Minimax Theory and its Applications
Volume 11 (2026), No. 1s, 589-600

Minimax-Based Intelligent Multimodal Surveillance for Real-Time Threat Detection: A Structured Review

Sumaiya Samreen[1], Dr. Priyadarshini C. Patil[2]

[1]Research Scholar, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi, Karnataka, India, affiliated with Visvesvaraya Technological University, Belagavi – 590018, Karnataka, India. Email: sumaiyasamreenriaz@gmail.com, ORCID: 0000-00001-6741-6703
[2] Associate Professor, Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi, Karnataka, India. Email: priyadarshinicpatil@gmail.com, ORCID: https://orcid.org/0000-0002-9465-5723

Abstract

Surveillance systems now generate far more video than any control room can watch, and the research response has been to add modalities. Audio, pose, optical flow, depth, language, hyperspectral bands and trajectory data have all been folded into threat-detection pipelines under the single banner of multimodal surveillance. This review argues that the banner conceals four largely separate research programmes with different assumptions, different benchmarks and different notions of what counts as success. It reads twenty-one studies published in IEEE journals between 2025 and 2026 against that split, grouping them by modality composition, by where fusion happens, and by supervision regime. Three findings come out of that reading. Audio, the modality most often invoked to justify the label, is genuine in only one public benchmark, and several leading systems synthesise it, which makes the reported audio contribution partly an artefact of the synthesiser. Real time is claimed by most of the reviewed work but measured at incompatible boundaries, with feature extraction excluded from the latency budget often enough that cross-paper timing comparison is close to meaningless. And modality degradation, the failure mode that determines whether a fused system survives deployment, is tested by two studies in twenty-one. The review organises the corpus into three comparative tables and closes on research gaps written as small concrete experiments, not as a call for a unified architecture.

Keywords**:** multimodal surveillance, video anomaly detection, audio-visual fusion, vision-language models, weakly supervised learning, edge inference, real-time threat detection, sensor fusion.

1. Introduction

A modern transport hub, airport apron or city centre is watched by hundreds of cameras and, in practice, by a handful of people. The arithmetic is unforgiving. Attention degrades within about twenty minutes of continuous monitoring, and the events that matter are short. A fight in a bus lasts a few seconds. Automated threat detection is not an efficiency measure here; it is the only way the installed camera base can be used at all.

The research response has moved through three phases. The first was reconstruction-based, treating anomaly as high reconstruction error under a model of normality [1], [2]. The second was weakly supervised, learning from video-level labels through multiple instance learning after Sultani et al. released UCF-Crime [3] and showed that frame-level annotation was not a prerequisite; robust temporal feature magnitude learning [4] and the magnitude-contrastive glance-and-focus design [5] refine that formulation. The third phase, where the current literature sits, adds modalities. Wu et al. released XD-Violence with audio tracks [6] and showed that sound carried information the visual stream did not. Transformer architectures [7], [8] made cross-modal attention cheap to write. Contrastive vision-language pretraining [9] made text usable without paired annotation, and VadCLIP [10] and open-vocabulary anomaly detection [11] carried that shift into surveillance.

The trouble is that the resulting body of work does not share a problem. A paper fusing audio and video on XD-Violence, a paper fusing RGB and point clouds on a Jetson board, a paper prompting a frozen language model to explain what a drone is looking at, and a paper fusing scream spectra with speech semantics in a railway station all call themselves multimodal surveillance. They answer different questions: whether an auxiliary signal improves a benchmark score, whether two sensor streams can be aligned within a power budget, whether a general-purpose model can explain a scene without training, and whether an alarm can be raised early enough to matter. Reading them as one literature gives a misleading picture of progress, because a gain in one programme says almost nothing about the others.

1.1 Motivation and Contribution

The immediate motivation is that the word multimodal has stopped discriminating between systems. It covers a network that converts audio into green-channel brightness modulation of an RGB frame before any learning happens

[12], a network that runs bidirectional prediction-level distillation between an audio branch and a visual branch [13], and a network whose three streams all derive from the same RGB video [14]. Those are three different claims about what fusion is for, not three variations on one technique.

A second motivation is less visible from inside any one paper. Almost every reviewed system claims real-time suitability, and almost none measures it the same way. Latency is reported on pre-extracted features in one study [15] and end to end on an embedded server in another [16]; frame rate covers the scoring head alone in one case [14] and the full pipeline in another [17]. Because the boundary is rarely stated in the abstract, a reader comparing sixteen milliseconds against three seconds is often comparing two different things.

A third motivation concerns evidence quality. Several of the strongest results here rest on private data [18], [19], [20], on closed commercial models used inside the method [21], or on synthesised inputs standing in for a missing modality [13], [15], and the papers vary in how plainly they say so.

The contributions of this review are five.

1. The reviewed methods are organised along three orthogonal axes, modality composition, fusion locus and supervision regime, in place of the single label the literature uses. Section 3 sets out the taxonomy and three comparative tables apply it.
2. Audio, language, depth and sensor-level modalities are treated as separate research programmes with their own benchmarks and failure modes, not as interchangeable additions to a visual backbone. That separation is what lets the review say which multimodal claims the evidence supports.
3. Deployment and latency evidence is compared explicitly, together with the boundary at which each study measured. Where a figure excludes feature extraction or came from a workstation GPU, this review says so instead of passing the number along.
4. The research gap is set out as specific measurement failures: synthesised audio, incompatible latency boundaries, untested modality degradation, benchmark monoculture, metrics that do not track operator utility, and unquantified hallucination in language-based systems.
5. Future directions are attached to each gap as small experiments, such as a modality-dropout protocol and a declared latency boundary. No integrated architecture is proposed.

2. Review Methodology

The corpus is twenty-one primary studies published in IEEE journals between January 2025 and August 2026, drawn from twelve titles including IEEE Access, the Transactions on Image Processing, on Circuits and Systems for Video Technology, on Intelligent Transportation Systems and on Industrial Informatics, together with the Internet of Things Journal and the Sensors Journal. Studies were included when they addressed detection, recognition or retrieval of threat-relevant events in a surveillance setting, reported results on a stated dataset, and either used more than one input modality or made an explicit real-time claim. Thirteen foundational works were added where the primary sources depended on them, taken from the reference sections of the corpus itself so that the background reflects what this literature builds on.

Each study was read for a fixed set of fields: the modalities genuinely used at inference, where fusion occurs, the supervision regime, the datasets, the headline metric with its exact value, every efficiency figure together with the hardware and the measurement boundary, and the limitations the authors acknowledge. Numbers quoted here come from the running text of the primary sources; where a paper reports inconsistent values in different sections, which happened in five cases, the results-section value is used.

3. A Taxonomy of Multimodal Surveillance for Threat Detection

The first axis is modality composition, and four families are visible. Audio-visual systems pair a visual backbone with an acoustic one, usually VGGish [22], and are almost entirely evaluated on XD-Violence [6]. Vision-derived multi-stream systems use several streams that all originate in the same RGB video: pose, optical flow, monocular depth, scene logits, frame differences. Vision-language systems introduce text, as CLIP-style class embeddings [9], as generated captions, or as the output of a large language model. Sensor-level systems use physically distinct sensors, such as an RGB-D camera, a hyperspectral imager, or a microphone array coupled to a speech 590ocalizati.

The second family deserves a note. A pose stream and a flow stream carry different information, but they carry it from the same photons, so if the camera is blinded by glare both fail together. An RGB-D or audio-visual pair does not have that property, which is precisely the robustness argument usually offered for multimodal design. Several papers in the second family make that argument anyway, and their evidence does not support it. Real time, meanwhile, is not an axis at all, and that is part of the problem: it gets asserted in the abstract when it should constrain all three choices.

The second axis is fusion locus, and four positions appear. Input-level fusion collapses modalities into one tensor before any learning, the clearest instance being MFCC energy encoded as green-channel brightness in an ordinary RGB frame [12]. Feature-level fusion, by far the most common, combines intermediate representations by concatenation, cross-attention or gating [13], [15], [23], [24]. Decision-level fusion combines per-modality scores, by a weighted sum [25] or a small learned combiner over thresholded confidences [20]. A fourth position is architectural, not representational: some systems split modality handling across edge and cloud, which turns the fusion decision into a decision about what to transmit [18], [19].

The third axis is supervision. Weakly supervised multiple instance learning on video-level labels dominates the anomaly-detection work [13], [14], [15], [26]. One-class classification, training only on normality, remains standard for pedestrian anomaly [25]. Full supervision is standard in the violence, activity-recognition and object-detection work [16], [17], [23], [24], [27], [28]. Self-supervised pretraining with weak fine-tuning appears once [29], and training-free operation, where a frozen model is prompted and never fitted, in the aerial anomaly work [21] and, in a purely statistical form, in hyperspectral detection [30].

4. Related Work

4.1 Audio-Visual Fusion Under Weak Supervision

The audio-visual line is the most technically mature part of this literature and the most exposed to a benchmark problem. Kim et al. propose ASTRA [13], which distils prediction-level anomaly confidence in three directions between a frozen CLIP visual branch and a VGGish audio branch, the fused-to-visual direction chosen deliberately because audio is the channel more likely to be noisy or absent. The design reaches 86.38% AP on XD-Violence. The visual-only inference variant reaches 85.94%, a drop of 0.44 points, and that small gap is the most informative number in the paper. It cuts both ways. The distillation genuinely transferred something into the visual branch, and the audio channel contributes very little at inference.

The same group's later work, SCV [15], reframes fusion as calibration, not mixing, admitting audio through a framewise similarity gate before reintegrating what the gate suppressed. Reported AP on XD-Violence is 88.80%. The category breakdown matters more than that aggregate. Audio alone reaches 87.7% AP on riot and 19.1% on abuse; fused, those become 97.5% and 73.4%. The gain sits almost entirely in acoustically distinctive categories, so the headline figure hides how uneven the benefit is.

Both systems hit the same obstacle. UCF-Crime [3] has no audio, ShanghaiTech [1] has none, and neither does CUHK Avenue [2]. XD-Violence [6] is effectively the only public surveillance benchmark with genuine 591ocalization sound. To evaluate cross-domain, both ASTRA and SCV synthesise audio for UCF-Crime, and both acknowledge the result is coarse and partly noisy, so the reported cross-domain performance is partly a measurement of the 591ocalizatio. ASTRA is unusually candid on a related point, stating it has no modality reliability estimation and no visual-only fallback, so mismatched audio degrades the fused prediction in some categories.

Jiang et al. take the opposite route [12]. No fusion network is built at all. Both modalities collapse into a single RGB image before the network sees anything. A shifted-frame edge difference encodes the video; averaged MFCC values modulate green-channel brightness to encode the audio. Reported AP on XD-Violence is 89.64%, the highest coarse-grained figure in this corpus, and AUC on UCF-Crime is 88.35%. The argument behind the design is that RGB colour actively misleads violence detectors, which attend to dark pixels instead of motion, and their class activation maps support this. Two caveats belong with the number. On UCF-Crime the multimodal claim does not hold, since there is no audio and the transformation runs in one stage. And despite framing the distilled student as deployment-ready, the paper reports no parameter count, no FLOPs and no latency.

4.2 Vision-Derived Multi-Stream Fusion

The second family builds several streams from one camera. Ahn et al. make the clearest argument for why [14]. Their position is that anomaly is a relational property of scene context and fine-grained action, not a property of the action on its own, so a model should answer where and what together. They combine SlowFast atomic action features, optical flow from a dual-stream I3D [31], and Places365 scene logits, attending within each stream and then across feature tokens, reaching 87.9% AUC on UCF-Crime and 98.2% on ShanghaiTech. The modality ablation is the useful part: action alone gives 80.6% on UCF-Crime, optical flow adds 4.5 points, and scene adds 0.6, though scene is what suppresses false positives from viewpoint change, which an aggregate AUC hides. Their 416 FPS covers the scoring head only; SlowFast, TV-L1 flow and the scene classifier all sit upstream.

Aung et al. push stream derivation further by discarding appearance altogether [27], combining pose skeletons re-rendered on a black background with frame differences to reach 95.05% on RWF-2000 with 118,755 parameters. The authors are honest that discarding appearance costs them the ability to separate violence from playful gestures, exactly the confusion an operator would most want resolved. Their 0.3325 s per instance excludes preprocessing, and preprocessing means running a pose estimator over fifty frames.

Wang et al. address the one-class setting, where only normal data is available at training time [25]. Velocity and appearance streams are contracted towards Gaussian-mixture density centres, scored separately and combined by a weighted sum, reaching 99.3% micro AUROC on Ped2, 93.7% on Avenue and 86.4% on ShanghaiTech. The velocity-only variant, which uses no RGB and is presented as privacy-preserving, reaches 99.3%, 89.1% and 85.2%, and that is the more interesting result: it shows how much of the task motion geometry carries alone. The paper reports no efficiency figures at all, despite a pipeline chaining Mask R-CNN, FlowNet2.0, ResNet-50, two contraction networks and two density estimators.

Three further studies here should be read with more care. Oubelkas et al. present SST-Net [29], which learns violence representations without labels or augmentations by autoregressively predicting latent frame transitions under causal masking, reporting 98.16% on a self-assembled corpus and 95.56% on 120 unseen real surveillance videos. The ablation should give a reader pause: removing the background suppression step drops accuracy to 76.4%, so a preprocessing heuristic is doing more work than the self-supervised objective. Mohamad et al. [26] decouple spatial

and temporal modelling at 7.9M parameters and 0.71 GFLOPs; the efficiency case is genuine, the evaluation is not, since headline figures differ across the abstract, contributions, results and conclusion, and the claimed fine-grained 592ocalization is never measured with a 592ocalization metric. Alshalawi et al. [32] report 99% accuracy for a vision transformer, but the evaluation is frame-level image classification on a Kaggle image set, so the number is not comparable to the video benchmarks the paper cites against it.

Table 1. Audio-visual and vision-derived multi-stream methods for threat detection.

Reference	Modalities	Methodology	Key reported result	Shortcomings
[12] Jiang et al., 2025	RGB video, audio (encoded into the image)	Audio and video collapsed into a single RGB image before learning; DenseNet extractor with transformer encoder; distilled into a two-convolution student; class activation maps for interpretability	89.64% AP on XD-Violence; 88.35% AUC on UCF-Crime	No parameter count, FLOPs or latency reported despite a deployment claim; on UCF-Crime the audio stage does not run, so the multimodal claim does not hold there
[13] Kim et al., 2026	RGB video, audio, text prompts	Frozen CLIP and VGGish encoders; local transformer plus residual GCN; three-way prediction-level knowledge distillation; contrastive semantic alignment across modalities	86.38% AP on XD-Violence; 85.94% AP visual-only; 87.71% AUC on UCF-Crime	Gain over strongest audio-visual baseline only 0.31 points; UCF-Crime audio synthesised; authors state there is no modality reliability estimation or fallback
[14] Ahn et al., 2025	Atomic action, optical flow, scene logits (all RGB-derived)	Per-stream temporal self-attention followed by semantic self-attention across feature tokens; segment-selective focal margin loss over top-K segments	87.9% AUC on UCF-Crime; 98.2% on ShanghaiTech; 51M parameters, 98.7 GFLOPS	416 FPS covers the scoring head only, excluding SlowFast, TV-L1 flow and the scene classifier; recall gain comes at a 12-point precision cost; all three streams share one camera
[15] Kim et al., 2026	RGB video, audio, text prompts	Two-stage adaptive fusion with a cosine-similarity admission gate followed by progressive audio reintegration; multiscale fusion prompts; hierarchical semantic alignment	88.80% AP and 35.40% average mAP on XD-Violence; 91.21% AUC cross-domain on refined UCF-Crime	Latency and FLOPs measured on pre-extracted features only; two underlying benchmarks; audio genuine in one of them; loses to a weaker baseline on the abuse category
[25] Wang et al., 2026	Optical flow velocity, RGB scene appearance	Gaussian mixture fitted to features; autoencoder contracts each feature to its nearest high-density centre; separate likelihood and nearest-neighbour scoring fused by a fixed weight	99.3% / 93.7% / 86.4% micro AUROC on Ped2, Avenue and ShanghaiTech; 99.3 / 89.1 / 85.2 velocity-only	No efficiency figures at all for a five-stage pipeline; mixture component count tuned on the test benchmarks; all three datasets are campus pedestrian scenes; Ped2 is saturated
[26] Mohamad et al., 2026	RGB video only	Motion-based frame selection; pretrained NASNetMobile turns each frame into one token; four-block vision transformer with query-adaptive pooling models time	91.55% AUC on UCF-Crime; 7.9M parameters, 0.71 GFLOPs	Headline figures inconsistent across abstract, contributions, results and conclusion; claimed fine-grained localisation never measured with a localisation metric; no training details given

Reference	Modalities	Methodology	Key reported result	Shortcomings
[27] Aung et al., 2025	Pose skeletons, frame differences	Dual-stream ConvLSTM; skeletons re-rendered on a black background; activation-weighted element-wise sum fusion with ReLU on pose and Tanh on motion	95.05% on RWF-2000; 98.50% Hockey; 118,755 parameters	Appearance discarded, so violence and play remain confusable by the authors' own admission; 0.33 s inference excludes running a pose estimator over fifty frames; one Crowd-dataset fold collapses to 82%
[29] Oubelkas et al., 2026	RGB video only	Gaussian mixture background suppression; ResNet-18 encoder; causally masked transformers predict latent frame transitions autoregressively; light classifier fine-tuned on frozen embeddings	98.16% on a custom corpus; 95.56% on 120 unseen real surveillance videos; 95 ms per ten-frame sequence	Headline result on a self-assembled non-public corpus with 162 test videos; removing the background-suppression heuristic drops accuracy to 76.4%, so preprocessing dominates
[32] Alshalawi et al., 2025	RGB still frames only	Comparative benchmark of VGG16, MobileNetV2, a feature-concatenation ensemble, YOLOv6 and a pretrained vision transformer on single frames	99% accuracy and F1 for the transformer on a violence image set; 98% on a composite road-anomaly set	Frame-level image classification, not video, despite the title; no standard video violence benchmark used; no measured latency, only a projection; composite anomaly set confounds domain with label

4.3 Language as a Modality

The vision-language line is the newest and most unsettled. Yuan et al. make the enabling contribution [33], arguing that surveillance video understanding has been confined to closed-set classification because no sentence-level annotated surveillance data existed. Their UCA dataset re-annotates 1,854 UCF-Crime videos with 23,542 queries timed to 0.1 s over 110.7 hours, and benchmarks roughly twenty models across five tasks. The headline finding is negative and useful: recall at one for an IoU of 0.3 stays below 10% for all seven grounding baselines. Models that do well on general video-language benchmarks collapse on surveillance footage, which the authors attribute to long durations, fixed viewpoints, low resolution and domain shift.

Li et al. show what a frozen model and careful prompting can do [21]. SemAero removes drone ego-motion through homography estimation, queries a frozen InternVL2-8B with a chain-of-thought prompt conditioned on the scene category, and refines the resulting scores over semantic embeddings and motion anchors from normal training video. Nothing is trained. Mean AUC-ROC reaches 77.99 on Drone-Anomaly and 71.07 on UIT-ADrone. The strongest compared baseline manages 71.22 and 69.56. Their ablation is the clearest demonstration in this corpus of how much prompting matters: the same frozen model with a generic prompt scores 51.84, and scene-specific prompting alone adds 7.33 over a prior-work prompt baseline. The authors concede that the framework does not resolve hallucination and that its cost suits offline monitoring, not onboard use. A third issue they do not stress is that the best variant generates its prompts with a closed commercial API at non-zero temperature, and the reproducible deterministic variant is one to three points worse.

Sadhwani et al. reframe the task itself [34], arguing that for safety-critical surveillance the question is not whether an event occurred but how severe it is. They grade fire, traffic and violence frames by severity using two representations from the same models, generated captions and dense pretext embeddings. Embeddings win consistently: BLIP reaches 94.70%, 98.04% and 86.63% against 85.37%, 94.01% and 76.80% for captions, a useful negative result about the human-readable output these models are usually valued for. The cost appears in their own timing table, where they take 28 to 33 ms per image against 2 to 3 ms for the convolutional baselines. The datasets are self-collected, unnamed and not public.

In two systems, language is a transport problem, not a perception problem. Onsu et al. [18] encode vehicle-centred crops into embeddings at the edge, transmit those over a noisy channel, reconstruct an image in the cloud, and have a tuned LLaVA answer traffic queries. This cuts data volume by roughly 99.9%, and query accuracy falls from 93.2% on original crops to 89% on reconstructed ones. Their finding that uniform quantisation survives channel noise far better than float encoding is worth carrying forward on its own. The framing is weaker than the engineering. The testbed is a synthetic digital twin with an idealised channel, the saving is computed against a raw bitmap and not any video codec, and the per-vehicle latency of 0.183 s scales linearly. Six vehicles exceed a second before the language model is invoked. Meng et al. [19] address retrieval, pushing a fusion model and a dynamic semantic index to the edge

and shipping only a small encrypted candidate set to the cloud, reaching 96% to 100% of brute-force nearest-neighbour accuracy. The 6G framing is not tested on a 6G radio; the testbed is a three-node gigabit local network.

4.4 Sensor-Level and Non-Visual Modalities

The fourth family uses physically distinct sensors, and is where the robustness argument is strongest. Biswas et al. [23] fuse 2-D RGB with voxelised 3-D point clouds from one RGB-D camera through a cross-modality attention module whose scaling factor is learned instead of fixed. They report 98.15% on their own IRMAD dataset under a random frame split. The leave-one-subject-out figure of 95.36% is the one to cite, since a random split over a 14,000-frame dataset places frames from the same recording on both sides. Their deployment numbers are among the most honest in the corpus: 124.51 frames per second on a workstation GPU, and 6.01 on a Jetson Orin Nano rising to 10.3 after conversion, a modest figure for a real-time claim that the paper does not disguise.

The same first author's edge-oriented work [24] fuses RGB appearance, optical flow and monocularly estimated depth by mid-level concatenation, reaching 98.10% on KTH and 77.60% on HMDB51 at a lowest reported 24 ms. Two things complicate the lightweight framing. The depth modality is estimated from RGB by a large transformer, so it is not an independent sensor signal and its complementarity with appearance is partly illusory. And the front end runs that transformer, dense optical flow and a CNN backbone per frame, which sits awkwardly with a 24 ms figure whose scope is not stated.

Zahra et al. [17] recognise individual and social activity in UAV footage simultaneously, fusing appearance descriptors with skeletal features by canonical correlation analysis and routing per-person embeddings through a relational graph network, reaching 83.6% and 91.5% on Okutama-Action and 85.7% and 93.2% on JRDB-Act. The deployment reality is the point of interest: about 3.04 s per execution on dual RTX 4090s, roughly 0.33 frames per second, which the authors correctly describe as post-mission review and not live monitoring. Nothing about onboard deployment is shown, and JRDB-Act is a ground-level robot dataset, not an aerial one.

Zhao et al. [20] present the most genuinely multimodal system in the corpus and the only one designed around anticipation, not detection. Five frozen detectors covering scream spectra, panic keyword semantics, running and falling kinematics, and wandering trajectory geometry each emit a scalar confidence, and a 0.3M-parameter combiner fuses them. A hinge term zeroes sub-threshold evidence per modality, so a weak channel cannot drag the score down. It is the closest thing to a principled degradation mechanism anywhere in this review. On their four-hour Shanghai Hongqiao corpus they report 84.9% F1, an 8% false-alarm rate and 1.18 s of average anticipation. The learned weights are the finding: audio at 0.44 and semantics at 0.39 carry roughly 83% of the model's sensitivity, meaning panic is audible before it is visible. Removing audio costs 9 points of recall and removing semantics 11, and in both cases the false-alarm rate more than quadruples. The structural risk is one the authors document without naming: their two dominant modalities are the two that fail together, and their own failure cases include a public-address announcement masking a shouted call for help.

Batri et al. [30] sit at the far edge of the taxonomy, addressing hyperspectral anomaly detection with a training-free statistical model. The claimed 14.6-fold to 585-fold improvement is in mean anomaly score. That is a scale-dependent statistic computed after per-scene normalisation, on five scenes with no ground truth. There is no AUC, no precision, no recall and no false-alarm rate, and the 95.52% consensus agreement is measured against a consensus built from the same baselines the method is compared with.

Table 2. Language-mediated and sensor-level multimodal systems for surveillance.

Reference	Modalities	Methodology	Key reported result	Shortcomings
[17] Zahra et al., 2026	Aerial RGB video, 2-D skeleton pose	Correlation-driven fusion using canonical correlation analysis with class-aware weights; relational graph convolutional network over per-person nodes; recurrent temporal layer with Markov decoding	83.6% individual and 91.5% social on Okutama-Action; 85.7% and 93.2% on JRDB-Act	About 3.04 s per execution on dual workstation GPUs, roughly 0.33 FPS, so no onboard deployment is shown; JRDB-Act is a ground-level robot dataset; two mutually inconsistent improvement claims
[18] Onsu et al., 2026	RGB camera frames, text queries, transmitted embeddings	Edge detection and segmentation with YOLOv11; vehicle-centred crops encoded to 768-dimensional ViT embeddings; transmitted over a noisy channel; cloud decoder feeds a LoRA-tuned LLaVA	About 99.9% reduction in transmitted data; query accuracy 93.2% on originals against 89% on reconstructions; uniform quantisation far more robust than float encoding	Synthetic digital-twin data and an idealised AWGN channel; saving computed against a raw bitmap, not a codec; latency scales at 0.183 s per vehicle before the language model runs

Reference	Modalities	Methodology	Key reported result	Shortcomings
[19] Meng et al., 2025	Text, images, video key frames, audio via speech recognition	Contrastive vision-language backbone fine-tuned on surveillance pairs; hierarchical semantic index maintained dynamically at the edge; only an encrypted top-K candidate set is sent to the cloud	96% to 100% of brute-force nearest-neighbour accuracy with retrieval time reduced by 1.2% to 45.2%	Headline results on a private corpus; the 6G framing is tested on a three-node gigabit local network; index parameters tuned on the evaluation data; language-specific backbone
[20] Zhao et al., 2026	Acoustic spectra, speech semantics, skeleton kinematics, trajectory	Five frozen unimodal detectors emit scalar confidences; a 0.3M-parameter combiner fuses them; a hinge term zeroes sub-threshold evidence so a weak channel cannot drag the score	84.9% F1 with an 8% false-alarm rate and 1.18 s average anticipation time; audio and semantics carry about 83% of sensitivity	Four-hour private corpus with 44 test clips and no data release; the two dominant modalities fail together under public-address noise, a case the authors document; proprietary speech recogniser
[21] Li et al., 2026	Aerial RGB video, text prompts, optical flow	Homography-based ego-motion removal with Kalman smoothing; frozen InternVL2-8B queried with scene-conditioned chain-of-thought prompts; dual-stream refinement over semantic embeddings and motion anchors	77.99 mean AUC-ROC on Drone-Anomaly and 71.07 on UIT-ADrone, against 71.22 and 69.56 for the best baseline; nothing is trained	Best variant generates prompts with a closed commercial API at non-zero temperature; deterministic reproducible variant is one to three points worse; offline use only; hallucination unresolved
[23] Biswas et al., 2026	2-D RGB, 2-D depth, 3-D point clouds	Dual-branch 2-D and 3-D convolutional network fused by cross-modality scaled dot-product attention with a learnable temperature that converges to about 0.72	98.15% on a random frame split and 95.36% leave-one-subject-out; 10.3 FPS on a Jetson Orin Nano after conversion, 124.5 FPS on a workstation GPU	Random frame split places frames from one recording on both sides; fifteen subjects and seven classes; controlled indoor capture only; no missing-modality test; no parameter or FLOP figures
[24] Biswas et al., 2026	RGB appearance, optical flow, monocular depth	Three modality-specific extractors concatenated at mid level; recurrent convolutional network with spatial attention gates; causally masked temporal self-attention with a Gaussian recency bias	98.10% KTH, 96.28% UCF-101, 83.47% JHMDB, 77.60% HMDB51; lowest reported inference time 24 ms	Depth estimated from RGB, so it is not an independent sensor; a large depth transformer runs per frame, which the 24 ms figure does not clearly include; two inconsistent sets of gain figures
[30] Batri et al., 2026	Hyperspectral bands (single sensor)	Training-free two-stage local and global statistics; three orthogonal features (Mahalanobis energy, spectral entropy, local-to-global divergence) fused with fixed weights held constant across scenes	Claimed 14.6-fold to 585-fold improvement in mean anomaly score over five baselines; consensus agreement 95.52%	No ground truth and therefore no AUC, precision, recall or false-alarm rate; the headline statistic is scale-dependent after per-scene normalisation; consensus built from the same baselines
[33] Yuan et al., 2025	RGB video, sentence-level text	Introduces the UCA benchmark re-annotating UCF-Crime at sentence level; evaluates about twenty models across grounding, captioning, dense captioning, anomaly detection and instruction-tuned language models	23,542 timed sentences over 110.7 hours; recall at one for IoU 0.3 stays generally below 10% for all grounding baselines	Derived from a single source dataset; no inter-annotator agreement reported; hallucination described qualitatively and never quantified; no closed-model baseline

Reference	Modalities	Methodology	Key reported result	Shortcomings
[34] Sadhwani et al., 2025	RGB frames, generated captions, vision-language embeddings	Frames extracted at one per second; captions and dense pretext embeddings produced by BLIP and Uform-Gen; separate severity classifiers trained on each representation	94.70 / 98.04 / 86.63% accuracy on fire, traffic and violence with embeddings, against 85.37 / 94.01 / 76.80% with captions	Datasets self-collected, unnamed and not public; severity labelling protocol never described; per-frame labels for an inherently temporal property; no end-to-end system latency

4.5 The Real-Time Constraint

Two studies treat deployment as the research problem and not as a closing paragraph, and they are the most instructive on what real time costs. Gallo et al. [16] build a violence detector for urban buses, six ceiling cameras feeding an embedded server, and argue that domain shift, not benchmark accuracy, is the bottleneck. Their evidence supports this. Models trained only on public data drop significantly on their proprietary bus test set, adding in-domain data recovers the loss across all three backbones, and removing Kinetics pretraining costs five to eight points. Their field trial on a real thirteen-metre bus over 53 experiments gives 86.4% accuracy with 0.84 precision and 0.81 recall, and end-to-end latency of 620 to 950 ms, with two to four seconds between the onset of an action and the availability of a score. Those numbers sit below most in this review and are considerably more believable, because the boundary is stated and the hardware is what would be installed. Their observation that daytime performs worse than nighttime, because natural light casts harsh shadows while onboard illumination is uniform, only emerges from a real installation. Zhang et al. [28] define airport surface detection as a distinct regime, on the grounds that more than 87% of annotated objects in their VASSO dataset occupy less than 0.58% of the frame. Their MFR-Net reaches 72.50% mAP and 56.22% on the small-object metric, against 50.30% for the closest competitor, and runs at 8.1 frames per second on a 48 GB datacentre GPU. That is not deployable on an operational apron, and the paper does not claim otherwise, but it shows how far the accuracy frontier sits from the latency frontier here.

Placed side by side, the deployment evidence is difficult to compare. Reported latency spans 12.5 ms for a scoring head over pre-extracted features [13], 16.3 ms for a fusion module excluding feature extraction [15], 24 ms for a recurrent action model [24], 95 ms for a ten-frame sequence [29], 97 ms on a Jetson after conversion [23], 620 to 950 ms end to end on an embedded server [16], 960 ms on a workstation GPU [20], and 3.04 s on dual workstation GPUs [17]. Frame rates span 8.1 [28], 10.3 and 124.5 [23], and 416 [14]. These figures are not on a common scale, and no paper states which boundary the field should adopt.

Table 3. Deployment characteristics and measurement boundaries of the reviewed systems.

Reference	Hardware	Reported latency or rate	What the measurement covers
[13] Kim et al., 2026	NVIDIA RTX 3090	12.5 ms; 23.49 GFLOPs	Scoring model over pre-extracted CLIP and VGGish features; encoder cost excluded
[14] Ahn et al., 2025	Intel i9-9900K and TITAN RTX	416 FPS; 98.7 GFLOPS	Anomaly-score computation only; SlowFast, TV-L1 flow and scene classifier upstream
[15] Kim et al., 2026	NVIDIA RTX A5000	16.3 ms; 25.80 GFLOPs	Fusion model only; the paper states feature extraction is excluded
[16] Gallo et al., 2026	Embedded server on a 13 m bus, six cameras	620-950 ms; 2-4 s from action onset to score	End to end on installed hardware, including the three-second buffer
[17] Zahra et al., 2026	Dual NVIDIA RTX 4090	3.04 s per execution, about 0.33 FPS	Full pipeline; authors describe it as post-mission, not live
[18] Onsu et al., 2026	Not stated for the edge node	0.183 s per detected vehicle	Region extraction plus embedding generation; language model inference excluded
[20] Zhao et al., 2026	NVIDIA RTX 4090	0.96 s per alarm; 1.18 s anticipation lead	Five detectors plus fusion head; stated 2 GB and 25 W design budget not measured on edge hardware
[21] Li et al., 2026	NVIDIA RTX 4090, Xeon Gold	Moderate runtime overhead; no absolute figure in text	Full pipeline; authors state it suits offline or near-online use, not onboard real time
[23] Biswas et al., 2026	Jetson Orin Nano; RTX A4000	10.3 FPS on Jetson; 124.5 FPS on workstation	Full model after TensorFlow Lite conversion; preprocessing not separately timed
[24] Biswas et al., 2026	Jetson (name inconsistent); RTX A4000	24 ms lowest reported	Scope not clearly stated; front end includes a depth transformer per frame
[27] Aung et al., 2025	NVIDIA A100	0.3325 s per instance	Model inference only; pose estimation over fifty frames excluded

Reference	Hardware	Reported latency or rate	What the measurement covers
[28] Zhang et al., 2026	NVIDIA L40S 48 GB	8.1 FPS	Full detector at 1280x720; no embedded platform evaluated
[29] Oubelkas et al., 2026	NVIDIA Tesla T4	95 ms per ten-frame sequence	End to end for a fixed ten-frame window at 64x64 resolution
[30] Batri et al., 2026	Laptop CPU, no GPU	About 25 minutes per scene	Whole training-free detector; no acceleration attempted
[34] Sadhwani et al., 2025	NVIDIA A40 48 GB	28-33 ms per image, 31-35 images/s	Model forward pass per image; no end-to-end system latency reported

4.6 Cross-Scheme Reading

Several patterns run across all three tables. Benchmark monoculture is the most visible: UCF-Crime [3] and XD-Violence [6] carry the great majority of the anomaly-detection evaluations, and a study needing a third dataset usually builds one. Nine of the twenty-one studies report their headline result on data they created, and six of those datasets are not public.

Then there is the audio problem. Audio is the modality most often used to justify the multimodal label, and it exists genuinely in one public benchmark, so every cross-domain audio-visual result in this review rests partly on synthesis [13], [15].

Fusion depth is the third pattern, and the least expected. The systems with the most elaborate fusion machinery are not reliably those with the largest gains. SCV's gated fusion adds 5.29 points of AP over its own baseline [15], which is substantial, but ASTRA's gain over the strongest previous audio-visual method is 0.31 points [13], and its visual-only variant loses only 0.44. Meanwhile a three-hundred-thousand-parameter combiner over five scalars produces a 3-point F1 gain and a 1.18-second anticipation lead that no visual system here offers [20].

Last, modality degradation is essentially untested. Two studies report what the trained fused system does at inference when a channel is withdrawn: ASTRA's visual-only variant [13] and the panic model's per-modality ablation [20]. A third, SCV [15], reports audio-only and fused category performance side by side, which shows where the acoustic signal lives but says nothing about how the fused model behaves once it is gone. Every other system here assumes all modalities are present and clean at inference. Cameras get occluded, microphones get wind noise, depth sensors fail in sunlight, and speech recognition fails under a public-address announcement, a failure the panic paper documents in its own data.

5. Research Gap and Future Directions

The module-level picture is healthy. Fusion mechanisms have become more sophisticated, weakly supervised training is well understood, and vision-language models have opened a genuinely new capability. The system-level picture is less flattering, and the gaps below are stated as measurement failures because that is what most of them are.

1. Audio is treated as a modality but functions as a benchmark artefact. One public benchmark carries genuine synchronised sound, so cross-domain evaluation depends on synthesised audio the papers themselves call coarse. Until a second audio-bearing benchmark exists, no claim about the general value of the acoustic channel can be tested, and the honest reading is that audio helps a great deal for acoustically distinctive categories such as riots and explosions and very little elsewhere.
2. Real time has no agreed definition and no agreed measurement boundary. Latency is reported over pre-extracted features, over a scoring head, over a fusion module, over a fixed-length sequence, and end to end on installed hardware, with the boundary usually buried in the experimental section and never in the abstract. This is the easiest gap to close.
3. Modality degradation is almost never tested, which undermines the main argument for multimodal design. If the point of adding a channel is robustness, the property to measure is the fused model's behaviour under channel loss, and it is not measured. Two studies in twenty-one report it, and reporting each branch in isolation, which several more do, does not substitute.
4. Benchmark monoculture and private data limit what the numbers can support. Nine studies report headline results on self-constructed datasets and six of those are unreleased, so several of the most operationally interesting findings rest on corpora no other group can access.
5. Metrics do not track operator utility. Area under the curve and average precision measure ranking quality over a whole test set. An operator cares how many false alarms arrive per hour and how long after an event the alert appears. Only the panic study reports anticipation time and a false-alarm rate together, and only the bus study reports precision and recall from a field trial on installed hardware.
6. Language-based systems are evaluated with metrics that cannot detect their characteristic failure. Hallucination is described qualitatively in at least two studies and quantified in none, while captioning metrics such as BLEU and CIDEr reward fluent output regardless of whether the described event occurred. For a system whose purpose is to tell a human what is happening, a confidently wrong sentence is worse than no sentence.
7. Internal reporting consistency is weaker than it should be. Five of the twenty-one papers report inconsistent headline numbers across their own abstract, results section and conclusion, and two report inconsistent improvement

claims against the same baseline, so a reader compiling a comparison table cannot tell which figure the authors stand behind.

The directions that follow are modest by design. A shared modality-dropout protocol, in which every system reports performance with each channel removed and with each degraded by a stated amount of noise, would cost nothing in model design and would reveal which systems are genuinely fused and which have a dominant stream carrying the result. A declared latency boundary, stated in the abstract alongside the hardware, would make deployment claims comparable for the first time. A second audio-bearing benchmark, even a small one, would let the acoustic channel be evaluated at all. Reporting anticipation time alongside detection accuracy would align the field's metrics with the reason surveillance automation exists. And a factuality check scoring whether a generated description names an event that actually occurred would begin to quantify the failure mode everyone acknowledges in prose.

6. Conclusion

This review has read intelligent multimodal surveillance as four separate research programmes instead of one field, and that framing is what makes the remaining problems visible. Audio-visual fusion under weak supervision is technically mature and constrained by a single benchmark. Vision-derived multi-stream fusion is productive, but it frequently makes a robustness argument its evidence does not support, since streams sharing a camera fail together. Vision-language systems have opened a capability that did not previously exist, the generation of an explanation and not merely a score, and they are judged by metrics that cannot tell a correct explanation from a fluent one. Sensor-level fusion offers the strongest robustness case and the weakest deployment numbers, and the most operationally valuable result in the corpus, a 1.18-second anticipation lead from acoustic and semantic cues, comes from a system whose data cannot be obtained.

The common thread is not a missing architecture. It is that the field measures what is easy to measure. Accuracy on two benchmarks is easy; latency at a declared boundary is harder and rarely done. Adding a modality is easy; showing the fused system still works once that modality fails is harder and has been attempted twice in twenty-one studies. None of these gaps requires a new model. Each requires a small addition to an existing evaluation protocol, and closing them would do more to move multimodal surveillance towards installations people actually trust than another fusion module.

References

[1] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection: A new baseline," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6536–6545.

[2] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2013, pp. 2720–2727.

[3] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[4] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 4975–4986.

[5] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 1, 2023, pp. 387–395.

[6] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in Proc. Eur. Conf. Comput. Vis. (ECCV), vol. 12375, 2020, pp. 322–339.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, pp. 1–21.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 139, 2021, pp. 8748–8763.

[10] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 6, 2024, pp. 6074–6082.

[11] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang, "Open-vocabulary video anomaly detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 18297–18307.

[12] W.-D. Jiang, C.-Y. Chang, M.-Y. Su, Y.-S. Lee, and D. S. Roy, "Toward interpretable multimodal violence detection with knowledge distillation and modality-aligned preprocessing," IEEE Trans. Syst., Man, Cybern., Syst., vol. 55, no. 9, pp. 6215–6228, Sep. 2025, doi: 10.1109/TSMC.2025.3576390.

[13] Y. Kim, T. Kim, W. Lee, and M. Cho, "ASTRA: Joint latent representation and knowledge transfer for multimodal anomaly detection," IEEE Access, vol. 14, pp. 109706–109725, 2026, doi: 10.1109/ACCESS.2026.3710522.

[14] D.-H. Ahn, Y.-J. Jo, D.-B. Kim, G. P. Nam, J.-H. Han, and H. Kim, "Where and what: Contextual dynamics-aware anomaly detection in surveillance videos," IEEE Trans. Image Process., vol. 34, pp. 6993–7007, 2025, doi: 10.1109/TIP.2025.3623392.

[15] Y. Kim, W. Lee, S. Kim, and M. Cho, "Selective audio–visual fusion via semantic calibration for video surveillance system," IEEE Trans. Ind. Informat., early access, 2026, doi: 10.1109/TII.2026.3709835.

[16] M. Gallo, M. G. Paganelli, P. R. Massenio, and D. Naso, "Real-time violence detection in urban bus environments," IEEE Access, vol. 14, pp. 55075–55089, 2026, doi: 10.1109/ACCESS.2026.3681675.

[17] I. Zahra, H. Alsolai, B. Alabdullah, F. Alhayan, A. Jalal, and H. Liu, "Multimodal neural network framework for individual and social activity recognition in UAV surveillance," IEEE Access, vol. 14, pp. 19628–19648, 2026, doi: 10.1109/ACCESS.2026.3658757.

[18] M. A. Onsu, P. Lohan, B. Kantarci, A. Syed, M. Andrews, and S. Kennedy, "Semantic edge–cloud communication for real-time urban traffic surveillance with ViT and LLMs over mobile networks," IEEE Trans. Netw. Sci. Eng., vol. 13, pp. 2503–2520, 2026, doi: 10.1109/TNSE.2025.3617881.

[19] L. Meng, G. Liu, and J. Luo, "DEVOLUTION: A symbiotic cloud–edge framework for real-time multimodal retrieval in 6G-based surveillance IoT system," IEEE Internet Things J., vol. 12, no. 20, pp. 41510–41525, Oct. 2025, doi: 10.1109/JIOT.2025.3601800.

[20] R. Zhao, L. Han, P. Jia, C. Han, B. Wei, C. Li, and P. Yan, "Pedestrian panic behavior recognition model based on multi-modal data fusion and experimental analysis in transit hubs," IEEE Trans. Intell. Transp. Syst., vol. 27, no. 5, pp. 6002–6015, May 2026, doi: 10.1109/TITS.2026.3651479.

[21] R. Li, X. Liu, Y. Hu, X. Zhu, and X. Cao, "Toward semantic-aware aerial video anomaly detection by exploiting multimodal large language model," IEEE Trans. Circuits Syst. Video Technol., vol. 36, no. 8, pp. 12182–12197, Aug. 2026, doi: 10.1109/TCSVT.2026.3686230.

[22] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 131–135.

[23] S. Biswas, S. Biswas, A. Biswas, A. Nandy, and A. K. Naskar, "Embedded AI-driven multimodal human activity recognition using RGB–D sensor data and cross-modality attention," IEEE Sensors J., vol. 26, no. 9, pp. 13818–13829, May 2026, doi: 10.1109/JSEN.2026.3673365.

[24] S. Biswas, A. Nandy, and A. K. Naskar, "Lightweight multimodal feature fusion and spatiotemporal learning for human action recognition on edge devices," IEEE Trans. Emerg. Topics Comput. Intell., vol. 10, no. 1, pp. 982–995, Feb. 2026, doi: 10.1109/TETCI.2025.3616054.

[25] X. Wang, X. Wang, H. Bai, E. G. Lim, and J. Xiao, "Intracluster contraction for pedestrian anomaly detection under one-class classification setting," IEEE Trans. Artif. Intell., vol. 7, no. 3, pp. 1466–1476, Mar. 2026, doi: 10.1109/TAI.2025.3601079.

[26] M. L. A. Mohamad, M. A. Abd Rahman, N. M. Shah, and A. K. Sangaiah, "Efficient NASNetMobile-enhanced vision transformer for weakly supervised video anomaly detection," IEEE Internet Things J., vol. 13, no. 1, pp. 536–548, Jan. 2026, doi: 10.1109/JIOT.2025.3625045.

[27] Gaddam, T., Pal, K., Rani, K., & Saharan, R. P. (2025). Effects of terminal heat stress on wheat yield and morpho-physiological traits in Ongole, Andhra Pradesh. Annals of Biology, 41(2), 289–298.

[28] S. T. Y. Aung, W. Kusakunniran, and Y. K. Hooi, "Leveraging fusion methods of human pose and motion dynamics for accurate violence detection in video surveillance," IEEE Access, vol. 13, pp. 168116–168125, 2025, doi: 10.1109/ACCESS.2025.3613765.

[29] S. Zhang, Y. Yang, Y. Zhu, S. Qian, X. Zhang, and K. Cai, "MFR-Net: Motion-guided feature refinement network for video small object detection in vision-based airport surveillance systems," IEEE Trans. Intell. Transp. Syst., vol. 27, no. 8, pp. 10041–10056, Aug. 2026, doi: 10.1109/TITS.2026.3686215.

[30] F. Oubelkas, A. Louja, L. Moumoun, and A. Jamali, "SST-Net: Self-supervised violence detection using spatiotemporal modeling and adaptive background suppression," IEEE Access, vol. 14, pp. 11185–11199, 2026, doi: 10.1109/ACCESS.2026.3655401.

[31] K. Batri, S. Lakshmi, T. R. Mahesh, S. B. Khan, A. Alshuhail, and A. Almusharraf, "STIM: A unified spatially informed model for robust hyperspectral anomaly detection," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 19, pp. 7204–7234, 2026, doi: 10.1109/JSTARS.2026.3656700.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.

[33] A. Alshalawi, W. Abdul, and G. Muhammad, "Advanced detection of violence from video: Performance evaluation of transformer and state of the art of convolution of neural network transformer," IEEE Access, vol. 13, pp. 74200–74216, 2025, doi: 10.1109/ACCESS.2025.3564435.

[34] T. Yuan, X. Zhang, B. Liu, K. Liu, J. Jin, and Z. Jiao, "Surveillance video-and-language understanding: From small to large multimodal models," IEEE Trans. Circuits Syst. Video Technol., vol. 35, no. 1, pp. 300–314, Jan. 2025, doi: 10.1109/TCSVT.2024.3462433.

[35] S. Sadhwani, J. A. Shamsi, M. B. Khan, N. Z. Bawany, and H. J. Syed, "Real-time detection of mixed-critical events using vision-language models," IEEE Access, vol. 13, pp. 181363–181384, 2025, doi: 10.1109/ACCESS.2025.3622638.