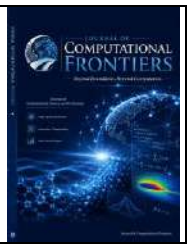
GYRES Scholarly Press

Journal of Computational Frontiers

Volume 01, Issue 01, 2026
Journal Web Page: https://gyrespress.com/jcf



Trustworthy Video Anomaly Detection: A Privacy-Aware, Explainable, and Efficient Deep Learning Framework

Mueen Ud Din
PhD Scholar, Department of Computer Science, Riphah International University, Pakistan

Article Info

Received: 20 August 2026
Received in revised: 25 August 2026
Accepted: 29 August 2026
Available online: 02 September 2026

Keywords

Video anomaly detection; trustworthy artificial intelligence; explainable AI; federated learning; uncertainty calibration; privacy-preserving learning

Abstract

Video anomaly detection (VAD) increasingly relies on high-capacity spatio-temporal models, yet accuracy-only evaluation is inadequate when automated alerts can affect safety, privacy, or human decision making. This paper introduces TAP-VAD, a trustworthy-AI design framework that treats detection quality, calibrated uncertainty, robustness, explainability, privacy, computational efficiency, and human oversight as coupled requirements rather than post-hoc additions. The architecture combines an adaptable video encoder, bounded-cost temporal memory, anomaly and open-vocabulary semantic heads, uncertainty calibration, layered spatial–temporal–semantic explanations, and a cross-silo federated learning option with secure aggregation. Its principal methodological contribution is an auditable evaluation protocol that specifies how each trust dimension should be measured, stress-tested, and reported on representative video anomaly benchmarks. The literature synthesis identifies three recurring gaps: trust properties are commonly evaluated in isolation, privacy mechanisms are rarely co-designed with explanation, and long-video efficiency is seldom linked to uncertainty-aware escalation. TAP-VAD therefore adds a policy layer that can abstain and route ambiguous cases to human review instead of converting every score into an automatic decision. The framework does not claim unperformed benchmark gains; it provides a falsifiable architecture, ablation plan, threat model, and reproducibility checklist for empirical validation. The resulting perspective connects computer-vision performance with epistemic and governance requirements for trustworthy anomaly detection.

1 INTRODUCTION

Video anomaly detection (VAD) seeks to identify events, actions, objects, or temporal patterns that depart from a model of normality. The task matters because modern video systems operate in settings where continuous manual monitoring is impractical, including transport, public-space surveillance, industrial safety, healthcare observation, and smart environments. Yet the central difficulty is not merely that anomalies are rare. Abnormality is contextual: the same action may be innocuous in one scene, suspicious in another, and ethically sensitive when the label is interpreted as a judgment about a person rather than a description of a visual pattern.

VAD has progressed through several distinct modeling assumptions. Early reconstruction and forecasting approaches learned the visual and temporal structure of normal video [1,2].

Corresponding Author:mueen80@gmail.com

Memory mechanisms were subsequently introduced to restrict a model's ability to reproduce abnormal inputs [3,4], whereas weak supervision enabled learning from long recordings labeled only at video level [5,6]. Current systems increasingly combine self-supervised pretraining, masked video modeling, transformers, vision-language representations, prompts, density estimation, and open-vocabulary inference [7-11]. Although these techniques expand modeling capacity, they also raise practical questions about interpretability, latency, memory demand, and data governance.

This creates a deployment gap. A VAD model can obtain a strong ranking metric while still be poorly calibrated, fragile under camera or scene shift, opaque about why a segment was flagged, expensive on long streams, or dependent on centralized sensitive video. In high-consequence use, these properties are not side constraints. A model that is accurate on average but confidently wrong under shift can induce inappropriate reliance; an explanation that looks plausible but is not faithful can create false reassurance; and federated training can reduce raw-data movement while remaining vulnerable to update leakage or poisoning [12-14].

The paper therefore reframes trustworthy VAD as a multi-objective system-design problem. We introduce TAP-VAD (Trustworthy, Accountable and Privacy-aware Video Anomaly Detection), a framework that integrates efficient temporal reasoning, calibrated uncertainty, explanation, privacy-preserving collaboration, robustness testing, and human escalation. The proposal builds on contemporary transformer and vision–language research [15,16], while explicitly avoiding the assumption that a larger backbone or a new prompt mechanism is sufficient for trustworthy deployment.

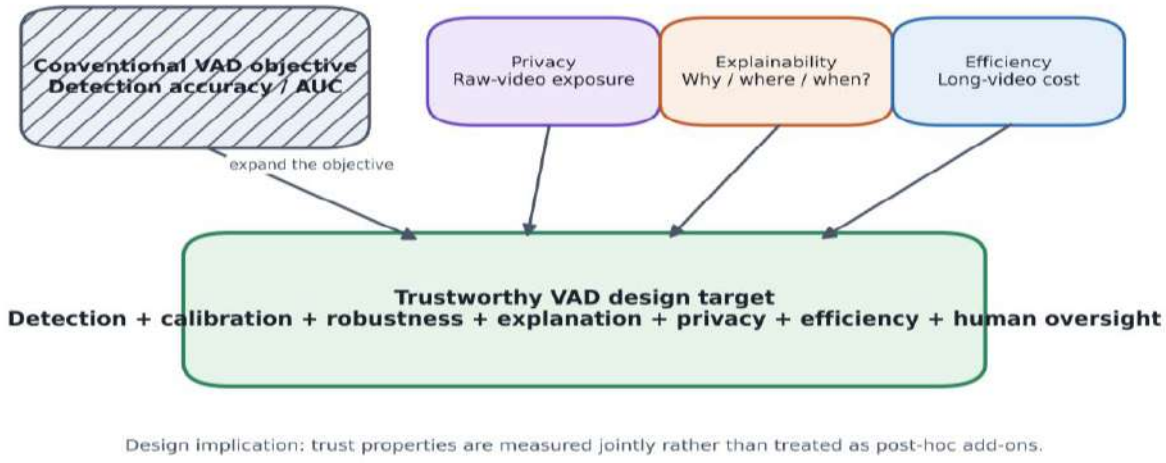


Fig. 1 From an accuracy-centered VAD objective to a multi-dimensional trustworthy design target. The figure illustrates the paper's central reframing: privacy, explanation, and efficiency are evaluated jointly with detection quality rather than appended after model training.

The research questions are:

- **RQ1:** How can long-range temporal context be modeled with bounded computation without erasing short, rare anomaly evidence?
- **RQ2:** How should anomaly confidence, predictive uncertainty, and explanatory evidence be combined so that the system can alert, abstain, or request human review appropriately?
- **RQ3:** What privacy–utility–communication trade-offs arise when VAD is trained across sites without centralizing raw video?
- **RQ4:** Which evaluation protocol can expose trade-offs among detection, calibration, robustness, explanation fidelity, privacy, and efficiency rather than optimizing one metric in isolation?

The contribution is methodological rather than a claim of completed benchmark superiority. Specifically, the paper (i) defines measurable trust requirements for VAD; (ii) proposes an end-to-end architecture that operationalizes those requirements; (iii) specifies an experimental, ablation, privacy, and stress-testing protocol; and (iv) connects technical evidence to appropriate human reliance. This distinction is important for scientific integrity: empirical claims should be added only after the corresponding experiments have been executed and reproduced.

2 RELATED WORK AND PROBLEM FRAMING

2.1 Video anomaly detection: from normality modeling to open vocabulary

A major line of deep VAD research estimates normal behavior and flags departures from it. Temporal-regularity learning and future-frame forecasting are representative early formulations [1,2]. Later object-centric and memory-based designs emphasized localized deviations and constrained overly expressive reconstruction models [17,3,4]. Self-supervised multi-task learning further diversified the training signal while avoiding dense anomaly annotation [7].

Weak supervision addresses a different practical constraint: abnormal events can be labeled at the video level while their temporal boundaries remain unknown. UCF-Crime popularized this setting for long real-world surveillance videos [5]. Subsequent work has tackled multiple-instance-learning bias, uncertain pseudo-labels, context–motion relations, scene awareness, caption-based semantics, prompt-based anomaly generation, and restoration-based objectives [6,18-23]. These methods show that the supervision regime changes what a score means: a one-class detector estimates deviation from learned normality, whereas a weakly supervised detector also absorbs label semantics and dataset-specific anomaly priors.

The most recent direction moves beyond fixed anomaly vocabularies. Wu et al. [26] formulated open-vocabulary VAD to detect and categorize both seen and unseen anomaly classes. Li et al. [11] further addressed ambiguity in novel anomaly detection and category confusion. Training-free VAD based on vision–language models and large language models has also emerged [24], while granular pipelines pursue object- and pixel-level localization [25]. In 2026, anomaly-aware CLIP adaptation and temporal graph reasoning continued this vision–language trend [16]. Open vocabulary is attractive for deployment, but it raises a trust question: semantic flexibility can improve coverage while making predictions sensitive to prompts, language priors, and the mismatch between textual categories and context-dependent visual abnormality.

Adjacent developments also strengthen the case for a broader trust protocol. Attention-based multi-task VAD adds complementary learning objectives [27], uncertainty-aware weak supervision questions the certainty of anomaly labels themselves [28], and large-language-model token alignment brings semantic reasoning deeper into recent VAD pipelines [29]. On the explanation side, more efficient concept-activation methods make concept-level auditing increasingly feasible at scale [30]. These advances improve individual components, but they do not remove the need to test how the components interact under one deployment-oriented protocol.

Table 1 Representative VAD paradigms and the trustworthiness questions they leave open.

Paradigm	Representative work	Strength	Unresolved trust issue
Normality reconstruction / prediction	Hasan et al. [1]; Liu et al. [2]; Gong et al. [3]; Park et al. [4]	Little or no anomaly annotation; intuitive normality model	Over-generalized reconstruction; scene specificity; weak semantics
Object-centric / pseudo-anomaly	Ionescu et al. [17]; Liu et al. [22]; Rai et al. [31]	Improves localized sensitivity; synthetic boundary learning	Pseudo-anomalies may encode design assumptions rather than real novelty
Self-supervised / masked modeling	Georgescu et al. [7]; Tong et al. [8]; Ristea et al. [32]; Gundavarapu et al. [33]	Efficient use of unlabeled video; transferable representations	Pretext success does not guarantee calibrated anomaly scores
Weakly supervised MIL	Sultani et al. [5]; Lv et al. [6]; Zhang et al. [18]; Cho et al. [19]	Scales to long videos with coarse labels	Pseudo-label noise; temporal localization bias; closed-set priors
Semantic / caption / prompt-guided	Chen et al. [21]; Yang et al. [34]; Zanella et al. [24]	Uses language to encode context and anomaly concepts	Prompt sensitivity; hallucinated or noisy textual evidence

Paradigm	Representative work	Strength	Unresolved trust issue
Density / score modeling	Micorek et al. [10]; Zhang et al. [35]	Directly models likelihood or score structure	Likelihood may be misaligned with semantic abnormality
Open-vocabulary / vision–language	Wu et al. [26]; Li et al. [11]; Kang et al. [16]	Detects and categorizes novel anomalies	Novel-class calibration, prompt bias, semantic ambiguity
Granular localization / mixture models	Huang et al. [25]; D'Amicantonio et al. [36]	Finer object evidence; specialist modeling	Additional complexity and evaluation burden
Deployment-oriented calibrated fusion	Mohanarangan & Palanisamy [37]	Explicit calibration and low-latency fusion	Clip-level assumptions can limit long-duration anomalies

2.2 Transformers, temporal scale, and efficiency

Transformers provide flexible long-range interaction through attention [38]. In vision, patch-token representations established the Vision Transformer [39], followed by video architectures that factorize space and time, learn multiscale hierarchies, or use shifted local windows [40-43]. Deformable attention and long–short contrastive learning further illustrate that temporal context can be structured rather than processed uniformly [44,44].

For VAD, this matters because a rare event may occupy only a few frames in a long sequence. Full self-attention over all spatio-temporal tokens can be computationally prohibitive, while aggressive sampling can delete the very evidence the model is expected to detect. Transformer-based VAD reviews therefore identify efficiency and long-sequence processing as persistent research concerns [15,9]. TAP-VAD treats temporal efficiency as an empirical trade-off: local context, selected salient tokens, and a compact recurrent memory are combined, then compared with full-attention and fixed-sampling baselines.

2.3 Explainability, uncertainty, and appropriate reliance

Post-hoc visual explanations can show where a model is sensitive, but sensitivity is not automatically a causal or faithful account of the decision. Grad-CAM, integrated gradients, SHAP, meaningful perturbation, LIME, concept activation vectors, and concept bottleneck models provide complementary views of attribution and concept-level reasoning [45-,51]. Sanity checks demonstrate that visually persuasive saliency maps can fail basic dependence tests [52]. For high-stakes decisions, this motivates either inherently interpretable structure or more demanding validation of explanations [53].

Explanations also have a social function. People seek contrastive, selective, and context-sensitive reasons rather than exhaustive descriptions of model internals [54]. Recent human-centered anomaly research similarly argues that explanations should be evaluated for the needs of the decision maker, not only for visual attractiveness [55,56]. TAP-VAD therefore separates explanation generation from explanation evaluation: a heatmap, temporal segment, concept label, or natural-language rationale is accepted only to the extent that perturbation, stability, localization, or human-usefulness tests support it.

Uncertainty provides the complementary control mechanism. Monte Carlo dropout and deep ensembles offer approximate predictive uncertainty, while post-hoc calibration can correct overconfident probabilities [57-59]. Under distribution shift, uncertainty quality must itself be tested rather than assumed [60]. In TAP-VAD, uncertainty is operational: it controls abstention and review thresholds, which makes reliability part of the human–AI interaction rather than a diagnostic statistic reported after deployment.

2.4 Privacy-preserving collaborative learning

Video is intrinsically privacy-sensitive because it can reveal faces, locations, routines, interactions, and contextual information that was never intended as a learning signal. Privacy and security concerns also extend to the video medium itself. Adjacent secure-video research has used randomized frame-selection steganography to conceal encrypted payloads within selected video frames while preserving visual fidelity [61]. Such data-hiding techniques address a different threat surface from privacy-preserving model training, but they reinforce the need to state explicitly what information a system protects and from whom. Cross-silo federated learning offers one way to train across organizations while retaining raw video locally [62]. Secure aggregation can prevent the server from inspecting individual client updates [63], while FedProx and SCAFFOLD address heterogeneous client optimization [64,65]. Federation, however, is not equivalent to privacy. Gradients or model updates may leak information, adversarial clients may poison a shared model, and non-identically distributed data can produce uneven performance [12-14,66,67]. Accordingly, TAP-VAD defines a threat model before selecting a privacy mechanism. Secure aggregation protects update visibility; client-level differential privacy can limit contribution disclosure; robust aggregation and anomaly checks target malicious updates; and governance rules still control capture, retention, access, and downstream use of video.

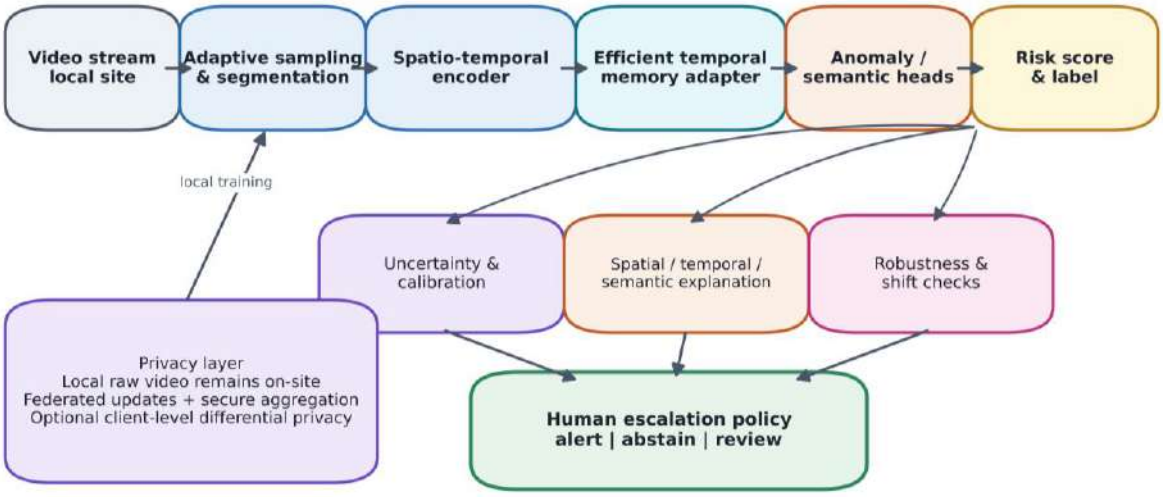


Fig. 2 TAP-VAD end-to-end architecture. A spatio-temporal detection path is coupled to an explicit trust layer for calibration, explanation, and robustness; a privacy layer governs collaborative training; and a human escalation policy converts uncertainty and evidence into alert, abstain, or review actions.

3 Trust requirements and system specification

The framework begins with operational requirements rather than a single benchmark objective. Each trust claim must map to at least one observable test. This prevents vague statements such as 'the model is explainable' or 'the method preserves privacy' from substituting for evidence.

Table 2 Operational definition of trustworthiness dimensions for TAP-VAD.

Dimension	Requirement	Primary measures	Reporting rule
Detection quality	Ranks and localizes anomalous events without excessive false alerts	Frame/clip AUC, AP, F1; temporal IoU; pixel/object overlap where annotations exist	Report per dataset and per anomaly type; confidence intervals
Calibration	Numerical confidence matches empirical correctness	ECE, Brier score, NLL, reliability diagram, risk–coverage/AURC	Measure before and after temperature scaling; repeat under shift
Robustness	Performance degrades gracefully under realistic change	Relative metric drop, worst-group score, corruption severity curve	Camera, compression, illumination, blur, occlusion, scene and domain shift

Dimension	Requirement	Primary measures	Reporting rule
Explainability	Evidence tracks the features or intervals that materially affect output	Deletion/insertion, perturbation fidelity, temporal localization, stability, concept tests	Separate visual plausibility from faithfulness; include failure cases
Privacy	Training minimizes exposure beyond the agreed threat model	Raw-data movement, secure-aggregation coverage, ε/δ if DP, leakage-attack success	State attacker capability and privacy mechanism explicitly
Efficiency	Model is feasible for long streams and stated hardware	Latency, FPS, peak memory, parameters, FLOPs, energy if measurable, communication/round	Measure end-to-end including sampling, explanation and privacy overhead
Human oversight	System supports appropriate reliance rather than automatic acceptance	Abstention rate, review precision, decision time, disagreement resolution	Predefine escalation rules and log evidence shown to reviewers

3.1 Formal task view

Let a video be partitioned into temporally ordered clips, with each clip mapped to a representation by a spatio-temporal encoder. The detector produces (a) a class-agnostic anomaly score, (b) an optional semantic distribution over anomaly concepts, and (c) an uncertainty estimate. The trusted output is not the anomaly score alone; it is a structured decision object containing the score, calibrated confidence, uncertainty, supporting spatial and temporal evidence, semantic evidence where applicable, and a policy action.

The optimization is therefore multi-objective. Detection loss remains necessary, but training and model selection also consider temporal consistency, representation regularization, calibration, privacy constraints in federated settings, and efficiency budgets. The paper deliberately avoids collapsing these quantities into one fixed weighted scalar because such a scalar can hide unacceptable regressions. Candidate models are compared on a Pareto frontier and subjected to minimum acceptance thresholds for calibration, privacy, and latency.

3.2 Threat model and ethical scope

The threat model distinguishes four surfaces: data capture, local training, communication, and decision use. Privacy-preserving learning addresses only part of this chain. A system that keeps raw video on-site may still be ethically problematic if cameras are deployed without lawful purpose, retention is excessive, or a model treats visually unusual behavior as evidence of wrongdoing. Therefore, anomaly scores are interpreted as signals for review, not as judgments about intent, identity, guilt, or riskiness of a person.

Table 3 Threat model, candidate controls, and residual risks.

Surface	Failure mode	Control in the protocol	Residual risk to disclose
Central raw-video collection	Unauthorized access, secondary use, identity/context exposure	Local retention; minimized capture; federated training	Residual risk remains at the camera/site
Model-update leakage	Reconstruction, membership or property inference	Secure aggregation; clipping; optional client-level DP; leakage tests	DP may reduce utility; secure aggregation does not prevent malicious local clients
Non-IID client data	Bias toward dominant sites; unstable convergence	FedProx/SCAFFOLD comparison; client-level metrics; worst-site reporting	Performance parity cannot be assumed
Poisoning/backdoor clients	Manipulated global behavior	Update anomaly checks; robust aggregation baseline; client quarantine policy	Robust defenses can reject benign rare clients

Surface	Failure mode	Control in the protocol	Residual risk to disclose
Prompt/semantic bias	Anomaly label influenced by wording or language priors	Prompt ensembles; neutral/scene-conditioned prompts; sensitivity analysis	Open-vocabulary labels remain context-dependent
Automation bias	Reviewer over-relies on score or attractive explanation	Abstention; confidence/evidence display; human protocol; audit log	Explanations can persuade even when unfaithful

4 Proposed TAP-VAD methodology

4.1 Data and preprocessing

The empirical study should use multiple benchmark families because a single dataset cannot represent open-world anomaly behavior. One-class datasets test normality modeling; multi-scene datasets test scene dependence; long weakly supervised datasets test sparse temporal localization; and open-vocabulary or granular settings test semantic generalization and fine localization. Splits must follow the original benchmark protocol unless a cross-dataset experiment is explicitly declared.

Preprocessing is intentionally conservative. Frames are resized and normalized according to the backbone, while temporal segmentation preserves timestamps so explanations can be mapped back to the original stream. Training augmentation may include photometric change, compression, mild blur, temporal jitter, and spatial crop, but the evaluation suite keeps a clean reference set separate from corruption stress tests. Sensitive raw video is not exported from a local site in the federated configuration.

Table 4 Benchmark families and their role in the proposed evaluation.

Dataset	Typical supervision	Why it is useful	Role in TAP-VAD
UCSD Ped2	One-class / normality modeling	Compact pedestrian-scene benchmark	Sanity check for reconstruction/density/localization; limited scene diversity
CUHK Avenue	One-class with localized events	Scene-specific abnormal actions	Temporal and spatial explanation tests
ShanghaiTech	Multi-scene semi/unsupervised VAD	Greater scene diversity	Cross-scene generalization and worst-scene reporting
UCF-Crime	Weakly supervised long untrimmed video	Real-world anomaly categories and sparse events	Long-video temporal localization, weak supervision, calibration
XD-Violence	Weak supervision; long events; multimodal source	Diverse violent/anomalous events	Visual-only and optional multimodal stress test; open-vocabulary comparison
UBnormal	Open-set abnormal-event benchmark	Diverse synthetic abnormal scenarios	Pseudo-anomaly transfer and open-set generalization
NWPU Campus	Large multi-scene benchmark with anomaly anticipation	43 scenes and broad anomaly taxonomy in the original benchmark	Scene shift, anticipation, and cross-scene evaluation [68]

4.2 Spatio-temporal representation

The visual encoder is treated as a replaceable component rather than a novelty claim. Candidate backbones include Video Swin, MViT, ViViT/TimeSformer-style factorized attention, and self-supervised VideoMAE features [8,41-43]. For open-vocabulary experiments, the visual representation can be aligned with CLIP-style text embeddings [69], optionally using frozen DINOv2 features as a robust visual baseline [70].

The key design choice is how temporal evidence is retained. TAP-VAD uses local attention windows for short-term motion, a salience selector to preserve informative tokens, and a compact recurrent summary memory for long-range context. The selector can be driven by motion magnitude, feature novelty, or learned attention, but all variants must be ablated against uniform sampling. This is important because an efficiency mechanism that systematically removes short anomalies would violate the system's purpose even if average latency improves.

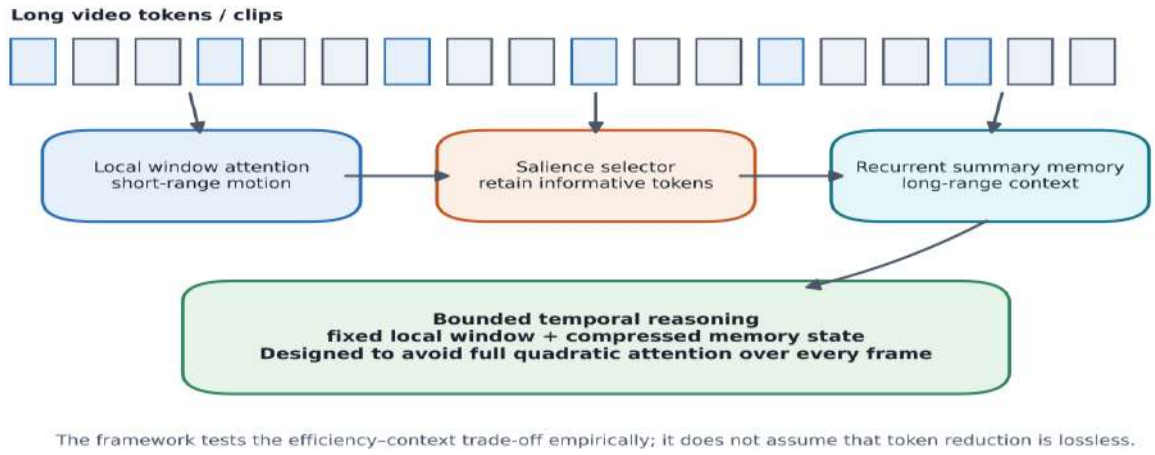


Fig. 3 Bounded-cost temporal reasoning. Local windows capture short-range motion; a salience selector retains informative evidence; and a compact memory summarizes long-range context. The efficiency mechanism is evaluated against full-attention and fixed-sampling baselines to quantify any loss of rare-event evidence.

4.3 Anomaly and semantic heads

A class-agnostic head produces a continuous anomaly score for each clip or frame. Its implementation depends on the supervision regime: a density or reconstruction residual can serve one-class learning; a multiple-instance objective can serve weak supervision; and a feature-distance or energy score can serve pretrained representations. The framework deliberately keeps the head modular so the trust protocol can compare whether uncertainty and explanation quality depend on the anomaly formulation.

The optional semantic head supports open-vocabulary anomaly categorization. Text prototypes are constructed for normal and anomalous concepts, then varied through prompt ensembles to measure prompt sensitivity. Class-agnostic detection and semantic classification are evaluated separately, following the principle that a system may correctly detect 'something unusual' while misnaming the event. This separation mirrors recent open-vocabulary work [26,11] and avoids treating a semantic misclassification as if no anomaly were present.

4.4 Calibration, uncertainty, and abstention

TAP-VAD distinguishes score magnitude from confidence. A validation set is used to fit temperature scaling or another monotonic calibration map; deep ensembles or Monte Carlo dropout can estimate epistemic uncertainty when compute permits [57-59]. Calibration is repeated under dataset shift because in-distribution reliability does not imply shift reliability [60].

The policy layer defines three actions. A low anomaly score with low uncertainty continues monitoring. A high anomaly score with acceptable uncertainty generates an alert together with evidence. High uncertainty, disagreement between model heads, or conflicting explanation evidence produces abstention and human review. Thresholds are selected on validation data under a stated cost model and are frozen before testing. This policy prevents uncertainty from being reported decoratively while every prediction still triggers the same downstream action.

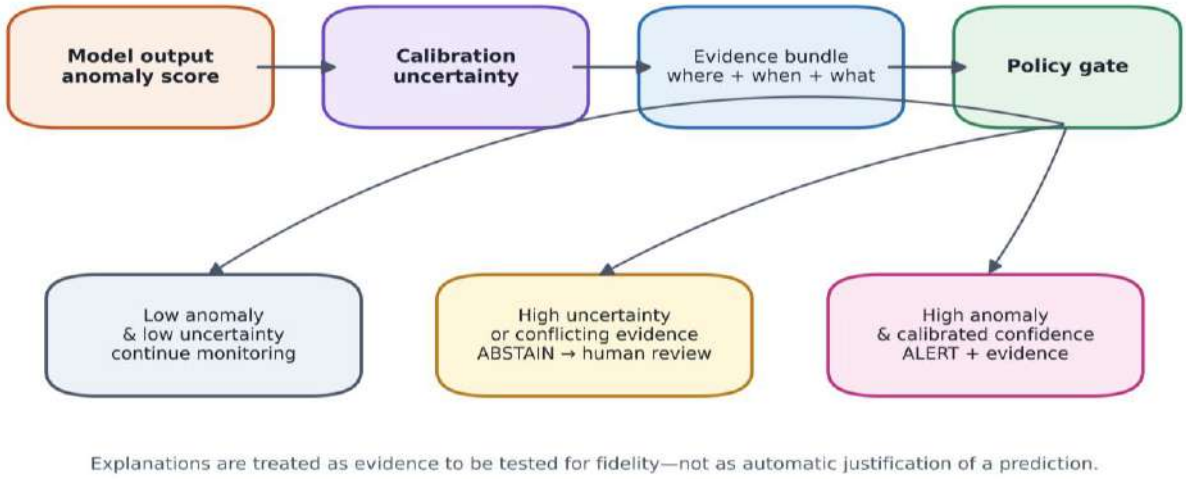


Fig. 5 Uncertainty-aware decision triage. Calibrated confidence and a multi-modal evidence bundle determine whether the system continues monitoring, abstains for human review, or raises an alert.

4.5 Explanation layer

The explanation layer has three levels. Spatial evidence identifies image regions that most affect the anomaly score, using gradient, perturbation, or concept-based methods. Temporal evidence identifies the clips whose removal or corruption materially changes the decision. Semantic evidence states which learned or text-aligned concepts are most associated with the score. A natural-language summary may be generated only from these structured signals and scene metadata; it should not introduce ungrounded details about intent or identity.

Faithfulness is tested rather than assumed. Saliency methods undergo parameter-randomization sanity checks inspired by Adebayo et al. [52]. Spatial evidence is evaluated against available boxes/masks or object tracks; temporal evidence is compared with annotated anomaly intervals; concept explanations are tested with TCAV-style sensitivity or concept bottlenecks [50,51]. If natural-language explanations are used, they are assessed for evidence coverage, unsupported claims, clarity, and usefulness to reviewers, following the human-centered emphasis in recent anomaly-explanation research [55].

Table 5 Explanation modalities, faithfulness tests, and intended questions.

Evidence type	Candidate method	Faithfulness / quality test	Question answered
Spatial saliency	Grad-CAM / Integrated Gradients / perturbation map	Region deletion/insertion; overlap with box/mask; randomization sanity check	Where in the frame affected the score?
Temporal attribution	Clip deletion, counterfactual masking, attention-flow comparison	Change in score after segment removal; temporal IoU; rank correlation	When did decisive evidence occur?
Concept explanation	TCAV, concept bottleneck, CLIP concept similarity	Concept sensitivity; completeness; counterfactual concept removal	What visual/semantic concept influenced the decision?
Natural-language summary	Template or local language model grounded only in structured evidence	Unsupported-claim rate; evidence coverage; reviewer usefulness	Can a reviewer understand the evidence without invented rationale?
Stability	Repeat explanation under mild input/model perturbation	Similarity/IoU/rank stability	Would an immaterial change produce a different story?

4.6 Federated privacy layer

The federated configuration is cross-silo: each participating site stores and preprocesses its own video, trains local model updates, and sends only protected updates to an aggregation service. FedAvg is the minimal baseline; FedProx and SCAFFOLD are included when client data are heterogeneous [62,64,65]. Secure aggregation is used when the threat model includes an honest-but-curious server that must not inspect individual updates [63].

Optional client-level differential privacy clips each client contribution and adds calibrated noise before or during aggregation. Because privacy noise may disproportionately affect rare anomalies or small clients, the experiment reports utility per site, privacy budget, convergence rounds, and communication overhead. Privacy stress tests include membership/property inference or gradient-reconstruction proxies where feasible [12,66], plus poisoned-client experiments inspired by federated backdoor research [13]. The objective is not to claim absolute privacy but to show what is protected under a stated attacker model.

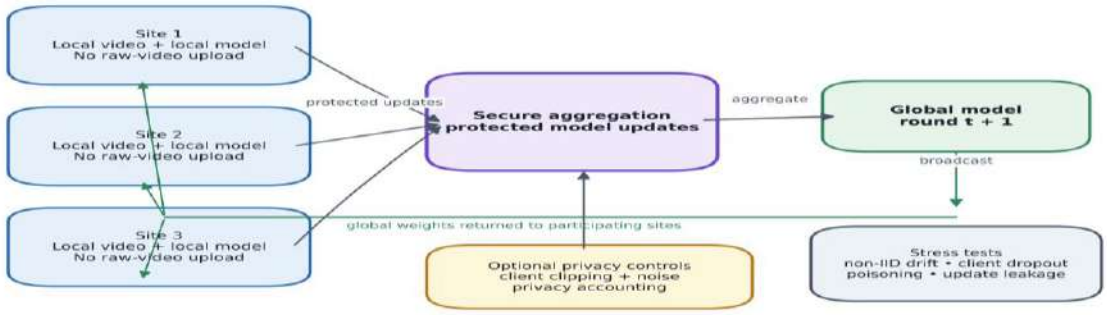


Fig. 4 Cross-silo federated learning for privacy-aware VAD. Raw video remains at each site; protected model updates are securely aggregated; optional clipping and differential privacy limit contribution exposure; and the evaluation includes non-IID, dropout, poisoning, and leakage stress tests.

4.7 Module specification and ablation variables

Table 6 TAP-VAD modules and planned ablations.

Component	Baseline	Proposed option	Required ablation
Video sampling	Uniform clips	Motion/novelty-aware adaptive selector	A1: uniform vs adaptive; A2: selector budget
Visual encoder	Video Swin or MViT	VideoMAE/DINOv2/CLIP-aligned features	A3: frozen vs fine-tuned; A4: backbone family
Temporal model	Full / windowed attention	Windowed attention + gated summary memory	A5: no memory; A6: no salience selector; A7: full attention
Anomaly head	One-class density or weakly supervised MIL	Dual class-agnostic + semantic heads	A8: remove semantic head; A9: head disagreement
Uncertainty	Raw score	Temperature scaling + ensemble/MC dropout	A10: uncalibrated; A11: calibration only; A12: ensemble
Explanation	Single heatmap	Spatial + temporal + concept evidence	A13: each modality alone; A14: explanation sanity check
Privacy	Central training / FedAvg	Secure aggregation + FedProx/SCAFFOLD; optional DP	A15: central vs FL; A16: heterogeneity optimizer; A17: DP sweep

Component	Baseline	Proposed option	Required ablation
Decision policy	Fixed score threshold	Alert / abstain / review based on calibrated risk and evidence	A18: no abstention; A19: uncertainty threshold sweep

5 EXPERIMENTAL PROTOCOL

5.1 Reproducible training and comparison

All comparisons should use a frozen protocol defined before the final test run. This includes dataset splits, frame rate or clip length, preprocessing, training epochs, optimizer, learning-rate schedule, initialization, model-selection criterion, anomaly threshold, calibration set, and prompt templates. At least three independent seeds are recommended for expensive video models; five seeds are preferred when feasible. Mean, standard deviation, and 95% confidence intervals should accompany primary metrics.

Baselines are grouped by capability rather than by publication date alone: reconstruction/prediction, memory-based normality, self-supervised/masked modeling, weakly supervised MIL, transformer-based VAD, density/score matching, vision–language/open-vocabulary methods, and deployment-oriented calibrated fusion. A new component is credited only when the corresponding ablation beats its controlled counterpart without unacceptable regressions in the trust dimensions. Paired bootstrap intervals or paired permutation tests can be used for test-set comparisons, with multiplicity control when many variants are compared.

5.2 Multi-dimensional evaluation

Table 7 Evaluation matrix and reporting cautions.

Dimension	Metrics	Test condition	Reporting caution
Detection	Frame/clip AUC; AP; F1; temporal IoU; localization overlap	Standard test split; per class/scene where possible	Ranking can look strong even when probabilities are unusable
Open vocabulary	Base/novel detection; category accuracy/mAP; harmonic mean	Prompt variants; seen/unseen split	Separate detection failure from naming failure
Calibration	ECE; Brier; NLL; reliability; AURC	Clean + shifted test sets	Report calibration set and binning; avoid ECE alone
Robustness	Relative AUC/AP drop; worst-group score	Compression, blur, illumination, occlusion, camera/scene/domain shift	Use realistic severity levels; retain clean reference
Explanation	Deletion/insertion; temporal IoU; stability; concept tests; human rating	Matched correct and incorrect predictions	Include explanations of errors, not only successes
Privacy	ε/δ if DP; leakage attack success; raw-data movement	FL with/without secure aggregation/DP	State attacker knowledge; privacy is threat-model specific
Efficiency	Latency; throughput; memory; parameters; FLOPs; energy; MB/round	Same hardware; batch 1 and deployment batch; long-video scaling	Include preprocessing, explanation, and communication overhead
Human oversight	Abstention rate; review precision/recall; time; override quality	Prospective or simulated triage study	Do not infer human benefit from model metrics alone

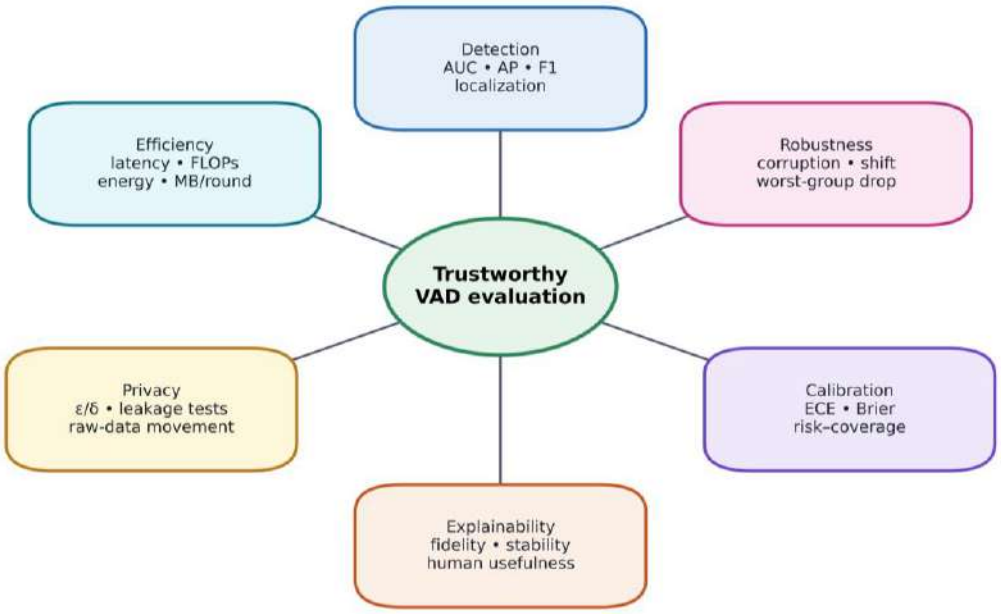


Fig. 6 Multi-dimensional evaluation scorecard. TAP-VAD reports dimension-wise evidence and Pareto trade-offs across detection, robustness, calibration, explanation, privacy, and efficiency instead of hiding competing objectives inside a single composite score.

5.3 Stress tests

Robustness tests should be separated into nuisance shift and semantic shift. Nuisance shift includes video compression, blur, noise, low light, frame dropping, mild camera shake, and partial occlusion. Semantic shift includes unseen scenes, new anomaly categories, benign rare events, and cross-dataset transfer. Common-corruption methodology provides a useful template for severity-controlled evaluation [71], while out-of-distribution baselines and uncertainty-under-shift work motivate explicit detection of unreliable cases [60,72].

Adversarial robustness is not the paper's central objective, but if adversarial claims are made they should be evaluated with strong parameter-free ensembles rather than a single weak attack [73]. More importantly for deployment, prompt sensitivity, client heterogeneity, calibration drift, and explanation instability should be treated as first-class stress tests because they arise directly from the proposed architecture.

5.4 Reproducibility and governance record

Table 8 Minimum reproducibility and governance information for an empirical TAP-VAD study.

Record	What must be documented	Status
Data provenance	Dataset version, license, official split, exclusions, frame extraction	Required
Model provenance	Backbone name/version, pretrained weights, license, parameter count	Required
Training	Seed, optimizer, schedule, epochs, batch size, augmentation, early stopping	Required
Calibration	Held-out calibration split, method, temperature/parameters, threshold selection	Required
Prompts	All text prompts, prompt ensemble rule, language, tokenizer/model version	Required for semantic head
Explanations	Method version, target layer/concept set, normalization, perturbation settings	Required
Federated setup	Number of clients, partition rule, rounds, local epochs, optimizer, aggregation	Required for FL

Record	What must be documented	Status
Privacy	Threat model, secure-aggregation setting, clipping/noise, ε/δ accountant, attack tests	Required if privacy is claimed
Compute	GPU/CPU model, software versions, latency protocol, energy tool if used	Required
Governance	Intended use, prohibited use, retention, human review, logging, failure escalation	Required for deployment claims

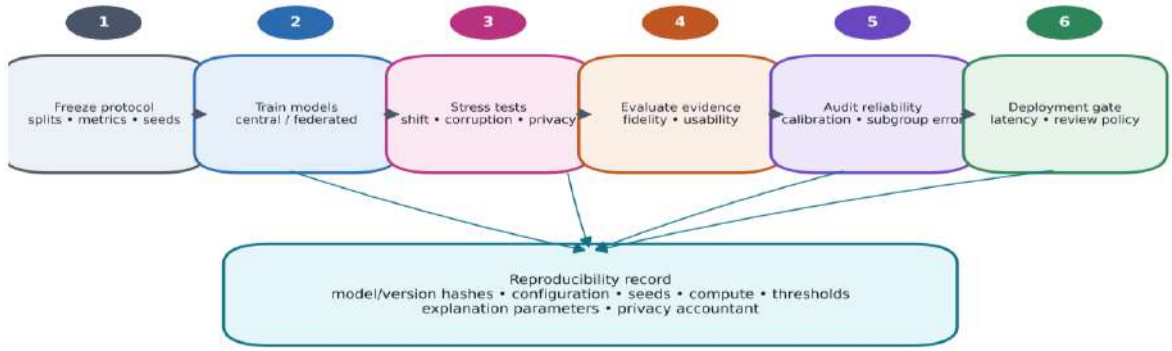


Fig. 7 Evaluation and governance lifecycle. The protocol is frozen before training; robustness, privacy, explanation, and calibration are audited before deployment; and all decisions are linked to a reproducibility record.

6 EXPECTED ANALYSES AND FALSIFIABLE OUTCOMES

Because this manuscript is derived from a research proposal rather than a completed benchmark study, this section states falsifiable expectations rather than fabricated results. The central architectural hypothesis is that windowed attention plus a compact temporal memory can approach full-attention detection quality at lower memory and latency, while adaptive salience preserves short anomaly evidence better than uniform down-sampling. The hypothesis is rejected if efficiency gains are accompanied by a statistically meaningful loss on short-duration or small-object anomalies.

The trust-layer hypothesis is that calibration and uncertainty-aware abstention will reduce high-confidence errors and improve risk–coverage behavior under scene shift. It is rejected if calibrated scores do not improve Brier/NLL/AURC or if abstention merely removes easy cases without enriching review with genuinely ambiguous events. Similarly, explanation fusion is justified only if temporal/spatial/concept evidence demonstrates fidelity and stability gains; attractive qualitative visualizations alone are insufficient.

For privacy, the testable expectation is not that federated learning will always outperform centralized learning. Rather, the framework asks whether acceptable utility can be retained under a stated privacy and communication budget, and how the trade-off changes across non-IID clients. A secure or private method that severely degrades performance on minority sites should be reported as such. The same principle applies to open-vocabulary semantics: improved novel-category coverage must be considered together with calibration, prompt sensitivity, and semantic confusion.

7 DISCUSSION

7.1 Why “anomaly” is not a neutral label

An anomaly is a deviation relative to a reference distribution, task definition, or social context. It is not equivalent to harm, criminality, or malicious intent. This distinction becomes especially important when semantic heads generate human-readable labels. A person running may be ordinary at a sports venue and unusual in a restricted corridor; a crowd may be expected during an event and anomalous at a closed facility. The technical system therefore needs scene context, but contextualization also increases the amount of information captured about people and places. Trustworthiness requires acknowledging this tension rather than presenting a universal anomaly taxonomy as objective ground truth.

For this reason, TAP-VAD treats model outputs as evidence for attention allocation. The system may state that a sequence differs from learned patterns, identify the frames and regions that contributed most, and report uncertainty. It should not infer intent beyond the available evidence. Human operators remain responsible for interpreting operational meaning under domain policy.

7.2 Explanation is evidence, not a guarantee

Explanations can fail in two directions. An unfaithful explanation can make a wrong prediction look reasonable; conversely, a technically faithful saliency map can be difficult for a human to interpret. The first problem concerns fidelity, the second usability. Existing attribution and concept methods address different parts of this space [45-47,49-50], while explanation research cautions against equating post-hoc narratives with transparent reasoning [53,54].

TAP-VAD therefore presents multiple evidence channels and explicitly includes explanations for errors. If an explanation method consistently highlights irrelevant scene regions, changes substantially under imperceptible perturbations, or narrates unsupported intent, the appropriate conclusion is that the explanation layer is inadequate—not that the detector has become trustworthy by virtue of having a visualization.

7.3 Privacy is broader than federated learning

Federated learning can reduce centralization of raw video, but it does not settle questions of whether recording was justified, who can access local footage, how long data are retained, or how alerts are used. Nor does it automatically prevent information leakage through updates. The privacy claim should therefore be phrased narrowly: the proposed training protocol minimizes raw-video movement and adds controls against specified update-level threats. Organizational governance remains necessary at the collection and decision layers.

7.4 Efficiency and sustainability

Computational efficiency is both an engineering and governance concern. A model that requires high-end hardware for every stream may be impractical for decentralized deployment and may encourage centralization of video solely for compute reasons. Masked pretraining, multiscale attention, token selection, lightweight fusion, and bounded memory offer alternative routes to efficiency [8,32,37,42]. TAP-VAD consequently reports end-to-end latency and memory—including explanation and privacy overhead—rather than only backbone FLOPs.

8 LIMITATIONS

First, TAP-VAD is a design and evaluation framework; it does not yet provide empirical benchmark gains. This is a deliberate limitation of the present manuscript and should be resolved through the protocol in Section 5 before any claim of state-of-the-art performance is made. Second, no finite benchmark can represent the open world of anomalous events. Cross-dataset and open-vocabulary tests reduce but do not eliminate this problem. Third, post-hoc explanations remain imperfect proxies for internal reasoning, and human usefulness studies can be affected by operator expertise and interface design.

Fourth, federated learning creates its own attack surface and communication cost. Differential privacy parameters that are acceptable for one cross-silo setting may be too restrictive or too weak for another. Fifth, trust dimensions can conflict: stronger privacy can reduce utility; more explanation can increase latency; greater abstention can reduce automated error but increase reviewer workload. These are not defects that should be hidden by a composite score. They are precisely the trade-offs the protocol is designed to expose.

9 CONCLUSION

Video anomaly detection is moving rapidly toward transformers, self-supervised video models, vision–language systems, and open-vocabulary reasoning. The next research problem is not only how to recognize more anomalies, but how to know when the system is uncertain, what evidence supports an alert, how sensitive the conclusion is to shift or prompt choice, whether sensitive video must be centralized, and whether the computational design is deployable. TAP-VAD addresses these questions in one architecture and one evaluation protocol. Its intended contribution is a testable standard for trustworthy VAD: detection quality remains central, but a deployable model must also earn confidence through calibration, robustness, faithful evidence, privacy accounting, efficiency measurement, and meaningful human oversight.

STATEMENTS AND DECLARATIONS

This anonymized review manuscript intentionally omits author-identifying declarations. Funding, competing interests, author contributions, correspondence details, and the disclosure of generative-AI assistance are provided on the separate title page in accordance with the journal's double-blind procedure. No new human-participant recruitment, animal research, or original empirical dataset is reported in this methodological manuscript. Any future collection of real-world video involving people must obtain the approvals and permissions required by the relevant institution and jurisdiction.

REFERENCES

[1]. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S., Learning temporal regularity in video sequences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 733–742, 2016.

[2]. Liu, W., Luo, W., Lian, D., & Gao, S., Future frame prediction for anomaly detection—A new baseline, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6536–6545, 2018.

[3]. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & van den Hengel, A., Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection, Proceedings of the IEEE/CVF International Conference on Computer Vision, 1705–1714, 2019.

[4]. Park, H., Noh, J., & Ham, B., Learning memory-guided normality for anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14372–14381, 2020.

[5]. Sultani, W., Chen, C., & Shah, M., Real-world anomaly detection in surveillance videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6479–6488, 2018.

[6]. Lv, H., et al., Unbiased multiple instance learning for weakly supervised video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8022–8031, 2023.

[7]. Georgescu, M.-I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M., Anomaly detection in video via self-supervised and multi-task learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12742–12752, 2021.

[8]. Tong, Z., Song, Y., Wang, J., & Wang, L., VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, Advances in Neural Information Processing Systems, 35, 2022.

[9]. Wu, P., Pan, C., Yan, Y., Pang, G., Yan, Q., Wang, P., & Zhang, Y., Deep learning for video anomaly detection: A review, IEEE Transactions on Neural Networks and Learning Systems, 37(7), 3010–3030. https://doi.org/10.1109/TNNLS.2025.3647892, 2026.

[10]. Micorek, J., Possegger, H., Narnhofer, D., Bischof, H., & Kozinski, M., MULDE: Multiscale log-density estimation via denoising score matching for video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18868–18877, 2024.

[11]. Li, F., Liu, W., Chen, J., Zhang, R., Wang, Y., Zhong, X., & Wang, Z., Anomize: Better open vocabulary video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 29203–29212, 2025.

[12]. Nasr, M., Shokri, R., & Houmansadr, A., Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, 2019 IEEE Symposium on Security and Privacy, 739–753. https://doi.org/10.1109/SP.2019.00065, 2019.

[13]. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V., How to backdoor federated learning, Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, PMLR 108, 2938–2948, 2020.

[14]. Uddin, M. P., et al., A systematic literature review of robust federated learning: Issues, solutions, and future research directions, ACM Computing Surveys, 57(10), 2025.

[15]. Dilek, E., & Dener, M., An overview of transformers for video anomaly detection, Neural Computing and Applications, 37, 17825–17857. https://doi.org/10.1007/s00521-025-11218-1, 2025.

[16]. Kang, Y., Hua, L., Fu, J., Yu, L., et al., ABD-CLIP: Anomaly-aware bidirectional CLIP with temporal graph-former for weakly supervised video anomaly detection, Complex & Intelligent Systems, 12, Article 230.

https://doi.org/10.1007/s40747-026-02349-6, 2026.

[17]. Ionescu, R. T., Khan, F. S., Georgescu, M.-I., & Shao, L., Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7842–7851, 2019.

[18]. Zhang, J., et al., Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16271–16280, 2023.

[19]. Cho, M., et al., Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12137–12146, 2023.

[20]. Sun, S., & Gong, X., Hierarchical semantic contrast for scene-aware video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22846–22856, 2023.

[21]. Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W., Qu, Y., Qiao, R., Ren, B., & Ma, L., TEVAD: Improved video anomaly detection with captions, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 5549–5559, 2023.

[22]. Liu, W., Chang, H., Ma, B., Shan, S., & Chen, X., Generating anomalies for video anomaly detection with prompt-based feature mapping, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 24500–24510, 2023.

[23]. Yang, Z., Liu, J., Wu, P., & Liu, X., Video event restoration based on keyframes for video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14592–14601, 2023.

[24]. Zanella, L., Menapace, W., Mancini, M., Wang, Y., & Ricci, E., Harnessing large language models for training-free video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18527–18536, 2024.

[25]. Huang, Y., Li, C., Zhang, H., Lin, Z., Lin, Y., Liu, H., Li, W., Liu, X., Gao, J., Huang, Y., Ding, X., & Yuan, Y., Track any anomalous object: A granular video anomaly detection pipeline, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8689–8699, 2025.

[26]. Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., & Zhang, Y., Open-vocabulary video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18297–18307, 2024.

[27]. Baradaran, M., & Bergevin, R., Multi-task learning based video anomaly detection with attention, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2886–2896, 2023.

[28]. Flaborea, A., et al. (2023). Are we certain it's anomalous? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2897–2907.

[29]. Chen, Z., et al., Aligning effective tokens with video anomaly in large language models, Proceedings of the IEEE/CVF International Conference on Computer Vision, 22695–22706, 2025.

[30]. Schmalwasser, D., et al., FastCAV: Efficient concept activation vectors for explainability, Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, 53316–53342, 2025.

[31]. Rai, A. K., Krishna, T., Hu, F., Drimbarean, A., Mcguinness, K., Smeaton, A. F., & O'Connor, N. E., Video anomaly detection via spatio-temporal pseudo-anomaly generation: A unified approach, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 3887–3899, 2024.

[32]. Ristea, N.-C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M., Self-distilled masked auto-encoders are efficient video anomaly detectors, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15984–15995, 2024.

[33]. Gundavarapu, N. B., et al., Extending video masked autoencoders to 128 frames, Advances in Neural Information Processing Systems, 37, 2024.

[34]. Yang, Z., Liu, J., & Wu, P., Text prompt with normality guidance for weakly supervised video anomaly detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18899–18908, 2024.

[35]. Zhang, H., Cao, C., Lv, Q., Min, L., & Zhang, Y., Autoregressive denoising score matching is a good video anomaly detector, Proceedings of the IEEE/CVF International Conference on Computer Vision, 12057–12067, 2025.

[36]. D'Amicantonio, G., Majhi, S., Kong, Q., Garattoni, L., Francesca, G., Bremond, F., &

Bondarev, E., Mixture of experts guided by Gaussian splatters matters: A new approach to weakly-supervised video anomaly detection, Proceedings of the IEEE/CVF International Conference on Computer Vision, 10275–10285, 2025.

[37]. Mohanarangan, K., & Palanisamy, P., Disagreement-aware late fusion of 3DCNN and MViT for clip-level video anomaly detection, Discover Computing, 29, 403. https://doi.org/10.1007/s10791-026-10197-8, 2026.

[38]. Vaswani, A., et al., Attention is all you need, Advances in Neural Information Processing Systems, 30, 2017.

[39]. Dosovitskiy, A., et al., An image is worth 16×16 words: Transformers for image recognition at scale, International Conference on Learning Representations, 2021.

[40]. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 813–824.

[41]. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C., ViViT: A video vision transformer, Proceedings of the IEEE/CVF International Conference on Computer Vision, 6836–6846, 2021.

[42]. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C., Multiscale vision transformers, Proceedings of the IEEE/CVF International Conference on Computer Vision, 6824–6835, 2021.

[43]. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H., Video Swin Transformer, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3202–3211, 2022.

[44]. Wang, Y., et al., Long-short temporal contrastive learning of video transformers, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14010–14020, 2022.

[45]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., Grad-CAM: Visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE International Conference on Computer Vision, 618–626, 2017.

[46]. Sundararajan, M., Taly, A., & Yan, Q., Axiomatic attribution for deep networks, Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 3319–3328, 2017.

[47]. Lundberg, S. M., & Lee, S.-I., A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 30, 2017.

[48]. Fong, R. C., & Vedaldi, A., Interpretable explanations of black boxes by meaningful perturbation, Proceedings of the IEEE International Conference on Computer Vision, 3429–3437, 2017.

[49]. Ribeiro, M. T., Singh, S., & Guestrin, C., 'Why should I trust you?' Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778, 2016.

[50]. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., & Sayres, R., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), Proceedings of the 35th International Conference on Machine Learning, PMLR 80, 2668–2677, 2018.

[51]. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P., Concept bottleneck models, Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 5338–5348, 2020.

[52]. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B., Sanity checks for saliency maps, Advances in Neural Information Processing Systems, 31, 2018.

[53]. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x, 2019.

[54]. Miller, T., Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007, 2019.

[55]. Padín-Torrente, H., Carneiro-Diaz, V., & Ortega-Fernandez, I., Toward human-centered explainability: Natural language explanations for anomaly detection, Information Systems Frontiers. https://doi.org/10.1007/s10796-026-10717-3, 2026.

[56]. van Mourik, F., et al. (2024). Tertiary review on explainable artificial intelligence: Where do we stand? Machine Learning and Knowledge Extraction, 6(3), 1997–2017.

[57]. Gal, Y., & Ghahramani, Z., Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, Proceedings of the 33rd International Conference on Machine Learning, PMLR 48, 1050–1059, 2016.

[58]. Lakshminarayanan, B., Pritzel, A., & Blundell, C., Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in Neural Information Processing Systems, 30, 2017.

[59]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q., On calibration of modern neural networks, Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330, 2017.

[60]. Ovadia, Y., et al., Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, Advances in Neural Information Processing Systems, 32, 2019.

[61]. Din, M. ud, Shoukat, I. A., Arif, E., Amjad, M., Mohsin, M., Mumtaz, M., & Rauf, M. A., Randomized frame selection based video steganography method for secure embedding of secret data, Journal of Computing & Biomedical Informatics, 8(2). https://doi.org/10.56979/802/2025, 2025.

[62]. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B., Communication-efficient learning of deep networks from decentralized data, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54, 1273–1282, 2017.

[63]. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K., Practical secure aggregation for privacy-preserving machine learning, Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191. https://doi.org/10.1145/3133956.3133982, 2017.

[64]. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V., Federated optimization in heterogeneous networks, Proceedings of Machine Learning and Systems, 2, 2020.

[65]. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T., SCAFFOLD: Stochastic controlled averaging for federated learning, Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 5132–5143, 2020.

[66]. Zhu, L., Liu, Z., & Han, S., Deep leakage from gradients, Advances in Neural Information Processing Systems, 32, 2019.

[67]. Kairouz, P., et al., Advances and open problems in federated learning, Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000083, 2021.

[68]. Cao, C., Zhang, Y., Lv, Q., Min, L., & Zhang, Y., A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20392–20401, 2023.

[69]. Radford, A., et al., Learning transferable visual models from natural language supervision, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 8748–8763, 2021.

[70]. Oquab, M., et al., DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research, 2024.

[71]. Hendrycks, D., & Dietterich, T., Benchmarking neural network robustness to common corruptions and perturbations, International Conference on Learning Representations, 2019.

[72]. Hendrycks, D., & Gimpel, K., A baseline for detecting misclassified and out-of-distribution examples in neural networks, International Conference on Learning Representations, 2017.

[73]. Croce, F., & Hein, M., Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2206–2216, 2020.