Video Anomaly Detection Beyond AUC: A Protocol-Aware Evidence Review of Methods, Benchmark Protocols, Reproducibility, and Deployment Readiness

Umar Shafiu Haruna[1], Abdulrauf Garba Sharifai[2], Aminu Abdulkadir Mahmoud[1], Lawan Garba[3], Muktar Danlami[1]

[1]Department of Cybersecurity, Northwest University, Kano State, Nigeria

[2]Department of Computer Science, Northwest University, Kano State, Nigeria

[3]Department of Software Engineering, Northwest University, Kano State, Nigeria

DOI: https://doi.org/10.51244/IJRSI.2026.1310000055

Received: 17 October 2026; Accepted: 22 October 2026; Published: 31 October 2026

ABSTRACT

Video anomaly detection (VAD) now spans normality modelling, weak supervision, multimodal and vision-language approaches, training-free inference, and video anomaly understanding (VAU). This diversity has made simple cross-paper ranking increasingly fragile because identical benchmark names can conceal differences in supervision, feature extraction, preprocessing, evaluation population, aggregation, post-processing, and model selection. This protocol-aware evidence review therefore treats comparability as a prerequisite for performance ranking. A bounded core corpus of 193 peer-reviewed VAD/VAU studies published from 2015 to 14 October 2026 was assembled through a documented structured-seed and official-source update process, with source-specific verification against major publisher and venue records. All core records received bibliographic, methodological-regime, provenance, verification, and public-code-link coding, while 12 studies underwent an extended primary-source protocol audit. The Protocol Compatibility Framework is demonstrated with study-level Tier A, B, and C cases. For weakly supervised UCF-Crime studies using I3D-RGB features, the audit separated preprocessing-specific comparisons: MIST and MSL formed a one-crop Tier-A pair (82.30% and 85.30% AUC), while RTFM, UR-DMU, and MGFN formed a ten-crop Tier-A stratum (84.30%–86.98%; median 86.97%). S3R was excluded from the strict quantitative summary because its crop setting was not explicit in the original S3R publication. Public code links were directly identifiable for 79 of 193 core records (40.9%), reported descriptively rather than as a population estimate. The review also introduces a worked Minimum VAD Reporting Standard (MVRS). Together, the framework and evidence register support more defensible benchmark synthesis by making protocol compatibility, traceability, reproducibility, and deployment context explicit before numerical claims are compared.

Keywords: video anomaly detection, video anomaly understanding, benchmark protocols, reproducibility, deployment readiness.

INTRODUCTION

Video anomaly detection (VAD) seeks to identify events, behaviours, or entities that deviate from what is expected in a video scene. The task is deceptively difficult because “anomaly” is not an intrinsic visual class. It is contextual, rare, often temporally extended, and potentially open-ended: running may be ordinary on a sports field but unusual in a restricted corridor; a bicycle can be normal on a road and anomalous in a pedestrian-only walkway. These properties have long motivated one-class and normality-modelling formulations in which systems learn regular patterns and flag deviations rather than enumerate every possible abnormal event [8], [33], [19].

The technical history of VAD reflects repeated attempts to represent normality more faithfully. Early work relied on dynamic textures, sparse reconstruction, local motion statistics, and dictionaries [33], [30]. Deep learning shifted the emphasis toward learned spatiotemporal representations and proxy objectives such as reconstruction and future-frame prediction [19], [31], [27]. Memory modules and object-centric designs later sought to prevent high-capacity models from reconstructing anomalies too well or from allowing background pixels to dominate the decision [22], [17], [38]. In parallel, large real-world datasets enabled weakly supervised multiple-instance learning, making it possible to learn from coarse video-level anomaly labels rather than dense temporal annotation [49], [9]. Recent work has expanded further toward training-free memory systems, efficient masked autoencoders, state-space models, multimodal fusion, frozen vision-language encoders, open-vocabulary recognition, and LLM-assisted reasoning and explanation [9], [60], [32], [62].

This rapid progress has created a second problem: it is increasingly difficult to determine whether two reported numbers measure the same thing. “Frame-level AUC on Avenue” may appear comparable across papers, yet preprocessing, frame rate, ground-truth handling, score smoothing, score normalisation, model selection, averaging, and even the evaluated frame population can differ. The difficulty is greater across supervision paradigms. A weakly supervised model trained with abnormal video-level labels, a one-class model trained only on normal clips, an open-vocabulary detector using language priors, and a training-free LLM pipeline may all report a scalar anomaly score, but they solve materially different learning problems. Street Scene explicitly argued that widely used VAD evaluation criteria could fail to reflect practical event detection [40], and recent deployment-oriented work has renewed attention to online inference, false alarms, and metrics beyond a single global AUC [61].

Existing surveys provide important foundations. Nayak et al. reviewed deep-learning VAD methods, datasets, infrastructure, and evaluation practice [35]. Ramachandra et al. organised single-scene VAD, including datasets, evaluation criteria, and best practices [41]. Samaila et al. systematically examined 530 papers published from 2003 to 2023 and synthesised methods, problem settings, datasets, and research prospects [45]. More recent reviews have broadened the taxonomy to multiple supervision levels and large models [54], focused specifically on privacy-preserving VAD [28], examined multimodal and weakly supervised systems [5], and surveyed foundation-model-driven video anomaly understanding (VAU) [48]. These works make another generic architecture taxonomy difficult to justify.

Comparison with recent VAD surveys indicates that protocol compatibility, reproducibility, and deployment readiness are widely acknowledged concerns, but they are not generally used as the primary structure for deciding whether published results are numerically comparable. This review addresses that gap by treating evaluation protocol as part of the result rather than as background metadata.

The paper makes five contributions. First, it provides a multidimensional evidence map spanning supervision, representation, model family, output task, benchmark, and deployment regime. Second, it introduces a Protocol Compatibility Framework and demonstrates it with explicit study-level Tier A, Tier B, and Tier C comparisons. Third, it reports a bounded core audited corpus of 193 peer-reviewed studies together with source provenance, verification records, and a 12-study extended primary-source protocol audit. Fourth, it provides a primary-study quantitative demonstration in two preprocessing-specific UCF-Crime Tier-A strata rather than pooling superficially similar I3D results indiscriminately. Fifth, it connects protocol compatibility to reproducibility, deployment evidence, and a worked Minimum VAD Reporting Standard (MVRS) example. The aim is not another architecture catalogue, but an evidence structure that makes clear which numerical comparisons are defensible and which should remain conditional or contextual.

The remainder of the paper is organised thematically. Background and Review Scope defines the VAD task and the scope of the review. Protocol-Aware Evidence Review Method describes acquisition, screening, coding, source verification, and protocol compatibility. The RQ1 and RQ2 sections examine methodological evolution and the benchmark landscape. RQ3 analyses evaluation practice and applies the Protocol Compatibility Framework to representative study-level cases. RQ4 reports protocol-compatible primary-study comparisons and the reproducibility audit, while RQ5 examines deployment readiness. The Emerging Frontier

section discusses the transition from VAD to open-vocabulary and explainable VAU, followed by research priorities, threats to validity, positioning against recent surveys, and the conclusion.

BACKGROUND AND REVIEW SCOPE

What constitutes a video anomaly?

Anomaly detection has a broad statistical history, commonly framed as the identification of patterns that do not conform to expected behaviour [8]. In video, however, the reference distribution is shaped by scene semantics as much as by visual frequency. Spatial location, object identity, direction of motion, time of day, interaction, and social context can all change whether an event should be considered abnormal. This is why the same visual event can receive different labels across datasets or camera views.

For review purposes, we treat VAD as a family of tasks in which the system must assign abnormality evidence to a video, clip, frame, object, spatial region, or event, with the central decision depending on deviation from scene- or dataset-defined normality. The scope includes one-class or semi-supervised normality modelling, unsupervised VAD, weakly supervised VAD, fully supervised open-set formulations, multimodal VAD, open-vocabulary VAD, anomaly anticipation, and VAU systems when anomaly detection or localisation remains an explicit component. Pure action recognition, violence classification on pre-segmented clips, generic image anomaly detection, and non-visual time-series anomaly detection are excluded unless the work explicitly evaluates a VAD setting.

Learning regimes

Normal-only or one-class VAD assumes that training data represent normal behaviour and that anomalies are unavailable or deliberately withheld. This setting underpins classical normality models and many reconstruction, prediction, memory, and distance-based methods [33], [19], [31], [27], [17], [38]. “Semi-supervised” is often used in the VAD literature for this normal-only training setting, although the terminology differs from conventional semi-supervised classification. This review records the authors’ terminology and maps every paper to an explicit supervision code so that naming differences do not obscure experimental differences.

Weakly supervised VAD uses coarse labels, typically a video-level indication that a training video contains an anomaly without identifying the anomalous frames. Sultani et al. popularised deep multiple-instance ranking for real-world surveillance videos [49], and MIL remains central to contemporary weakly supervised work [9], [59], [67]. Fully supervised and open-set settings provide stronger labels but can preserve open-world testing by separating anomaly classes between train and test, as exemplified by UBnormal [3]. Unsupervised methods remove labels entirely and must infer both normality and contamination structure from the observed corpus [63]. Open-vocabulary systems extend the task further by attempting to detect and name anomaly categories that may not have been seen during training [55].

From anomaly detection to anomaly understanding

Traditional VAD typically produces a scalar anomaly score and, in some cases, spatial or temporal localisation. Foundation models are shifting the objective toward anomaly understanding: the system may be expected to identify what happened, why it is unusual in the scene, what category it belongs to, and how confident the explanation should be. VadCLIP illustrates the use of frozen vision-language alignment for weakly supervised VAD [56]; LAVAD uses VLM captions and an LLM for training-free anomaly scoring [64]; AnomalyRuler induces and applies scene-specific normality rules with an LLM [60]; VERA explicitly targets explainable VAD through verbalised learning of guiding questions [62]. A dedicated foundation-model survey formalises this transition in terms of anomaly perception and anomaly cognition [48].

This transition creates new evaluation requirements. A frame-level ROC curve cannot evaluate the factuality of a textual explanation, whether the correct anomaly class was named, or whether a model’s rationale is faithful to the evidence. Recent VAU work therefore proposes anomaly-focused semantic evaluation and

meta-benchmarks rather than relying only on generic language similarity [24]. For this reason, the present review treats detection and understanding as connected but distinct layers of evidence.

PROTOCOL-AWARE EVIDENCE REVIEW METHOD

Reporting framework and review period

The review follows PRISMA 2020 principles for transparent identification, eligibility assessment, screening, and synthesis [36], while deliberately calibrating its claim of coverage. The evidence base was constructed through a structured-review seed plus targeted official-source updating rather than a de novo export from every bibliographic database. It is therefore treated as a protocol-aware evidence review with a bounded core audited corpus, not as a complete bibliometric census of VAD. The review period spans 1 January 2015 to 14 October 2026, covering the modern deep-learning era while retaining earlier landmark studies in the Background when historically necessary. The quantitative core contains peer-reviewed journal and full-conference studies; two recent preprints are retained only as emerging-frontier context and are excluded from core corpus counts.

The 2015 starting point captures the transition from hand-crafted and shallow representations toward learned spatiotemporal modelling while keeping the evidence set suitable for protocol-level coding. Studies are included when anomaly detection, localisation, recognition, anticipation, explanation, or understanding is an explicit video-level objective. The cut-off date was fixed before quantitative synthesis so that the reported corpus composition and reference set correspond to a defined review window. Figure 1 summarises the multidimensional coding structure used throughout the review.

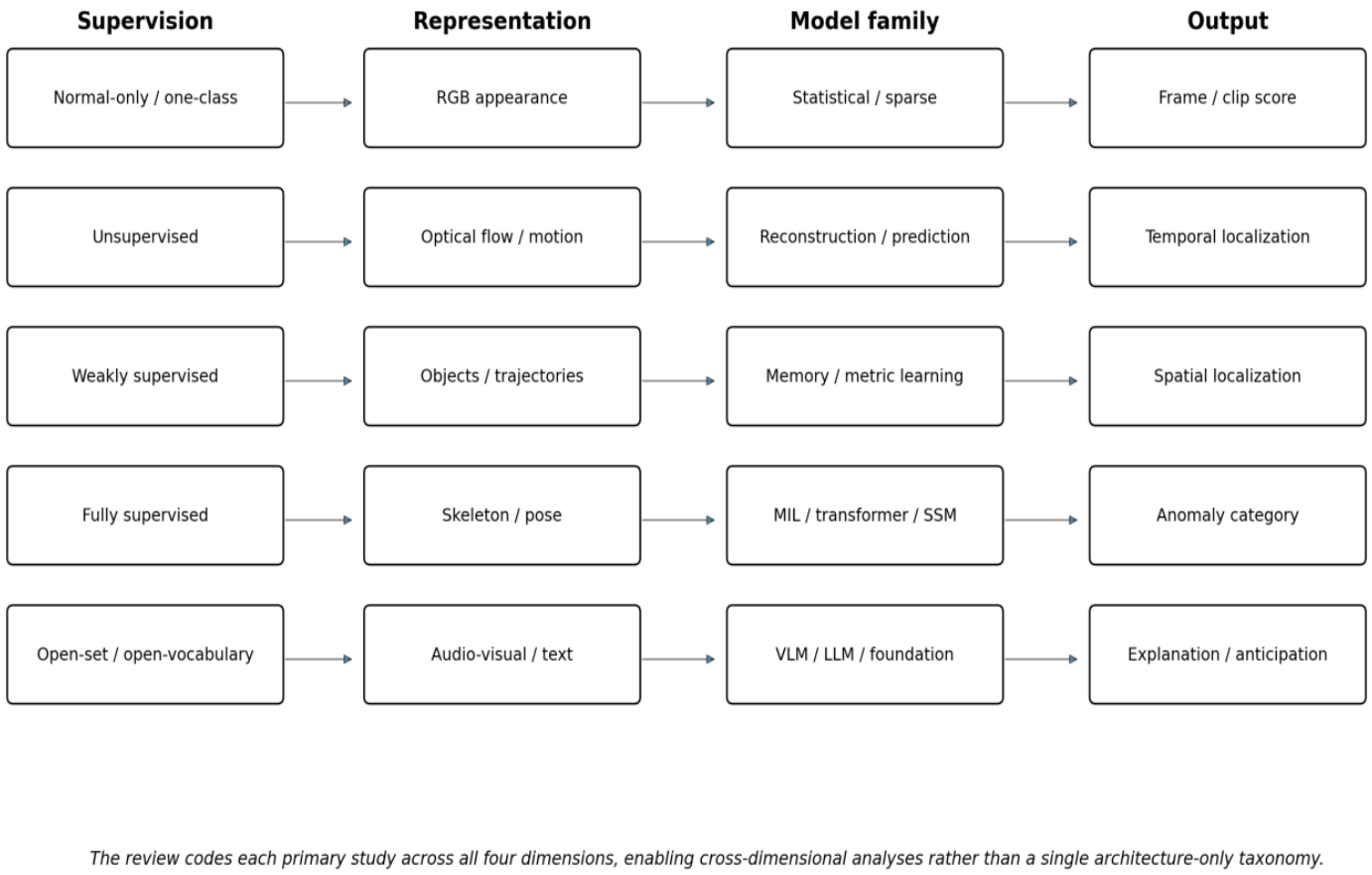


Figure 1. Multidimensional evidence map used to code VAD studies in this review.

Research questions

Table 1. Research questions and corresponding extraction fields.

Research question	Question	Primary evidence extracted
RQ1 — Research landscape	How has VAD evolved in learning paradigm, architecture, representation, anomaly task, and application domain?	Year; supervision; representation; model family; task output; domain.
RQ2 — Datasets and benchmarks	Which datasets, anomaly types, scenes, modalities, and annotation granularities are used, and how representative are they of deployment conditions?	Dataset; split; scenes/cameras; anomaly classes; modality; annotation; access/licence.
RQ3 — Evaluation practice	Which metrics, aggregation procedures, train/test protocols, score post-processing, and model-selection practices are used?	Metric definition; micro/macro; frame/clip/event level; smoothing; score normalisation; thresholding; selection protocol.
RQ4 — Performance and reproducibility	Within protocol-compatible strata, what performance levels are reported and how reproducible are the experiments?	Scores; uncertainty; seeds; code; weights; evaluation script; preprocessing; hardware; compute.
RQ5 — Deployment readiness	How well do current approaches address latency, causality, privacy, domain shift, explainability, false alarms, and open-world operation?	FPS/latency; causal/offline; memory; privacy; cross-domain tests; human-facing outputs; failure analysis.

Information sources and evidence-acquisition strategy

Evidence acquisition used a two-stage seed-and-update design. Stage 1 used the structured peer-reviewed bibliography accompanying the DeepVAD review [54] as a high-recall evidence seed and retained records meeting the eligibility criteria. Stage 2 updated the recent frontier through official publisher and venue records for 2025–14 October 2026, including CVF, AAAI, PMLR/ICML, IEEE/WACV, and Elsevier journal pages. The frozen core register contains 171 eligible seed studies and 22 additional peer-reviewed update studies after title, venue, year, DOI, and duplicate checks, for 193 unique records. Official publisher and venue pages were used as the preferred verification sources. Figure 2 reports this documented acquisition pathway; it is not presented as an exhaustive database-export PRISMA flow.

The search logic was method-agnostic. Core task terms were “video anomaly detection”, “abnormal event detection”, and “video anomaly understanding”, supplemented by terms for weak supervision, open vocabulary, training-free inference, multimodality, foundation models, causal or streaming operation, and privacy-preserving learning. As a coverage-sensitivity check, source-specific verification searches were conducted on 22 October 2026 while retaining 14 October 2026 as the substantive eligibility cut-off. The verification covered CVF Open Access, AAAI Proceedings, PMLR/ICML, Elsevier/ScienceDirect, Springer Nature, and backward checking from recent VAD surveys and protocol-focused papers. Details of the search source, search terms, search date, review cut-off and search purpose are provided in the publicly archived companion evidence register available on Zenodo at https://doi.org/10.5281/zenodo.23579051 and are summarised in Table 2. This verification is used to calibrate coverage claims and to check the primary protocol case studies; it is not used to reconstruct unrecorded historical screening counts.

Table 2. Source-specific search and verification record.

Source	Search terms / approach	Date and eligibility cut-off	Role in evidence review
CVF Open Access (CVPR/WACV/ICCV)	“video anomaly detection”; “video anomaly understanding”; weak-supervision/open-vocabulary variants	22 Oct 2026; include only records published by 14 Oct 2026	Frontier coverage sensitivity; title, venue and primary-protocol verification

AAAI Proceedings	"video anomaly detection"; weakly supervised; "video anomaly understanding"	22 Oct 2026; cut-off 14 Oct 2026	Official article/PDF verification and recent-study sensitivity
PMLR / ICML	"video anomaly"; "video anomaly understanding"	22 Oct 2026; cut-off 14 Oct 2026	Conference-frontier verification
Elsevier / ScienceDirect	"video anomaly detection" plus 2025/2026 filters	22 Oct 2026; cut-off 14 Oct 2026	Journal-frontier verification; post-cut-off records excluded
Springer Nature	"video anomaly detection" plus 2025/2026 filters	22 Oct 2026; cut-off 14 Oct 2026	Journal-frontier verification and publication-date checks
Backward checking	Recent VAD surveys and references in evaluation/protocol papers	22 Oct 2026; cut-off 14 Oct 2026	Terminology variants, comparator tracing and primary-source recovery

Eligibility criteria

Table 3. Inclusion and exclusion criteria.

Criterion	Include	Exclude
Publication type	Peer-reviewed journal or full conference paper in the primary corpus; benchmark/dataset paper directly enabling VAD evaluation.	Editorials, posters without sufficient method detail, theses, slides, non-peer-reviewed material in primary corpus.
Task	Detection, localisation, recognition, anticipation, explanation, or understanding of anomalous events in video with an explicit anomaly objective.	Generic action recognition, ordinary event classification, image-only anomaly detection, non-video time series.
Input	Video or temporally ordered visual observations; multimodal video is eligible.	Purely textual or tabular anomaly tasks; binary/opcode security anomalies.
Evidence	Sufficient description of dataset, supervision, and evaluation to code at least the core protocol fields.	Results that cannot be associated with a defined dataset/protocol or lack evaluable experimental details.
Period	Primary corpus: 2015–14 Oct 2026; earlier landmark papers for historical background.	Primary studies outside the main period, except clearly identified historical context.
Language	English-language scholarly publications.	Non-English records when an English full text sufficient for extraction is unavailable.

Screening, deduplication, and evidence verification

Records in the structured seed and targeted update were normalised by title, venue, publication year, and DOI where available. Repeated appearances across taxonomy categories were collapsed to a single study record, and update records were added only after duplicate checking against the seed. The frozen core register therefore contains 193 unique peer-reviewed studies. Screening proceeded in two sequential passes: task/publication eligibility followed by metadata and primary-source verification for the fields required in corpus-level analysis. Ambiguous records were checked against official publisher, conference, dataset, or project pages where available. The full included-record register, together with acquisition-provenance information, verification status and the coding dictionary, is publicly available in the companion evidence register archived on Zenodo at https://doi.org/10.5281/zenodo.23579051.

The original structured-seed acquisition did not preserve a candidate-level log of every excluded record and exclusion reason. Those historical counts are therefore not reconstructed retrospectively, because doing so would create false precision. Instead, the companion evidence package provides the complete 193-record included register, the documented two-stage acquisition pathway, source-specific verification logs, and study-

level primary-source audit records. The 22 October 2026 coverage-sensitivity check reinforces that the frozen register should be interpreted as a bounded evidence set under the documented procedure rather than as an exhaustive census. The two context-only preprints [29], [1] remain outside all core corpus statistics.

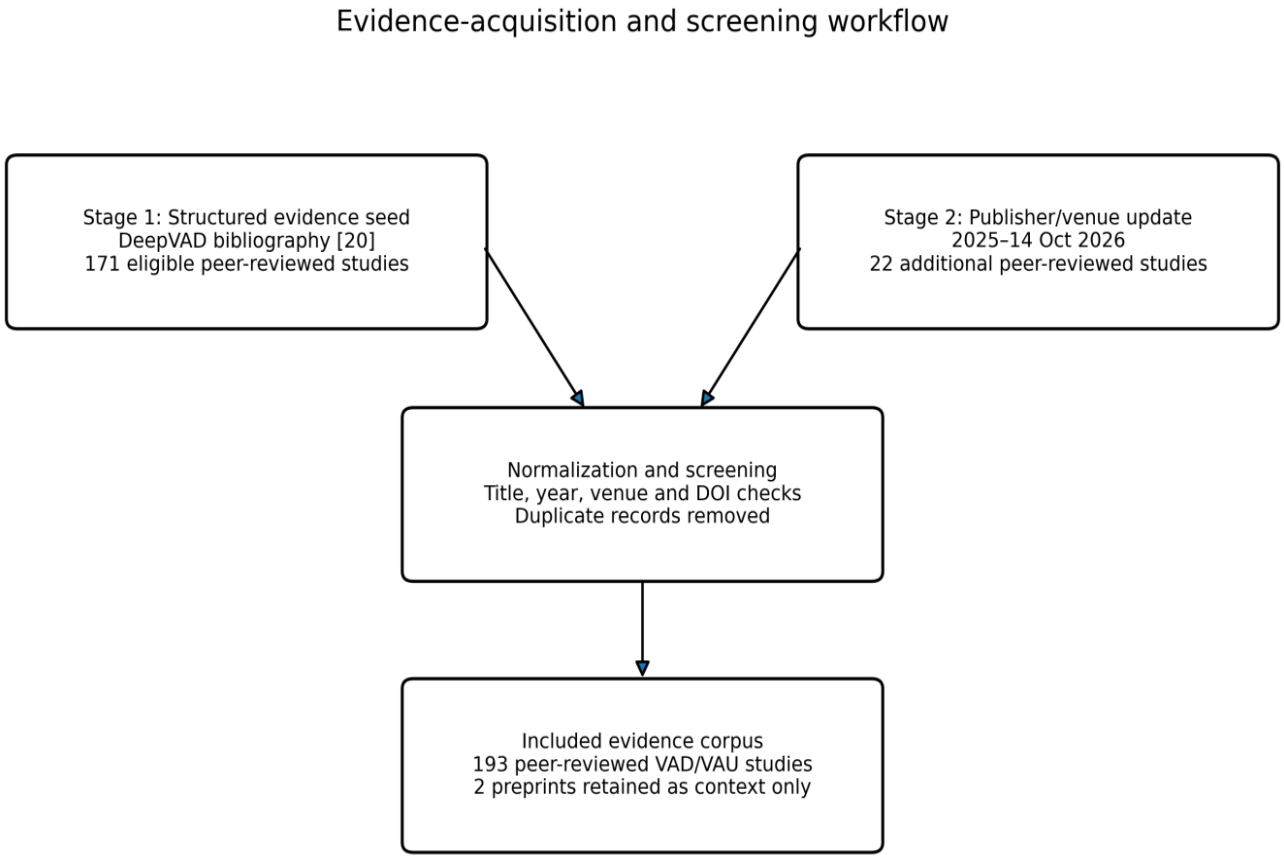


Figure 2. Documented seed-and-update acquisition workflow used to construct the bounded 193-study core audited corpus.

Data extraction and coding

Table 4. Core and extended data-extraction schema.

Domain	Fields to extract	Coverage
Bibliographic	Title/study identifier, year, venue, acquisition provenance, verification status, inclusion status.	Core: all 193 core audited studies
Learning setup	Primary methodological regime: normal-only/one-class, unsupervised, weakly supervised, fully supervised, open-set/few-shot, semantic/training-free/VAU.	Core: all 193 core audited studies
Representation	RGB, optical flow, object crops, skeleton/pose, trajectory, audio, text, multimodal, pretrained/foundation representation.	Extended: 12-study primary-source audit
Method	Backbone; temporal model; memory/statistical module; MIL; transformer/SSM; VLM/LLM; training objective; post-processing.	Extended: 12-study primary-source audit
Dataset protocol	Dataset version; train/validation/test split; sampling rate/stride; resized resolution; anomaly labels used; scene-specific processing.	Extended: 12-study primary-source audit
Evaluation	Metric name and implementation; frame/clip/video/pixel/event level; micro/macro aggregation; smoothing; score normalisation; threshold selection.	Extended: 12-study primary-source audit

Reproducibility	Code; weights; split files; evaluation code; hyperparameters; random seeds; repetitions/uncertainty; environment/hardware.	Core code-link field for all 193; extended details for 12-study audit
Deployment	Causal/offline; latency/FPS; model size/FLOPs; peak memory; edge test; domain shift; privacy protection; human explanation.	Extended: 12-study primary-source audit
Outcome	Reported score(s), uncertainty, comparator, ablation, failure cases, claimed limitations.	Extended: 12-study primary-source audit

All 193 core studies received coding for publication year, methodological regime, venue, public-code-link status, acquisition provenance, verification status, and inclusion status. Extended protocol extraction was then applied to 12 primary studies (12/193 = 6.2%) selected under explicit framework-demonstration criteria rather than informal representativeness. The audit included one foundational UCF-Crime weak-supervision anchor; six UCF-Crime I3D-RGB candidates selected for detailed compatibility testing because they share weak video-level supervision and frame-level ROC-AUC reporting; and five contrast cases selected because they alter the evaluation population, information available to the model, real-time temporal context, or operational metric interpretation. Primary-source checking then separated the six quantitative candidates by preprocessing: MIST and MSL use one-crop I3D-RGB, whereas RTFM, UR-DMU, and MGFN use ten-crop I3D-RGB. S3R was retained in the extended audit but excluded from the strict Tier-A quantitative summary because its crop setting was not explicit in the primary paper. Each extended record was checked against the primary paper and, where necessary, supplementary material or an official project page. Table 5 states the coverage and selection logic. This purposive audit is designed to demonstrate the framework and does not claim to be a random sample of all 193 studies.

Table 5. Extended primary-source protocol audit: coverage and selection criteria.

Audit component	Studies	n	Selection rationale
Historical protocol anchor	Sultani et al. [49]	1	Foundational UCF-Crime weak video-level supervision and canonical benchmark context
Quantitative compatibility candidates	MIST [69], RTFM [70], MSL [71], S3R [76], UR-DMU [72], MGFN [73]	6	Same benchmark family, weak video-level supervision, I3D-RGB and frame-level ROC-AUC; audited to determine whether strict preprocessing-matched Tier-A strata could be formed.
Protocol-contrast cases	UMIL [75], REWARD [74], VadCLIP [56], PI-VAD [34], Acharya et al. [2]	5	Selected to expose altered evaluation population, bounded real-time context, external semantic/poly-modal information, and operational metric differences
Total extended audit	Primary-source audited studies	12 (6.2%)	Framework-demonstration subset; not a probability or representative sample of the full corpus.

Study-quality and claim audit

The audit distinguishes reporting transparency from architectural novelty. For the 12-study extended subset, the extraction asks whether the exact benchmark and split are identifiable; whether supervision and external information access are stated; whether evaluation granularity, metric population and aggregation are clear; whether score post-processing and model selection can be traced; whether repeated-run uncertainty is reported; and whether code, weights, evaluation scripts, hardware, latency or deployment constraints are verifiable. These items are used to explain why a result can or cannot support a particular comparison, not to reward one architecture family over another.

No single reporting-quality score is used to exclude studies or to rank scientific merit. The analysis applies the protocol vector and compatibility rules directly to the study-level cases, while the broader reporting-quality,

reproducibility and deployment constructs are retained as structured audit templates rather than presented as corpus-wide empirical scores. A worked MVRS application later in the paper shows how these criteria are used without collapsing heterogeneous evidence into a potentially misleading single number. Independent duplicate human coding was not conducted; accordingly, no inter-rater agreement statistic is reported.

Protocol compatibility framework

Table 6. Protocol Compatibility Framework used for quantitative synthesis.

Tier	Definition	Permitted use in synthesis
Tier A — Directly comparable	Same benchmark version and canonical or explicitly identical split; compatible supervision; same annotation granularity; same metric definition and aggregation; materially equivalent evaluation population.	May be placed in the same quantitative stratum and summarised statistically.
Tier B — Conditionally comparable	Same benchmark, but one or more material differences in preprocessing, supervision, aggregation, score post-processing, model selection, or annotation handling.	Compare narratively or in separate sub-strata; do not present a single rank order without qualification.
Tier C — Context only	Custom subset/split, altered evaluation population, non-equivalent metric, materially different task, or insufficient protocol detail.	Use for methodological context only; exclude from pooled benchmark statistics and direct SOTA ranking.

A result should therefore be ranked only after its comparability with other reported results has been established. Sharing the same dataset name does not, by itself, demonstrate that two results address the same experimental question.

Mathematical coding and evidence-synthesis framework

To make comparison rules explicit, each extended-audit result is represented as a structured protocol record rather than as a single AUC value. The record includes dataset/version, split and supervision, evaluation granularity, metric definition, aggregation population, feature or representation source, preprocessing and post-processing, reproducibility evidence, and deployment context. Equation (1) formalises this protocol vector. In the present analysis, this vector and the pairwise compatibility logic are applied directly to the worked study comparisons. The reporting-quality, reproducibility, and deployment expressions below are retained as audit definitions only; they are not presented as corpus-wide numerical outcomes.

$$\mathbf{z}_i = (d_i, s_i, g_i, m_i, a_i, f_i, p_i, r_i, e_i) \qquad (1)$$

Here, d denotes the dataset/version, s the split and supervision specification, g the evaluation granularity, m the metric definition, a the aggregation rule, f the feature or representation source, p the post-processing description, r the reproducibility record, and e the deployment-evidence record. This representation makes protocol differences explicit before numerical comparison is attempted.

Public-code linkage is treated only as a descriptive property of the bounded core corpus. Let x denote the number of core records for which a public repository link is directly identifiable in the audited metadata and n the number of core records. The reported proportion is therefore:

$$P_{code} = x / n = 79 / 193 = 0.409 \qquad (2)$$

Thus, 79 of 193 core records (40.9%) had a directly identifiable public-code link. No confidence interval is used to infer a field-wide population proportion because the 193 studies do not constitute a probability sample of all VAD research. Absence of a linked repository in the audited metadata is also not treated as proof that no implementation exists elsewhere.

For the study-quality audit, $q_{ij} \in \{0, 0.5, 1\}$ denotes whether study i omits, partially reports, or fully reports quality item j. With prespecified weights w_j (equal weights by default), reporting-quality completeness is defined as

$$Q_i = \frac{\sum_{j=1}^{J} w_j q_{ij}}{\sum_{j=1}^{J} w_j}, \quad 0 \le Q_i \le 1 \qquad (3)$$

Q_i is retained only as a transparent reporting-completeness template for extended audits and future corpus updates. It is not aggregated across all 193 core records and is not used to rank model quality. The worked MVRS example applies the underlying reporting items directly rather than presenting an unsupported global Q_i distribution.

Pairwise protocol compatibility is coded over L protocol dimensions. For studies i and j, let $(\cdot)$ be the indicator function and $z_{i\ell}$ the value of protocol dimension ℓ. A descriptive compatibility proportion is

$$C_{ij} = \frac{\sum_{\ell=1}^{L} w_\ell \mathbb{I}(z_{i\ell}=z_{j\ell})}{\sum_{\ell=1}^{L} w_\ell}, \quad 0 \le C_{ij} \le 1 \qquad (4)$$

The scalar C_{ij} supports the study-level case audit but does not determine the compatibility tier by itself. Tier A additionally requires agreement on mandatory dimensions: benchmark version, test split, supervision regime, evaluation granularity, metric definition, and aggregation population. A high descriptive overlap cannot compensate for a different test split, different metric population, or a different meaning of "AUC." Tier B permits limited material differences that require qualification; Tier C denotes results that answer a materially different empirical question or are insufficiently specified for direct numerical ranking.

Reproducibility and deployment evidence can likewise be represented as completeness constructs when the necessary fields are extracted uniformly. In the present review, however, R_i and E_i are used as structured checklists within the extended audit and MVRS example rather than as corpus-wide scores, because the core 193-record register does not contain all required fields uniformly. This distinction prevents formal notation from implying a level of quantitative coverage that the evidence register does not support.

$$R_i = \frac{1}{M}\sum_{m=1}^{M} r_{im}, \qquad E_i = \frac{1}{H}\sum_{h=1}^{H} e_{ih} \qquad (5)$$

Accordingly, only P_{code} is reported corpus-wide, while the protocol vector and pairwise compatibility logic are demonstrated on the extended primary-source subset. Q_i, R_i and E_i remain transparent audit definitions and are not interpreted as model-performance measures or population estimates.

Table 7. Mathematical quantities used in the protocol-aware review synthesis.

Symbol	Definition	Use in the review
P_{code}	Descriptive public-code-link proportion	Computed only for the bounded 193-record core corpus: 79/193 = 40.9%; no population-level CI.
Q_i	Reporting-quality completeness	Audit template only; not aggregated across the full core corpus.
C_{ij}	Pairwise protocol compatibility proportion	Applied to study-level Tier A/B/C case comparisons; not averaged globally.
R_i	Reproducibility evidence completeness	Structured audit criterion for the extended subset; not a corpus-wide empirical score.
E_i	Deployment-evidence completeness	Structured deployment criterion for the extended subset; not a corpus-wide empirical score.

Included evidence corpus and descriptive audit

The documented acquisition process yielded a bounded core register of 193 unique peer-reviewed VAD/VAU studies: 171 retained from the structured DeepVAD seed [54] and 22 added through the targeted 2025–2026

official-source update. Corpus-level percentages in this paper describe that audited register only. They are used for methodological-regime composition, publication-year description, and metadata-verifiable public-code linkage; they are not extrapolated to all VAD publications. Two recent preprints [29], [1] remain contextual frontier evidence and are excluded from all core counts.

Table 8. Composition of the bounded core audited corpus after seed acquisition, targeted updating, deduplication, and verification.

Methodological regime	Studies,n (%)	Identifiable public-code links, n (%)
Normal-only / semi-supervised	105 (54.4%)	39 (37.1%)
Weakly supervised	55 (28.5%)	20 (36.4%)
Unsupervised	14 (7.3%)	7 (50.0%)
Open-set / few-shot	10 (5.2%)	5 (50.0%)
Semantic / training-free / VAU	9 (4.7%)	8 (88.9%)
Total	193 (100.0%)	79 (40.9%)

Normal-only/semi-supervised VAD is the largest methodological family in the included corpus (54.4%), followed by weak supervision (28.5%). Open-set/few-shot and semantic/training-free/VAU categories are smaller in absolute terms but become increasingly visible in the 2024–2026 frontier. The code-link column records whether a public repository link was directly identifiable in the audited metadata; absence of a linked repository is therefore not interpreted as proof that no implementation exists.

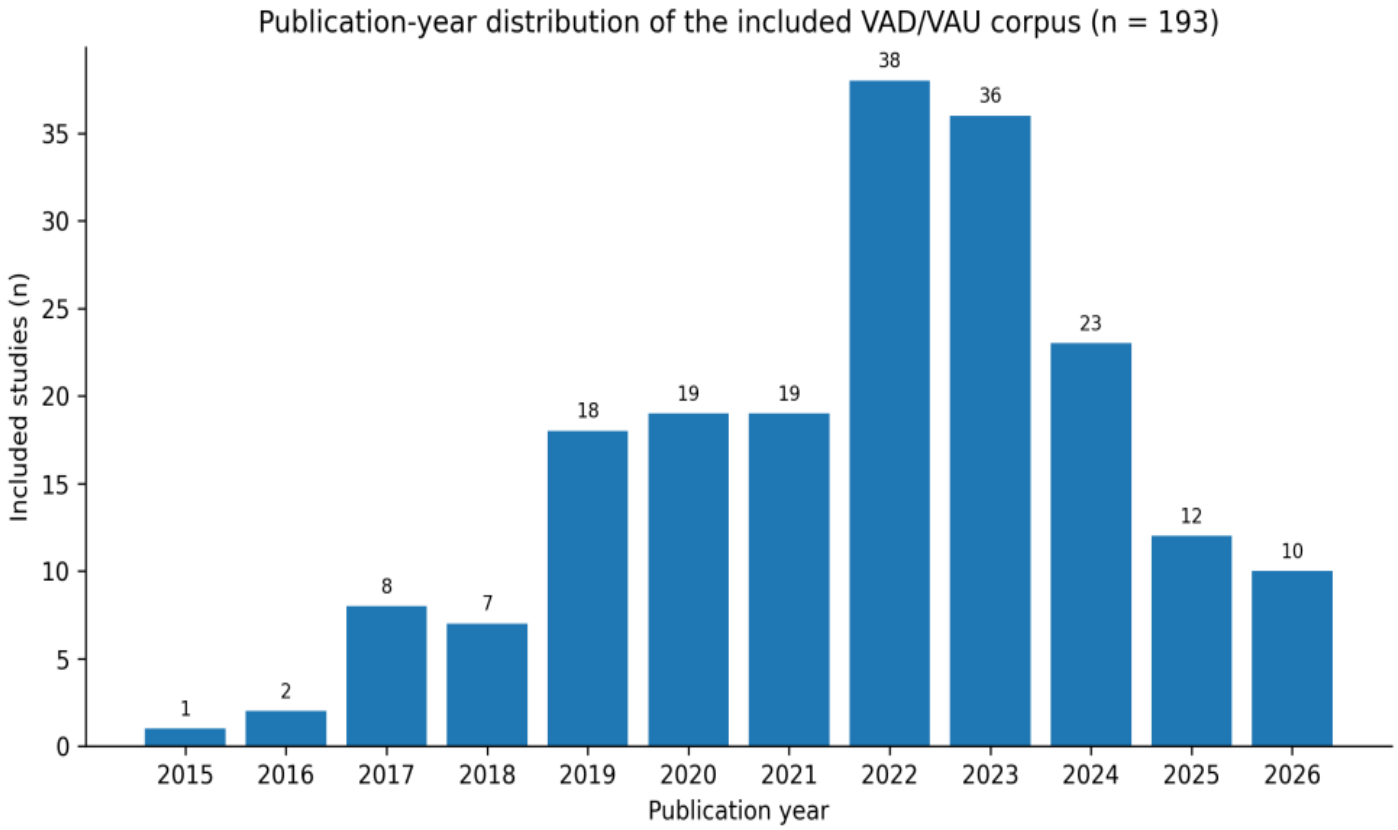


Figure 3. Publication-year distribution of the bounded 193-study core audited VAD/VAU corpus.

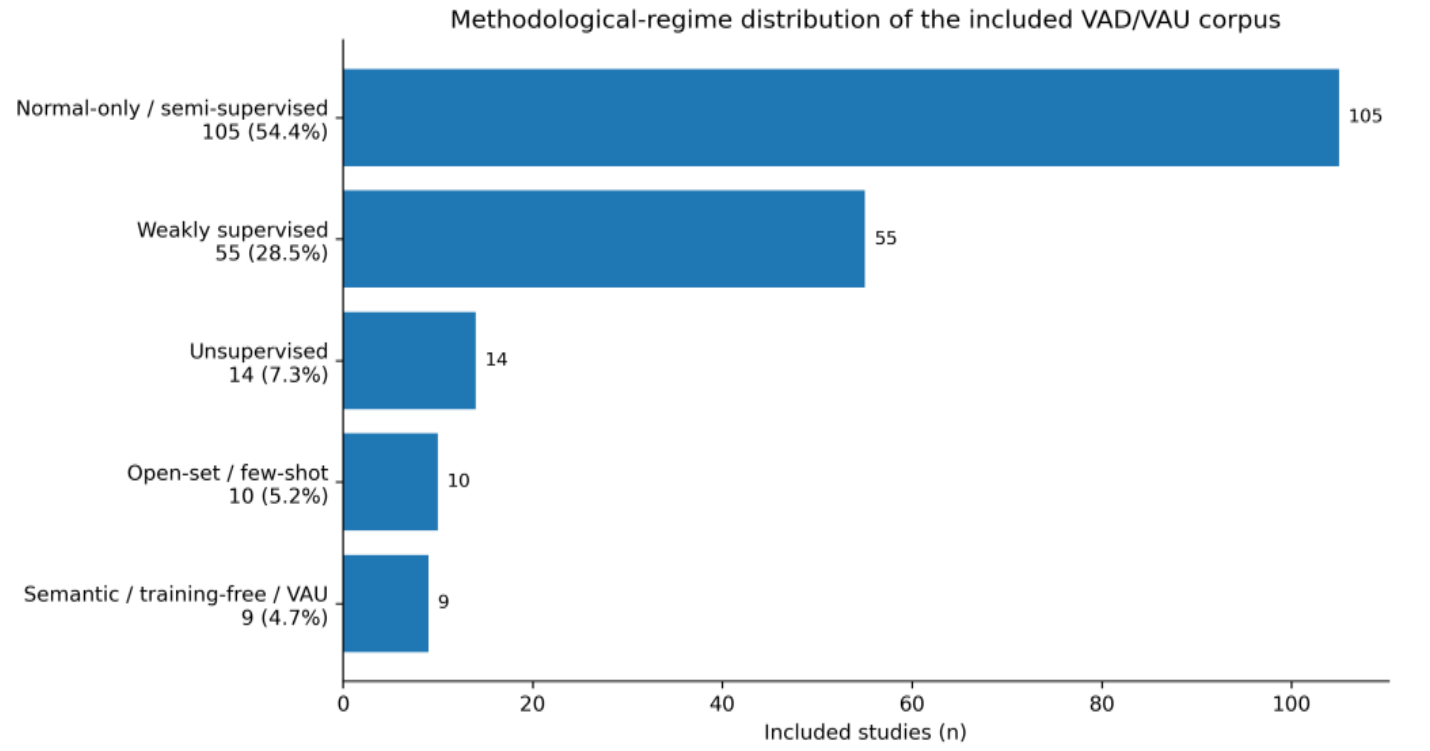


Figure 4. Methodological-regime distribution of the bounded core audited VAD/VAU corpus (n = 193).

EVOLUTION OF VIDEO ANOMALY DETECTION (RQ1)

Statistical, sparse, and dictionary-based normality

Before deep representation learning became dominant, VAD commonly modelled appearance and motion with local descriptors, dynamic textures, sparse dictionaries, and reconstruction cost. Mahadevan et al. combined appearance and dynamics through mixtures of dynamic textures and introduced the UCSD crowded-scene benchmark [33]. Lu et al. emphasised computational efficiency through sparse combination learning and reported 140–150 frames per second on a desktop MATLAB implementation, an early reminder that anomaly detection is a systems problem as well as an accuracy problem [30].

The enduring contribution of these approaches is conceptual: normal behaviour is represented compactly, and anomaly is operationalised as poor fit to that representation. Later deep methods often replace the handcrafted representation but retain the same logic. This continuity matters when interpreting claims of novelty; many modern memory, metric, and one-class systems can be understood as more expressive mechanisms for storing or estimating the support of normality.

Reconstruction, prediction, and temporal regularity

Autoencoder-based VAD made it possible to learn both the representation and the reconstruction function from video data. Hasan et al. used autoencoders to learn temporal regularity with little or no supervision [19]. Luo et al. connected temporally coherent sparse coding with a stacked recurrent network and introduced the large ShanghaiTech benchmark, bringing multi-scene complexity to the centre of VAD evaluation [31]. Liu et al. shifted the proxy task from reconstructing the observed frame to predicting a future frame, arguing that abnormal events should be harder to predict under a model trained on normal dynamics [27].

These methods established a powerful template but also exposed a structural weakness: a high-capacity decoder can sometimes reconstruct or predict anomalous content surprisingly well. Reconstruction error is therefore not a guaranteed monotonic measure of abnormality. This motivated memory restrictions, adversarial constraints, object-centric decomposition, synthetic anomalies, feature-level one-class objectives, and alternative score geometries.

Memory, object-centric, and one-class formulations

Ionescu et al. moved from whole-frame reasoning toward object-centric representations and used normality clusters as "dummy anomalies" in a one-versus-rest formulation [22]. Gong et al. augmented an autoencoder with an explicit memory of prototypical normal patterns, while Park et al. introduced memory-guided normality with compactness and separateness constraints [17], [38]. Wu et al. developed a deep one-class neural formulation for complex scenes, illustrating the broader movement from pixel reconstruction toward discriminative normality boundaries [57].

Memory and object-centric methods address different failure modes. Memory limits the representational freedom available for anomalous inputs, while object-centric analysis reduces dilution by static background. Their comparison should therefore consider not only AUC but also object detector dependencies, memory size, inference complexity, and the kinds of anomalies each representation can express.

Weak supervision and real-world long videos

The UCF-Crime benchmark changed the scale and supervision pattern of the field. Sultani et al. introduced a dataset of 1,900 long untrimmed surveillance videos spanning approximately 128 hours and used video-level labels within a multiple-instance ranking framework [49]. Weak supervision made large real-world datasets feasible but introduced label ambiguity: an "abnormal" bag can contain many normal snippets, while the temporal boundary of an event remains unknown.

This noisy-label structure remains a central research problem. Prompt-enhanced MIL injects semantic priors to distinguish abnormal events and normal context [9]. More recent approaches explicitly model weakly labelled information or noisy supervision. TLMA uses hard normal anchors, contrastive learning, and motion-aware enhancement [59], while D2MIL removes suspected noisy instances and uses a vision-language model to recover difficult abnormal samples [67]. These developments suggest that the next gains in weak supervision may depend as much on label-noise management and boundary quality as on larger feature encoders.

Unsupervised, efficient, and training-free directions

Unsupervised VAD attempts to remove even the assumption of a clean normal-only training set. Zaheer et al. proposed generative cooperative learning in which generator and discriminator components cross-supervise without labelled anomalies [63]. Efficiency has become a parallel concern. Ristea et al. reported a lightweight self-distilled masked autoencoder that processed 1,655 FPS while remaining competitive on four benchmarks [43]. VideoPatchCore removes training altogether and organises normal video features in tailored memories [4]. VADMamba explores state-space modelling for fast temporal processing, combining frame prediction and optical-flow reconstruction [32].

These papers make clear that "best" VAD depends on the operating objective. A model that gains a small AUC increment at ten times the latency may be inappropriate for edge surveillance; conversely, a fast detector may be unsuitable when precise spatial localisation or explanation is required. The synthesis therefore avoids collapsing accuracy, speed, memory, and annotation cost into a single ranking.

Table 9. Methodological evolution and the evaluation questions that accompany each family.

Era / family	Representative studies	Core idea	Evaluation issue that remains relevant
Statistical / sparse	Mahadevan et al. [33]; Lu et al. [30]	Model compact normal appearance-motion structure; detect poor fit.	Scene specificity, handcrafted features, operational speed.
Deep regularity / prediction	Hasan et al. [19]; Luo et al. [31]; Liu et al. [27]	Learn normal reconstruction or future dynamics.	Powerful models may reconstruct anomalies; score normalisation varies.

Memory / object / one-class	Ionescu et al. [22]; Gong et al. [17]; Wu et al. [57]; Park et al. [38]	Restrict normal prototypes, localise objects, learn compact normal boundary.	Detector dependencies, memory design, object-vs-frame scoring.
Weak supervision / MIL	Sultani et al. [49]; Chen et al. [9]; Xu et al. [59]; Zhao et al. [67]	Learn temporal anomaly localisation from coarse video labels.	Label noise, bag construction, temporal boundary ambiguity.
Efficiency / training-free	Ristea et al. [43]; Ahn et al. [4]; Lyu et al. [32]	Reduce training or inference cost while retaining accuracy.	Hardware dependence, inconsistent FPS reporting, offline components.
Vision-language / reasoning	Wu et al. [55], [56]; Zanella et al. [64]; Yang et al. [60]; Ye et al. [62]	Use language-aligned representations, open vocabulary, reasoning, explanation.	Prompt dependence, semantic evaluation, model version drift, cost.

DATASET AND BENCHMARK LANDSCAPE (RQ2)

VAD benchmarks do more than supply data; they define what "anomaly" means and therefore shape the method families that appear successful. Small fixed-camera datasets reward precise modelling of a narrow normal distribution. Long untrimmed crime datasets reward weak supervision and temporal localisation. Multiscene benchmarks expose domain variation, while open-set and open-vocabulary benchmarks test whether systems can go beyond a predefined anomaly list. A review that ignores these differences risks interpreting dataset properties as model properties.

UCSD Ped2, CUHK Avenue, and ShanghaiTech remain influential in normal-only VAD. Ped2 contains 16 training and 12 testing sequences of a pedestrian walkway, with anomalous objects such as bicycles and vehicles appearing in test data [33], [51]. Avenue contains 16 training and 21 testing clips (30,652 frames in total), with rectangular ground-truth annotations for abnormal events [30], [12]. ShanghaiTech was introduced to increase scale and scene diversity and is commonly reported with 330 training and 107 testing videos across 13 scenes [31], [46]. These datasets are valuable for continuity, but their extensive reuse also raises the possibility that benchmark optimisation outpaces real-world generalisation.

UCF-Crime shifted the field toward long, real surveillance footage and weak labels [49]. XD-Violence extended this trajectory with 4,754 untrimmed videos, 217 hours of content, and audio signals, enabling audio-visual modelling under weak supervision [58]. Street Scene was created specifically in response to small datasets and evaluation criteria that the authors argued were poorly aligned with practical anomaly detection; it introduced event-oriented criteria intended to better reflect real deployment [40]. Cross-modal weakly supervised work also illustrates how modality fusion changes the information available to the detector and should therefore be coded explicitly [16].

Newer benchmarks deliberately target shortcomings of older formulations. UBnormal provides a supervised open-set setting with pixel-level anomaly annotations during training while keeping train and test anomaly classes disjoint [3]. NWPU Campus contains 43 scenes, 28 abnormal-event classes, and 16 hours of video and explicitly models scene-dependent anomalies; it also supports anomaly anticipation [7]. These benchmarks make it possible to ask whether a method understands scene context rather than memorising an anomaly appearance.

Table 10. Representative VAD benchmarks and the evaluation questions they enable.

Benchmark	Scale / setting	Supervision / annotation relevance	Why it matters for this review
UCSD Ped2	16 train, 12 test sequences; fixed pedestrian walkway.	Normal-only training in canonical use; frame-level test annotations; some spatial masks in UCSD resources.	Classic small-scene continuity; susceptible to saturation and protocol variation.
CUHK Avenue	16 train, 21 test clips; 30,652 frames; fixed campus avenue.	Normal training; test anomalies; rectangular spatial ground truth.	Widely used single-scene benchmark; useful for analysing spatial vs frame evaluation.

ShanghaiTech	Commonly 330 train, 107 test videos; 13 scenes.	Normal-only training in classic protocol; frame/pixel anomaly annotations in test.	Cross-scene variability; protocol and aggregation details materially affect interpretation.
UCF-Crime	1,900 long untrimmed videos; ~128 h; 13 anomaly categories.	Video-level labels support weak supervision; temporal localisation evaluated on test.	Large real-world benchmark; central to MIL and long-video VAD.
XD-Violence	4,754 untrimmed videos; 217 h; audio-visual.	Weak labels; multimodal audio + visual signals.	Tests multimodal and violence-oriented anomaly detection.
Street Scene	Large, varied street-surveillance dataset.	Event-based evaluation criteria introduced with dataset.	Highlights mismatch between frame metrics and practical event detection.
UBnormal	Synthetic multi-scene benchmark with disjoint anomaly types.	Pixel-level anomaly labels at training; supervised open-set design.	Enables direct comparison of one-class and supervised open-set learning.
NWPU Campus	43 scenes; 28 abnormal classes; 16 h.	Scene-dependent anomalies; detection and anticipation.	Tests context sensitivity, scene diversity, and proactive anomaly modelling.

Results obtained from different benchmarks should be interpreted with caution. A higher score on a narrow fixed-camera dataset does not necessarily indicate that a method is stronger than another method evaluated on a more heterogeneous long-video benchmark. The two settings may differ substantially in scene complexity, anomaly diversity, temporal structure, and evaluation difficulty, and therefore pose different statistical and semantic challenges.

Dataset reuse, saturation, and hidden comparability problems

Repeated benchmark use is valuable because it supports longitudinal comparison, yet it can also encourage implicit tuning to test sets. This risk is particularly acute when papers report many architectural variants on a fixed public test set without a separate validation protocol. The extraction therefore distinguishes "validation-selected final test" from "descriptive test-set model comparison." The latter can be useful for ablation, but should not be presented with the same evidential strength as a held-out final result.

Another comparability problem is that a dataset label may hide changed input representations. For example, a paper may use RGB frames, precomputed optical flow, object crops from a detector, skeleton trajectories, or frozen CLIP features on the same benchmark. These are legitimate choices, but they alter information content, computational dependencies, and sometimes the effective sample unit. The review records representation and preprocessing alongside every score rather than treating the benchmark name as a sufficient descriptor.

EVALUATION PRACTICE: BEYOND A SINGLE AUC (RQ3)

Evaluation granularity

VAD metrics operate at different granularities. Video-level detection asks whether a video contains an anomaly; clip- or frame-level detection assigns temporal scores; pixel- or object-level localisation asks where the abnormal event occurs; event-level evaluation asks whether distinct anomalous episodes were detected; VAU additionally evaluates category and explanation. A method can perform well at one granularity and poorly at another. Scores are therefore not pooled across granularities even when the metric name is identical.

Street Scene is an important precedent because its authors explicitly argued that conventional criteria were flawed for practical anomaly detection and proposed evaluation measures that better reflect event detection [40]. More broadly, metric-taxonomy research shows that anomaly-detection measures encode different assumptions about granularity, thresholding, ranking, and class imbalance [47]. This motivates a general principle: metric choice should follow the operational question, not merely historical convention.

ROC-AUC, precision–recall, and prevalence

ROC-AUC remains common in VAD because it is threshold independent and measures ranking across positive and negative observations. Precision–recall analysis emphasises the quality of positive predictions and is often recommended for highly skewed problems [15], [44]. At the same time, more recent analysis has clarified that ROC-AUC itself is invariant to class prevalence under fixed score distributions, whereas PR-AUC changes with prevalence [42]. These perspectives are not contradictory when interpreted carefully: ROC and PR answer different questions, and neither removes the need to state the evaluated population.

The practical implication is not to replace one universal metric with another. ROC-AUC and AP/PR-AUC are extracted separately; the exact positive/negative population and aggregation are recorded; and direct comparisons are restricted to compatible prevalence and sampling regimes where appropriate. For deployment-oriented studies, thresholded measures such as false-alarm rate, event recall, time-to-detection, or precision at an operating point can be more actionable than either area-under-curve summary [61].

Mathematical interpretation of common VAD metrics

The phrase "frame-level AUC" is often treated as self-explanatory, but its meaning depends on the observations included in the positive and negative populations. Let $y_t \in \{0,1\}$ be the ground-truth label for frame t, s_t a continuous anomaly score, and τ a decision threshold. The basic operating-point quantities are

$$\mathrm{TPR}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau)+\mathrm{FN}(\tau)}, \qquad \mathrm{FPR}(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau)+\mathrm{TN}(\tau)} \quad (6)$$

ROC-AUC can then be viewed either geometrically as the area under the ROC curve or probabilistically as the chance that an anomalous observation receives a higher score than a normal observation (with half credit for ties):

$$\mathrm{AUC} = P(s_A > s_N) + \frac{1}{2}P(s_A = s_N) \quad (7)$$

This probabilistic interpretation exposes an important issue for pooled frame-level evaluation. If all anomalous and normal frames across all test videos are pooled, the reported AUC contains every cross-video positive-negative pair:

$$\mathrm{AUC}_{\mathrm{pool}} = \frac{1}{N_A N_N}\sum_{i:y_i=1}\sum_{j:y_j=0}\left[I(s_i > s_j) + \frac{1}{2}I(s_i = s_j)\right] \quad (8)$$

Consequently, a model can receive credit for separating videos or scenes that differ globally even when within-video temporal localisation is weaker. This is not an argument that pooled AUC is invalid; it is an argument that its comparison population must be stated. Caetano et al. [6] reached a related conclusion from benchmark experiments, showing that commonly used reference metrics can materially alter conclusions about which supervision paradigm appears strongest. More recently, Acharya et al. [2] showed that high AUROC in weakly supervised VAD can coexist with poor recall at practically meaningful false-positive rates.

Precision and recall emphasise positive predictions. At threshold τ, precision $P(\tau)=TP/(TP+FP)$ and recall $R(\tau)=TP/(TP+FN)$. Average precision can be expressed as a weighted sum over successive recall increments:

$$\mathrm{AP} = \sum_{n=1}^{N}(R_n - R_{n-1})\,P_n \quad (9)$$

For highly imbalanced anomaly populations, AP is often more sensitive than ROC-AUC to false positives, but it is also prevalence dependent. Therefore, AP values are compared only after verifying the evaluated positive prevalence and sampling construction. Where event localisation is reported, temporal intersection-over-union (IoU) is also extracted:

$$\mathrm{IoU}(E,\hat{E}) = \frac{|E\cap\hat{E}|}{|E\cup\hat{E}|} \quad (10)$$

A detected event can then be counted as a true positive only when its IoU with a ground-truth event exceeds a declared threshold. This makes the operating definition explicit and helps distinguish "ranking frames well" from "detecting incidents as incidents." Two recent evaluation-focused preprints, retained in the emerging-frontier evidence set rather than peer-reviewed corpus counts, further motivate this distinction: Liu et al. propose annotation-averaged AUC/AP and latency-aware AP [29], while Abdulaziz and Bondarev audit pooled AUC on UCF-Crime and report rank changes under stricter within-video comparisons [1].

Micro, macro, frame, clip, and event aggregation

Two papers can report "AUC" while aggregating scores differently. Micro aggregation pools observations across videos and therefore weights long videos more strongly. Macro aggregation computes per-video or per-scene metrics before averaging and gives each unit more equal influence. Frame-level pooling can also make long normal stretches dominate the denominator, whereas event-level metrics emphasise whether distinct incidents are found. The aggregation rule is therefore recorded explicitly and unspecified aggregation is treated as a reporting deficiency.

For V videos, macro aggregation can be written as an unweighted mean of per-video metrics, whereas micro aggregation first pools all frame-level predictions. For AUC, the macro form is

$$\mathrm{AUC}_{\mathrm{macro}} = \frac{1}{V}\sum_{v=1}^{V} \mathrm{AUC}_v \qquad (11)$$

The two summaries answer different questions. Micro AUC gives greater influence to videos contributing more frames or more positive-negative pairs; macro AUC gives each video equal weight. The evidence tables therefore treat the aggregation rule as part of the metric definition rather than as a footnote.

Score normalisation and smoothing are equally important. Min–max normalisation per video can alter cross-video calibration; temporal moving averages can improve apparent frame-level smoothness without changing the underlying detector; post-processing with ground-truth-informed parameters can create leakage. Consequently, the extraction form treats post-processing as part of the evaluated method, not as an incidental implementation detail.

Table 11. Evaluation details required for defensible cross-paper comparison.

Evaluation element	Common variants	Risk if omitted	Review handling
Granularity	video, clip, frame, object, pixel, event, semantic explanation	Same metric name can measure different tasks.	Separate strata; no cross-granularity pooling.
ROC / PR	ROC-AUC, AP, PR-AUC, partial AUC	Different prevalence sensitivity and interpolation conventions.	Record exact implementation; AP not silently equated with trapezoidal PR-AUC.
Aggregation	micro, macro, per-scene, pooled frames	Long videos/scenes may dominate pooled result.	Tier downgrade when aggregation differs or is unspecified.
Post-processing	smoothing, normalisation, calibration, thresholding	Can materially change ranking or false-alarm burden.	Treat as part of protocol; extract parameters and fitting data.
Model selection	validation-based, test-set descriptive, best-of-many	Test reuse inflates confidence in "best" model.	Separate held-out final results from exploratory test comparisons.
Uncertainty	single run, multi-seed SD/CI, bootstrap	Tiny score differences may not be meaningful.	Flag single-seed learnable results; prioritise uncertainty-aware evidence.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

Protocol tiers in practice

The framework is applied at the level of a comparison, not as a permanent label attached to an entire paper. Tier A requires two reported results to answer substantially the same empirical question: the same benchmark version and test split, compatible supervision, the same evaluation granularity and metric population, and no material difference in preprocessing or aggregation that changes the evaluated result. Tier B captures comparisons that remain informative but differ in a material dimension such as crop/preprocessing, representation, or external information access. Tier C covers results for which the evaluated population, temporal context, task, or metric meaning changes enough that a direct numerical rank would be misleading.

Table 12 demonstrates these rules using primary studies. The first two cases show strict Tier-A comparisons after preprocessing is made explicit: a one-crop I3D-RGB pair and a ten-crop I3D-RGB stratum. A third case shows why one-crop and ten-crop I3D results should not be silently merged. The remaining examples show how external semantic priors, a changed evaluation population, and a bounded real-time decision window alter the meaning of a headline AUC. The purpose of a Tier downgrade is therefore not to penalise methodological innovation, but to limit the strength of the numerical claim that can be made across incompatible protocols.

Table 12. Worked applications of the Protocol Compatibility Framework.

Comparison	Shared protocol elements	Material difference	Tier / synthesis consequence
MIST [69] 82.30% vs MSL [71] 85.30%, UCF-Crime	Same canonical benchmark/test set; weak video-level supervision; one-crop I3D-RGB; frame-level ROC-AUC	Architecture/training objective differs, which is the object of comparison	Tier A: direct one-crop comparison
RTFM [70] 84.30% vs UR-DMU [72] 86.97% vs MGFN [73] 86.98%, UCF-Crime	Same benchmark family; weak video-level supervision; ten-crop I3D-RGB; frame-level ROC-AUC	Architecture/training objective differs	Tier A: direct ten-crop stratum
MSL [71] 85.30% vs UR-DMU [72] 86.97%, UCF-Crime	Same benchmark; weak supervision; I3D-RGB; frame-level ROC-AUC	One-crop versus ten-crop preprocessing	Tier B: numerical comparison requires preprocessing qualification
RTFM [70] 84.30% vs VadCLIP [56] 88.02%, UCF-Crime	Same benchmark; weak supervision; overall frame-level ROC-AUC	I3D visual representation versus frozen CLIP plus language-image semantic prior	Tier B: qualify information access and representation
VadCLIP [56] overall AUC 88.02% vs anomaly-video-only AUC 70.23%	Same method and benchmark	Positive/negative evaluation population changes	Tier C: values answer different questions and are not interchangeable
MGFN [73] offline 86.98% vs REWARD-window re-evaluation of MGFN [74] 81.76% at 8.533 s	Same UCF-Crime family and AUC concept	Bounded real-time decision window replaces offline temporal context	Tier C: operational-context comparison only

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

PROTOCOL-COMPATIBLE PERFORMANCE AND REPRODUCIBILITY (RQ4)

Quantitative synthesis and benchmark context

The extended audit did not support treating all six candidate I3D-RGB UCF-Crime results as one strict Tier-A stratum. Instead, preprocessing separated the directly comparable results into two sub-strata. MIST [69] and MSL [71] both report one-crop I3D-RGB under weak video-level supervision and frame-level ROC-AUC, with AUCs of 82.30% and 85.30%, respectively, an absolute difference of 3.00 percentage points. RTFM [70], UR-DMU [72], and MGFN [73] use ten-crop I3D-RGB under the same benchmark family, supervision regime, and frame-level ROC-AUC, with AUCs of 84.30%, 86.97%, and 86.98%; the median is 86.97% and the range is 84.30%–86.98%. S3R [76] reports 85.99% with I3D-RGB but is not included in the strict quantitative summary because the crop setting is not explicit in the primary paper. Table 13 reports these study-level protocol decisions. The resulting summaries are descriptive protocol-compatible comparisons, not a conventional meta-analysis: the studies reuse the same fixed benchmark test set, do not report comparable sampling variances, and may share pretrained features or evaluation conventions.

Table 13. Primary-study UCF-Crime protocol-compatible quantitative strata.

Study	Year	I3D preprocessing / supervision	Frame ROC-AUC (%)	Compatibility status
MIST [69]	2021	One-crop I3D-RGB; weak video-level labels	82.3	Tier A-1 (one-crop)
MSL [71]	2022	One-crop I3D-RGB; weak video-level labels	85.3	Tier A-1 (one-crop)
RTFM [70]	2021	Ten-crop I3D-RGB; weak video-level labels	84.3	Tier A-2 (ten-crop)
UR-DMU [72]	2023	Ten-crop I3D-RGB; weak video-level labels	86.97	Tier A-2 (ten-crop)
MGFN [73]	2023	Ten-crop I3D-RGB; weak video-level labels	86.98	Tier A-2 (ten-crop)
S3R [76]	2022	I3D-RGB; crop setting not explicit in primary paper	85.99	Extended audit only; excluded from strict Tier-A summary
Summary		Tier A-1: n=2; Tier A-2: n=3	Tier A-1 difference = +3.00 pp; Tier A-2 median = 86.97, range = 84.30–86.98	Descriptive within-protocol comparisons; no meta-analysis

No raw AUC values are pooled across Tier B or Tier C comparisons. VAD studies often reuse the same datasets, pretrained backbones, public features, code bases, and evaluation scripts, so their reported scores are neither statistically independent nor guaranteed to measure an identical population. The Tier-A sub-strata above are therefore described with simple within-protocol summaries rather than a pooled inferential effect. Where preprocessing or another protocol component changes materially, the comparison is retained as qualified or contextual evidence.

The review also avoids defining a mean-plus-standard-deviation threshold as “state of the art.” Such a threshold can look precise while mixing supervision, preprocessing, aggregation, model-selection, and test-population differences. Within a verified Tier-A stratum, transparent within-protocol summaries and clearly defined performance differences are more interpretable. Outside such strata, numerical results are used to map the literature rather than to construct a universal league table.

For any Tier-A stratum containing results $x_1,\ldots,x_n$ for the same metric, robust summaries are preferred because AUC/AP values are bounded and often concentrated near the upper end of the scale. Central tendency and spread are therefore represented by

$$\tilde{x} = \text{median}(x_1, \ldots, x_n), \qquad \text{IQR} = Q_3 - Q_1 \qquad (12)$$

A descriptive upper-performance band is reported as [Q_3, max(x)] when a stratum is sufficiently populated. This is not labelled a universal "state-of-the-art threshold." For controlled comparisons against a common baseline b within the same protocol, absolute performance change is recorded as

$$\Delta_{i,b} = x_i - x_b \qquad (13)$$

The primary-study protocol audit provides the main quantitative demonstration of these rules. Secondary-source performance tables accompanying Wu et al. [54] remain useful only as broader benchmark context. Under normal-only/semi-supervised reporting, Ped2 is concentrated near the ceiling (median 96.9% AUC across 29 tabulated results), whereas ShanghaiTech has a substantially broader and lower distribution (median 77.0% across 31 results). Under weak supervision, the secondary UCF-Crime median is 86.10% (n = 19). These values are retained to illustrate saturation and heterogeneity, not to override the primary protocol audit or to imply that all tabulated scores are Tier-A compatible.

Table 14. Secondary-source benchmark-context distributions derived from the performance tables accompanying Wu et al. [54]. These summaries are diagnostic context only and are not Tier-A pooled estimates.

Paradigm	Benchmark	n	Median AUC (%)	IQR (%)	Range (%)
Normal-only / semi-supervised	UCSD Ped2	29	96.9	95.40–98.30	90.00–99.80
Normal-only / semi-supervised	Avenue	33	90	86.00–91.50	72.00–93.70
Normal-only / semi-supervised	ShanghaiTech	31	77	73.65–80.95	68.00–90.20
Normal-only / semi-supervised	UBnormal	5	59.3	58.50–62.70	56.40–71.80
Weakly supervised	UCF-Crime	19	86.1	83.67–86.90	75.40–88.08
Weakly supervised	XD-Violence	12	80.78	78.40–82.17	75.41–88.05
Weakly supervised	ShanghaiTech	11	97.48	96.02–97.71	84.44–98.35
Weakly supervised	TAD	2	92.3	—	91.66–92.93

Table 14 reinforces why a single cross-dataset ranking is inappropriate. Small fixed-camera datasets can exhibit near-saturated AUC distributions while larger multi-scene benchmarks retain substantially more headroom. The weakly supervised tables also mix datasets with different anomaly definitions and temporal annotation practices. The values therefore map reported performance levels; only the preprocessing-matched primary-study comparisons in Table 13 are treated as directly comparable.

Reproducibility is more than an open repository

Code availability is useful but insufficient as a reproducibility indicator. Within the bounded core register, a public repository link is directly identifiable in the audited metadata for 79 of 193 studies (40.9%). This number is reported as a descriptive property of the assembled evidence set, not as an estimate of the proportion of all VAD publications with code. The review therefore removes population-style confidence-interval language from the headline result. Repositories may exist even when they are not linked from the audited source, and links may also disappear or change over time.

The audit consequently separates artefact linkage from experimental completeness. Within the core register, identifiable code links occur in 39/105 normal-only/semi-supervised records, 20/55 weakly supervised records, 7/14 unsupervised records, 5/10 open-set/few-shot records, and 8/9 semantic/training-free/VAU records. These within-corpus proportions are descriptive only. Reproducibility assessment additionally considers source code, pretrained weights, split files, evaluation implementation, preprocessing detail,

hyperparameters, random seeds, software versions, and repeated-run uncertainty where those fields are available in the extended audit.

Table 15. Reproducibility and reporting audit used in RQ4.

Dimension	Minimum evidence	Stronger evidence
Code	Implementation or sufficient algorithmic detail.	Versioned repository with environment file and release tag.
Weights / features	Checkpoint or reproducible training procedure.	Public weights and exact feature extraction pipeline.
Data split	Exact train/validation/test definition.	Machine-readable split files and dataset version/hash.
Evaluation	Metric name and aggregation described.	Released evaluation script with post-processing parameters.
Randomness	Seed or deterministic claim.	Multiple seeds with SD/CI and statistical comparison where relevant.
Compute	Hardware family reported.	Latency/FPS, batch size, precision, memory, model size/FLOPs under stated hardware.
Model selection	Validation procedure described.	Pre-registered/explicit selection criterion with untouched final test.
Failure analysis	Qualitative limitations.	Per-class/per-scene errors, false-alarm examples, calibration and domain-shift analysis.

Evidence patterns in the bounded core corpus and representative literature

The bounded core corpus makes several shifts visible without requiring protocol-incompatible score pooling. Normal-only/semi-supervised VAD is the largest coded family (105/193), followed by weak supervision (55/193). Unsupervised, open-set/few-shot, and semantic/training-free/VAU studies together account for 33 core records. These values describe the composition of the audited register, not the prevalence of each regime in the entire VAD literature. Figure 3 is therefore interpreted as a distribution of the reviewed core evidence rather than a complete bibliometric census.

Representative methods also confirm that “supervision” now includes information beyond anomaly labels. VadCLIP exploits frozen vision-language priors [56]; Open-Vocabulary VAD introduces semantic knowledge and synthetic unseen-anomaly reasoning [55]; LAVAD uses captioning and LLM reasoning without target-domain training [64]; and VERA moves toward explanation-aware VAU [62]. These systems cannot be ranked fairly against classical normal-only models without recording external pretraining, prompts, language-model access, and output task. The corpus-level taxonomy therefore treats information access as part of the protocol rather than as an implementation footnote.

DEPLOYMENT READINESS (RQ5)

Causality and temporal context

A model intended to raise an alarm in a live stream cannot use future frames that have not yet arrived. Yet many offline VAD methods employ bidirectional temporal context, whole-video normalisation, or smoothing windows centred on the target frame. These choices can be valid for forensic review or delayed monitoring, but they should not be described as strict real-time inference without a causal implementation. The review therefore codes each method as causal/online, bounded-delay, or offline/retrospective.

Anomaly anticipation introduces a related but distinct requirement: the system must provide useful warning before or near event onset. NWPU Campus explicitly supports anticipation as well as detection [7]. Future deployment claims should therefore distinguish detection latency from computational latency: a fast model can still be operationally late if it requires many future frames.

Computational cost and edge feasibility

Runtime claims are hardware dependent. FPS without image resolution, batch size, hardware, numerical precision, decoding cost, and preprocessing can mislead. The review extracts these contextual factors and does not directly rank FPS from incompatible platforms. Model parameters and FLOPs are more portable but still incomplete because memory movement, feature extraction, object detection, optical flow, and external APIs can dominate end-to-end latency.

Efficient models deserve separate attention because surveillance deployments may require dozens or hundreds of concurrent streams. Lu et al. made speed an explicit design goal as early as 2013 [30], and recent work such as Ristea et al. and VADMamba shows renewed interest in high-throughput VAD [43], [32]. A citation-worthy review should therefore treat efficiency as a first-class axis rather than an afterthought.

To make runtime reporting operationally interpretable, the review decomposes end-to-end latency rather than treating the neural-network forward pass as the whole system:

$$T_{\mathrm{E2E}} = T_{\mathrm{decode}} + T_{\mathrm{pre}} + T_{\mathrm{feature}} + T_{\mathrm{model}} + T_{\mathrm{post}} + T_{\mathrm{I/O}} \quad (14)$$

For a live stream with input rate FPS_stream and measured end-to-end throughput FPS_E2E, a simple throughput adequacy ratio is

$$\rho = \frac{\mathrm{FPS}_{\mathrm{E2E}}}{\mathrm{FPS}_{\mathrm{stream}}} \quad (15)$$

A value $\rho \geq 1$ is a necessary throughput condition for keeping pace with a single stream, but it is not sufficient for low-latency alerting because buffering, batching, future-frame context, and queueing can still delay decisions. Similarly, if a model contains P stored parameters at b bits per parameter, parameter memory alone is

$$M_{\mathrm{weights}} = \frac{P\,b}{8} \text{ bytes} \quad (16)$$

This excludes activations, feature caches, video buffers, detector/optical-flow modules, and framework overhead. The distinction is important for edge deployment, where memory movement and preprocessing may dominate. A 2026 edge-oriented VAD survey explicitly analyses RGB, skeleton, and multimodal systems using latency, model size, FLOPs, and FPS, and identifies a persistent gap between benchmark accuracy and deployable edge solutions [23]. The present review extends that concern by requiring the context in which each efficiency number was obtained.

Privacy, domain shift, and multi-camera operation

Video surveillance can expose faces, identities, locations, and behavioural patterns. Privacy-preserving VAD has become sufficiently developed to warrant a dedicated survey, covering privacy risks and techniques at the data, feature, and system levels [28]. The present review does not duplicate that taxonomy; instead, it codes whether primary VAD studies make a privacy claim, whether privacy preservation is evaluated quantitatively, and whether the method requires transmitting raw video to a remote service.

Domain shift remains another major barrier. A model calibrated on one camera may face different viewpoints, illumination, crowd density, cultural context, or anomaly base rates elsewhere. Multi-scene datasets such as ShanghaiTech and NWPU Campus partially expose this issue [31], [7], but cross-dataset and cross-camera evaluation remains a stronger test of deployment robustness. The synthesis distinguishes within-dataset testing, cross-scene evaluation, cross-dataset transfer, and online adaptation where reported.

Human factors, false alarms, and actionable explanations

A deployed VAD system is part of a decision process. Excessive false alarms increase operator workload and can cause alarms to be ignored. Conversely, an opaque high anomaly score may be difficult to act on.

Deployment-oriented evaluation should therefore report the false-alarm burden at useful operating points and, where possible, time-to-detection, event recall, calibration, and explanation quality rather than only ranking metrics.

Foundation-model systems create an opportunity to produce human-readable rationales, but explanation introduces new failure modes. Plausible language is not necessarily faithful to visual evidence. VERA attempts to guide VLM reasoning with learned questions [62], while recent VAU evaluation research explicitly targets semantic validity and robustness of anomaly descriptions [24]. Human trust should therefore be earned through grounded evidence, not assumed from fluent text.

EMERGING FRONTIER: FROM VAD TO OPEN-WORLD VIDEO ANOMALY UNDERSTANDING

Vision-language adaptation

Vision-language models change the representation layer by aligning visual content with semantic concepts learned from large-scale pretraining. VadCLIP demonstrates that frozen CLIP representations can be adapted to weakly supervised VAD through visual classification and language-image alignment branches [56]. This creates fine-grained semantic outputs without training a visual backbone from scratch, but it also introduces dependence on the external pretraining corpus and text prompts.

Open-Vocabulary VAD takes a further step by asking the system to detect and categorise seen and unseen anomalies. Wu et al. decouple class-agnostic detection from class-specific classification and use semantic knowledge and generated pseudo-unseen anomalies to extend category coverage [55]. Such methods challenge the conventional binary normal/anomalous evaluation and require category-aware metrics and careful handling of “unknown” events.

Training-free language reasoning

LAVAD uses an off-the-shelf visual captioner to describe frames, cleans captions through cross-modal similarity, aggregates temporal descriptions with an LLM, and derives anomaly scores without target-domain training [64]. AnomalyRuler instead infers normality rules from few normal examples and applies those rules during anomaly reasoning [60]. These approaches are attractive where retraining is expensive, but reproducibility now depends on prompts, model provider/version, decoding parameters, API changes, and sometimes non-deterministic inference.

The review therefore extends the reproducibility checklist for foundation-model VAD to include model name and exact checkpoint/API version; system and user prompts; temperature and decoding settings; frame/caption sampling; context-window construction; external knowledge; cost and latency; and whether outputs were cached. Without these details, a nominally “training-free” pipeline may be difficult to reproduce months later.

Explainable VAU and trustworthy evaluation

Explainable VAD aims to move from a number to an account of what is unusual and why. VERA uses verbalised learning to optimise guiding questions for frozen VLMs and reports both detection and explanation capability [62]. The broader foundation-model VAU literature now distinguishes anomaly perception from anomaly cognition [48]. This distinction is useful because a system can localise an anomaly correctly while explaining it incorrectly, or produce a plausible description despite missing the event.

Leng et al. explicitly target this evaluation problem with a class-guided chain-of-evaluation metric and an anomaly-focused meta-benchmark designed to assess semantic validity and robustness to descriptive style [24]. Their work signals an important direction for the field: explanation quality requires diagnostic, anomaly-specific evaluation rather than generic text similarity alone.

Semantic weak supervision in 2026: From coarse labels to structured anomaly knowledge

Recent weakly supervised work increasingly treats the weakness of video-level labels as a semantic problem rather than only a multiple-instance optimisation problem. LAS-VAD groups temporally related frames through anomaly-connected components and introduces intention-aware reasoning to separate visually similar normal and abnormal behaviours [53]. In parallel, the Mixture of Memory Experts framework combines shared and expert-specific memories and uses language-derived anomaly prototypes for semantic routing, explicitly modelling both common and diverse anomaly patterns [50].

These developments reinforce a central theme of this review: the information available to a method must be coded before its numerical result is ranked. A weakly labelled detector using only video-level binary labels, a method augmented with large-language-model semantic prototypes, and a training-free VLM/LLM system may all be evaluated on UCF-Crime, yet they do not consume equivalent prior knowledge. The evidence map therefore records external semantic knowledge as a separate axis rather than absorbing it into the generic label "weak supervision."

Verified 2025–2026 frontier update: Granularity, streaming, privacy, and training-free reasoning

The targeted update confirms that the recent frontier is not moving in only one architectural direction. Open-vocabulary capability is being refined rather than merely introduced: Anomize addresses ambiguity in detecting and categorising novel anomaly classes [25]. Granularity is also expanding. Track Any Anomalous Object reframes anomaly localisation around granular anomalous-object tracking [21], while Holmes-VAU introduces hierarchical clip-, event-, and video-level anomaly understanding with more than 70,000 multi-granular annotations [65]. In weak supervision, π-VAD injects pose, depth, panoptic, optical-flow, and language-derived cues during training [34], and DAKD distils multi-backbone knowledge into a single student model [14]. These works show why the extraction schema must distinguish modality availability during training from inference-time computational cost.

Normality modelling is likewise diversifying. Noise-resistant predictive coding with dynamic memory explicitly targets background interference and heterogeneous normal patterns [20], while a motion- and appearance-guided patch diffusion model treats local reconstruction through conditional diffusion [68]. Privacy becomes a direct methodological constraint rather than a separate deployment discussion: federated weakly supervised VAD with multimodal prompting is evaluated without centralising institutional anomaly data [52]. Efficiency and streaming are also explicit design targets. Park et al. combine shortcut exploration with abnormal-feature-distance learning in a fast normality model [37], and StreamVAD develops progressive online context integration for long-term dependencies under a streaming constraint [18].

The 2026 update further strengthens the shift toward target-domain-light or semantic reasoning. LAVIDA uses pseudo anomalies and an MLLM in a zero-shot VAD setting [13]; VADER integrates relation features and large-language-model reasoning for causal VAU [11]; PRISM performs training-free open-set detection through intrinsic statistical modelling on pretrained multimodal embeddings [10]; and DR-VAD uses definition-guided LLM reasoning without target-domain training [39]. Federated privacy-preserving VAD has also progressed beyond the weak-supervision setting [26], while Fine-VAD explicitly targets fine-grained anomaly categorisation rather than only binary abnormality scoring [66]. Together with recent semantic weak-supervision methods [59], [53], [50], these studies justify retaining open-set capability, external model dependence, causal/streaming operation, privacy regime, granularity, and explanation output as first-class review variables.

RESEARCH GAPS AND STANDARDISED RESEARCH ADVANCEMENTS

The purpose of a high-impact review should be to reduce avoidable ambiguity for the next generation of studies. The recommendations below are deliberately operational. They are not intended to impose one model family or metric, but to make evidence easier to compare and reproduce.

Table 16. Proposed Minimum VAD Reporting Standard (MVRS).

Area	Minimum recommendation	Why it matters
Dataset identity	State dataset version/source, split, excluded videos, frame rate/stride, resolution, and any relabelling.	Prevents hidden changes to the evaluated population.
Supervision	State exactly which labels are available during training, validation, selection, calibration, and test.	Separates one-class, weak, supervised, open-set, and training-free claims.
Model selection	Use a validation set or nested protocol; keep final test untouched for the main claim.	Reduces test-set overfitting and "best-of-many" bias.
Metrics	Report metric implementation, granularity, aggregation, and post-processing. Prefer multiple complementary metrics.	Makes AUC/AP/event scores interpretable and reproducible.
Uncertainty	For stochastic learnable models, report multiple runs or justified uncertainty estimates when feasible.	Prevents over-interpreting tiny score differences.
Efficiency	Report end-to-end latency/FPS with hardware, batch size, precision, input resolution, and preprocessing; add parameters/FLOPs where useful.	Supports deployment comparison rather than headline FPS.
Reproducibility	Release code, evaluation script, split files, environment, and weights/features where licensing permits.	Reduces hidden implementation variance.
Deployment scope	Declare causal/online, bounded-delay, or offline use; report false-alarm burden and time-to-detection where relevant.	Aligns benchmark claims with real operational use.
Privacy / governance	State whether raw identifiable video leaves the device/site and how privacy is protected.	Makes deployment implications visible.
Foundation models	Report exact model/checkpoint/API version, prompts, decoding, context construction, caching, cost, and external services.	Makes training-free and VLM/LLM results reproducible despite model drift.
VAU explanations	Ground explanations in temporal/spatial evidence; evaluate semantic correctness and faithfulness separately from fluency.	Reduces persuasive but unsupported explanations.
Cross-paper claims	Use direct "outperforms SOTA" language only for Tier-A-compatible comparisons; otherwise say "contextual comparison."	Prevents protocol differences from being mistaken for model gains.

Table 17. Worked application of the Minimum VAD Reporting Standard to the MGFN UCF-Crime result [73].

MVRS item	Evidence identified in MGFN	Status	Interpretive consequence
Dataset identity	UCF-Crime and XD-Violence are explicitly identified; the selected result is UCF-Crime.	Reported	Benchmark identity is traceable.
Supervision	Weak video-level supervision is explicit.	Reported	The learning regime is interpretable.
Model selection	Ablations and hyperparameters are reported, but a fully isolated held-out selection protocol is not clearly documented.	Partial	Possible test-selection ambiguity remains.
Metric / population	Frame-level ROC-AUC is reported for UCF-Crime; the full micro/macro population construction is not versioned as an executable protocol.	Partial	Canonical convention is plausible, but exact implementation remains a reproducibility dependency.
Post-processing	Input/features and model settings are described; not all score normalisation/smoothing choices are provided as a versioned evaluation script.	Partial	Implementation variance may remain.

Uncertainty	No repeated-seed SD or CI for the headline UCF-Crime AUC was identified.	Not reported	Small score differences should not be over-interpreted.
Reproducibility	Public code is directly linked and important feature/hyperparameter information is reported.	Reported / partial	Strong artefact evidence, but code linkage is not identical to full reproducibility.
Efficiency / deployment	No complete end-to-end latency, FPS, hardware, and causal deployment profile is established for the full pipeline.	Not reported	The AUC should not be read as evidence of real-time deployability.
Cross-paper claims	MGFN compares against published benchmark results.	Partial	The protocol framework permits unqualified rank claims only against Tier-A-compatible comparators.

The worked example illustrates the intended role of the MVRS: it identifies which parts of a result are sufficiently documented for reproduction and direct comparison without converting reporting completeness into a model-quality score. The same model can therefore be scientifically strong while some benchmark claims remain difficult to reproduce or compare exactly. Reporting-quality items are used as transparent audit evidence rather than collapsed into a corpus-wide scalar ranking.

Benchmark protocols should be versioned artefacts

A benchmark should include not only videos and labels but executable evaluation code, canonical split files, a version identifier, and a machine-readable protocol. This would allow papers to state, for example, “Avenue Protocol v1.2, frame-level micro ROC-AUC, no per-video score normalisation” rather than relying on prose. Versioned protocols would also make historical results traceable when annotations or scripts are corrected.

Benchmark saturation should trigger transfer tests

When many methods achieve near-ceiling performance on a small benchmark, another decimal point of AUC becomes less informative. The community should pair saturated datasets with cross-scene, cross-dataset, or domain-shift tests and report calibration and false alarms. The objective is not to abandon historical benchmarks but to stop treating them as sufficient evidence of generalisation.

Weak supervision needs noise-aware evaluation

Weakly supervised VAD is attractive because dense temporal annotation is expensive, but video-level labels create unavoidable instance noise. Recent work directly targets weakly labelled information and noisy supervision [59], [67]. Future studies should report not only global AUC/AP but also temporal boundary quality, class-specific failure, and sensitivity to the proportion of normal content inside abnormal bags.

Efficiency should be end-to-end

Feature extraction, optical flow, object detection, VLM captioning, and LLM calls can dominate cost outside the core anomaly head. End-to-end profiling should therefore include these components. A lightweight scoring head attached to an expensive external encoder is not necessarily a lightweight VAD system. Conversely, frozen or cached features may be appropriate for forensic batch analysis. Reporting should make the intended operating mode explicit.

Explanation requires evidence and accountability

VAU systems should separate detection confidence, anomaly category confidence, and explanation confidence. Explanations should cite the video segment or object evidence that supports the claim, and evaluation should test factual consistency under perturbations rather than reward verbosity. As foundation models become more central, benchmark design must also address prompt sensitivity, external knowledge leakage, and model-version drift [48], [24].

THREATS TO VALIDITY

Selection and coverage bias. The strongest remaining limitation is coverage. The 193-study register was created through a high-recall structured seed [54] and targeted official-source updating, not through a contemporaneous de novo export from every bibliographic database. The 22 October 2026 source-specific verification scan, conducted against the fixed 14 October eligibility cut-off, was added to make this limitation testable and to verify the primary protocol cases. It also confirmed that the frozen 193-study register should be interpreted as a bounded audited evidence set rather than a census. Corpus composition and code-link percentages should therefore be interpreted as descriptive of the assembled register rather than as population-level estimates. The complete included-record register, together with the source-specific search and verification records, is publicly available in the companion evidence register archived on Zenodo at https://doi.org/10.5281/zenodo.23579051.

Screening-record limitation and publication bias. The historical structured-seed stage did not preserve a candidate-level record of every excluded item and exclusion reason; those numbers are not reconstructed retrospectively. Peer-reviewed publication is used for the core quantitative corpus to improve evidential stability, but this may underrepresent innovations that first appear as preprints. Two highly relevant preprints are therefore described only in a separately labelled frontier set and excluded from core statistics.

Extraction and coding error. VAD protocols are often distributed across the main paper, supplementary files, repositories, and dataset pages. The 12-study extended audit was checked against primary sources, and the companion file preserves source URLs and protocol notes. Independent duplicate human coding was not conducted; accordingly, no inter-rater agreement statistic is reported. This is treated as a limitation of the review.

Protocol classification subjectivity. Tier assignment still requires judgement about whether a difference changes the empirical question. Subjectivity is reduced by (i) specifying mandatory Tier-A dimensions, including material preprocessing, (ii) applying the framework to concrete study comparisons in Table 12, and (iii) providing the underlying protocol fields in the companion audit. The descriptive compatibility proportion C_{ij} cannot override a mismatch in benchmark version, test split, supervision, preprocessing, evaluation granularity, metric population, or aggregation.

Rapid model drift. Foundation-model systems depend on checkpoints and APIs that can change over time. The extraction therefore records model identity and source information where relevant. Results that cannot be tied to a stable model identity should be interpreted cautiously, particularly when external proprietary services contribute to the reported performance.

Performance-synthesis bias. Even the preprocessing-matched Tier-A results are not statistically independent effect estimates: the studies share the same benchmark test set and may share pretrained features, evaluation conventions, or code lineage. Table 13 therefore provides protocol-compatible descriptive comparisons rather than a meta-analysis. S3R is intentionally withheld from the strict summary because a key preprocessing detail is not explicit in the primary paper. The broader secondary-source distributions in Table 14 remain diagnostic context only. This distinction avoids assigning inferential precision to repeated measurements of the same benchmark population.

POSITIONING AGAINST EXISTING VAD SURVEYS

The strongest reason for this review to exist is not breadth alone. Recent surveys already cover deep-learning taxonomies, single-scene methods, systematic trends, privacy-preserving VAD, multimodal and weakly supervised approaches, broad supervision regimes, and foundation-model-driven VAU [35], [41], [45], [28], [5], [54], [48]. The present review is designed to complement rather than replicate them by treating protocol comparability and reproducibility as primary analytical objects.

Table 18. Positioning of this review relative to recent VAD surveys.

Review	Scope	Systematic / scale	Primary emphasis	Gap addressed by present review
Nayak et al. (2021) [35]	Deep-learning VAD.	Narrative comprehensive review.	Definitions, models, datasets, infrastructure, metrics, challenges.	Update to 2026; explicit protocol tiers, reproducibility and deployment audit.
Ramachandra et al. (2022) [41]	Single-scene VAD.	Survey with extensive benchmark comparison.	Problem formulations, datasets, evaluation criteria, best practices.	Cross-setting protocol compatibility and modern weak/open-world/foundation methods.
Samaila et al. (2024) [45]	Broad VAD, 2003–2023.	Systematic; 530 identified/evaluated papers reported.	Methods, problem settings, anomaly types, datasets, prospects.	Protocol-level score comparability, reproducibility, deployment, post-2023 frontier.
Liu et al. (2026) [28]	Privacy-preserving VAD.	Dedicated survey.	Privacy leakage and protection across data/features/systems.	Integrate privacy as one deployment axis without replacing privacy-specific depth.
Wu et al. (2026) [54]	Deep VAD across supervision levels.	Comprehensive review.	Semi/weak/full/ unsupervised/open-set; large models; datasets, code, metrics.	Protocol-aware pooling rules, reproducibility completeness, operational readiness.
Barbosa et al. (2026) [5]	Multimodal and weakly supervised VAD.	56-page survey.	Audio/visual/text fusion, weak supervision, robustness, interpretability.	Broader evidence map plus protocol and deployment audit across paradigms.
Su et al. (2026/2027) [48]	Foundation-model VAU.	Dedicated emerging-field survey.	Perception–cognition taxonomy, VAU benchmarks and metrics.	Connect foundation VAU to legacy VAD protocols and reproducibility/deployment evidence.
Caetano et al. [6] (2023)	Benchmark-based comparison of deep VAD models	Experimental benchmark review	Metric choice and model operating points	Adds protocol compatibility, reproducibility, deployment, and full evidence-map coding.
Le et al. [23] (2026)	Edge-based IoT VAD, modalities, real-time applications	Focused survey	Latency, model size, FLOPs, FPS, edge readiness	Present review integrates edge evidence with benchmark protocol and reproducibility rather than treating deployment separately.
This review	VAD/VAU, 2015–14 Oct 2026.	PRISMA-informed protocol-aware evidence review; bounded core register of 193 peer-reviewed studies plus a 12-study primary-source protocol audit.	Protocol compatibility, study-level Tier cases, primary Tier-A synthesis, reproducibility, deployment readiness, and modern VAD/VAU.	Determine which results are directly comparable, demonstrate downgrade reasons, and connect reporting quality to reproducibility/deployment interpretation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue X October 2026

CONCLUSION

VAD has reached a stage at which methodological diversity is both a strength and a source of evidential ambiguity. The field now spans normal-only reconstruction and prediction, memory and object-centric models, weak supervision, multimodal and vision-language adaptation, open-set and training-free inference, streaming operation, privacy-preserving learning, fine-grained localisation, and anomaly understanding. A review that simply places their headline AUC values in one table would obscure rather than clarify progress.

This protocol-aware evidence review uses a bounded core audited register of 193 peer-reviewed VAD/VAU studies from 2015 to 14 October 2026 and adopts a simple principle: comparability should be established before performance is ranked. Within that register, normal-only/semi-supervised studies form the largest coded group (54.4%), weakly supervised studies account for 28.5%, and directly identifiable public-code links occur in 79/193 records (40.9%). These are descriptive properties of the audited evidence set, not population estimates. The source-specific verification log, companion register, and primary-source protocol audit make those claims traceable without implying exhaustive literature coverage.

The main contribution is demonstrated through study-level protocol auditing rather than assumed from shared dataset names. Apparently similar weakly supervised UCF-Crime I3D-RGB results separate into distinct preprocessing strata: MIST and MSL form a one-crop Tier-A comparison, while RTFM, UR-DMU, and MGFN form a ten-crop Tier-A stratum. S3R remains useful contextual evidence but is excluded from the strict quantitative summary because the crop setting is not explicit in the primary paper. Vision-language priors, poly-modal training information, anomaly-only evaluation populations, and bounded real-time windows further illustrate why other headline AUC values require qualified or context-only comparison. The worked MVRS example connects protocol transparency to reproducibility and deployment interpretation.

The field is simultaneously moving from scalar anomaly detection toward open-vocabulary, training-free, causal, privacy-aware, fine-grained, and explanation-oriented video anomaly understanding. Future progress should therefore prioritise versioned benchmark protocols, machine-readable split and evaluation artefacts, cross-scene and cross-dataset testing, uncertainty-aware reporting, end-to-end deployment measurement, and evidence-grounded explanation evaluation. Protocol-aware comparison cannot remove all uncertainty from VAD research, but it can prevent incompatible experiments from being mistaken for direct evidence of superiority.

APPENDIX A: EVIDENCE-EXTRACTION FORM

Table A1. Evidence-extraction form used for corpus coding.

Field	Coding instruction / allowed values
Study ID	Unique ID; DOI preferred.
Peer review	Journal / full conference / context-only emerging-frontier preprint.
Supervision	normal-only; unsupervised; weak; full; open-set; open-vocabulary; training-free.
Representation	RGB; flow; object; skeleton; trajectory; audio; text; multimodal; other.
Temporal model	none; RNN/LSTM; TCN; transformer; state-space; bidirectional; LLM temporal reasoning; other.
Dataset + version	Exact name, source, version/release if available.
Split	Train/val/test units and counts; official/custom; cross-scene/cross-dataset.
Metric	Name, implementation, granularity, micro/macro, AP vs trapezoidal PR-AUC, event rule.
Post-processing	Normalisation, smoothing, calibration, threshold, fitting data.
Selection	Validation-selected / fixed / test-descriptive / unclear.
Reproducibility	Code, weights, splits, eval script, seed, runs, uncertainty, environment.

Compute	Parameters, FLOPs, FPS, latency, memory, hardware, batch, precision.
Deployment	Causal/online; bounded-delay; offline; privacy; domain shift; human explanation.
Protocol tier	A / B / C with explicit downgrade reason.
Evidence source	Page/table/figure/repository path used for each extracted value.
Protocol vector z_i	Record dataset/version, split/supervision, granularity, metric, aggregation, feature source, post-processing, reproducibility, and deployment evidence.
Metric population	Record exactly which observations form positives and negatives; note pooled-across-video, per-video, per-scene, event-level, or other construction.
Protocol compatibility	Assign Tier A/B/C only after checking mandatory core dimensions; retain pairwise C_{ij} as a descriptive audit value.
Quality completeness Q_i	Code each reporting-quality item as 0, 0.5, or 1; use only for sensitivity analysis, not as a model-quality ranking.
Reproducibility completeness R_i	Verify code, weights, split files, preprocessing, hyperparameters, seeds/runs, environment, and evaluation implementation.
Deployment evidence E_i	Verify causality, end-to-end latency/throughput context, memory/compute, privacy, calibration, domain-shift evaluation, and human factors.
External semantic knowledge	Record use of CLIP/VLM/LLM/foundation-model priors, prompts, captioners, external APIs, synthetic semantic data, or language-derived prototypes.
Evaluation uncertainty	Record number of runs, SD/SE/CI, bootstrap method, statistical test, and whether uncertainty covers model stochasticity, data sampling, or both.

APPENDIX B: PROTOCOL-COMPATIBILITY CHECKLIST FOR BENCHMARK TABLES

A result is eligible for a direct Tier-A numerical comparison only when the core protocol conditions are compatible and no undocumented difference materially changes the empirical question. The checklist below is intended for review coding and for authors constructing benchmark comparison tables.

1) Dataset identity and version are the same.

2) Test split and exclusion rules are the same.

3) Supervision regime and access to anomaly labels are comparable.

4) Feature source and external pretrained knowledge are stated.

5) Evaluation granularity (video/clip/frame/object/pixel/event) is the same.

6) Metric definition and implementation are compatible.

7) Aggregation population (pooled/micro, macro, per-scene, per-video) is stated.

8) Score normalisation, smoothing, calibration, and threshold selection are stated.

9) Any validation-guided model selection is separated from final test evaluation.

10) Runtime and deployment claims include hardware, batch size, input resolution, preprocessing, and causal context.

APPENDIX C: COMPANION EVIDENCE REGISTER

The companion evidence register supporting this review is publicly available on Zenodo at https://doi.org/10.5281/zenodo.23579051. It contains the bounded 193-study core register, including study

identifiers, publication year, methodological regime, venue, public-code-link status, acquisition provenance, verification status, inclusion status and supporting notes. The repository also includes the source-specific verification log, coding dictionary, 12-study extended protocol audit, protocol-compatible UCF-Crime comparison strata, worked Tier A/B/C protocol-classification cases, and an implementation example of the Minimum VAD Reporting Standard (MVRS). These materials provide a transparent and traceable record of the evidence base used in the review and enable readers to examine the corpus-level descriptive analyses, study-level protocol classifications and framework applications reported in the manuscript.

REFERENCES

1. S. Abdulaziz and E. Bondarev, "Auditing Frame-Level AUC in Weakly Supervised Video Anomaly Detection: Granularity, Resolution, and Scene Bias," arXiv:2608.11985, 2026. [Preprint; included only in the emerging-frontier evidence set.]
2. A. Acharya, H. Sapkota, Q. Yu, and X. Liu, "The Road Less Seen: Segment Exploration for Weakly Supervised Video Anomaly Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 14147–14156.
3. A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "UBnormal: New benchmark for supervised open-set video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20143–20153, https://doi.org/10.1109/CVPR52688.2022.01951.
4. S. Ahn, Y. Jo, K. Lee, and S. Park, "VideoPatchCore: An effective method to memorize normality for video anomaly detection," in Proc. Asian Conf. Computer Vision (ACCV), 2024, pp. 2179–2195.
5. R. Z. Barbosa, H. S. Oliveira, and J. M. R. S. Tavares, "A survey on multi-modal and weakly supervised approaches for robust anomaly detection in video data," Information Fusion, vol. 126, Art. 103388, 2026, https://doi.org/10.1016/j.inffus.2025.103388.
6. F. Caetano, P. Carvalho, and J. S. Cardoso, "Unveiling the performance of video anomaly detection models — A benchmark-based review," Intelligent Systems with Applications, vol. 18, Art. 200236, 2023, https://doi.org/10.1016/j.iswa.2023.200236.
7. C. Cao, Y. Lu, P. Wang, and Y. Zhang, "A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20392–20401, https://doi.org/10.1109/CVPR52729.2023.01953.
8. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, Art. 15, 2009, https://doi.org/10.1145/1541880.1541882.
9. J. Chen, L. Li, L. Su, Z.-J. Zha, and Q. Huang, "Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18319–18329, https://doi.org/10.1109/CVPR52733.2024.01734.
10. Y. Chen, Z. Ding, and Y. Shang, "PRISM: Training-free video anomaly detection via intrinsic statistical modeling," in Proc. 43rd Int. Conf. Machine Learning (ICML), Proc. Machine Learning Research, vol. 306, pp. 14037–14049, 2026.
11. Y. Cheng, Y.-H. Lin, M.-H. Chen, F.-E. Yang, and S.-H. Lai, "VADER: Towards causal video anomaly understanding with relation-aware large language models," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2026, pp. 7301–7311.
12. Chinese University of Hong Kong, "Avenue Dataset for Abnormal Event Detection." [Online]. Available: https://www.cse.cuhk.edu.hk/~leojia/projects/detectabnormal/dataset.html. Accessed: Oct. 14, 2026.
13. Z. Dai, K. Li, J. Liu, J. Yang, and Y. Qiao, "No Need For Real Anomaly: MLLM empowered zero-shot video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 35648–35658.
14. J. Dalvi, A. Dabouei, G. Dhanuka, and M. Xu, "Distilling aggregated knowledge for weakly-supervised video anomaly detection," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2025, pp. 5439–5448.
15. J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning (ICML), 2006, pp. 233–240, https://doi.org/10.1145/1143844.1143874.

16. A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 1965–1974, https://doi.org/10.1109/CVPRW63382.2024.00202.
17. D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1705–1714, https://doi.org/10.1109/ICCV.2019.00179.
18. L. Han, G. Liang, P. Wang, D. Liu, and K. Zhao, "StreamVAD: A streaming framework with progressive context integration for multi-temporal scale video anomaly detection," Neurocomputing, vol. 658, Art. 131669, 2025, https://doi.org/10.1016/j.neucom.2025.131669.
19. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 733–742.
20. H. Hu, W. Du, P. Liao, B. Wang, and S. Fan, "Noise-resistant video anomaly detection via RGB error-guided multiscale predictive coding and dynamic memory," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 19109–19119.
21. Y. Huang, C. Li, H. Zhang, Z. Lin, Y. Lin, H. Liu, W. Li, X. Liu, J. Gao, Y. Huang, X. Ding, and Y. Yuan, "Track Any Anomalous Object: A granular video anomaly detection pipeline," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 8689–8699.
22. R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7842–7851, https://doi.org/10.1109/CVPR.2019.00803.
23. H. Le, C.-K. Lu, and C.-C. Hsu, "Video anomaly detection for edge-based IoT systems: A survey of input modalities and real-time applications," Intelligent Systems with Applications, vol. 29, Art. 200635, 2026, https://doi.org/10.1016/j.iswa.2026.200635.
24. J. Leng, Z. Huang, M. Tan, Z. Wu, and X. Gao, "Towards trustworthy video anomaly understanding: A class-guided chain-of-evaluation metric and an anomaly-focused meta-benchmark," in Proc. 43rd Int. Conf. Machine Learning (ICML), PMLR, vol. 306, pp. 65554–65574, 2026.
25. F. Li, W. Liu, J. Chen, R. Zhang, Y. Wang, X. Zhong, and Z. Wang, "Anomize: Better open vocabulary video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 29203–29212.
26. S. Li, F. Liu, L. Jiao, J. Wang, Y. Ma, X. Wang, Y. Liu, Q. Bao, X. Liu, L. Li, and P. Chen, "Privacy-preserving video anomaly detection via federated learning," Knowledge-Based Systems, vol. 349, Art. 116454, 2026, https://doi.org/10.1016/j.knosys.2026.116454.
27. W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection — A new baseline," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6536–6545, https://doi.org/10.1109/CVPR.2018.00684.
28. Y. Liu, S. Liu, X. Zhu, H. Yang, J. Li, J. Guo, L. Teng, D. Yang, Y. Wang, and J. Liu, "Privacy-preserving video anomaly detection: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 37, no. 1, pp. 2–21, 2026, https://doi.org/10.1109/TNNLS.2025.3600252.
29. Z. Liu, X. Wu, W. Li, L. Yang, and S. Wang, "Rethinking Metrics and Benchmarks of Video Anomaly Detection," arXiv:2505.19022, 2025. [Preprint; included only in the emerging-frontier evidence set.]
30. C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2013, pp. 2720–2727, https://doi.org/10.1109/ICCV.2013.338.
31. W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 341–349, https://doi.org/10.1109/ICCV.2017.45.
32. J. Lyu, M. Zhao, J. Hu, X. Huang, Y. Chen, and S. Du, "VADMamba: Exploring state space models for fast video anomaly detection," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2025, pp. 1–6, https://doi.org/10.1109/ICME59968.2025.11209020.
33. V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1975–1981, https://doi.org/10.1109/CVPR.2010.5539872.

34. S. Majhi, G. D'Amicantonio, A. Dantcheva, Q. Kong, L. Garattoni, G. Francesca, E. Bondarev, and F. Bremond, "Just Dance with π! A poly-modal inductor for weakly-supervised video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 24265–24274.
35. R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," Image and Vision Computing, vol. 106, Art. 104078, 2021, https://doi.org/10.1016/j.imavis.2020.104078.
36. M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. n71, 2021, https://doi.org/10.1136/bmj.n71.
37. C. Park, D. Kim, M. Cho, M. Kim, M. Lee, S. Park, and S. Lee, "Fast video anomaly detection via context-aware shortcut exploration and abnormal feature distance learning," Pattern Recognition, vol. 157, Art. 110877, 2025, https://doi.org/10.1016/j.patcog.2024.110877.
38. H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14372–14381.
39. Y. Pei, M. Hao, Y. Huang, and Q. Zou, "DR-VAD: Definition-guided reasoning for training-free video anomaly detection," Neurocomputing, vol. 674, Art. 132836, 2026, https://doi.org/10.1016/j.neucom.2026.132836.
40. B. Ramachandra and M. Jones, "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2020, pp. 2569–2578, https://doi.org/10.1109/WACV45572.2020.9093457.
41. B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2293–2312, 2022, https://doi.org/10.1109/TPAMI.2020.3040591.
42. E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," Patterns, vol. 5, no. 6, Art. 100994, 2024, https://doi.org/10.1016/j.patter.2024.100994.
43. N.-C. Ristea, F.-A. Croitoru, R. T. Ionescu, M. Popescu, F. S. Khan, and M. Shah, "Self-distilled masked auto-encoders are efficient video anomaly detectors," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15984–15995, https://doi.org/10.1109/CVPR52733.2024.01513.
44. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, Art. e0118432, 2015, https://doi.org/10.1371/journal.pone.0118432.
45. Y. A. Samaila, P. Sebastian, N. S. S. Singh, A. N. Shuaibu, S. S. A. Ali, T. I. Amosa, G. E. M. Abro, and I. Shuaibu, "Video anomaly detection: A systematic review of issues and prospects," Neurocomputing, vol. 591, Art. 127726, 2024, https://doi.org/10.1016/j.neucom.2024.127726.
46. ShanghaiTech SVIP Lab, "ShanghaiTech Campus Dataset." [Online]. Available: https://svip-lab.github.io/dataset/campus_dataset.html. Accessed: Oct. 14, 2026.
47. S. Sørbø and M. Ruocco, "Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series," Data Mining and Knowledge Discovery, vol. 38, pp. 1027–1068, 2024, https://doi.org/10.1007/s10618-023-00988-8.
48. W. Su, K. Tang, D. Liu, P. Wu, Q. Yan, and G. Pang, "A survey on foundation model-driven video anomaly understanding," Pattern Recognition, vol. 183, Art. 114890, 2027 (online 2026), https://doi.org/10.1016/j.patcog.2026.114890.
49. W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6479–6488, https://doi.org/10.1109/CVPR.2018.00678.
50. B. Sun, J. Chen, Z. Wu, F. Gao, F. Yang, L. Su, and Y. Wang, "Joint Learning of General and Diverse Patterns with Mixture of Memory Experts for Weakly-Supervised Video Anomaly Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 35638–35647.
51. UCSD Statistical Visual Computing Laboratory, "UCSD Anomaly Detection Dataset." [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm. Accessed: Oct. 14, 2026.
52. B. Wang, C. Huang, J. Wen, W. Wang, Y. Liu, and Y. Xu, "Federated weakly supervised video anomaly detection with multimodal prompt," Proc. AAAI Conf. Artificial Intelligence, vol. 39, no. 20, pp. 21017–21025, 2025, https://doi.org/10.1609/aaai.v39i20.35398.

53. Y. Wang and S. Zhao, "Weakly Supervised Video Anomaly Detection with Anomaly-Connected Components and Intention Reasoning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 28545–28556.
54. P. Wu, C. Pan, Y. Yan, G. Pang, Q. Yan, P. Wang, and Y. Zhang, "Deep learning for video anomaly detection: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 37, no. 7, pp. 3010–3030, 2026, https://doi.org/10.1109/TNNLS.2025.3647892.
55. 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. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18297–18307, https://doi.org/10.1109/CVPR52733.2024.01732.
56. 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," Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 6, pp. 6074–6082, 2024, https://doi.org/10.1609/aaai.v38i6.28423.
57. P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 7, pp. 2609–2622, 2020, https://doi.org/10.1109/TNNLS.2019.2933554.
58. 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 Computer Vision – ECCV 2020, 2020, pp. 322–339, https://doi.org/10.1007/978-3-030-58577-8_20.
59. R. Xu, R. Wang, Y. Zhang, T. Tao, X. Li, and L. Jing, "TLMA: Mitigating the impact of weakly labeled information for video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 35597–35606.
60. Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, "Follow the rules: Reasoning for video anomaly detection with large language models," in Computer Vision – ECCV 2024, 2024, https://doi.org/10.1007/978-3-031-73004-7_18.
61. S. Yao, G. A. Noghre, A. D. Pazho, and H. Tabkhi, "Evaluating the effectiveness of video anomaly detection in the wild: Online learning and inference for real-world deployment," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 4832–4841.
62. M. Ye, W. Liu, and P. He, "VERA: Explainable video anomaly detection via verbalized learning of vision-language models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 8679–8688.
63. M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative cooperative learning for unsupervised video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14744–14754, https://doi.org/10.1109/CVPR52688.2022.01433.
64. L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing large language models for training-free video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18527–18536, https://doi.org/10.1109/CVPR52733.2024.01753.
65. H. Zhang, X. Xu, X. Wang, J. Zuo, X. Huang, C. Gao, S. Zhang, L. Yu, and N. Sang, "Holmes-VAU: Towards long-term video anomaly understanding at any granularity," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 13843–13853.
66. M. Zhang, Y. Zhu, P. Ren, H. Sun, Q. Qi, Z. Zhuang, H. Wang, L. Zhang, J. Liao, and J. Wang, "Fine-VAD: Towards fine-grained video anomaly detection via progressive cross-granularity learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 35514–35523.
67. Y. Zhao, Y. Wang, W. Guo, S. Xu, X. Cai, X. Lin, Y. Zhang, and X. Yuan, "Learning from noisy supervision: A denoising-debiasing framework for weakly supervised video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2026, pp. 21326–21335.
68. H. Zhou, J. Cai, Y. Ye, Y. Feng, C. Gao, J. Yu, Z. Song, and W. Yang, "Video anomaly detection with motion and appearance guided patch diffusion model," Proc. AAAI Conf. Artificial Intelligence, vol. 39, no. 10, pp. 10761–10769, 2025, https://doi.org/10.1609/aaai.v39i10.33169.
69. J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14009–14018, https://doi.org/10.1109/CVPR46437.2021.01379.

70. 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. Computer Vision (ICCV), 2021, pp. 4975–4986, https://doi.org/10.1109/ICCV48922.2021.00493.
71. S. Li, F. Liu, and L. Jiao, "Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection," Proc. AAAI Conf. Artificial Intelligence, vol. 36, no. 2, pp. 1395–1403, 2022, https://doi.org/10.1609/aaai.v36i2.20028.
72. H. Zhou, J. Yu, and W. Yang, "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection," Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 3, pp. 3769–3777, 2023, https://doi.org/10.1609/aaai.v37i3.25489.
73. 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," Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 1, pp. 387–395, 2023, https://doi.org/10.1609/aaai.v37i1.25112.
74. H. Karim, K. Doshi, and Y. Yilmaz, "Real-Time Weakly Supervised Video Anomaly Detection," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2024, pp. 6848–6856, https://doi.org/10.1109/WACV57701.2024.00670.
75. H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8022–8031, https://doi.org/10.1109/CVPR52729.2023.00775.
76. J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-Supervised Sparse Representation for Video Anomaly Detection," in Computer Vision – ECCV 2022, pp. 729–745, 2022, https://doi.org/10.1007/978-3-031-19778-9_42.