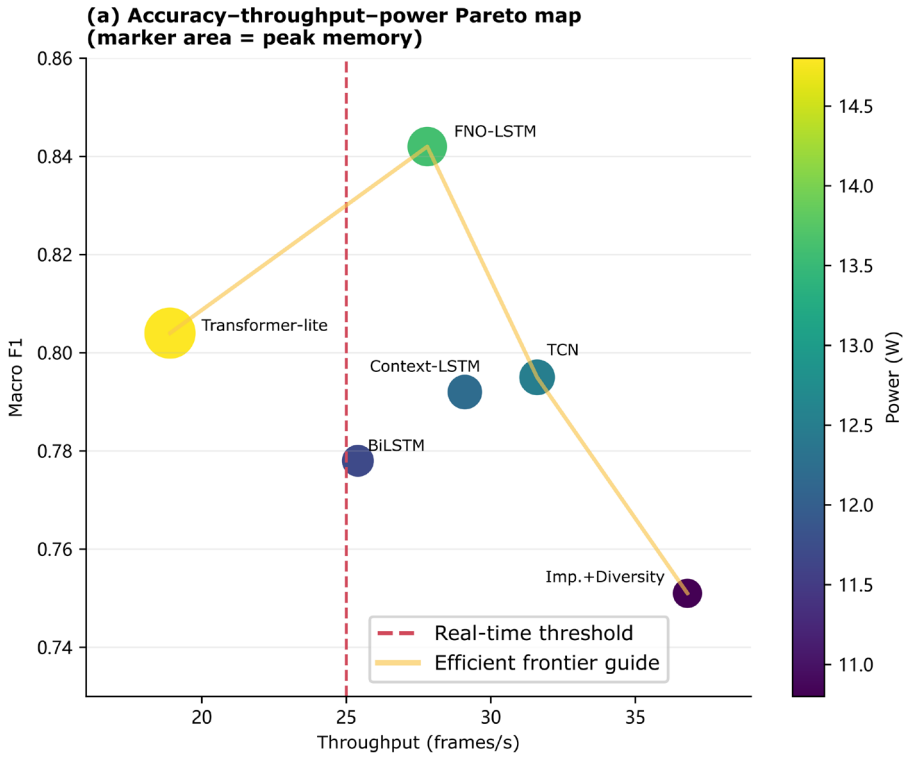

Figure 6. Accuracy–throughput–power Pareto analysis with memory-scaled markers, real-time threshold, and efficient frontier.

Figure 7(a) shows summary compression by scene: transmitted video was reduced by 90.8% for traffic, 92.4% for campus, 89.7% for warehouse, and 93.3% for perimeter data. Figure 7(b) gives summary coverage as the clip budget increases. Coverage rose from 0.71 at six clips to 0.88 at twelve clips and 0.92 at eighteen clips, while redundancy increased sharply beyond twelve. Figure 7(c) reports median and 95th-percentile delay. Normal mode measured 286 and 428 ms; reduced-mode operation measured 264 and 401 ms; recurrent fallback measured 238 and 365 ms; cloud round-trip processing measured 691 and 1,184 ms. Local inference therefore avoided tail-delay exposure to link variability [23].

Figure 7(d) presents energy across a 24 h workload. Adaptive scheduling consumed 116 Wh, compared with 143 Wh for always-normal operation and 101 Wh for recurrent-only inference. Figure 7(e) gives boundary error distributions: the proposed median absolute error was 0.43 s, compared with 0.61 s for transformer-lite and 0.68 s for temporal convolution. Figure 7(f) plots bootstrap effect sizes and confidence intervals. Improvements over all baselines were positive; the F1 difference from transformer-lite was 0.038 with a 95% interval of [0.027, 0.050]. A paired permutation test gave $p<0.001$ after Holm correction. Statistical resampling followed recommendations for correlated video segments [24], and energy results were repeated across five runs as advised for edge benchmarks [25].

Budget Sensitivity also changed the kind of retained evidence. Six clips per hour were selected based on event peaks and coverage of rare classes to achieve 0.71 moments of coverage and 0.09 redundancy. At twelve clips, the onset and resolution segments of multi-stage incidents were kept, raising coverage to 0.88 with a redundancy of 0.14. Eighteen clips increased coverage by only four percentage points but doubled redundancy to 0.28 and expanded transmitted volume by 47%. Thus, the chosen budget was not the maximum-accuracy one but rather the knee of the operating curve. Warehouse scenes needed slightly longer clips because the loading activities occurred around occlusions, and perimeter summaries used shorter intrusion clips. The selector met the same duration budget by changing the clip composition instead of using a fixed-length one. The behavior of the camera queue is used to alter the temporal structure of nominal events in different scenes.

The load controller entered reduced-mode operation for 7.8% of the windows and recurrent fallback for 1.6%. No omissions in the summary; 0.34% of the clips showed a one-window delay during rapid temperature fluctuations. At 40°C in the environment, the throughput dropped by 8.7% but was still over 25 fps. Network

replay of the same 312-h stream showed that the local summaries used 38.2 GB instead of the 454.7 GB of transmitted video. These figures represent model-side savings and exclude protocol overhead. Privacy risk is also reduced as the non-selected background video remains in the local rolling cache, although the event clips still require access control and retention policies [26]. There is no general performance data for moving cameras or audio-dependent events in the experiment; instead, a uniform engineering trade-off across the four fixed-camera environments is used.

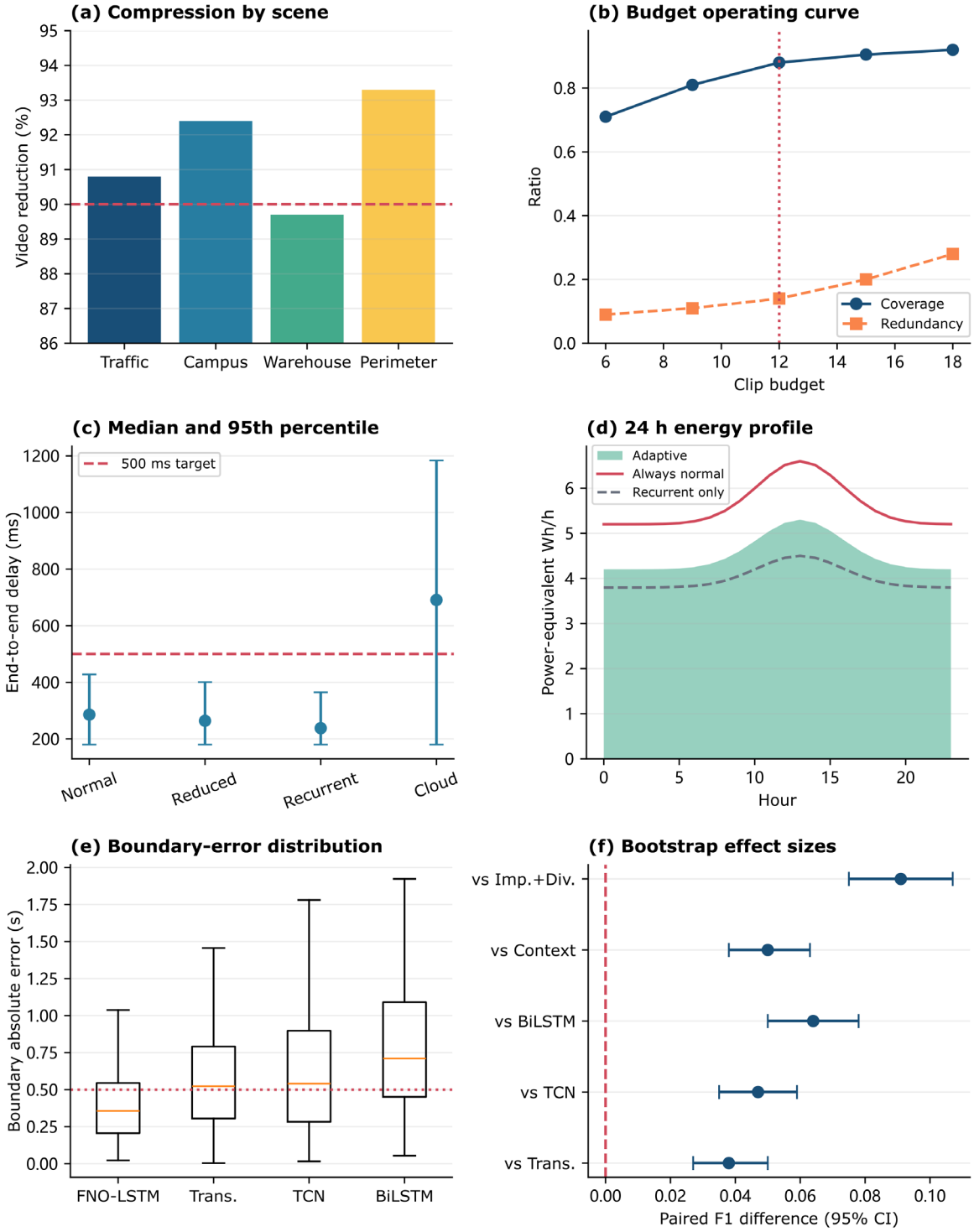

Figure 7. Operational evaluation: (a) compression by scene; (b) budget–coverage–redundancy relation; (c) latency distributions; (d) daily energy profile; (e) boundary-error distributions; (f) bootstrap effect sizes.

Conclusion

This paper proposes a context-aware FNO-LSTM model for event summarization on edge cameras. One branch is for capturing the long-term temporal structure; the other is a recurrent branch that keeps track of local order and persistence; and a reliability gate can adapt to various illumination conditions and track interruptions in illumination and device load. A diversity-aware causal selector generated compact event clips from the predictions, and quantized inference and load-aware scheduling supported continuous embedded operation. The three parts of the architecture are representation learning, clip-budget control and device scheduling in a causal pipeline. Therefore, the two goals of summary quality and deployability are treated as joint Design objectives, and edge compression is not performed until after a cloud-oriented model has been trained.

The study uses fixed cameras, a single detector family, and only four types of operating scenes. The Fourier window has a bounded look-ahead, and very short events may still be mistaken for transient noise. Context reliability labels were partially generated by simulated corruption and may not cover compound sensor faults or gradual calibration drift. The same main accelerator platform was chosen for the experiment, so the spectral core efficiency, memory traffic, and thermal behavior will differ from other processors. Human review was only tested as a small auxiliary measure, and it could not be shown that it improved all control-room workflows.

Future work will explore adaptive spectral windows, cross-camera event aggregation and uncertainty-aware human review. Online calibration can customize gating for the camera without the need to centrally store the original video, and continuous learning can be limited to preserving previously validated event boundaries. Add more trials to cover moving platforms, audio-visual incidents, adverse weather, and extended device-aging cycles. Organized privacy accounting, formal verification of load controllers, and standardized energy reporting can also be used for safety-critical and regulated deployments. A fleet-level scheduler could also coordinate summary budgets for cameras covering the same incident from multiple views.

Author Contributions

Marko Kovač contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Tomislav Petrović contributes to analysis, investigation, data collection, draft preparation. Stjepan Vuković contributes to methodology, visualization. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Zhu, W., Lu, J., Li, J., & Zhou, J. (2020). DSNet: A Flexible Detect-to-Summarize Network for Video Summarization. IEEE Transactions on Image Processing, 30, 948-962. https://doi.org/10.1109/tip.2020.3039886

[2] Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S. W., & De Albuquerque, V. H. C. (2021). A comprehensive survey of multi-view video summarization. Pattern Recognition, 109, 107567. https://doi.org/10.1016/j.patcog.2020.107567

[3] Ji, Z., Zhao, Y., Pang, Y., Li, X., & Han, J. (2020). Deep Attentive Video Summarization With Distribution Consistency Learning. IEEE Transactions on Neural Networks and Learning Systems, 32(4), 1765-1775. https://doi.org/10.1109/tnnls.2020.2991083

[4] Meena, P., Kumar, H., & Yadav, S. K. (2023). A review on video summarization techniques. Engineering Applications of Artificial Intelligence, 118, 105667. https://doi.org/10.1016/j.engappai.2022.105667

[5] Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., & Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. Pattern Recognition, 111, 107677. https://doi.org/10.1016/j.patcog.2020.107677

[6] Lin, J., Zhong, S. H., & Fares, A. (2022). Deep hierarchical LSTM networks with attention for video summarization. Computers & Electrical Engineering, 97, 107618. https://doi.org/10.1016/j.compeleceng.2021.107618

[7] Rafiq, M., Rafiq, G., Agyeman, R., Choi, G. S., & Jin, S. I. (2020). Scene Classification for Sports Video Summarization Using Transfer Learning. Sensors, 20(6), 1702. https://doi.org/10.3390/s20061702

[8] Zhao, B., Gong, M., & Li, X. (2021). AudioVisual Video Summarization. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 5181-5188. https://doi.org/10.1109/tnnls.2021.3119969

[9] Liu, T., Meng, Q., Huang, J. J., Vlontzos, A., Rueckert, D., & Kainz, B. (2022). Video Summarization Through Reinforcement Learning With a 3D Spatio-Temporal U-Net. IEEE Transactions on Image Processing, 31, 1573-1586. https://doi.org/10.1109/tip.2022.3143699

[10] Ji, Z., Fang, J., Pang, Y., & Shao, L. (2020). Deep attentive and semantic preserving video summarization. Neurocomputing, 405, 200-207. https://doi.org/10.1016/j.neucom.2020.04.132

[11] Zhu, W., Han, Y., Lu, J., & Zhou, J. (2022). Relational Reasoning Over Spatial-Temporal Graphs for Video Summarization. IEEE Transactions on Image Processing, 31, 3017-3031. https://doi.org/10.1109/tip.2022.3163855

[12] Zhu, W., Lu, J., Han, Y., & Zhou, J. (2021). Learning multiscale hierarchical attention for video summarization. Pattern Recognition, 122, 108312. https://doi.org/10.1016/j.patcog.2021.108312

[13] Zhao, B., Li, X., & Lu, X. (2020). TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization. IEEE Transactions on Industrial Electronics, 68(4), 3629-3637. https://doi.org/10.1109/tie.2020.2979573

[14] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2020). AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. IEEE Transactions on Circuits and Systems for Video Technology, 31(8), 3278-3292. https://doi.org/10.1109/tcsvt.2020.3037883

[15] Tiwari, V., & Bhatnagar, C. (2021). A survey of recent work on video summarization: approaches and techniques. Multimedia Tools and Applications, 80(18), 27187-27221. https://doi.org/10.1007/s11042-021-10977-y

[16] Wu, J., Zhong, S. H., & Liu, Y. (2020). Dynamic graph convolutional network for multi-video summarization. Pattern Recognition, 107, 107382. https://doi.org/10.1016/j.patcog.2020.107382

[17] Yang, L., Peng, H., Zhang, D., Fu, J., & Han, J. (2020). Revisiting Anchor Mechanisms for Temporal Action Localization. IEEE Transactions on Image Processing, 29, 8535-8548. https://doi.org/10.1109/tip.2020.3016486

[18] Xia, H., & Zhan, Y. (2020). A Survey on Temporal Action Localization. IEEE Access, 8, 70477-70487. https://doi.org/10.1109/access.2020.2986861

[19] Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., & Gan, C. (2021). Graph Convolutional Module for Temporal Action Localization in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 6209-6223. https://doi.org/10.1109/tpami.2021.3090167

[20] Yang, L., Han, J., Zhao, T., Lin, T., Zhang, D., & Chen, J. (2021). Background-Click Supervision for Temporal Action Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 9814-9829. https://doi.org/10.1109/tpami.2021.3132058

[21] Liu, Y., Wang, L., Wang, Y., Ma, X., & Qiao, Y. (2022). FineAction: A Fine-Grained Video Dataset for Temporal Action Localization. IEEE Transactions on Image Processing, 31, 6937-6950. https://doi.org/10.1109/tip.2022.3217368

[22] Lin, J., Yang, P., Zhang, N., Lyu, F., Chen, X., & Yu, L. (2022). Low-Latency Edge Video Analytics for On-Road Perception of Autonomous Ground Vehicles. IEEE Transactions on Industrial Informatics, 19(2), 1512-1523. https://doi.org/10.1109/tii.2022.3181986

[23] Barragan, D. R., Guim, F., Polo, J., Silva, P. M., Berral, J. L., & Carrera, D. (2022). Towards automatic model specialization for edge video analytics. Future Generation Computer Systems, 134, 399-413. https://doi.org/10.1016/j.future.2022.03.039

[24] Sun, H., Yu, Y., Sha, K., & Zhong, H. (2022). EdgeEye: A Data-Driven Approach for Optimal Deployment of Edge Video Analytics. IEEE Internet of Things Journal, 9(19), 19273-19295. https://doi.org/10.1109/jiot.2022.3166896

[25] Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of Edge Computing and Deep Learning: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 22(2), 869-904. https://doi.org/10.1109/comst.2020.2970550

[26] Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., & Benson, S. M. (2022). U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow. Advances in Water Resources, 163, 104180. https://doi.org/10.1016/j.advwatres.2022.104180