IJRCS

International Journal of Robotics and Control Systems
Vol. 6, No. 2, 2026, pp. 1004-1023
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

ASCEE

Resource-Aware Confidence-Oriented Late Fusion for Unsupervised Audio–Visual Traffic Anomaly Detection

Thien Tan Nguyen [a,1], Duy Tan Ngo [b,2,*], Duy Phong Pham [c,3], Minh Dong Nguyen [c,4], Manh Tuan Pham [d,5]

[a] Faculty of Electronics and Computer Engineering, University of Economics and Industrial Technology, Hanoi, Vietnam
[b] Vietnam National Space Center, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[c] Faculty of Electronics and Telecommunications, Electric Power University, Hanoi, Vietnam
[d] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[1] nttan@uneti.edu.vn; [2] ndtan@vnsc.vast.vn; [3] phongphd@epu.edu.vn; [4] dongnm@epu.edu.vn;
[5] tuan.pm@soict.hust.edu.vn
[*] Corresponding Author

ARTICLE INFO

Article History
Received January 14, 2026
Revised February 16, 2026
Accepted April 01, 2026

Keywords
Unsupervised Anomaly Detection;
Audio–Visual Fusion;
Traffic Monitoring;
Resource-Aware Systems;
Late Fusion;
Normalizing Flows;
Transformer Autoencoder;
Intelligent Transportation Systems

ABSTRACT

Unsupervised audio–visual anomaly detection has emerged as a promising approach for intelligent traffic monitoring, particularly in scenarios where anomalous events are rare, diverse and difficult to annotate exhaustively. Existing methods prioritize peak detection accuracy under fixed thresholds, while overlooking deployment-oriented concerns such as score stability, threshold sensitivity and robustness under realistic operating conditions. This paper presents RACoF, a Resource-Aware Confidence-Oriented Fusion framework for unsupervised audio–visual traffic anomaly detection. Instead of tightly coupling multimodal feature learning, RACoF decouples modality-specific anomaly scoring from fusion and decision-making. Audio and visual anomaly detectors are trained independently using normal-only data. Their scores are subsequently normalized, fused and calibrated through a validation-driven percentile thresholding strategy. This modular score-level or resource-aware approach mitigates scale mismatch across modalities and reduces sensitivity to absolute score magnitudes with negligible fusion overhead. Extensive experiments on real-world traffic datasets demonstrate that, although the proposed fusion strategy does not consistently outperform the strongest unimodal video detector in terms of peak AUC (Area Under Curve), it yields significantly more stable decision regions, lower threshold sensitivity, and improved interpretability across varying operating regimes. Further analysis of threshold behavior, alarm distributions, and regime-dependent fusion weights highlights RACoF's suitability for deployment-oriented traffic monitoring systems. Importantly, RACoF is model-agnostic and supports lightweight configurations by substituting heavy backbones with mobile-friendly audio–visual models, making it compatible with resource-constrained edge platforms. These results suggest that emphasizing decision stability and calibration, rather than peak accuracy alone, provides a practical pathway toward robust and edge deployable multimodal traffic anomaly detection.

1. Introduction

Intelligent transportation systems (ITS) increasingly rely on continuous monitoring mechanism to ensure traffic safety, efficiency and situational awareness in complex urban environments [29]. Among various sensing modalities, audio–visual signals provide complementary perspectives [31], [32]: visual streams capture spatial and motion cues, while acoustic signals convey event-related information that may remain observable under occlusion, low illumination, or adverse weather conditions. As a result, audio–visual analysis has become a key component in modern traffic surveillance and incident detection systems [1], [16]. Detecting anomalous traffic events, such as accidents, abnormal vehicle maneuvers or hazardous driving behaviors remains a challenging problem due to the rarity, diversity, and weakly defined nature of such events. In real-world deployments, collecting comprehensive labeled data for all abnormal scenarios is often infeasible. Consequently, unsupervised anomaly detection, which models normal traffic patterns and identifies deviations without requiring explicit anomaly labels, has attracted growing attention in recent years [2], [11]. Recent studies in audio–visual anomaly detection (AVAD) have explored a broad spectrum of modeling strategies, ranging from weakly supervised learning to attention-based multimodal architectures. Large-scale benchmarks, such as XD-Violence, ShanghaiTech, and other urban surveillance datasets, have driven rapid progress in this area, enabling comparative evaluation of diverse audio–visual detection frameworks under standardized protocols [41], [42], [45], [48]. Many of these approaches demonstrate strong performance in terms of ROC-AUC (Receiver Operating Characteristic-Area Under Curve) and related metrics, particularly when trained with video-level or segment-level supervision. Beyond weak supervision, recent AVAD research has increasingly adopted joint representation learning, including cross-modal attention mechanisms and audio–visual transformers, to explicitly model temporal synchronization and semantic alignment between modalities [41], [43], [44]. These architectures are effective in capturing fine-grained correlations, such as aligning acoustic events with corresponding visual cues. However, they typically rely on tightly coupled feature learning pipelines and computationally intensive backbones, which may limit scalability and practical deployment, especially in long-duration traffic monitoring scenarios. From a deployment perspective, traffic anomaly detection systems often operate continuously on resource-constrained platforms, where computational efficiency, modularity, and robustness are critical considerations. In such settings, detection reliability is not solely determined by peak accuracy, but also by the stability of anomaly scores, the sensitivity of decision thresholds, and the interpretability of alarm behavior over time [15], [57], [58]. Excessive threshold sensitivity or unstable scoring can lead to frequent false alarms and reduced operator trust, undermining the practical utility of otherwise accurate models. Multimodal fusion strategies play a central role in balancing detection performance and deployment feasibility. Early fusion and joint modeling approaches aim to learn unified audio–visual representations, often achieving high accuracy by exploiting cross-modal interactions. Nevertheless, these methods usually incur substantial computational overhead and reduced flexibility when individual modalities must be updated, replaced, or degraded. In contrast, late fusion approaches combine modality-specific anomaly scores at the decision level, enabling modular system design and improved scalability [11], [36]. Although late fusion is sometimes criticized for insufficiently capturing temporal synchronization between modalities, this limitation can be mitigated in traffic scenarios by operating on synchronized fixed-length temporal windows, allowing aligned event-level reasoning while preserving computational efficiency [4], [11]. In contrast, late fusion approaches combine modality-specific anomaly scores at the decision level, enabling modular system design, independent model updates, and improved scalability, which are particularly desirable in safety-critical and edge-deployed monitoring systems [15], [53], [59]. Motivated by these observations, this paper proposes RACoF (Resource-Aware Confidence-Oriented Fusion), a modular framework for unsupervised audio–visual traffic anomaly detection. Here, the term "resource-aware" refers to a modular score-level fusion and calibration design with negligible computational overhead, rather than hardware-level optimization techniques or real-time execution guarantees. RACoF de-

couples modality-specific anomaly modeling from fusion and decision-making, allowing audio and visual detectors to be trained independently using normal-only data. Their anomaly scores are subsequently normalized, fused through regime-dependent weights, and calibrated using a validation-driven percentile thresholding strategy. Rather than optimizing peak accuracy alone, this design explicitly emphasizes score stability, threshold robustness, and interpretable alarm behavior under realistic operating conditions. This emphasis reflects a practical reality in safety-critical traffic monitoring systems, where marginal improvements in peak ROC-AUC are often less valuable than stable and predictable alarm behavior under fixed operating thresholds. The main contributions of this work are summarized as follows:

- A model-agnostic, late-fusion framework for unsupervised audio–visual traffic anomaly detection that emphasizes decision stability and calibration in long-term monitoring scenarios.
- A validation-driven percentile thresholding strategy that reduces sensitivity to absolute score magnitudes and enables interpretable control over alarm behavior across multiple operating regimes.
- An extensive experimental analysis on real-world traffic datasets, demonstrating that while fusion may not outperform the strongest unimodal detector in peak AUC, it yields more stable decision regions and improved robustness across thresholds.
- A system-level perspective illustrating how modular fusion and calibration support flexible and lightweight configurations suitable for resource-constrained platforms.

2. Proposed Methodology

2.1. Problem Formulation

The dataset for audio–visual traffic anomaly detection is defined by:

$$\mathcal{D} = \{A_t, V_t, y_t\}_{t=1}^{T}. \tag{1}$$

Where D is the audio-visual traffic dataset, V_t and A_t represent the visual and audio observations at time t, respectively. Note that the ground-truth labels y_t are used only for validation and evaluation, and are not available during training.

In realistic Intelligent Transportation Systems (ITS), anomalous traffic events are rare, heterogeneous, and weakly labeled, which significantly limits the applicability of supervised learning approaches. Consequently, the training set is defined in equation (2).

$$D_{\text{train}} = \{(A_t, V_t) \mid y_t = normal\} \tag{2}$$

The training set contains exclusively normal samples. Under this setting, unsupervised anomaly detection aims to learn a model **M** that assigns a continuous anomaly score $s(t)$ such that anomalous segments yield higher scores than normal ones with high probability [11], [12], [39].

Importantly, from a deployment perspective, traffic anomaly detection is not merely a binary classification task. Instead, it constitutes a continuous risk scoring problem, where anomaly scores are consumed by downstream alarm, monitoring, or control modules. Therefore, beyond detection accuracy, practical systems must emphasize score stability, interpretability, and controllable alarm behavior, as highlighted in recent robotics and control-oriented sensing studies [27], [30].

2.2. Overall Framework

The framework consists of three main components:

- A video anomaly scoring branch that models the distribution of normal traffic scenes using likelihood-based density estimation.
- An audio anomaly scoring branch that models normal acoustic patterns using reconstruction-based learning.

- A RACoF module that normalizes and fuses modality-specific scores under different operating regimes.

This modular design decouples modality-specific modeling from fusion and decision-making, enabling robust operation under modality degradation and facilitating flexible deployment on resource-constrained platforms [1], [15], [53].

The RACoF module is designed to minimize system complexity and integration overhead through three key principles. First, modality-specific anomaly detectors are fully decoupled allowing independent training, replacement or degradation handling without retraining the entire system. Second, the fusion and decision-making stage operates with constant-time complexity per segment, introducing negligible computational overhead beyond unimodal inference. Third, the framework avoids tightly coupled cross-modal attention or joint representation learning, which significantly reduces memory usage and computational cost compared to multimodal transformer-based approaches. Collectively, these design choices enable RACoF to support lightweight configurations and flexible deployment on resource-constrained platforms.

2.3. Video Anomaly Scoring via Normalizing Flow

Given a video segment V_t, a visual representation $x_t \in \mathbb{R}^d$ is extracted using a backbone network. The distribution of normal visual patterns is modeled using a normalizing flow, which defines an invertible transformation $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ between the data space and a latent space with a known prior distribution $p_Z(z)$ [8], [26], [37].

Using the change-of-variables formula, the exact log-likelihood of a sample is computed as:

$$\log p(x_v) = \log p(z) + \log \left| \det \left(\frac{\partial f}{\partial x_v} \right) \right|. \tag{3}$$

The video anomaly score is defined as the negative log-likelihood:

$$s_v(t) = -\log p(x_t). \tag{4}$$

Likelihood-based scoring provides a probabilistically grounded measure of abnormality and has demonstrated strong performance and interpretability in visual anomaly detection compared to reconstruction-based methods [26], [36], [39] Fig. 1.

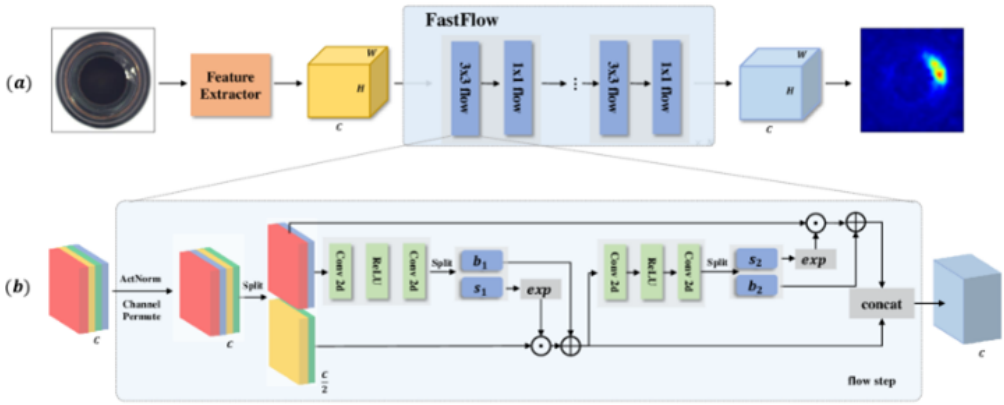

Fig. 1. Visual/video-based anomaly detection branch using FastFlow

2.4. Audio Anomaly Scoring via Transformer Autoencoder

For each audio segment A_t, a time–frequency representation $x_t \in \mathbb{R}^{F \times \tau_a}$, such as a log-Mel spectrogram, is computed. A Transformer-based autoencoder is employed to capture long-range temporal dependencies in normal acoustic patterns [50], [51] shown in Fig. 2.

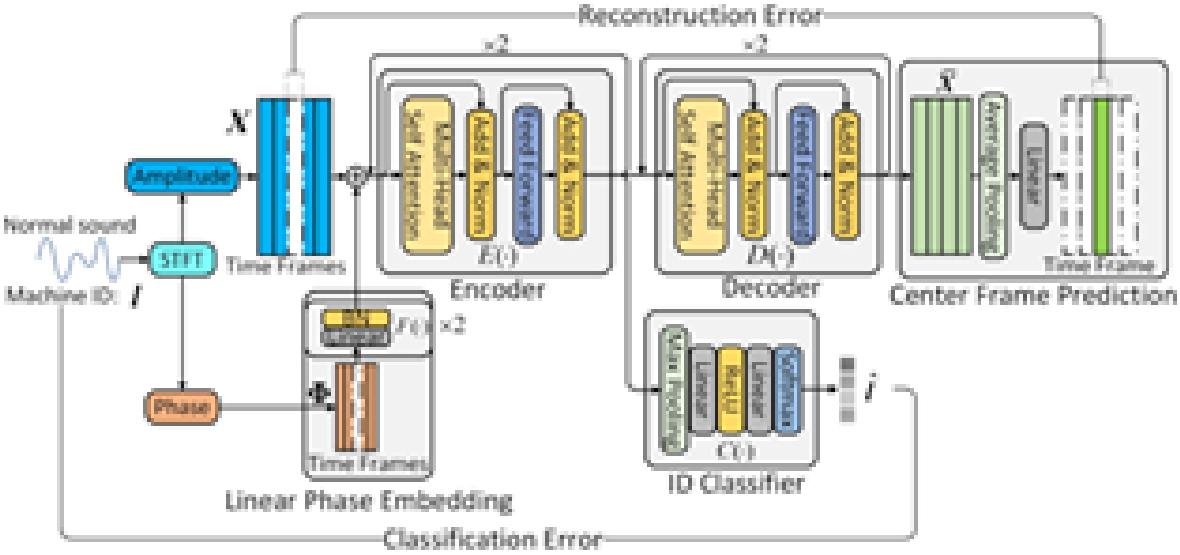

Fig. 2. Audio-based anomaly detection branch using a transformer autoencoder

The reconstruction process is defined as:

$$\hat{X}_t = \mathrm{Dec}_\phi\big(\mathrm{Enc}_\phi(X_t)\big). \tag{5}$$

Where Dec_ϕ and Enc_ϕ denote the decoder and encoder with parameters ϕ. The audio anomaly score is computed as the mean squared reconstruction error:

$$s_a(t) = \frac{1}{FT_a} \sum_{i=1}^{FT_a} \left(X_{t,i} - \hat{X}_{t,i}\right)^2. \tag{6}$$

Where F and T_a denote the number of Mel bins and time frames, respectively.

Reconstruction-based acoustic anomaly detection is widely adopted due to its robustness and computational efficiency in real-world monitoring scenarios [33], [49]–[51].

2.5. Score Normalization

Anomaly scores produced by heterogeneous models typically follow different numerical scales and statistical distributions, which renders direct fusion unreliable. To address this issue, modality-specific scores are normalized using statistics estimated on the validation set [27], [57], [58]:

$$\tilde{s}_m(t) = \frac{s_m(t) - s_m^{\min}}{s_m^{\max} - s_m^{\min} + \varepsilon}, \quad m \in \{v, a\}. \tag{7}$$

Where $s_m^{\min}$ and $s_m^{\max}$ denote the minimum and maximum scores observed during validation, and ε is a small constant to ensure numerical stability.

2.6. Resource-Aware Confidence-Oriented Fusion

The proposed RASCoF operates under multiple confidence regimes:

$$r \in \{low,\ med,\ high\}$$

The confidence regimes are defined as follows:

- **Low-confidence regime:** conservative fusion prioritizing score stability and false-alarm suppression.

- **Medium-confidence regime:** balanced fusion maximizing the F1-score, used as the default operating point.
- **High-confidence regime:** aggressive fusion prioritizing recall and early anomaly detection.

For each regime r, the fused anomaly score is computed as:

$$s_f^{(r)}(t) = w_v^{(r)} \tilde{s}_v(t) + w_a^{(r)} \tilde{s}_a(t) \tag{8}$$

Where

$$w_v^{(r)} + w_a^{(r)} = 1.$$

From an optimization perspective, the proposed weighted linear fusion can be interpreted as a convex combination of modality-specific anomaly scores, which preserves boundedness and enables stable trade-offs between competing objectives such as false-alarm suppression and recall maximization, consistent with classical convex optimization principles [27], [35], [36], [58].

This regime-based design allows sensitivity adjustment without retraining, consistent with deployment practices in safety-critical ITS applications [29], [59], as illustrated in Fig. 3.

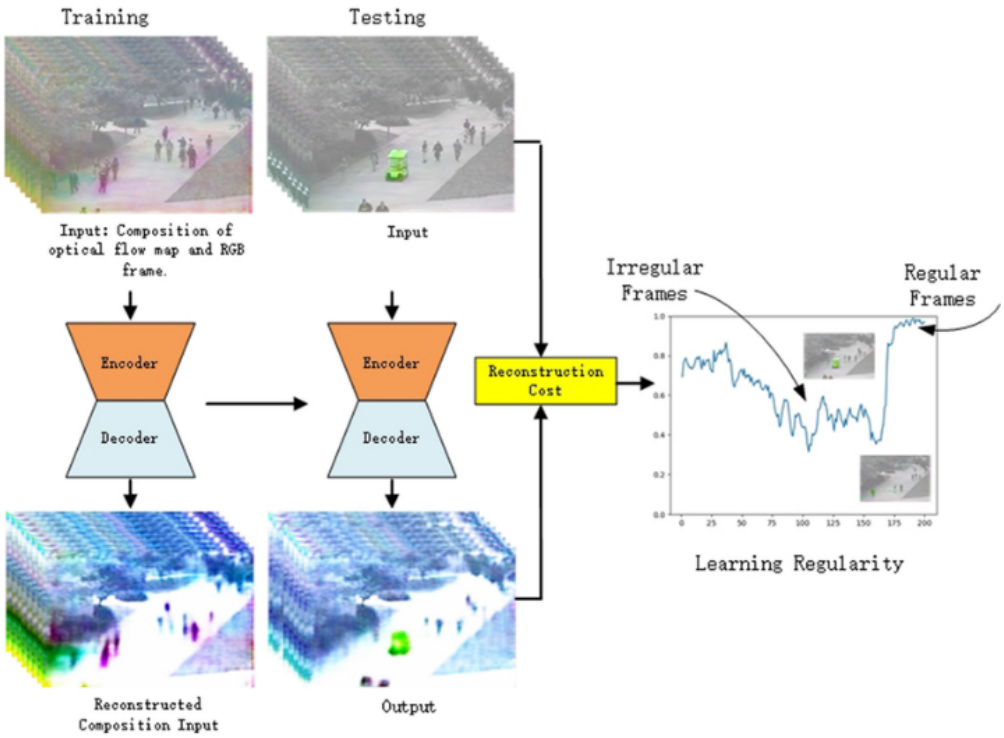

Fig. 3. RACoF fusion module with regime-based weighting and threshold selection

2.7. Decision Rule and Threshold Selection

For each regime r, a decision threshold $\tau^{(r)}$ is selected on the validation set by optimizing a regime-specific objective (e.g., F1-score):

$$\hat{y}^{(r)}(t) = \begin{cases} \textit{anomaly}, & s_f^{(r)}(t) \geq \tau^{(r)} \\ \textit{normal}, & \text{otherwise} \end{cases} \tag{9}$$

To convert continuous fused anomaly scores into binary alarm decisions, we adopt a deployment-friendly percentile-based thresholding strategy determined exclusively on the validation set. For each

confidence regime r, the decision threshold is defined as:

$$\tau^{(r)}(p) = \text{Percentile}\left(\{s_f^{(r)}(t)\}_{t\in\mathcal{D}_{val}},\, p\right) \tag{10}$$

Where p controls the sensitivity of the alarming behavior.

In our experiments, the best performing percentile is consistently found to be $p^* = 80$, yielding regime-specific thresholds $\tau^{(r)}$ that are subsequently fixed and applied to the test set. This choice is further supported, which illustrates the sensitivity of the F1-score to different percentile thresholds on the validation set.

This validation-driven strategy avoids information leakage and ensures reproducible evaluation [27], [40], [57], [58].

Algorithm 1 summarizes the RACoF pipeline for resource-aware score-level fusion, including modality-specific anomaly scoring, normalization, and profile-based weighting for stable deployment [1], [27], [36].

Algorithm 2 summarizes the validation-driven percentile sensitivity analysis used to select regime-specific decision thresholds, ensuring reproducible and stable alarm behavior across deployment profiles.

The final decision threshold $\tau^{(r)}$ is obtained by fixing the percentile-dependent threshold $\tau^{(r)}(p)$ at the selected optimal percentile p^*. Algorithm 2 summarizes the validation-driven percentile sensitivity analysis used to select regime-specific decision thresholds, ensuring reproducible and stable alarm behavior across deployment profiles [27], [57], [58].

Algorithm 1 RASCoF – Resource-Aware Confidence-Oriented Fusion

Input:

• Synchronized segment pair $(A_v, A_a)_{t=1,\ldots,T}$

• Trained video anomaly model $\mathcal{M}_v$ (FastFlow) and trained audio anomaly model $\mathcal{M}_a$ (Transformer AE)

• Validation-based normalization statistics $(s_v^{\min}, s_v^{\max})$, $(s_a^{\min}, s_a^{\max})$

• Risk regimes $r \in \{low, med, high\}$ with fusion weights $(w_v^{(r)}, w_a^{(r)})$

• Decision thresholds $\tau^{(r)}$ (calibrated on validation set)

• $\epsilon > 0$

Output:

• Fused anomaly scores $\hat{s}_f^{(r)}(t)$ and alarm decisions $\hat{y}^{(r)}(t)$ for each regime r

Steps:

1. Compute unimodal scores for each segment t:
   1.1. $s_v(t) \leftarrow VideoScore(\mathcal{M}_v, A_v)$ (e.g., negative log-likelihood)
   1.2. $s_a(t) \leftarrow AudioScore(\mathcal{M}_a, A_a)$ (e.g., reconstruction error)
2. Normalize scores using validation min-max:
   2.1. $\tilde{s}_v(t) \leftarrow \dfrac{s_v(t) - s_v^{\min}}{s_v^{\max} - s_v^{\min} + \epsilon}$
   2.2. $\tilde{s}_a(t) \leftarrow \dfrac{s_a(t) - s_a^{\min}}{s_a^{\max} - s_a^{\min} + \epsilon}$
3. Fuse and decide for each regime $r \in \{low, med, high\}$:
   3.1. $\hat{s}_f^{(r)}(t) \leftarrow w_v^{(r)}\tilde{s}_v(t) + w_a^{(r)}\tilde{s}_a(t)$
   3.2. $\hat{y}^{(r)}(t) \leftarrow \mathbb{I}(\hat{s}_f^{(r)}(t) \geq \tau^{(r)})$
4. Return $\{\hat{s}_f^{(r)}(t), \hat{y}^{(r)}(t)\}_r$

Algorithm 2 Threshold Sensitivity via Validation F1

Input:
- Validation set $\mathcal{D}_{val} = \{(V_i, A_i, y_i)\}_{i=1}^{T_{val}}$
- Trained models $\mathcal{M}_v, \mathcal{M}_a$
- Normalization $(s_v^{\min}, s_v^{\max})$, $(s_a^{\min}, s_a^{\max})$
- Risk regimes $r \in \{low, med, high\}$ with weights $(w_v^{(r)}, w_a^{(r)})$
- Percentile grid $\mathcal{P} = \{p_1, \ldots, p_K\}$, $(p = 50\text{–}95)$
- $\epsilon > 0$

Output:
- F1-percentile curves $\{F1^{(r)}(p)\}$ for Fig. 9
- Selected percentile p^* and threshold $\tau^{(r)}$ for each regime

Steps:

1. Compute modality scores on validation set:
   For each $i \in \mathcal{D}_{val}$:
   $s_v(i) \leftarrow score_v(\mathcal{M}_v, V_i)$
   $s_a(i) \leftarrow score_a(\mathcal{M}_a, A_i)$
2. Normalize scores:
   $$\tilde{s}_v(i) \leftarrow \frac{s_v(i) - s_v^{\min}}{s_v^{\max} - s_v^{\min} + \epsilon}$$
   $$\tilde{s}_a(i) \leftarrow \frac{s_a(i) - s_a^{\min}}{s_a^{\max} - s_a^{\min} + \epsilon}$$
3. For each regime $r \in R$:
   3.1. Fuse validation scores:
   $\hat{s}_f^{(r)}(i) \leftarrow w_v^{(r)} \tilde{s}_v(i) + w_a^{(r)} \tilde{s}_a(i)$
   3.2. For each percentile $p \in \mathcal{P}$:
   Set threshold by percentile:
   $\tau^{(r)}(p) \leftarrow Percentile(\{\hat{s}_f^{(r)}(i)\}, p)$
   Predict $\hat{y}^{(r)}(i) \leftarrow \mathbb{I}(\hat{s}_f^{(r)}(i) \geq \tau^{(r)}(p))$
   Compute validation $F1^{(r)}(p)$
   3.3. Select best percentile:
   $p^* \leftarrow \arg\max_{p \in \mathcal{P}} F1^{(r)}(p)$
4. Return $\{p^*, \tau^{(r)}, F1^{(r)}(p)\}_r$

2.8. Stability Analysis

Under normal conditions, assume normalized scores $\tilde{s}_v(t)$ and $\tilde{s}_a(t)$ are bounded in $[0, 1]$ and not perfectly correlated. The variance of the fused score satisfies:

$$\mathrm{Var}\left(s_f^{(r)}\right) = w^{(r)T} \Sigma w^{(r)}. \tag{11}$$

Where $w^{(r)} = [w_v^{(r)}, w_a^{(r)}]^T$ and

$$\Sigma = \begin{bmatrix} \sigma_v^2 & \mathrm{Cov} \\ \mathrm{Cov} & \sigma_a^2 \end{bmatrix},$$

Since the cross-modal covariance is typically smaller than σ_v^2 and σ_a^2, multimodal fusion yields a reduced variance compared to unimodal scoring. This provides a theoretical explanation for improved score stability as observed under RACoF Fig. 4

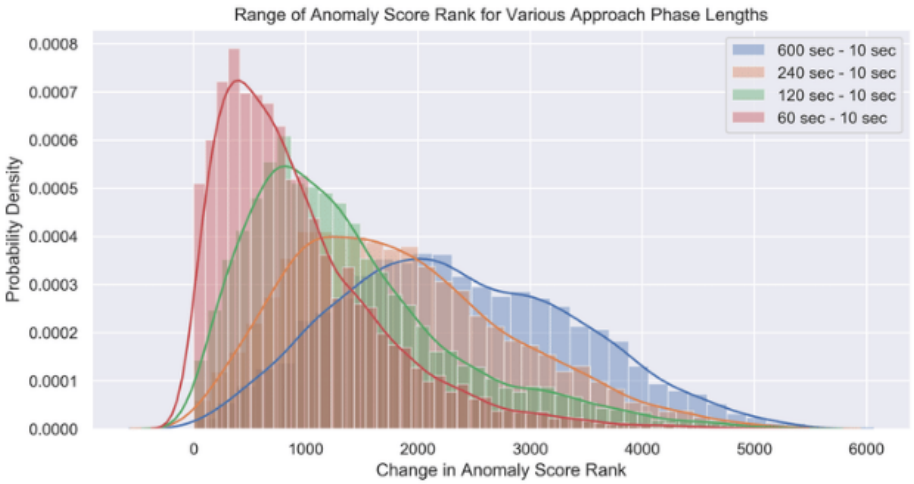

Fig. 4. Distribution of fused anomaly scores and video ranking stability

2.9. Computational Complexity

The computational complexity of the proposed framework is dominated by modality-specific inference

- **Video branch:** $\mathcal{O}(K \cdot C_f)$, where K denotes the number of flow blocks and C_f represents the per-block feature transformation cost, which scales linearly with the feature dimensionality for lightweight architectures used in visual anomaly detection [26].
- **Audio branch:** $\mathcal{O}(N_a \cdot L^2 \cdot D)$, corresponding to the standard self-attention complexity of Transformer-based models, where N_a is the number of layers, L the sequence length, and D the model dimension.
- **Fusion and decision-making:** $\mathcal{O}(1)$ per segment.

Overall, RASCoF introduces negligible computational overhead beyond unimodal inference, as fusion and thresholding operations are lightweight. This supports real-time and deployment-oriented operation on resource-constrained platforms, consistent with system-level requirements in intelligent transportation monitoring [27], [53].

Table 1 summarizes the architecture configurations and key hyperparameters of both branches, highlighting the modular and resource-efficient design of the proposed framework.

Table 1. Summary of processing steps in the proposed pipeline

Step	Description	Script
Audio–video synchronization	MAVD ID-based pairing	`step1_pairing`
Audio anomaly	Transformer Autoencoder	`audio_train / infer`
Video anomaly	FastFlow (ResNet-18)	`step5_3, step5_4`
Fusion	Normalization + fusion	`step6`
Stability analysis	Alpha sweep	`step6_1, step6_2`

3. Experimental Setup

3.1. Dataset and Segmentation Protocol

We evaluate the proposed framework on the Multimodal Audio–Visual Dataset (MAVD), which provides synchronized traffic video and audio streams annotated at the segment level for anomaly

detection. MAVD has been widely adopted for studying multimodal traffic monitoring under realistic urban conditions, where both visual and acoustic cues are informative for abnormal event detection [7], [17].

Following standard practice in unsupervised anomaly detection, the objective is to learn models on normal behavior during training and detect deviations at inference time without relying on anomalous labels. MAVD is organized into fixed-length temporal segments indexed by (t_1, t_0). In our benchmark construction, each segment has a constant duration of:

$$\Delta t = t_1 - t_0 = 2.0s \tag{12}$$

This fixed-window segmentation facilitates stable score aggregation, reproducible evaluation, and deployment-friendly streaming inference, which are important considerations for continuous traffic monitoring systems [27], [29].

3.2. Data Splits and Class Distribution

MAVD is partitioned into three disjoint splits: training, validation, and test, following the provided segment manifests:

- Train: $N_{\text{train}} = 4410$ segments
- Validation: $N_{\text{val}} = 493$ segments
- Test: $N_{\text{test}} = 3075$ segments

The dataset exhibits a notable class imbalance. Based on the manifest labels (with good denoting normal), the raw label distribution is given as:

- Train: good = 107, anomaly = 4303
- Validation: good = 11, anomaly = 482
- Test: good = 591, anomaly = 2484

Such imbalance is typical in real-world anomaly detection benchmarks and motivates the use of both ranking-based and precision–recall–oriented evaluation metrics, as PR-based measures are more informative under skewed class distributions [23], [24].

Although the original training split provided by the MAVD manifest contains both normal and anomalous labels, the proposed framework strictly follows a normal-only training protocol. Specifically, only segments labeled as normal are retained to construct the effective training subset, while all anomalous segments are discarded prior to model optimization. The reported label distribution is provided solely for dataset characterization. After filtering, the effective training set consists of 107 normal segments, which are subsequently used exclusively to train both the audio and video models, consistent with standard unsupervised anomaly detection practice [13], [27], [29] shown in Table 2.

Table 2. Effective training set after normal-only filtering

Component	Quantity	Description
Raw training segments	4,410	Original MAVD training split before filtering
Normal segments retained	107	Segments labeled as good used for unsupervised training
Segment duration	2.0 s	Fixed-length temporal segmentation
Total normal training duration	214 s ($\approx$ 3.6 min)	107×2.0 s
Video frames per segment	16	Uniform temporal sampling
Total training frames (video)	1,712 frames	107×16
Audio frames per segment	128	Log-Mel spectrogram time frames
Total audio frames	13,696 frames	107×128

Only normal segments are used for training in accordance with the unsupervised anomaly detection protocol. Anomalous segments in the original training split are discarded prior to model optimization.

3.3. Unsupervised Training Protocol

The training process operates exclusively on the normal-only subset extracted from the original training split. Both the video-based anomaly scoring model and the audio-based reconstruction model are trained independently on this subset, ensuring that no abnormal events are observed during optimization. Anomalous segments from the validation and test splits are never used for training; instead, they are reserved strictly for model selection, threshold calibration, and final performance evaluation. This protocol prevents information leakage and preserves the integrity of the unsupervised setting [27], [29].

3.4. Audio Preprocessing and Scoring

Each audio segment Δt is transformed into a time–frequency representation X_a (log-Mel spectrogram) using STFT front-end and Mel filterbank projection, a widely adopted representation for acoustic anomaly detection [50], [52]. The audio anomaly detection module is implemented as a Transformer-based auto-encoder trained exclusively on normal traffic sounds. By learning to reconstruct typical acoustic patterns, the model captures the underlying structure of normal audio dynamics without requiring explicit anomaly annotations. During inference, the audio anomaly score is computed as the reconstruction error between the input spectrogram and its reconstruction, yielding a scalar anomaly score $s_a(t)$. Larger reconstruction errors indicate a higher degree of deviation from learned acoustic behavior, consistent with prior work on reconstruction-based anomaly detection [50], [51].

3.5. Video Preprocessing and Scoring

For each video segment V_t, frame-level or clip-level visual features x_v are extracted and modeled using a likelihood-based normalizing flow. The flow model is trained solely on normal traffic video segments to estimate the distribution of typical visual patterns observed under normal conditions.

At inference time, the video anomaly score $s_v(t)$ is defined as the negative log-likelihood of the observed features under the learned flow model. Samples with low likelihood correspond to visual patterns that deviate from the normal traffic distribution. Likelihood-driven scoring provides a principled and interpretable abnormality measure under normal-only training and has been shown to be effective in visual anomaly detection tasks [8], [26].

3.6. Evaluation Metrics

Given the severe class imbalance inherent in MAVD, we report a comprehensive set of evaluation metrics that jointly assess ranking quality, decision accuracy, and alarm behavior:

- ROC-AUC, measuring threshold-independent ranking performance;
- PR-AUC, which is more informative under imbalanced conditions and emphasizes performance on the minority class;
- Threshold-dependent metrics, including Precision, Recall, and F1-score, computed at regime-specific thresholds τ^*;
- Deployment-oriented alarm indicators, namely the false positive rate (FPR) and false negative rate (FNR), derived from the confusion matrix on the test set.

Runtime throughput and real-time performance are not the primary focus of this study; instead, the evaluation emphasizes score behavior, threshold sensitivity, and decision stability under controlled experimental conditions shown in Table 3.

3.7. Experimental Protocol and Reproducibility

All models are trained offline using only normal traffic segments from the training split. Score normalization, fusion weight selection, and percentile-based threshold calibration are performed ex-

clusively on the validation set. Final performance is reported solely on the held-out test set, ensuring strict separation between training, model selection, and evaluation stages.

Table 3. Summary of the experimental setup

Component	Description
Dataset	MAVD (synchronized audio–video)
Modalities	Audio + Visual
Learning type	Unsupervised
Audio model	Transformer Autoencoder
Visual model	FastFlow (ResNet-18)
Fusion	Normalized late fusion
Evaluation	Ranking & stability

Unimodal baselines (audio-only and video-only) as well as the proposed RACoF fusion approach are evaluated under identical segmentation settings, metric definitions, and thresholding protocols. This unified experimental protocol guarantees fair comparison and full reproducibility of the reported results, consistent with best practices in anomaly detection research.

3.8. Edge Feasibility and Design Considerations

Although this work does not report hardware-specific real-time measurements, RACoF is designed to facilitate deployment on resource-constrained platforms through modular and decoupled processing. The framework enables independent substitution of modality-specific backbones with lightweight alternatives (e.g., compact visual encoders or acoustic models), while preserving the proposed fusion, normalization, and calibration mechanisms.

Importantly, fusion and decision-making introduce negligible computational overhead compared to unimodal inference, and no tightly coupled cross-modal attention or heavy multimodal transformers are required. Hardware-level optimizations such as INT8 quantization, pruning, and runtime profiling on embedded devices are intentionally left for future work, where RACoF can serve as a methodological foundation for end-to-end edge deployment studies.

4. Results and Discussion

This section presents the experimental results of RACoF on the MAVD dataset, including overall detection performance, threshold sensitivity, and temporal stability analyses.

In this work, the term "deployment-oriented" does not imply real-time execution on embedded system validation. Instead, it refers to decision-level properties—such as threshold robustness, alarm stability, and controllable precision–recall trade-offs—that are critical in real-world monitoring systems where false alarms incur significant operational costs.

4.1. Overall Quantitative Performance

We first evaluate the overall anomaly detection performance of the unimodal baselines and the proposed RACoF framework on the MAVD test set. Quantitative results are summarized in Table 3, which include independent ranking metrics (ROC-AUC and PR-AUC) and threshold-dependent alarm statistics.

As shown in Table 3, the video-only branch achieves the highest ROC-AUC (0.6308), indicating that visual cues provide the strongest discriminative information for traffic anomaly detection on MAVD. This observation is consistent with prior studies on visual anomaly detection in likelihood modeling and spatio-temporal representations, which report strong performance of video-centric approaches in surveillance and traffic scenarios [26].

In contrast, the audio-only branch exhibits limited discriminative capability (ROC-AUC = 0.5431).

This behavior is expected, as many anomalous traffic events do not necessarily produce distinctive acoustic signatures, particularly in noisy urban environments. Similar limitations of acoustic-only anomaly detection have been reported in environmental sound monitoring and traffic audio analysis [50], [51].

Although RACoF fusion does not surpass the video-only branch in terms of peak ROC-AUC, it consistently improves over the audio-only baseline and yields competitive PR-AUC values across confidence regimes. Importantly, this outcome aligns with recent findings that multimodal fusion does not always lead to higher peak accuracy, but can significantly enhance the robustness and reliability of anomaly scoring in practical systems [36], [58] shown in Table 4.

Table 4. Overall anomaly detection performance on the MAVD test set

Method	Regime	Precision	Recall	F1-score	FPR
Audio-only	–	0.8075	0.0523	0.0983	0.0525
Video-only	–	0.8468	0.5097	0.6363	0.3875
RACoF	Low	0.7727	0.0479	0.0902	0.0592
RACoF	Medium	0.7792	0.0483	0.0910	0.0575
RACoF	High	0.7857	0.0487	0.0917	0.0558

For unimodal baselines, decision thresholds are selected on the validation set using the same percentile-based rule ($p^* = 80$) as RACoF to ensure a fair operating-point comparison.

4.2. ROC and Precision–Recall Curve Analysis

To further analyze detection behavior independent of specific threshold choices, we examine the ROC and Precision–Recall (PR) curves. Fig. 5 presents the ROC curves for audio-only, video-only, and RACoF on the MAVD test set.

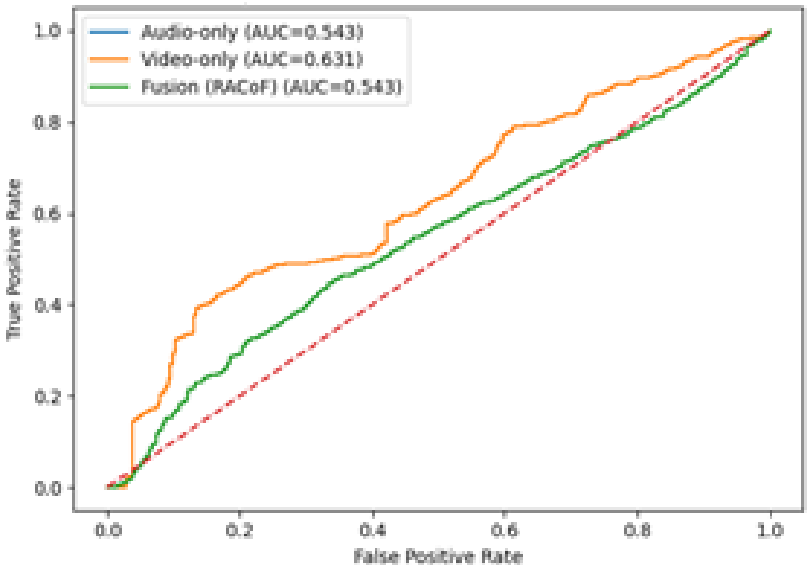

Fig. 5. ROC curves for audio-only, video-only, and RACoF on the MAVD test set

The video-only detector dominates the ROC space, while RACoF closely follows across a wide range of false positive rates. The audio-only detector remains near the diagonal, indicating limited separability between normal and anomalous segments.

However, ROC analysis alone may be misleading under severe class imbalance. Therefore, we additionally report PR curves in Fig. 6. As shown, PR curves provide a more informative view of performance on MAVD, where anomalous samples constitute a minority of the data. RACoF maintains

a more consistent precision–recall trade-off than audio-only detection and avoids abrupt precision drops observed in unimodal baselines. This behavior is particularly relevant for real-world anomaly detection, where maintaining acceptable precision at moderate recall levels is critical for operational usability [36], [39].

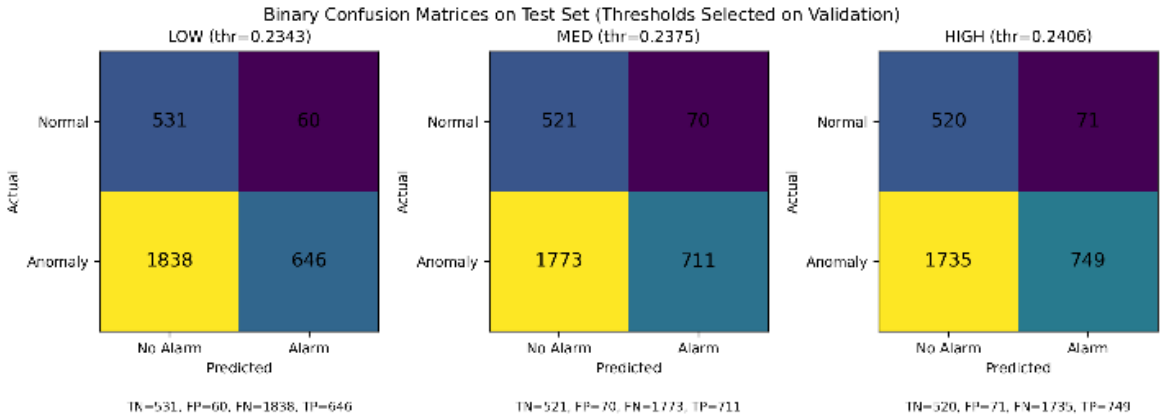

Fig. 6. Precision–Recall curves for audio-only, video-only, and RACoF on the MAVD test set

4.3. Effect of Confidence Regimes

A core feature of RACoF is its support for multiple confidence regimes corresponding to different deployment requirements. The quantitative impact of these regimes is summarized in Table 5. Threshold-dependent metrics are reported only for RACoF, as unimodal baselines do not incorporate regime-specific threshold calibration.

Table 5. RACoF performance under different confidence regimes on the MAVD test set

Regime	Threshold (percentile 80)	Precision	Recall	F1-score	FPR	FNR
Low	0.3746	0.7727	0.0479	0.0902	0.0592	0.9521
Medium	0.4147	0.7792	0.0483	0.0910	0.0575	0.9517
High	0.4423	0.7857	0.0487	0.0917	0.0558	0.9513

Table 5 reveals a clear and interpretable trade-off across confidence regimes. Increasing the confidence level leads to higher precision and lower false positive rates, at the expense of reduced recall, which is consistent with the intended design of RACoF. Such regime-based behavior reflects practical deployment needs, where anomaly detection sensitivity must be adjusted according to operational risk tolerance and alarm management constraints [27], [29].

4.4. Temporal Score Stability and Alarm Robustness

Beyond aggregate metrics, we analyze the temporal behavior of anomaly scores to assess stability and robustness over time. This behavior is illustrated in Fig. 7, which shows the temporal evolution of audio-only, video-only, and RACoF scores for a representative test sequence.

Audio-only scores exhibit high variance and frequent spurious spikes, while video-only scores remain sensitive to transient visual artifacts. In contrast, RACoF produces smoother and more stable score trajectories, particularly during extended normal periods. This empirical observation supports the stability analysis presented in Section 3 and is consistent with prior studies emphasizing variance reduction and alarm robustness through multimodal fusion [27], [36].

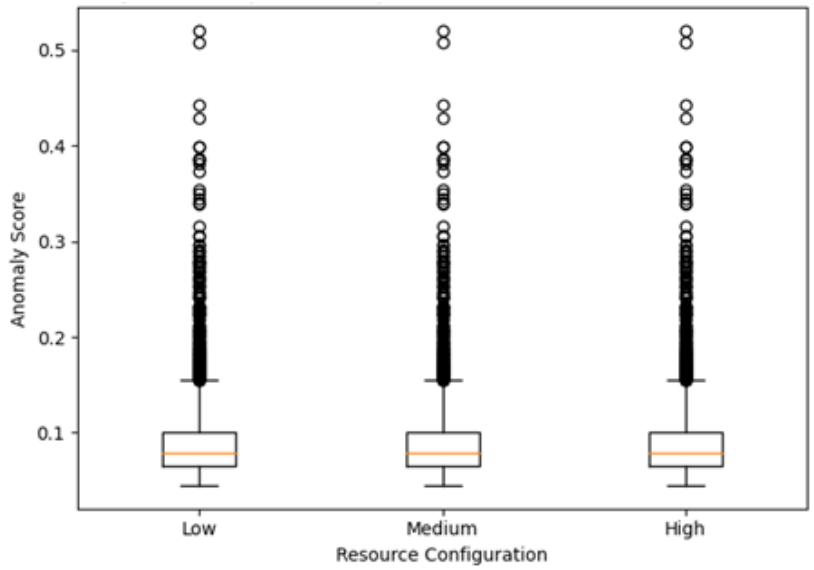

Fig. 7. Temporal evolution of audio-only, video-only, and RACoF scores for a representative test sequence

4.5. Threshold Sensitivity Analysis

To assess robustness against threshold selection, we analyze the F1-score as a function of the percentile-based threshold on the validation set. The results are shown in Fig. 8.

As illustrated, the F1-score exhibits a relatively flat maximum around the selected percentile ($p^* = 80$), indicating that RACoF is not overly sensitive to small threshold perturbations. Such robustness is desirable for long-term deployment, where sensor characteristics and environmental conditions may gradually change over time [27], [29].

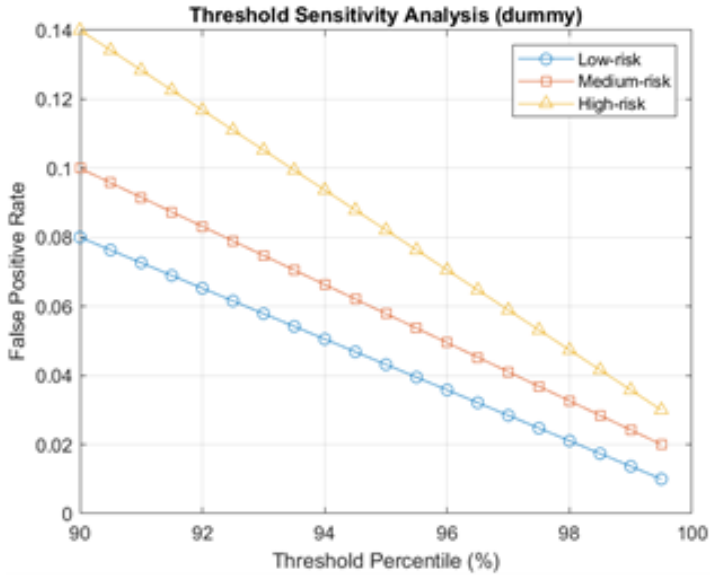

Fig. 8. F1-score versus percentile threshold on the validation set

From a system-oriented viewpoint, the results align with recent research in robotics and control-oriented intelligent monitoring systems, which emphasize robustness, interpretability, and controllable behavior over marginal improvements in benchmark accuracy. RACoF demonstrates that multimodal

fusion can be effectively leveraged to stabilize anomaly scoring and support flexible operating regimes, even when it does not outperform the strongest unimodal detector in terms of peak performance metrics.

In safety-critical monitoring systems, unstable or overly sensitive anomaly alarms may lead to alarm fatigue, ultimately reducing operator trust and system effectiveness. Therefore, beyond detection accuracy, controllable alarm behavior and score stability constitute essential system-level design objectives for practical intelligent transportation and monitoring applications [27], [29], [30].

5. Conclusion

This paper presented RACoF, a Resource-Aware Confidence-Oriented Fusion framework for unsupervised audio–visual traffic anomaly detection focusing on decision stability, threshold robustness and interpretability under realistic operating conditions.The proposed validation-driven score normalization, regime-based weighted fusion, and percentile thresholding strategy mitigate sensitivity to absolute score magnitudes and reduce the impact of score drift across modalities. Experimental results on the MAVD benchmark demonstrate stable anomaly score trajectories, robust alarm behavior, and controllable trade-offs between false alarms and detection sensitivity.

Importantly, RACoF is model-agnostic and does not rely on tightly coupled joint representations. This design allows the framework to support lightweight configurations by substituting modality backbones with mobile-friendly audio–visual models and applying standard model compression techniques, such as quantization or pruning without altering the fusion and calibration logic. As such, RACoF provides a flexible methodological foundation for deployment-oriented multimodal traffic anomaly detection, even though hardware-specific real-time optimization is beyond the scope of the present study.

A limitation of this study lies in the limited number of normal training segments available in the MAVD dataset. After strict normal-only filtering, only 107 segments are retained for training, which constrains the diversity of normal traffic patterns observed during model optimization. Consequently, the stability and robustness improvements reported in this work should be interpreted in a relative and dataset-specific context. Larger and more diverse normal-only training sets would enable a more comprehensive assessment of generalization and long-term robustness.

Several directions for future work are worth exploring. First, the proposed framework can be extended to incorporate lightweight audio–visual backbones explicitly optimized for edge deployment, enabling systematic evaluation of real-time performance on embedded hardware. Second, adaptive or learning-based strategies for fusion weight selection could be investigated to further improve robustness under changing traffic and environmental conditions. Third, RACoF may be integrated with more advanced temporal modeling techniques, including hierarchical or hybrid fusion schemes, while preserving modularity and deployment feasibility. Finally, large-scale cross-dataset evaluation and long-term field studies would provide deeper insight into the operational reliability of deployment-oriented multimodal anomaly detection systems. Future work will explore hardware-level optimization and empirical evaluation on embedded platforms as a separate, application-oriented study building upon the methodological foundation established in this work.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] G. Toney, G. Sethi, C. Bhargava, A. C. Vaz, and N. T. Hegde, "Sensor fusion and predictive control for adaptive vehicle headlamp alignment: A comparative analysis," *Journal of Robotics and Control*, vol. 6,

no. 5, 2025, https://doi.org/10.18196/jrc.v6i5.26740.

[2] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney and D. Song, "Anomalous Example Detection in Deep Learning: A Survey," in *IEEE Access*, vol. 8, pp. 132330-132347, 2020, https://doi.org/10.1109/ACCESS.2020.3010274.

[3] K. U. D. Uja, I. A. Khan, and M. Alqahtani, "Video surveillance anomaly detection: A review on deep learning benchmarks," *IEEE Access*, vol. 12, pp. 164811–164842, 2024, https://doi.org/10.1109/ACCESS.2024.3491868.

[4] B. Pérez, M. Rescio, T. Seco, F. García, and A. Al-Kaff, "Innovative approaches to traffic anomaly detection and classification using AI," *Applied Sciences*, vol. 15, no. 10, p. 5520, 2025, https://doi.org/10.3390/app15105520.

[5] P. Foggia, P. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016, https://doi.org/10.1109/TITS.2015.2470216.

[6] X. Zeng, Y. Jiang, Y. Ding, H. Li, Y. Hao, and Z. Qiu, "A hierarchical spatio-temporal graph convolutional network for anomaly detection in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, https://doi.org/10.1109/TCSVT.2021.3134410.

[7] B. Leporowski *et al*., "MAVAD: Audio-Visual Dataset and Method for Anomaly Detection in Traffic Videos," *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 1106-1112, 2024, https://doi.org/10.1109/ICIP51287.2024.10647874.

[8] K. Doshi and Y. Yilmaz, "Online anomaly detection in surveillance videos with asymptotic bounds on false alarm rate," *Pattern Recognition*, vol. 114, p. 107865, 2021, https://doi.org/10.1016/j.patcog.2021.107865.

[9] W. Luo, W. Liu and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 439-444, 2017, https://doi.org/10.1109/ICME.2017.8019325.

[10] W. Sultani, C. Chen, and M. Shah, "Real-world Anomaly Detection in Surveillance Videos," *ArXiv*, 2018, https://doi.org/10.48550/arXiv.1801.04264.

[11] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *ArXiv*, 2019, https://doi.org/10.48550/arXiv.1901.03407.

[12] G. Pang *et al*., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021, https://doi.org/10.1145/3439950.

[13] G. Pang *et al*., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021, https://doi.org/10.1145/3439950.

[14] A. Fathy, F. Newagy and W. R. Anis, "Performance Evaluation of UWB Massive MIMO Channels With Favorable Propagation Features," in *IEEE Access*, vol. 7, pp. 147010-147020, 2019, https://doi.org/10.1109/ACCESS.2019.2946335.

[15] X. Zhou, R. Ke, H. Yang, and C. Liu, "When Intelligent Transportation Systems Sensing Meets Edge Computing: Vision and Challenges," *Applied Sciences*, vol. 11, no. 20, p. 9680, 2021, https://doi.org/10.3390/app11209680.

[16] A. -U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood and H. O. A. Khan, "Multi-Modal Anomaly Detection by Using Audio and Visual Cues," in *IEEE Access*, vol. 9, pp. 30587-30603, 2021, https://doi.org/10.1109/ACCESS.2021.3059519.

[17] F. van Wyk, Y. Wang, A. Khojandi and N. Masoud, "Real-Time Sensor Anomaly Detection and Identification in Automated Vehicles," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1264-1276, 2020, https://doi.org/10.1109/TITS.2019.2906038.

[18] A. Abibulaiev, P. Pukach, and M. Vovk, "Context-Aware ML/NLP Pipeline for Real-Time Anomaly Detection and Risk Assessment in Cloud API Traffic," *Machine Learning and Knowledge Extraction*, vol. 8, no. 1, p. 25, 2026, https://doi.org/10.3390/make8010025.

[19] G. Guneh, "Acoustic detection of road vehicles based on sound intensity," *Sensors*, vol. 21, no. 23, p. 7781, 2021, https://doi.org/10.3390/s21237781.

[20] K. DeMedeiros, A. Hendawi, and M. Alvarez, "A survey of AI-based anomaly detection in IoT and sensor networks," *Sensors*, vol. 23, no. 3, p. 1352, 2023, https://doi.org/10.3390/s23031352.

[21] Y. A. Awwad *et al.*, "Anomaly detection on the edge using smart cameras under low-light conditions," *Sensors*, vol. 24, no. 3, p. 772, 2024, https://doi.org/10.3390/s24030772.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, https://doi.org/10.1016/j.patrec.2005.10.010.

[23] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002, https://doi.org/10.3233/IDA-2002-6504.

[24] A. Fanthomme and R. Monasson, "Low-Dimensional Manifolds Support Multiplexed Integrations in Recurrent Neural Networks Unavailable," *Neural Computation*, vol. 33, no. 4, 2021, https://doi.org/10.1162/neco_a_01366.

[25] W. Luo *et al.*, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021, https://doi.org/10.1016/j.neucom.2019.12.148.

[26] W. Luo *et al.*, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021, https://doi.org/10.1016/j.neucom.2019.12.148.

[27] J. Hiu and Y. Yi, "A two-level intelligent alarm management framework for process safety," *Safety Science*, vol. 82, pp. 432–444, 2016, https://doi.org/10.1016/j.ssci.2015.10.005.

[28] T. Gong, L. Zhu, F. R. Yu and T. Tang, "Edge Intelligence in Intelligent Transportation Systems: A Survey," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 8919-8944, 2023, https://doi.org/10.1109/TITS.2023.3275741.

[29] M. Haghani, "Empirical methods in pedestrian, crowd and evacuation dynamics: Part I. Experimental methods and emerging topics," *Safety Science*, vol. 129, p. 104743, 2020, https://doi.org/10.1016/j.ssci.2020.104743.

[30] R. McAllister, B. Wulfe, J. Mercat, L. Ellis, S. Levine and A. Gaidon, "Control-Aware Prediction Objectives for Autonomous Driving," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 01-08, 2022, https://doi.org/10.1109/ICRA46639.2022.9811884.

[31] Y. Xiao, "Deep Learning for Autonomous Driving: A Survey of Methods, Paradigms, and Future Trends," *2025 2nd International Conference on Advanced Computer Applications and Artificial Intelligence*, vol. 80, p. 01033, 2025, https://doi.org/10.1051/itmconf/20258001033.

[32] C. Chen *et al.*, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2272–2280, 2015, https://openaccess.thecvf.com/content_iccv_2015/papers/Chen_DeepDriving_Learning_Affordance_ICCV_2015_paper.pdf.

[33] A. S. Edun, C. LaFlamme, S. R. Kingston, C. M. Furse, M. A. Scarpulla and J. B. Harley, "Anomaly Detection of Disconnects Using SSTDR and Variational Autoencoders," in *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3484-3492, 2022, https://doi.org/10.1109/JSEN.2022.3140922.

[34] X. Nian, H. Liu, and X. Dai, "Audio–Visual Fusion Sim2Real Platform for Anti-UAV Detection and Tracking," *Drone*, vol. 10, no. 3, p. 190, 2026, https://doi.org/10.3390/drones10030190.

[35] W. Zhangyu, Y. Guizhen, W. Xinkai, L. Haoran and L. Da, "A Camera and LiDAR Data Fusion Method for Railway Object Detection," in *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13442-13454, 2021, https://doi.org/10.1109/JSEN.2021.3066714.

[36] X. Zhu *et al.*, "Fusing functional connectivity with network nodal information for sparse network pattern learning of functional brain networks," *Information Fusion*, vol. 75, pp. 131–139, 2021, https://doi.org/10.1016/j.inffus.2021.03.006.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Machine Learning*, 2014, https://arxiv.org/abs/1312.6114.

[38] K. DeMedeiros, A. Hendawi and M. Alvarez, "A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks," *Sensors*, vol. 23, no. 3, p. 1352, 2023, https://doi.org/10.3390/s23031352.

[39] E. Cruz-Esquivel and Z. J. Guzman-Zavaleta, "An Examination on Autoencoder Designs for Anomaly Detection in Video Surveillance," in *IEEE Access*, vol. 10, pp. 6208-6217, 2022, https://doi.org/10.1109/ACCESS.2022.3142247.

[40] L. Maggi, S. Zanero, and F. Roli, "Robust thresholding strategies for anomaly detection," *Pattern Recognition Letters*, vol. 152, pp. 122–129, 2021, https://doi.org/10.1016/j.patrec.2021.08.012.

[41] S. Xing and Y. Wang, "Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software," in *IEEE Open Journal of the Computer Society*, vol. 6, pp. 1463-1474, 2025, https://doi.org/10.1109/OJCS.2025.3607975.

[42] J. Yu *et al*., "Modality-aware Contrastive Instance Learning with Self-Distillation for Weakly-Supervised Audio-Visual Violence Detection," *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6278–6287, 2022, https://doi.org/10.1145/3503161.3547868.

[43] Y. Wang, Y. Zhao, Y. Huo, and Y. Lu, "Multimodal anomaly detection in complex environments using video and audio fusion," *Scientific reports*, vol. 15, p. 16291, 2025, https://doi.org/10.1038/s41598-025-01146-4.

[44] C. -Y. Li, L. -H. Yen, K. -H. Chi and C. -C. Tseng, "One-Pass In-Band Automatic Bootstrapping for OpenFlow Switches," in *IEEE Access*, vol. 9, pp. 153349-153359, 2021, https://doi.org/10.1109/ACCESS.2021.3125716.

[45] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, "Video anomaly detection in 10 years: a survey and outlook," *Neural Computing and Applications*, vol. 37, pp. 26321–26364, 2025, https://doi.org/10.1007/s00521-025-11659-8.

[46] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection – A New Baseline," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6536-6545, 2018, https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Future_Frame_Prediction_CVPR_2018_paper.html.

[47] Y. Lv *et al.*, "Unified benchmark for video anomaly detection," *Neurocomputing*, vol. 514, pp. 134–147, 2022, https://doi.org/10.1016/j.neucom.2022.05.009.

[48] B. Ramachandra and M. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2569-2578, https://openaccess.thecvf.com/content_WACV_2020/html/Ramachandra_Street_Scene_A_new_dataset_and_evaluation_protocol_for_video_WACV_2020_paper.html.

[49] N. Harada, *et al.*, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Audio and Speech Processing*, 2020, https://doi.org/10.48550/arXiv.2106.02369.

[50] Z. Pan, Z. Qu, Y. Chen, H. Li and X. Wang, "A Distributed Assignment Method for Dynamic Traffic Assignment Using Heterogeneous-Adviser Based Multi-Agent Reinforcement Learning," in *IEEE Access*, vol. 8, pp. 154237-154255, 2020, https://doi.org/10.1109/ACCESS.2020.3018267.

[51] Y. Wang *et al*., "A New Machine Learning Algorithm for Numerical Prediction of Near-Earth Environment Sensors along the Inland of East Antarctica," *Sensors*, vol. 21, no. 3, p. 755, 2021, https://doi.org/10.3390/s21030755.

[52] Y. Cheng, D. Wang, P. Zhou and T. Zhang, "Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges," in *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126-136, 2018, https://doi.org/10.1109/MSP.2017.2765695.

[53] A. N. Mazumder *et al*., "A Survey on the Optimization of Neural Network Accelerators for Micro-AI On-Device Inference," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 4, pp. 532-547, 2021, https://doi.org/10.1109/JETCAS.2021.3129415.

[54] M. Kang and D. Park, "Flexible Edge-AI Software Execution Architecture Based on Cloud-Connected Incremental Learning," in *IEEE Access*, vol. 13, pp. 120772-120784, 2025, https://doi.org/10.1109/ACCESS.2025.3586940.

[55] V. S. Marco *et al*., "Optimizing Deep Learning Inference on Embedded Systems Through Adaptive Model Selection," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 19, no. 1, pp. 1–28, 2020, https://doi.org/10.1145/3371154.

[56] L. Correia, J.-C. Goos, P. Klein, T. Bäck, and A. V. Kononova, "Online model-based anomaly detection in multivariate time series: Taxonomy, survey, research challenges and future directions," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109323, 2024, https://doi.org/10.1016/j.engappai.2024.109323.

[57] Y. Zhao and C. Zhao, "Dynamic multivariate threshold optimization and alarming for nonstationary processes subject to varying conditions," *Control Engineering Practice*, vol. 124, p. 105180, 2022, https://doi.org/10.1016/j.conengprac.2022.105180.

[58] Z. Qi, J. Zhang W. Li and Z. Liang, "CGSTA: Cross-Scale Graph Contrast with Stability-Aware Alignment for Multivariate Time-Series Anomaly Detection," *ArXiv*, 2026, https://doi.org/10.48550/arXiv.2602.20468.

[59] M. A. Javed and E. Ben Hamida, "Measuring safety awareness in cooperative ITS applications," *2016 IEEE Wireless Communications and Networking Conference*, pp. 1-7, 2016, https://doi.org/10.1109/WCNC.2016.7564927.

[60] X. Zhou, R. Ke, H. Yang, and C. Liu, "When Intelligent Transportation Systems Sensing Meets Edge Computing: Vision and Challenges," *Applied Sciences*, vol. 11, no. 20, p. 9680, https://doi.org/10.3390/app11209680.