LOOK INSIDE EACH VIDEO: RETHINKING HOW VIDEO ANOMALY DETECTION IS EVALUATED

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly detection (VAD) aims to localize anomalous events by identifying when they occur. A common evaluation pools frame-level anomaly scores across all test videos and measures whether anomalous frames receive higher scores than normal frames (Micro-AUROC). This pooling creates comparisons both within the same video and across different videos. However, the pooled evaluation protocol commonly used in current VAD benchmarks makes Micro-AUROC heavily dependent on cross-video comparisons, which do not directly test whether anomalous frames rank above normal frames within the same video. To quantify this imbalance, we decompose Micro-AUROC into within-video and cross-video comparisons. Across four widely used weakly supervised VAD benchmarks, within-video pairs account for only 0.071–0.388% of all anomalous–normal frame pairs. This imbalance can make Micro-AUROC a poor indicator of within-video ranking. Video-level classifiers that assign a single score to each video achieve 81.40–97.18% Micro-AUROC while obtaining 50% Within-AUROC. For existing weakly supervised VAD detectors, replacing all frame scores in each video with their mean sets Within-AUROC to 50%. Across 120 runs, this replacement retains a median 98.8% of the original Micro-AUROC margin above chance. Because the decomposition depends on the evaluation protocol rather than the training supervision, the same distinction applies whenever VAD frame scores are pooled across videos. High Micro-AUROC alone therefore does not establish temporal localization. Our analysis reveals a limitation of pooled VAD evaluation and motivates reporting within-video ranking alongside Micro-AUROC.

1 INTRODUCTION

Weakly supervised video anomaly detection (WSVAD) uses video-level supervision to localize anomalous events by identifying when they occur (Wu et al., 2022b; Song et al., 2025). At test time, WSVAD methods assign an anomaly score to each frame (Tian et al., 2021). Micro-AUROC evaluates these scores by pooling all frames across the test set, whereas Macro-AUROC averages per-video AUROC over videos containing both anomalous and normal frames. Despite this distinction, WSVAD benchmarks predominantly report pooled metrics, such as Micro-AUROC and Average Precision (AP) (Sultani et al., 2018; Wu et al., 2020). By pooling frames across videos, these metrics assess how anomalous and normal frames are ranked across the test set rather than specifically within individual videos. We examine what this pooled evaluation captures by decomposing the anomalous–normal frame pairs underlying Micro-AUROC into two groups (Figure 1(a)): within-video pairs and cross-video pairs. Within-video pairs directly test whether anomalous frames rank above normal frames in the same video. In contrast, cross-video pairs compare frames from different videos and may be ranked correctly even when a scorer fails to distinguish anomalous from normal frames within a video.

This distinction raises a basic question: *how much does Micro-AUROC depend on distinguishing anomalous from normal frames within the same video?* We answer this question in three steps.

First, we decompose Micro-AUROC into *Within-AUROC*, computed over within-video pairs, and *Cross-AUROC*, computed over cross-video pairs. Across four widely used WSVAD benchmarks (ShanghaiTech, XD-Violence, UCF-Crime, and MSAD) within-video pairs account for only 0.071%–0.388% of all anomalous–normal frame pairs. Because within-video pairs constitute such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

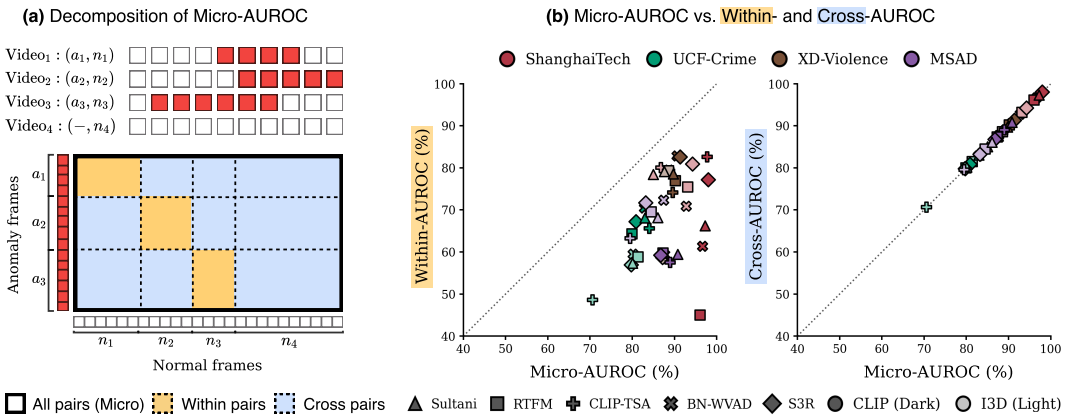

Figure 1: **Micro-AUROC closely matches Cross-AUROC.** (a) Anomalous–normal frame pairs are divided into within-video and cross-video comparisons. (b) Micro-AUROC is compared with Within-AUROC and Cross-AUROC across four benchmarks and five WSVAD methods. Detectors with similar Micro-AUROC can have substantially different Within-AUROC.

a small fraction of all pairs, even increasing Within-AUROC from chance to perfection can raise Micro-AUROC by at most 0.036–0.194 percentage points, with Cross-AUROC held fixed. Consistent with this pair-count analysis, Figure 1(b) shows that detector-level Micro-AUROC closely tracks Cross-AUROC, while Within-AUROC can differ substantially. We further derive an upper bound on the within-video pair fraction that decreases inversely with the number of videos when frames of each class remain broadly distributed across videos. We call this effect *temporal dilution*.

Second, this is not merely a theoretical possibility. We train video-level classifiers using only video-level labels and assign each video's predicted score to all of its frames. All video-constant outputs have Within-AUROC of 50% by construction. The best classifier achieves a Micro-AUROC of 81.40% to 97.18% across the benchmarks. Thus, high Micro-AUROC does not necessarily indicate that a model distinguishes anomalous from normal frames within individual videos.

Finally, we test whether existing detectors retain high Micro-AUROC after removing their within-video ranking. We replace every frame score in a video with that video's mean, making all anomalous–normal frame pairs within the video tied. If Micro-AUROC substantially depended on within-video ranking, this replacement should cause a substantial decrease. Instead, across 120 reproduced WSVAD runs, *video-mean replacement* preserves a median 98.8% of the original Micro-AUROC margin above chance.

Taken together, these results answer our question: on the benchmarks studied, Micro-AUROC depends only weakly on distinguishing anomalous from normal frames within the same video. High Micro-AUROC therefore does not by itself establish temporal localization. The same decomposition applies whenever VAD frame scores are pooled across videos for Micro-AUROC evaluation. We therefore argue that evidence of temporal localization must come from evaluating within-video ranking, not from pooled metrics alone.

2 Related Work

Video anomaly detection. Video anomaly detection assigns anomaly scores over time to identify unusual events in video. Existing methods are commonly studied under different supervision settings, including semi-supervised, weakly supervised, and training-free VAD, with different benchmarks and training assumptions (Tian et al., 2021; Zanella et al., 2024; Song et al., 2025; Li et al., 2026a; Song & Lee, 2026a). Despite these differences, these settings share a common evaluation target: a detector assigns anomaly scores over time, and performance is commonly measured by how these scores rank anomalous frames above normal frames (Appendix I). Our analysis concerns this shared evaluation step rather than a particular training paradigm. We use WSVAD as our primary empirical setting because its standard evaluation pools frame scores across many videos, making it a direct test of our question. We then extend the analysis to semi-supervised and training-free VAD.

Evaluation in video anomaly detection. Frame-level VAD evaluation typically pools predictions across the test set and summarizes how well anomaly scores separate anomalous from normal frames using AUROC or AP (Li et al., 2013; Lu et al., 2013; Liu et al., 2018; Sultani et al., 2018; Wu et al., 2020; Zhu et al., 2024). Subsequent work has changed different parts of this evaluation. One approach removes all-normal videos and evaluates only anomaly-containing videos, reducing the influence of predictions on videos that contain no anomaly (Lv et al., 2021; 2023). Another computes AUROC separately for each mixed-label video and averages across videos, thereby removing comparisons between frames from different videos (Acsintoae et al., 2022; Ristea et al., 2024). Complementary evaluations ask different localization questions. Temporal detection mAP evaluates predicted event intervals and, in fine-grained settings, anomaly categories (Wu et al., 2024c), while fixed-operating-point evaluation measures detection performance at specified false-positive rates rather than averaging over all thresholds (Acharya et al., 2026). These alternatives change the evaluated videos, the aggregation unit, the localization target, or the operating point. Our analysis instead leaves pooled Micro-AUROC unchanged and asks how its anomalous–normal comparisons are divided between frames from the same video and frames from different videos.

Decomposing pooled AUROC. The pairwise view of AUROC makes it possible to separate an aggregate ranking according to group membership. Prior work in fairness and bipartite ranking uses this view to distinguish within-group from cross-group positive–negative comparisons and to show that their relative weights depend on group and class composition (Borkan et al., 2019; Kallus & Zhou, 2019; Vogel et al., 2021). A related analysis in open-world VAD separates comparisons according to the queried anomaly definition, showing that one type of comparison can dominate the reported score (Song & Lee, 2026b). We instead partition pairs by video identity. With this grouping, within-group comparisons are exactly anomalous–normal pairs from the same video, while cross-group comparisons come from different videos. This lets us quantify how benchmark composition determines the direct weight of within-video ranking in Micro-AUROC and characterize how high Micro-AUROC can remain when every frame in a video receives the same score.

3 What does Micro-AUROC measure?

3.1 Within-video and cross-video comparisons

Micro-AUROC is computed after pooling the frame scores from all test videos. Equivalently, it measures how often an anomalous frame is ranked above a normal frame, with tied pairs receiving half credit (Song & Lee, 2026c). Each anomalous–normal pair either comes from the same video or from two different videos. We separate Micro-AUROC according to these two pair populations.

Let video i contain a_i anomalous frames and n_i normal frames, and let

$$A = \sum_i a_i, \qquad N = \sum_i n_i, \qquad W = \sum_i a_i n_i. \tag{1}$$

The pooled evaluation contains AN anomalous–normal pairs in total, of which W come from the same video. We define *Within-AUROC* as the average pairwise correctness over these W within-video pairs and *Cross-AUROC* as the average over the remaining $AN - W$ cross-video pairs.

Proposition 1 (Exact decomposition by video identity)**.** *When both pair populations are nonempty,*

$$\text{Micro-AUROC} = w\,\text{Within-AUROC} + (1-w)\,\text{Cross-AUROC}, \qquad w = \frac{W}{AN}. \tag{2}$$

The weight w depends only on the evaluated frame labels and video boundaries, not on the detector scores. With AUROC on the $[0, 1]$ scale, Equation (2) gives

$$|\text{Micro-AUROC} - \text{Cross-AUROC}| = w\,|\text{Within-AUROC} - \text{Cross-AUROC}| \le w. \tag{3}$$

Micro-AUROC must therefore be close to Cross-AUROC when w is small, regardless of the detector's within-video ranking. Appendix A proves the decomposition by partitioning anomalous–normal pairs according to video identity.

3.2 TEMPORAL DILUTION

The within-video pair fraction becomes small when anomalous and normal frames are spread across many videos. To make this precise, define

$$p_i^+ = \frac{a_i}{A}, \qquad p_i^- = \frac{n_i}{N}, \qquad w = \sum_i p_i^+ p_i^-. \tag{4}$$

Here, w is the probability that a randomly sampled anomalous frame and normal frame come from the same video.

This probability satisfies

$$w \le \min\left\{\max_i p_i^+, \max_i p_i^-\right\}. \tag{5}$$

If the largest fraction of anomalous frames and the largest fraction of normal frames contributed by any single video both scale as $O(1/V)$, where V is the number of videos, then $w = O(1/V)$. We call this decrease in the within-video pair fraction *temporal dilution*. The number of videos alone does not determine w, because the distribution of each class across videos also matters.

Corollary 1 (Maximum direct contribution of Within-AUROC)**.** *If Cross-AUROC is fixed, improving Within-AUROC from* 50% *to* 100% *can increase Micro-AUROC by at most* $w/2$*.*

Table 1 shows that only 0.071%–0.388% of anomalous–normal pairs come from the same video on the four WSVAD benchmarks. Holding Cross-AUROC fixed, improving Within-AUROC from 50% to 100% can therefore increase Micro-AUROC by at most 0.036–0.194 percentage points. With such small within-video pair fractions, Micro-AUROC is numerically close to Cross-AUROC, as shown in Figure 1(b). The same figure shows that detectors with similar Micro-AUROC can nevertheless have substantially different Within-AUROC.

Table 1: Within-video pair fraction and maximum direct contribution to Micro-AUROC.

Benchmark	V	w	Max. gain (pp)
ShanghaiTech	199	0.140%	0.070
UCF-Crime	290	0.388%	0.194
XD-Violence	800	0.071%	0.036
MSAD	240	0.206%	0.103

The four WSVAD benchmarks are not isolated cases. Figure 2 shows the within-video pair fraction across a broader set of 13 VAD benchmarks. Appendix B reports the statistics for each test split. UCSD Ped2 and Avenue have within-video fractions of several percent, whereas UBnormal and XD-Violence fall below 0.2%. The relationship is not determined by the number of videos alone, since the distribution of anomalous and normal frames also affects w. The broader pattern is consistent with temporal dilution.

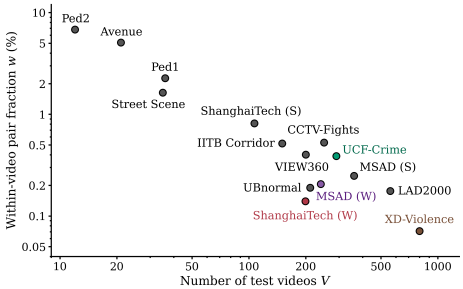

Figure 2: Within-video pair fractions across VAD benchmarks. Colored points denote the four WSVAD benchmarks used in our main experiments.

The decomposition therefore shows that pooled Micro-AUROC can assign very little weight to within-video comparisons, both on the WSVAD benchmarks and more broadly across VAD benchmarks.

4 VIDEO-LEVEL ANOMALY CLASSIFICATION

The previous section shows that only a small fraction of the comparisons in Micro-AUROC involve anomalous and normal frames from the same video. This raises a natural question: can a video-level anomaly classifier achieve high Micro-AUROC without distinguishing anomalous from normal frames within the same video? We test the simplest case by assigning one anomaly score to each video and repeating that score across all of its frames. Every within-video anomalous–normal pair is then tied, so Within-AUROC is 50% by construction.

Table 2: Micro-AUROC (%) from video-level anomaly classifiers. Each classifier predicts a single anomaly score per video and assigns that score to all frames in the video. Results are averaged over three runs. The oracle uses frame-level test labels to order videos by their ground-truth fraction of anomalous frames. Within-AUROC is 50% for every row.

Classifier	Feature	ShanghaiTech	UCF-Crime	XD-Violence	MSAD
(a) Mean-pool	CLIP	96.87±0.00	78.50±0.04	88.76±0.02	**89.76**±0.01
(b) Sequence-linear	CLIP	96.60±0.01	80.03±0.21	**90.78**±0.04	89.34±0.02
(c) Bi-GRU	CLIP	97.16±0.01	77.84±0.92	89.92±0.20	88.99±0.50
(d) Transformer	CLIP	**97.18**±0.15	80.40±0.87	88.13±1.77	88.82±0.27
(a) Mean-pool	I3D	92.06±0.04	79.21±0.10	89.22±0.02	86.20±0.05
(b) Sequence-linear	I3D	91.18±0.07	**81.40**±0.06	89.33±0.20	86.47±0.08
(c) Bi-GRU	I3D	91.10±0.48	77.26±0.09	90.45±0.64	84.63±0.28
(d) Transformer	I3D	85.34±2.12	79.07±0.77	87.30±1.42	85.17±0.81
Oracle using test labels		98.35	92.95	95.46	96.90

4.1 Maximum Micro-AUROC

We first characterize the maximum Micro-AUROC possible when every frame in a video receives the same anomaly score. In this setting, the ordering between videos determines all cross-video comparisons. Consider two videos i and j. Assigning a higher anomaly score to video i correctly orders $a_i n_j$ cross-video pairs, while assigning a higher score to video j correctly orders $a_j n_i$ pairs. The first ordering is preferable exactly when

$$a_i n_j \geq a_j n_i \qquad \Longleftrightarrow \qquad \frac{a_i}{a_i + n_i} \geq \frac{a_j}{a_j + n_j}. \tag{6}$$

Proposition 2 (Optimal ordering with one anomaly score per video)**.** *If every frame in a video receives the same anomaly score, Micro-AUROC is maximized by ordering videos according to their ground-truth fraction of anomalous frames $a_i/(a_i + n_i)$, with arbitrary tie breaking among equal fractions.*

The pairwise preferences above are transitive, so ordering videos by their fraction of anomalous frames maximizes the number of correctly ordered cross-video pairs. Within-AUROC remains fixed at 50%, so the same ordering maximizes Micro-AUROC by Equation (2). This ordering is also consistent with likelihood-ratio ordering in bipartite ranking and ROC analysis (Fawcett & Niculescu-Mizil, 2007; Clémençon et al., 2008; Menon & Williamson, 2016). Appendix A gives the proof.

Using the frame-level test labels to construct this ordering gives 98.35% Micro-AUROC on ShanghaiTech, 92.95% on UCF-Crime, 95.46% on XD-Violence, and 96.90% on MSAD (Table 2). Thus, assigning only one anomaly score to each video can yield 92.95%–98.35% Micro-AUROC on these benchmarks, even though Within-AUROC is 50%.

4.2 Training Video-Level Anomaly Classifiers

The maximum above uses frame-level test labels to determine the ordering between videos. We therefore next test whether video-level labels alone are sufficient to produce high Micro-AUROC. We train four video-level anomaly classifiers using the same video-level supervision available in WSVAD. Each classifier predicts one anomaly score for an entire video. For Micro-AUROC evaluation, we repeat this video-level anomaly score across all frames of the video. Within-AUROC is therefore 50% for every classifier.

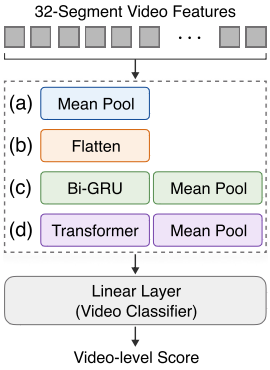

Figure 3: Architecture of the four video classifiers.

As illustrated in Figure 3 the classifiers differ in how they process the video features before predicting the video-level anomaly score. Mean-pool averages the features over time before classification. Sequence-linear, bi-GRU, and Transformer classifiers instead process the standard

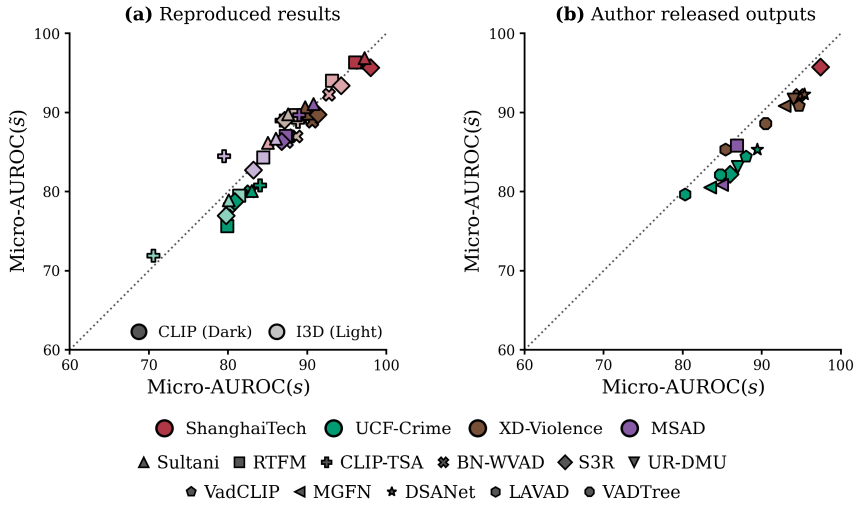

Figure 4: **Micro-AUROC remains high after video-mean replacement.** All frame scores in each video are replaced with that video's mean score. The horizontal axis shows the original Micro-AUROC, and the vertical axis shows Micro-AUROC after replacement. (a) Each point represents one of 40 combinations of method, feature, and benchmark, averaged over three training runs. (b) Each point represents one of 13 released results. The diagonal denotes no change.

32-segment feature sequence (Sultani et al., 2018) before predicting one anomaly score for the video. We evaluate each classifier with CLIP and I3D features over three runs.

Table 2 shows that video-level anomaly classifiers achieve high Micro-AUROC across all four benchmarks. The best classifier reaches 97.18% on ShanghaiTech, 81.40% on UCF-Crime, 90.78% on XD-Violence, and 89.76% on MSAD. On ShanghaiTech with CLIP features, the video-level Transformer achieves 97.18%, compared with 97.74% for CLIP-TSA. Their Within-AUROC values are 50% and 82.59%, respectively (Appendix Table 9). Thus, similar Micro-AUROC values can accompany very different within-video rankings even when both models use video-level supervision and the same feature representation.

5 Within-Video Ranking in Existing Detectors

We next examine whether high Micro-AUROC in existing detectors reflects within-video ranking. We evaluate the MIL ranking method of Sultani et al. (2018), RTFM (Tian et al., 2021), S3R (Wu et al., 2022b), CLIP-TSA (Joo et al., 2023), and BN-WVAD (Zhou et al., 2024) with I3D and CLIP features on ShanghaiTech, UCF-Crime, XD-Violence, and MSAD. Three training runs per combination give 120 runs. We separately evaluate 13 results from author-released checkpoints or frame-level anomaly scores. Appendix C gives the experimental details.

5.1 Video-Mean Replacement

We replace all frame scores in each video with that video's mean and evaluate the original and replaced scores on the same frames and annotations. Given frame-level scores $s = \{s_{it}\}_{i,t}$ and T_i evaluated frames in video i, *video-mean replacement* gives

$$\tilde{s}_{it} = \bar{s}_i = \frac{1}{T_i} \sum_{u=1}^{T_i} s_{iu}. \tag{7}$$

Every within-video anomalous–normal pair is tied, so Within-AUROC is 50%. Replacement can also change cross-video comparisons, so its effect on Micro-AUROC must be measured rather than inferred from Equation (2).

Table 3: Retained Micro-AUROC margin above chance after video-mean replacement. Each reproduced run or released result contributes one R_{mean} value as defined in Equation (8). The final two columns count results retaining at least 85% and 90% of the original margin, respectively. Reproduced runs and released results are summarized separately.

Benchmark	n	Median R_{mean}	Minimum R_{mean}	$R_{\text{mean}} \geq 85\%$	$R_{\text{mean}} \geq 90\%$
Reproduced runs					
ShanghaiTech	30	99.3%	95.0%	$30/30$	$30/30$
UCF-Crime	30	92.6%	78.1%	$29/30$	$26/30$
XD-Violence	30	100.0%	94.3%	$30/30$	$30/30$
MSAD	30	100.1%	95.5%	$30/30$	$30/30$
All	120	98.8%	78.1%	$119/120$	$116/120$
Released checkpoints and scores					
All	13	91.4%	88.3%	$13/13$	$9/13$

With AUROC on the $[0, 1]$ scale, the retained margin above chance is

$$R_{\text{mean}} = \frac{\text{Micro-AUROC}(\tilde{s}) - 0.5}{\text{Micro-AUROC}(s) - 0.5}. \tag{8}$$

We compute this ratio for each reproduced run or released result before aggregation. Values of 100% indicate unchanged Micro-AUROC, and larger values indicate an increase.

Video-mean replacement preserves a median 98.8% of the Micro-AUROC margin across 120 reproduced runs, with 116 retaining at least 90% (Figure 4 and table 3). UCF-Crime has the largest decreases, with median retention of 92.6% and a minimum of 78.1%. The 13 released results retain a median 91.4% and a minimum of 88.3%. High Micro-AUROC thus persists despite chance within-video ranking. Appendix E gives the complete results.

For frame-level AP, we compute retention above the anomalous-frame fraction rather than 50%. Median AP retention is 94.6% on XD-Violence and 93.0% on MSAD, compared with 80.8% on ShanghaiTech and 72.1% on UCF-Crime across the reproduced runs. These are empirical results, not consequences of the AUROC decomposition (Appendix E.3).

5.2 Evaluation Beyond Weak Supervision

Cross-video comparisons remain dominant beyond weak supervision, but video-mean replacement need not preserve high Micro-AUROC. The decomposition applies regardless of training supervision, whereas replacement can also change cross-video score orderings. Table 4 compares semi-supervised and training-free detectors on the same frames before and after replacement. For TrajVAD-P (Song & Lee, 2026a) on ShanghaiTech, Micro-AUROC is 88.59% and Cross-AUROC is 88.57%, even though Within-AUROC is 91.58%. Only 0.814% of anomalous–normal pairs come from the same video, so cross-video comparisons dominate despite strong within-video ranking.

Replacement has little effect on LAVAD (Zanella et al., 2024) and VADTree (Li et al., 2026a) on XD-Violence. LAVAD changes from 85.40% to 85.32% Micro-AUROC, whereas TrajVAD-P on ShanghaiTech decreases from 88.59% to 61.92%. The test splits also permit different maximum values with one anomaly score per video. The optimal ordering in Proposition 2 gives 98.35% Micro-AUROC on ShanghaiTech's weakly supervised split but 75.78% on its semi-supervised split (Table 6). The value attained after replacement further depends on how the detector means order the videos.

Large decreases are not confined to semi-supervised methods. On the 44 anomaly-containing ShanghaiTech videos shared by the weakly supervised and semi-supervised evaluations, CLIP-TSA with CLIP features decreases from 68.58% to 45.22% Micro-AUROC after replacement. TrajVAD-P also decreases on these frames, from 89.66% to 63.16%. Both evaluations have $w = 1.891\%$. A small within-video pair fraction therefore does not by itself imply high retention after replacement. Appendix F reports the complete results and common-test comparison.

Table 4: **Cross-video comparisons remain dominant beyond weak supervision.** We report the same detectors across multiple benchmarks. AUROC values and the within-video pair fraction w are percentages. The final column gives Micro-AUROC after video-mean replacement, which sets Within-AUROC to 50% for every row.

Benchmark	Method	w (%)	Original scores			After replacement
			Micro-AUROC	Cross-AUROC	Within-AUROC	Micro-AUROC
Semi-supervised						
ShanghaiTech	TrajVAD-P	0.814	88.59	88.57	91.58	61.92
ShanghaiTech	UniVAD	0.814	89.32	89.30	92.22	61.30
Avenue	Jigsaw-VAD	5.079	90.32	90.21	92.40	65.27
UCSD Ped2	Jigsaw-VAD	6.802	86.41	85.96	92.57	61.72
Training-free						
UCF-Crime	LAVAD	0.388	80.28	80.36	61.75	79.66
UCF-Crime	VADTree	0.388	84.74	84.78	74.17	82.10
XD-Violence	LAVAD	0.071	85.40	85.42	64.78	85.32
XD-Violence	VADTree	0.071	90.49	90.50	80.80	88.62

Table 5: **Changing the ranking metric changes detector comparisons.** For each benchmark, 5 methods evaluated with 2 feature representations give 10 configurations, which are ranked by their 3-run mean performance. We compare all 45 unique pairs of configurations under Micro-AUROC and each alternative metric. A reversal occurs when the two metrics give opposite orderings for a pair. Ties are excluded from the reversal count. ρ is Spearman's rank correlation.

	Within-AUROC		Macro-AUROC		AUC_A	
Benchmark	Reversals	ρ	Reversals	ρ	Reversals	ρ
ShanghaiTech	23	-0.10	21	0.10	19	0.20
UCF-Crime	9	0.78	10	0.68	13	0.58
XD-Violence	15	0.44	18	0.26	25	-0.08
MSAD	31	-0.53	28	-0.30	27	-0.35
All (180 pairs)	78 (43.3%)	–	77 (42.8%)	–	84 (46.7%)	–

5.3 Detector Comparisons Across Ranking Metrics

Micro-AUROC can favor a detector with worse within-video ranking. We compare detector configurations using their original frame-level anomaly scores under Micro-AUROC, Within-AUROC, and Macro-AUROC. Within-AUROC weights each mixed-label video by its number of anomalous–normal pairs, whereas Macro-AUROC gives each such video equal weight.

On ShanghaiTech with CLIP features, S3R has higher Micro-AUROC than CLIP-TSA, at 97.99% versus 97.74%. Within-AUROC instead favors CLIP-TSA, at 82.59% versus 77.21% for S3R. Macro-AUROC gives the same ordering, at 81.85% versus 80.02%. Thus, the detector favored by pooled evaluation is worse under both within-video metrics.

These reversals occur across benchmarks and feature representations. Table 5 shows opposite orderings to Micro-AUROC for 78 of 180 configuration pairs under Within-AUROC and 77 pairs under Macro-AUROC. On every benchmark, both within-video metrics reverse at least one Micro-AUROC ordering among methods using CLIP features and among methods using I3D features. The disagreement therefore persists when the feature representation is held fixed.

Even when the metrics agree on detector ordering, they can show very different performance gaps. On ShanghaiTech with CLIP features, CLIP-TSA exceeds RTFM by 1.69 percentage points in Micro-AUROC but by 37.60 points in Within-AUROC. Micro-AUROC alone therefore establishes neither the ordering nor the size of performance differences under within-video evaluation. Appendix Table 9 reports the configuration-level scores.

Removing all-normal videos does not isolate within-video ranking. $\mathrm{AUC_A}$ removes all-normal videos and pools frames from the remaining videos (Lv et al., 2021; 2023). Using the benchmark annotations, within-video pairs still account for only 0.24%–1.90% of the remaining comparisons. Ordering videos by their ground-truth anomaly fractions gives 77.68%–85.20% $\mathrm{AUC_A}$ with one score per video (Table 17). Anomaly-only pooling changes detector rankings, but the remaining cross-video comparisons prevent it from replacing within-video evaluation (Table 5).

6 Implications for VAD Evaluation

Report within-video ranking alongside pooled metrics. The ranking reversals in Table 5 show that higher Micro-AUROC need not indicate better within-video ranking. This distinction is rarely reflected through Macro-AUROC in the VAD papers we surveyed: 112 of 113 report frame-level AUROC or AP, whereas only 10 report Macro-AUROC, and none of the 59 WSVAD papers do so (Table 22). We therefore recommend reporting Micro-AUROC together with Within-AUROC and Macro-AUROC to distinguish pooled from within-video ranking. Within-AUROC weights each mixed-label video by its number of anomalous–normal pairs, whereas Macro-AUROC gives each mixed-label video equal weight. Cross-AUROC and the within-video pair fraction w explain how the two pair populations contribute to Micro-AUROC under Equation (2).

Report the composition of the evaluated split. The within-video pair fraction depends on the evaluated frames and video boundaries, not on a dataset name alone. For example, ShanghaiTech has $w = 0.140\%$ on its weakly supervised split and $w = 0.814\%$ on its semi-supervised split (Table 6). Removing all-normal videos does not isolate within-video ranking: after anomaly-only pooling, within-video pairs still account for only 0.24%–1.90% of comparisons on the four benchmarks (Table 17). Evaluation results should therefore identify the tested split and, when pooled AUROC is reported, its within-video pair fraction.

Scope of within-video evaluation. Within-AUROC and Macro-AUROC measure ranking within mixed-label videos. They do not evaluate false alarms on all-normal videos or temporal interval boundaries. These properties require separate evaluation when they are part of the claim being made (Wu et al., 2024c; Acharya et al., 2026). Appendix H examines interval-based evaluation.

7 Limitations

Statistical uncertainty in detector comparisons. Our comparisons of Micro-AUROC, Within-AUROC, and Macro-AUROC report point estimates of performance differences between detectors. We do not estimate uncertainty in these differences. Because frames within the same video are dependent, such uncertainty should be estimated using video-level resampling or a variance estimator that accounts for within-video dependence (Appendix G,Obuchowski (1997)).

Dependence on frame-level annotations. Multi-round re-annotation of UCF-Crime and XD-Violence documents substantial annotator disagreement (Liu et al., 2025). The decomposition and bound on w hold for any fixed labeling, but the numerical values of w and Within-AUROC depend on the annotated event boundaries. Comparisons between detectors under within-video evaluation therefore also depend on these annotations.

8 Conclusion

Micro-AUROC is widely used to compare VAD model performance. However, high Micro-AUROC does not by itself establish that a detector ranks anomalous frames above normal frames within the same video. We examine this gap by decomposing Micro-AUROC into within-video and cross-video comparisons. Video-level classifiers and video-mean replacement then show that Micro-AUROC can remain high even when within-video ranking is at chance. Evaluating the original detector scores within videos further reveals differences in within-video performance that Micro-AUROC can obscure, including reversals in detector rankings. VAD evaluations should therefore distinguish pooled ranking, within-video ranking, and interval detection, and support claims about each with the corresponding evidence.

AI USE STATEMENT

In this work, we used generative AI tools to implement isolated utility functions in inference code, and these functions were verified and tested for correctness by the authors. Additionally, we used generative AI tools for proofreading, grammar checking, and typographical checks during writing. We have reviewed all AI-assisted work, and every reported number was produced by the deterministic evaluation pipeline of the Reproducibility Statement. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes existing video anomaly detection benchmarks. We do not collect or annotate new videos, nor do we redistribute the benchmark videos. Our analysis reports aggregate evaluation results and does not involve person-level decisions.

REPRODUCIBILITY STATEMENT

Appendix C specifies the benchmark splits and annotations, the evaluated frame coverage, the method–feature–benchmark training grid and seed aggregation, and the handling of author-released checkpoints and scores. It also describes the video-level classifier training, AUROC tie convention, and checks for video-mean replacement. Appendix E reports the method-level results, while Appendix A gives the proofs of the AUROC decomposition and video-constant optimum.

REFERENCES

Anusha Acharya, Hitesh Sapkota, Qi Yu, and Xumin Liu. The road less seen: Segment exploration for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14147–14156, 2026.

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20143–20153, 2022.

Sunghyun Ahn, Youngwan Jo, Kijung Lee, and Sanghyun Park. Videopatchcore: An effective method to memorize normality for video anomaly detection. In *Asian Conference on Computer Vision*, pp. 312–328. Springer, 2024.

Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. Anyanomaly: Zero-shot customizable video anomaly detection with lvlm. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3026–3035. IEEE, 2026.

Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, François Brémond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10275–10285, 2025.

Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742*, 2021a.

Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 207–214. IEEE, 2021b.

Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Pseudobound: Limiting the anomaly reconstruction capability of one-class classifiers using pseudo anomalies. *Neurocomputing*, 534:147–160, 2023.

Mohammad Baradaran and Robert Bergevin. Multi-task learning based video anomaly detection with attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2886–2896. IEEE, 2023.

Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, Jacob Dueholm, Bharathkumar Ramachandra, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229:103656, 2023.

Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Limitations of pinned auc for measuring unintended bias. *arXiv preprint arXiv:1903.02088*, 2019.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20392–20401. IEEE, 2023.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18319–18329. IEEE, 2024.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection, 2022.

Kai Cheng, Yaning Pan, Yang Liu, Xinhua Zeng, and Rui Feng. Denoising diffusion-augmented hybrid video anomaly detection via reconstructing noised frames. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 695–703, 2024.

MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129: 108703, 2022.

MyeongAh Cho, Taeoh Kim, Minho Shim, Dongyoon Wee, and Sangyoun Lee. Towards multi-domain learning for generalizable video anomaly detection. *Advances in Neural Information Processing Systems*, 37:50256–50284, 2024.

Chengxi Chu, Nurul Japar, and Chee Kau Lim. Scene-dependent video anomaly detection: A benchmark and weakly supervised model. *Alexandria Engineering Journal*, 133:477–486, 2025.

Chengxi Chu, Nurul Japar, and Chee Kau Lim. Reba: Residual mixture-of-experts and bidirectional video-text alignment for better fine-grained weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8280–8290, June 2026.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. 2008.

Jash Dalvi, Ali Dabouei, Gunjan Dhanuka, and Min Xu. Distilling aggregated knowledge for weakly-supervised video anomaly detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5439–5448. IEEE, 2025.

Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68 (1):97–106, 2007.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14004–14013. IEEE, 2021.

Yue Feng, Yingshi Liu, Jingrou Zhang, Jie Qin, et al. Monitor: Exploiting large language models with instruction for online video anomaly detection. *Advances in Neural Information Processing Systems*, 38:35851–35871, 2026.

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12742–12752, 2021.

Ayush Ghadiya, Purbayan Kar, Vishal Chudasama, and Pankaj Wasnik. Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1965–1974. IEEE, 2024.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1705–1714, 2019.

Satoshi Hashimoto, Tatsuya Konishi, Tomoya Kaichi, Kazunori Matsumoto, and Mori Kurokawa. Cade: Continual weakly-supervised video anomaly detection with ensembles. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 708–717. IEEE, 2026.

Ping He, Huibin Li, and Miaolin Han. A self-paced multiple instance learning framework for weakly supervised video anomaly detection. *Applied Sciences*, 15(3):1049, 2025.

Han Hu, Wenli Du, Peng Liao, Bing Wang, and Siyuan Fan. Noise-resistant video anomaly detection via rgb error-guided multiscale predictive coding and dynamic memory. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19109–19119. IEEE, 2025.

Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 7834–7843. IEEE, 2019.

Yashika Jain, Ali Dabouei, and Min Xu. Cross-domain learning for video anomaly detection with limited supervision. In *European Conference on Computer Vision*, pp. 468–484. Springer, 2024.

Yalong Jiang. Local patterns generalize better for novel anomalies. In *International Conference on Learning Representations*, volume 2025, pp. 49969–49995, 2025.

Yalong Jiang and Liquan Mao. Vision-language models assisted unsupervised video anomaly detection. *arXiv preprint arXiv:2409.14109*, 2024.

Wenping Jin, Li Zhu, and Jing Sun. Aligning first, then fusing: A novel weakly supervised multimodal violence detection method. *Knowledge-Based Systems*, 322:113709, 2025.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3230–3234. IEEE, 2023.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6834–6842. IEEE, 2024.

Demetris Lappas, Vasileios Argyriou, and Dimitrios Makris. Dynamic distinction learning: adaptive pseudo anomalies for video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3961–3970. IEEE, 2024.

Junyeop Lee, Hyunbon Koo, Seongjun Kim, and Hanseok Ko. Cognitive refined augmentation for video anomaly detection in weak supervision. *Sensors*, 24(1):58, 2023.

Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, and Xinbo Gao. Beyond euclidean: Dual-space representation learning for weakly supervised video violence detection. *Advances in Neural Information Processing Systems*, 37:17373–17397, 2024.

Min Li, Jing Sang, Yuanyao Lu, and Lina Du. Wsvad-clip: Temporally aware and prompt learning with clip for weakly supervised video anomaly detection. *Journal of Imaging*, 11(10):354, 2025.

Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1395–1403, 2022.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *Advances in Neural Information Processing Systems*, 38:148372–148404, 2026a.

Zhaoyi Li, Chenrui Shi, Che Sun, and Yuwei Wu. Robust video anomaly detection via causal feature-guided data augmentation. *Computer Vision and Image Understanding*, pp. 104671, 2026b.

Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *Advances in Neural Information Processing Systems*, 38:29713–29744, 2026.

Hao Liu, Lijun He, Miao Zhang, and Fan Li. Vadiffusion: Compressed domain information guided conditional diffusion for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8398–8411, 2024.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Zihao Liu, Xiaoyu Wu, Wenna Li, Linlin Yang, and Shengjin Wang. Rethinking metrics and benchmarks of video anomaly detection. *arXiv preprint arXiv:2505.19022*, 2025.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2013.

Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *European conference on computer vision*, pp. 125–141. Springer, 2020.

Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021.

Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031. IEEE, 2023.

Jingke Meng, Huilin Tian, Ge Lin, Jian-Fang Hu, and Wei-Shi Zheng. Audio-visual collaborative learning for weakly supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2025.

Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18868–18877, 2024.

Mengjingcheng Mo, Jiankang Zheng, Jiaxu Leng, and Xinbo Gao. Retrieval-guided contextual inference for training-free video anomaly detection in low-light scenarios. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 930–939, 2026.

Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11988–11996. IEEE, 2019.

Furkan Mumcu, Michael J Jones, Anoop Cherian, and Yasin Yilmaz. Leveraging multimodal llm descriptions of activity for explainable semi-supervised video anomaly detection. *arXiv preprint arXiv:2510.14896*, 2025.

Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1273–1283, 2019.

Nancy A Obuchowski. Nonparametric analysis of clustered roc curve data. *Biometrics*, pp. 567–578, 1997.

Aniello Panariello, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Consistency-based self-supervised learning for temporal anomaly localization. In *European Conference on Computer Vision (ECCV) Workshops*, 2022.

Wenfeng Pang, Wei Xie, Qianhua He, Yanxiong Li, and Jichen Yang. Audiovisual dependency attention for violence detection in videos. *IEEE Transactions on Multimedia*, 25:4922–4932, 2023.

Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1908–1918. IEEE, 2022.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14360–14369. IEEE, 2020.

Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 33:4923–4936, 2024.

Helei Qiu and Biao Hou. Temporal context and representative feature learning for weakly supervised video anomaly detection. *Journal of Information and Intelligence*, 2025.

Helei Qiu, Biao Hou, and Yanyu Cui. Learning opposite prompts for weakly supervised video anomaly detection. *Knowledge-Based Systems*, 324:113600, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ayush K Rai, Tarun Krishna, Feiyan Hu, Alexandru Drimbarean, Kevin McGuinness, Alan F Smeaton, and Noel E O'connor. Video anomaly detection via spatio-temporal pseudo-anomaly generation: A unified approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3887–3899, 2024.

Praboda Rajapaksha, Noel Crespi, Xuechen Zhao, Mengzhu Wang, Nan Yin, Xinwang Liu, Zhigang Luo, et al. Consistency-constrained unsupervised video anomaly detection framework based on co-teaching. *Neurocomputing*, 610:128589, 2024.

Tal Reiss and Yedid Hoshen. An attribute-based method for video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13566–13576. IEEE, 2022.

Nicolae-Catalin Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, and Mubarak Shah. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15984–15995, 2024.

Debi Prasad Senapati, Santosh Kumar Pani, Santos Kumar Baliarsingh, Prabhu Prasad Dev, and Ramasubbareddy Somula. Poinclip-vad: a hyperbolic cross-modal fusion framework for video anomaly detection. *Scientific Reports*, 16(1):20341, 2026.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025.

Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10296–10306. IEEE, 2023.

Inpyo Song and Jangwon Lee. Bounding-box trajectories matter for video anomaly detection. *arXiv preprint arXiv:2605.21957*, 2026a.

Inpyo Song and Jangwon Lee. Rethinking open-world video anomaly detection: Diagnosing definition blindness. *arXiv preprint arXiv:2607.20780*, 2026b.

Inpyo Song and Jangwon Lee. A vlm answer is not an anomaly score: Rank compression in training-free video anomaly detection. *arXiv preprint arXiv:2608.21244*, 2026c.

Inpyo Song, Sanghyeon Lee, Minjun Joo, and Jangwon Lee. Anomaly detection for people with visual impairments using an egocentric 360-degree camera. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2828–2837. IEEE, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Bo Sun, Junxi Chen, Zhe Wu, Feng Gao, Fan Yang, Li Su, and Yaowei Wang. Joint learning of general and diverse patterns with mixture of memory experts for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35638–35647, June 2026a.

Shengyang Sun and Xiaojin Gong. Event-driven weakly supervised video anomaly detection. *Image and Vision Computing*, 149:105169, 2024.

Shengyang Sun, Jiashen Hua, Junyi Feng, and Xiaojin Gong. Enhancing weakly supervised multimodal video anomaly detection through text guidance. *IEEE Transactions on Multimedia*, 2026b.

Wenwen Sun, Lin Cao, Yanan Guo, and Kangning Du. Multimodal and multiscale feature fusion for weakly supervised video anomaly detection. *Scientific Reports*, 2024.

Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. X-man: Explaining multiple sources of anomalies in video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3218–3226. IEEE, 2021.

Weijun Tan, Qi Yao, and Jingfeng Liu. Overlooked video classification in weakly supervised video anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 212–220. IEEE, 2024.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4975–4986, 2021.

Robin Vogel, Aurélien Bellet, and Stephan Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International conference on artificial intelligence and statistics*, pp. 784–792. PMLR, 2021.

Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pp. 494–511. Springer, 2022.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. *arXiv preprint arXiv:2603.00550*, 2026.

Zhiqiang Wang, Xiaojing Gu, Jingyu Hu, and Xingsheng Gu. Ensemble anomaly score for video anomaly detection using denoise diffusion model and motion filters. *Neurocomputing*, 553: 126589, 2023.

Chongke Wu, Sicong Shao, Cihan Tunc, Pratik Satam, and Salim Hariri. An explainable and efficient deep learning framework for video anomaly detection. *Cluster computing*, 25(4):2715–2737, 2022a.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pp. 729–745. Springer, 2022b.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pp. 322–339. Springer, 2020.

Peng Wu, Xiaotao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685, 2022c.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307. IEEE, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9301–9310, 2024b.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6074–6082, 2024c.

Chen Xu, Chunguo Li, and Hongjie Xing. Discriminative score suppression for weakly supervised video anomaly detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9587–9596. IEEE, 2025.

Rong Xu, Runqi Wang, Yingjun Zhang, Tao Tao, Xiaomeng Li, and Liping Jing. Tlma: Mitigating the impact of weakly labeled information for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35597–35606, 2026.

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5504–5514. IEEE, 2023.

Chen Yang, Xue-song Tang, Kuangrong Hao, Yubing Li, and Zhiqi Wang. Explicit knowledge-structured weakly supervised video anomaly detection. *Digital Signal Processing*, pp. 106088, 2026.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pp. 304–322. Springer, 2024a.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14592–14601. IEEE, 2023.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18899–18908. IEEE, 2024b.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688. IEEE, 2025.

Wenti Yin, Huaxin Zhang, Xiang Wang, Yuqing Lu, Yicheng Zhang, Bingquan Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12027–12035, 2026.

Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 583–591, 2020.

Guang Yu, Junhua Xi, Siqi Wang, Zhiping Cai, Kefeng Deng, and Kaijun Ren. Effective video anomaly detection by step-constrained diffusion model. *Pattern Recognition*, pp. 114012, 2026.

Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 6278–6287, 2022.

Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*, pp. 358–376. Springer, 2020.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536. IEEE, 2024.

Xianlin Zeng, Yalong Jiang, Yufeng Wang, Qiang Fu, and Wenrui Ding. Progressive prediction: Video anomaly detection via multi-grained prediction. *IET Image Processing*, 18(10):2568–2583, 2024.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16271–16280. IEEE, 2023.

Hanwen Zhang, Congqi Cao, Qinyi Lv, Lingtong Min, and Yanning Zhang. Autoregressive denoising score matching is a good video anomaly detector. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12057–12067. IEEE, 2025.

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17385–17394. IEEE, 2024.

Menghao Zhang, Yiyan Zhu, Pengfei Ren, Haifeng Sun, Qi Qi, Zirui Zhuang, Huazheng Wang, Lei Zhang, Jianxin Liao, and Jingyu Wang. Fine-vad: Towards fine-grained video anomaly detection via progressive cross-granularity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35514–35523, 2026.

Yaxin Zhao, Yang Wang, Wenya Guo, Sihan Xu, Xiangrui Cai, Xi Lin, Ying Zhang, and Xiaojie Yuan. Learning from noisy supervision: A denoising-debiasing framework for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21326–21335, 2026.

Hantao Zheng, Ning Han, Yawen Zeng, Hegui Zhu, and Hao Chen. Logical consistency optimization for few-shot weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9045–9054, 2026.

Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1237–1246. IEEE, 2019.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3769–3777, 2023.

Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10761–10769, 2025.

Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Heng Tao Shen. Batchnorm-based weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13642–13654, 2024.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

A PROOFS FOR AUROC ATTRIBUTION

Tie-aware pair contribution. For scores x and y, define

$$\tau(x, y) = \mathbf{1}[x > y] + \tfrac{1}{2}\mathbf{1}[x = y].$$

Let P_i and N_i be the positive- and negative-frame index sets of video i. Micro-AUROC is

$$\text{Micro-AUROC} = \frac{1}{AN} \sum_i \sum_{t \in P_i} \sum_j \sum_{u \in N_j} \tau(s_{it}, s_{ju}).$$

Let $\mathcal{M} = \{i : a_i > 0 \text{ and } n_i > 0\}$ denote the mixed-label videos. Videos outside $\mathcal{M}$ have no same-video positive–negative pairs, but their frames remain in the cross-video terms above.

Proof of Proposition 1. Partition the outer sum into terms with $i = j$ and terms with $i \neq j$. Define

$$S_{\text{within}} = \sum_{i \in \mathcal{M}} \sum_{t \in P_i} \sum_{u \in N_i} \tau(s_{it}, s_{iu}), \qquad S_{\text{cross}} = \sum_{i \neq j} \sum_{t \in P_i} \sum_{u \in N_j} \tau(s_{it}, s_{ju}).$$

Under the conditions of Proposition 1, the corresponding pair counts are $W = \sum_{i \in \mathcal{M}} a_i n_i > 0$ and $AN - W > 0$. Therefore

$$\text{Within-AUROC} = \frac{S_{\text{within}}}{W}, \qquad \text{Cross-AUROC} = \frac{S_{\text{cross}}}{AN - W},$$

and

$$\text{Micro-AUROC} = \frac{S_{\text{within}} + S_{\text{cross}}}{AN} = \frac{W}{AN}\text{Within-AUROC} + \frac{AN - W}{AN}\text{Cross-AUROC}.$$

This is Equation 2 with $w = W/(AN)$. At the excluded boundaries, one normalized component is undefined. If $W = 0$, Micro-AUROC reduces to Cross-AUROC. If $AN - W = 0$, it reduces to Within-AUROC.

Proof of Equation 5 and Corollary 1. Since p^+ and p^- are probability vectors,

$$w = \sum_i p_i^+ p_i^- \leq \|p^+\|_\infty \sum_i p_i^- = \|p^+\|_\infty.$$

Exchanging the two classes gives $w \leq \|p^-\|_\infty$, which proves Equation 5. If no video captures more than C_+/V of the positive frames or more than C_-/V of the negative frames, then $w \leq \min(C_+, C_-)/V$.

For the exact replication statement, make k disjoint copies of every video. Each copy contributes $a_i n_i$ same-video pairs, so the numerator of w becomes $k \sum_i a_i n_i$. The total positive and negative frame counts each grow by k, so the denominator becomes $k^2 AN$. Hence the new weight is w/k.

Finally, with Cross-AUROC fixed, changing $\mathrm{Within\text{-}AUROC}$ from 0.5 to 1 changes Micro-AUROC by $w(1 - 0.5) = w/2$. This proves Corollary 1.

Proof of Proposition 2. Assigning one anomaly score to every frame in a video ties all within-video anomalous–normal pairs, so $\mathrm{Within\text{-}AUROC} = 0.5$. Consider two distinct videos i and j. If $c_i > c_j$, the ordering correctly ranks the $a_i n_j$ pairs whose positive frame is in i and negative frame is in j. Reversing the order instead correctly ranks $a_j n_i$ pairs. Therefore placing i above j is weakly preferable exactly when

$$a_i n_j \geq a_j n_i.$$

For nonempty videos, this condition is equivalent to

$$\frac{a_i}{a_i + n_i} \geq \frac{a_j}{a_j + n_j}.$$

The pairwise preferences are thus induced by one scalar and are transitive. Sorting videos by anomaly fraction maximizes every pairwise exchange and therefore the total number of concordant cross-video pairs. Equal fractions can be tied or ordered arbitrarily without changing AUROC.

B Benchmark composition across VAD settings

The decomposition in Equation (2) depends on frame labels and video boundaries, not on the supervision used to train a detector. Table 6 gives the numerical benchmark statistics for Figure 2 and records the common ShanghaiTech subset used in Appendix F.3. The table covers 13 benchmarks and distinguishes their test splits. ShanghaiTech has separate 199-video weakly-supervised and 107-video semi-supervised splits. MSAD Protocols II and I contain 240 and 360 test videos, respectively.

Table 6: Test-set composition across VAD settings. V_+ counts videos containing an anomaly. w is the within-video pair fraction. Max. gain is the Micro-AUROC increase in percentage points from improving Within-AUROC from 50% to 100% while holding Cross-AUROC fixed. The final column gives the maximum Micro-AUROC (%) with one anomaly score per video.

Benchmark	V	V_+	Frames	w	Max. gain (pp)	Maximum Micro-AUROC
Weakly-supervised test splits						
ShanghaiTech	199	44	142,912	0.140%	0.070	98.35
UCF-Crime	290	140	1,114,144	0.388%	0.194	92.95
XD-Violence	800	500	2,330,384	0.071%	0.036	95.46
MSAD (Protocol II)	240	120	146,410	0.206%	0.103	96.90
LAD2000	560	280	628,569	0.176%	0.088	95.42
NTU CCTV-Fights	250	250	438,345	0.526%	0.263	78.92
VIEW360	200	105	153,641	0.401%	0.200	88.21
Semi-supervised test splits						
ShanghaiTech	107	107	40,791	0.814%	0.407	75.78
Avenue	21	21	15,324	5.079%	2.540	74.81
UCSD Ped1	36	36	7,200	2.267%	1.134	74.58
UCSD Ped2	12	12	2,010	6.802%	3.401	83.96
UBnormal	211	158	92,640	0.190%	0.095	93.86
MSAD (Protocol I)	360	240	227,524	0.248%	0.124	94.07
Street Scene	35	35	146,410	1.631%	0.816	88.42
IITB Corridor	150	140	181,567	0.517%	0.259	86.90
Common test subset						
ShanghaiTech (common44)	44	44	17,472	1.891%	0.945	78.21

Within-video pairs account for 6.802% of comparisons on UCSD Ped2 and 5.079% on Avenue, compared with 0.190% on UBnormal and 0.071% on XD-Violence. Cross-video pairs therefore account for most comparisons in each listed test split, although the within-video fraction varies substantially. These differences also affect the maximum Micro-AUROC with one anomaly score per video. On ShanghaiTech, this maximum is 98.35% on the weakly-supervised split and 75.78% on the semi-supervised split. Thus, the same metric structure need not permit equally high scores under different test-set compositions.

The table uses the full annotations for each listed protocol. Detector results use the frames for which their anomaly scores are available. The corresponding frame coverage is described in Appendix C and the notes accompanying the supplementary results.

C Experimental setup

C.1 Methods and training protocol

The main WSVAD experiments evaluate five methods including Sultani, RTFM, CLIP-TSA, BN-WVAD, and S3R (Sultani et al., 2018; Tian et al., 2021; Joo et al., 2023; Zhou et al., 2024; Wu et al., 2022b) with CLIP and I3D features (Radford et al., 2021; Carreira & Zisserman, 2017) on ShanghaiTech, UCF-Crime, XD-Violence, and MSAD (Liu et al., 2018; Sultani et al., 2018; Wu et al., 2020; Zhu et al., 2024). We use the official implementation of each method and follow the hyperparameters reported in the corresponding paper. When a setting is not specified in the paper, we use the default provided by the official implementation. Each method, feature, and benchmark combination is trained with three random seeds, giving 30 runs per benchmark and 120 detector runs in total. Performance tables report the mean over the three runs, while retention is computed for each run before aggregation.

Detector metrics are computed on the frames for which anomaly scores are available and use the corresponding official frame annotations. Ground-truth-only analyses instead use the full benchmark annotations. For MSAD, this gives 143,920 evaluated frames for detector results, compared with 146,410 frames in the full protocol. The video-level anomaly classifiers are evaluated with both CLIP and I3D features on all four benchmarks, with three runs for each configuration.

Mean-pool classifier training. The mean-pool classifier is trained on the standard training pool with video labels only. The training configuration uses BCE, Adam with learning rate 10^{-3}, batch size 128, 2,000 updates, and final-iterate reporting. There is no frame-label input, validation split, checkpoint selection, or test-time fitting. Pooling is performed before the head, so the model receives one vector per video and cannot recover segment order.

Tie convention. All AUROCs use mid-ranks, so a tied positive–negative pair contributes 0.5 (Song & Lee, 2026c). An anomaly score that is constant within each video gives every mixed-label video's per-video AUROC a value of 0.5, so Macro-AUROC and Within-AUROC are both 0.5. Pure-label videos have no defined per-video AUROC, so Macro-AUROC excludes them. They also contribute no pair to Within-AUROC. This equality provides a deterministic validation check. We also verify that all frames in a video receive the same anomaly score after video-mean replacement.

Released outputs. Released denotes author-provided checkpoints or scores, including checkpoints for which we run inference. Metrics use the recorded features and official GT. Feature substitutions are identified explicitly: MGFN on UCF-Crime uses released weights with substitute features and is not the original feature configuration. MSAD RTFM and MGFN use released checkpoints with official I3D features.

D Additional video-level anomaly classification results

One-bit video-composition control. This analytic control assigns one to every frame of a video containing any anomaly and zero to every frame of an all-normal video. If N_{-}^{anom} is the number of negative frames inside anomaly-containing videos and N_{-} is the total negative-frame count, every

Table 7: Ground-truth-based controls with one anomaly score per video. One-bit uses the binary video label, and the optimum orders videos by their fraction of anomalous frames. MSAD uses the full benchmark annotations here. On the frames used for detector evaluation, its two values are 89.57 and 96.90.

Benchmark	Videos	Frames	One-bit AUROC	Constant optimum
ShanghaiTech	199	142912	96.32	98.35
XD-Violence	800	2330384	84.93	95.46
UCF-Crime	290	1114144	81.57	92.95
MSAD	240	146410	89.53	96.90

Table 8: Frame-level AP (%) of video-level anomaly classifiers. Entries are means with sample standard deviations over three training runs.

Benchmark	Features	Mean-pool	Seq.-linear	Bi-GRU	Transformer
ShanghaiTech	CLIP	50.39±0.01	45.61±0.07	53.94±0.45	54.45±2.18
ShanghaiTech	I3D	42.26±0.17	36.62±0.10	41.12±0.81	34.80±2.19
XD-Violence	CLIP	65.36±0.09	70.73±0.04	69.76±0.47	57.03±6.44
XD-Violence	I3D	67.24±0.04	65.92±0.48	69.68±2.23	56.47±7.29
UCF-Crime	CLIP	17.50±0.06	20.23±0.11	19.48±1.12	20.35±1.83
UCF-Crime	I3D	19.59±0.25	21.19±0.14	17.77±0.08	19.18±1.98
MSAD	CLIP	63.34±0.05	65.31±0.22	67.46±0.99	63.78±1.46
MSAD	I3D	64.31±0.15	65.81±0.16	62.78±1.19	55.11±2.74

positive has score one and

$$\mathrm{AUROC}_{\mathrm{binary}} = 1 - \frac{1}{2}\frac{N_{-}^{\mathrm{anom}}}{N_{-}}.$$

This gives 96.32%, 84.93%, 81.57%, and 89.53% Micro-AUROC on ShanghaiTech, XD-Violence, UCF-Crime, and MSAD protocol GT, respectively (Table 7). It uses test video labels as an analytic composition control, not as a trainable method.

The mean-pool classifier averages features before classification. The other three classifiers process the feature sequence before predicting one anomaly score for the video. The sequence-linear head maps a fixed 32-position sequence to one logit. The bi-GRU and Transformer encode that sequence and pool to one logit. These classifiers can use the temporal order of the features, but each assigns its predicted video-level anomaly score to every frame for evaluation.

Table 2 reports the Micro-AUROC of the video-level anomaly classifiers. Table 8 reports standard AP on all four benchmarks in a separate table; AP remains XD-Violence's official metric and does not admit the AUROC pair decomposition.

E Detailed detector results

This section reports the numerical results underlying Section 5. The 120 detector training runs and the 13 results from released checkpoints or scores remain separate throughout.

E.1 Micro-AUROC decomposition

Table 9 gives Micro-AUROC, Cross-AUROC, Within-AUROC, and Macro-AUROC for the original predictions, together with Micro-AUROC after video-mean replacement. It includes all 40 combinations of method, features, and benchmark from the training experiments. Semi-supervised and training-free AUROC results are reported in Appendix F and are not included in these WSVAD retention summaries.

Table 9: Detector AUROC results (%). Mean Micro is Micro-AUROC after video-mean replacement. † denotes the mean over three training runs. Training-free AUROC results are in Table 14.

Benchmark	Method	Features	Micro	Mean Micro	Cross	Within	Macro
Three-run averages							
ShanghaiTech	Sultani†	CLIP	97.24	96.87	97.29	66.24	66.33
ShanghaiTech	Sultani†	I3D	85.00	86.16	85.01	78.43	77.49
ShanghaiTech	RTFM†	CLIP	96.05	96.32	96.13	44.99	44.34
ShanghaiTech	RTFM†	I3D	93.14	94.01	93.17	75.45	75.07
ShanghaiTech	CLIP-TSA†	CLIP	97.74	95.93	97.76	82.59	81.85
ShanghaiTech	CLIP-TSA†	I3D	86.76	88.98	86.77	80.06	79.45
ShanghaiTech	BN-WVAD†	CLIP	96.61	96.27	96.66	61.37	61.82
ShanghaiTech	BN-WVAD†	I3D	92.74	92.27	92.77	70.90	72.28
ShanghaiTech	S3R†	CLIP	97.99	95.65	98.02	77.21	80.02
ShanghaiTech	S3R†	I3D	94.26	93.41	94.27	80.92	80.11
UCF-Crime	Sultani†	CLIP	82.98	80.08	83.04	68.11	66.47
UCF-Crime	Sultani†	I3D	80.08	78.89	80.17	57.30	67.46
UCF-Crime	RTFM†	CLIP	79.91	75.63	79.97	64.33	64.84
UCF-Crime	RTFM†	I3D	81.39	79.48	81.48	58.83	65.77
UCF-Crime	CLIP-TSA†	CLIP	84.03	80.77	84.10	65.60	68.25
UCF-Crime	CLIP-TSA†	I3D	70.58	71.90	70.67	48.60	54.49
UCF-Crime	BN-WVAD†	CLIP	82.84	80.04	82.89	70.45	68.69
UCF-Crime	BN-WVAD†	I3D	80.19	78.16	80.27	59.51	67.00
UCF-Crime	S3R†	CLIP	80.84	78.73	80.89	67.23	68.55
UCF-Crime	S3R†	I3D	79.73	76.95	79.82	56.95	63.90
XD-Violence	Sultani†	CLIP	89.72	90.68	89.73	78.58	73.02
XD-Violence	Sultani†	I3D	87.58	89.79	87.58	79.11	74.70
XD-Violence	RTFM†	CLIP	90.21	89.86	90.22	76.97	73.40
XD-Violence	RTFM†	I3D	88.56	89.65	88.57	79.33	76.74
XD-Violence	CLIP-TSA†	CLIP	89.59	89.46	89.60	74.15	73.35
XD-Violence	CLIP-TSA†	I3D	88.82	88.83	88.82	79.33	77.92
XD-Violence	BN-WVAD†	CLIP	90.60	88.84	90.61	82.84	75.49
XD-Violence	BN-WVAD†	I3D	88.65	86.99	88.65	79.52	74.43
XD-Violence	S3R†	CLIP	91.34	89.74	91.34	82.61	76.77
XD-Violence	S3R†	I3D	87.15	89.02	87.17	58.69	59.78
MSAD	Sultani†	CLIP	90.79	91.07	90.86	59.46	63.57
MSAD	Sultani†	I3D	86.05	86.70	86.08	68.19	73.28
MSAD	RTFM†	CLIP	87.30	87.03	87.36	59.76	65.51
MSAD	RTFM†	I3D	84.45	84.30	84.48	69.56	68.22
MSAD	CLIP-TSA†	CLIP	88.99	89.66	89.05	57.54	56.78
MSAD	CLIP-TSA†	I3D	79.50	84.47	79.54	63.24	58.14
MSAD	BN-WVAD†	CLIP	87.65	86.94	87.71	59.41	63.81
MSAD	BN-WVAD†	I3D	87.42	86.30	87.46	72.34	80.03
MSAD	S3R†	CLIP	86.75	86.36	86.81	59.28	69.20
MSAD	S3R†	I3D	83.19	82.71	83.21	71.70	71.83
Released checkpoints and scores							
ShanghaiTech	S3R	I3D	97.39	95.75	97.42	82.38	82.29
UCF-Crime	S3R	I3D	85.99	82.20	86.08	62.97	71.40
UCF-Crime	DSANet	CLIP	89.44	85.34	89.50	73.72	70.01
UCF-Crime	MGFN	I3D	83.44	80.54	83.51	63.03	64.36
UCF-Crime	UR-DMU	I3D	86.96	83.14	87.03	68.30	74.17
UCF-Crime	VadCLIP	CLIP	88.01	84.44	88.07	72.93	66.62
XD-Violence	BN-WVAD	I3D	94.71	92.24	94.72	86.36	78.54
XD-Violence	DSANet	CLIP	95.40	92.29	95.41	85.48	79.85
XD-Violence	MGFN	I3D	92.88	90.85	92.89	81.84	75.43

Continued on the next page

Table 9 (continued)

Benchmark	Method	Features	Micro	Mean Micro	Cross	Within	Macro
XD-Violence	UR-DMU	I3D	94.02	91.67	94.02	85.44	77.91
XD-Violence	VadCLIP	CLIP	94.68	90.84	94.69	84.26	75.29
MSAD	RTFM	I3D	86.88	85.83	86.91	73.36	75.98
MSAD	MGFN	I3D	84.96	80.88	84.98	79.14	79.18

E.2 VIDEO-MEAN REPLACEMENT

Table 10 extends the run-level summary in Table 3 to the released results. Retention is computed before any averaging across runs. The median is 98.8% over the 120 training runs and 91.4% over the 13 released results. The minimum values are 78.1% and 88.3%, respectively.

Table 10: Micro-AUROC retention after video-mean replacement. Each training run contributes one R_{mean} value as defined in Equation (8). Results from released checkpoints and scores are summarized separately. The final two columns count results retaining at least 85% and 90%, respectively.

Benchmark	n	Median R_{mean}	Min. R_{mean}	$R_{\text{mean}} \geq 85\%$	$R_{\text{mean}} \geq 90\%$
Detector training runs					
ShanghaiTech	30	99.3%	95.0%	30/30	30/30
UCF-Crime	30	92.6%	78.1%	29/30	26/30
XD-Violence	30	100.0%	94.3%	30/30	30/30
MSAD	30	100.1%	95.5%	30/30	30/30
All	120	98.8%	78.1%	119/120	116/120
Released checkpoints and scores					
ShanghaiTech	1	96.5%	96.5%	1/1	1/1
UCF-Crime	5	89.7%	89.5%	5/5	2/5
XD-Violence	5	94.5%	91.4%	5/5	5/5
MSAD	2	92.7%	88.3%	2/2	1/2
All	13	91.4%	88.3%	13/13	9/13

E.3 FRAME-LEVEL AP

For frame-level AP, we subtract the anomalous-frame fraction rather than 50% when calculating retention. Let π denote this fraction and let s and $\tilde{s}$ denote the original and replaced anomaly scores. On the $[0, 1]$ scale, AP retention is

$$\frac{\mathrm{AP}(\tilde{s}) - \pi}{\mathrm{AP}(s) - \pi}.$$

Table 11 summarizes this ratio when the original AP exceeds π. One UCF-Crime training run does not satisfy this condition, so it remains in the run count but is excluded from the ratio summary.

Across the training runs, median AP retention is 80.8% on ShanghaiTech, 72.1% on UCF-Crime, 94.6% on XD-Violence, and 93.0% on MSAD. These results show that the response of frame-level AP depends on the benchmark. They are empirical comparisons rather than consequences of the Micro-AUROC decomposition.

Table 11: Frame-level AP retention after video-mean replacement. n counts evaluated results, and Defined n counts those whose original AP exceeds the anomalous-frame fraction. The final three columns count increases in AP and retention of at least 85% and 90%, respectively.

Benchmark	Source	n	Defined n	Median	Minimum	AP increases	Retention $\geq 85\%$	Retention $\geq 90\%$
ShanghaiTech	Training runs	30	30	80.8%	54.9%	3/30	11/30	9/30
ShanghaiTech	Released	1	1	67.3%	67.3%	0/1	0/1	0/1
UCF-Crime	Training runs	30	29	72.1%	50.0%	1/30	5/29	2/29
UCF-Crime	Released	5	5	63.9%	54.2%	0/5	0/5	0/5
XD-Violence	Training runs	30	30	94.6%	82.6%	6/30	25/30	22/30
XD-Violence	Released	5	5	85.6%	81.9%	0/5	4/5	0/5
MSAD	Training runs	30	30	93.0%	82.5%	3/30	26/30	21/30
MSAD	Released	2	2	87.9%	85.2%	0/2	2/2	1/2

Table 12 reports the original and replaced AP values for every WSVAD method and feature configuration. The training-run rows average AP over three runs. The retention summaries in Table 11 instead use the ratio calculated separately for each run.

Table 12: Frame-level AP (%) before and after video-mean replacement. † denotes the mean over three training runs. These are the same WSVAD predictions as in Table 9.

Benchmark	Method	Features	Original AP	Video mean AP
Three-run detector averages				
ShanghaiTech	Sultani†	CLIP	55.83	50.69
ShanghaiTech	Sultani†	I3D	43.07	40.23
ShanghaiTech	RTFM†	CLIP	45.48	48.24
ShanghaiTech	RTFM†	I3D	47.34	39.38
ShanghaiTech	CLIP-TSA†	CLIP	63.24	42.55
ShanghaiTech	CLIP-TSA†	I3D	49.74	38.84
ShanghaiTech	BN-WVAD†	CLIP	49.03	45.82
ShanghaiTech	BN-WVAD†	I3D	51.42	44.40
ShanghaiTech	S3R†	CLIP	65.61	39.18
ShanghaiTech	S3R†	I3D	54.41	37.30
UCF-Crime	Sultani†	CLIP	22.88	17.99
UCF-Crime	Sultani†	I3D	20.88	17.00
UCF-Crime	RTFM†	CLIP	21.19	15.82
UCF-Crime	RTFM†	I3D	21.66	17.70
UCF-Crime	CLIP-TSA†	CLIP	24.79	18.80
UCF-Crime	CLIP-TSA†	I3D	15.07	14.17
UCF-Crime	BN-WVAD†	CLIP	22.12	18.80
UCF-Crime	BN-WVAD†	I3D	21.03	19.26
UCF-Crime	S3R†	CLIP	20.15	17.17
UCF-Crime	S3R†	I3D	24.36	16.85
XD-Violence	Sultani†	CLIP	72.47	69.76
XD-Violence	Sultani†	I3D	68.61	67.68
XD-Violence	RTFM†	CLIP	65.26	67.41
XD-Violence	RTFM†	I3D	67.81	64.11
XD-Violence	CLIP-TSA†	CLIP	64.89	64.10
XD-Violence	CLIP-TSA†	I3D	69.93	62.57
XD-Violence	BN-WVAD†	CLIP	67.86	63.71
XD-Violence	BN-WVAD†	I3D	69.45	62.17
XD-Violence	S3R†	CLIP	66.85	65.70
XD-Violence	S3R†	I3D	54.96	63.58
MSAD	Sultani†	CLIP	70.74	69.02
MSAD	Sultani†	I3D	67.58	63.84
MSAD	RTFM†	CLIP	64.30	62.02

Continued on the next page

Table 12 (continued)

Benchmark	Method	Features	Original AP	Video mean AP
MSAD	RTFM†	I3D	62.98	59.85
MSAD	CLIP-TSA†	CLIP	66.97	66.55
MSAD	CLIP-TSA†	I3D	57.19	58.85
MSAD	BN-WVAD†	CLIP	65.31	61.91
MSAD	BN-WVAD†	I3D	64.89	58.91
MSAD	S3R†	CLIP	62.79	59.16
MSAD	S3R†	I3D	60.79	55.89
Released WSVAD checkpoints and scores				
ShanghaiTech	S3R	I3D	59.10	41.61
UCF-Crime	S3R	I3D	31.50	20.55
UCF-Crime	DSANet	CLIP	37.37	25.57
UCF-Crime	MGFN	I3D	23.54	18.75
UCF-Crime	UR-DMU	I3D	31.21	26.52
UCF-Crime	VadCLIP	CLIP	33.55	24.17
XD-Violence	BN-WVAD	I3D	84.93	76.13
XD-Violence	DSANet	CLIP	86.96	77.43
XD-Violence	MGFN	I3D	80.12	73.22
XD-Violence	UR-DMU	I3D	81.65	73.21
XD-Violence	VadCLIP	CLIP	84.52	73.37
MSAD	RTFM	I3D	66.29	62.24
MSAD	MGFN	I3D	63.54	57.54
Supplementary training-free methods				
XD-Violence	LAVAD	VLM	62.05	60.15
XD-Violence	VADTree	VLM	67.95	61.92

F Additional VAD Results

This appendix reports supplementary semi-supervised and VLM-based VAD results. For each method, we evaluate the original and replaced anomaly scores on the same frames and annotations. These results remain separate from the main WSVAD retention summaries.

F.1 Semi-supervised anomaly detection

Table 13 reports results on Avenue, UCSD Ped2, and the 107-video ShanghaiTech split used for semi-supervised evaluation. The methods include frame-prediction, object, pose, and trajectory-based approaches, with their original feature representations. Accurate-Interpretable-VAD produces identical anomaly scores across its three recorded runs. MNAD-Pred excludes the first four frames of each test video, which slightly changes its within-video pair fraction.

Micro-AUROC remains numerically close to Cross-AUROC in these results. For example, TrajVAD-P on ShanghaiTech has Micro-AUROC of 88.59% and Cross-AUROC of 88.57%, with a within-video pair fraction of 0.814%. Its Within-AUROC is 91.58%. Cross-video comparisons can therefore dominate the pooled metric even when a detector distinguishes anomalous from normal frames well within individual videos.

Video-mean replacement has a different effect from that observed in our main WSVAD experiments. TrajVAD-P's Micro-AUROC decreases from 88.59% to 61.92%. On Avenue, the three-run Jigsaw-VAD result decreases from 90.32% to 65.27%. On UCSD Ped2, Accurate-Interpretable-VAD decreases from 99.06% to 48.51%. These changes show that a small within-video pair fraction does not by itself imply that Micro-AUROC will remain high after replacement. The decomposition fixes the weight of each pair population, whereas replacement also changes the cross-video comparisons.

Table 13: Semi-supervised detector AUROC (%). Mean Micro is Micro-AUROC after video-mean replacement, and w is the within-video pair fraction (%). † denotes three-run means. For these rows, Micro and Mean Micro include sample standard deviations.

Method	w	Micro	Mean Micro	Cross	Within	Macro
ShanghaiTech (107-video test split)						
TrajVAD-P	0.814	88.59	61.92	88.57	91.58	91.45
TrajVAD-T	0.814	84.89	63.98	84.89	84.98	87.00
MoCoDAD	0.814	76.68	58.77	76.66	79.32	77.52
SPARTA	0.814	84.83	56.57	84.80	87.72	89.69
STG-NF	0.814	85.94	56.64	85.92	87.63	89.20
UniVAD	0.814	89.32	61.30	89.30	92.22	91.01
Avenue (21-video test split)						
Accurate-Interpretable-VAD†	5.079	93.33±0.00	63.15±0.00	93.19	95.91	95.85
Jigsaw-VAD†	5.079	90.32±0.37	65.27±1.04	90.21	92.40	88.94
MNAD-Pred†	5.099	86.95±1.45	63.49±1.82	86.81	89.59	87.87
Jigsaw-VAD	5.079	92.10	63.75	92.04	93.19	91.47
UCSD Ped2 (12-video test split)						
Accurate-Interpretable-VAD†	6.802	99.06±0.00	48.51±0.00	99.05	99.14	99.46
Jigsaw-VAD†	6.802	86.41±1.57	61.72±2.78	85.96	92.57	95.33
MNAD-Pred†	6.870	92.99±1.68	61.26±3.61	92.87	94.59	95.66

F.2 VLM-Based Anomaly Detection

Table 14 reports additional VLM-based results. LAVAD and VADTree perform detection without task-specific training. VERA instead learns guiding questions from video-level labels while keeping the VLM weights fixed (Ye et al., 2025). We also evaluate two AnomalyRuler configurations. AnomalyRuler-base applies keyword rules and exponential averaging to frame descriptions. AnomalyRuler-Full adds Mistral deduction and uses the implementation's keyword fallback.

Table 14: Additional VLM-based detector AUROC (%) on the evaluated test splits. Mean Micro is Micro-AUROC after video-mean replacement, and w is the within-video pair fraction (%). Each row evaluates one set of anomaly scores.

Method	w	Micro	Mean Micro	Cross	Within	Macro
UCF-Crime						
LAVAD	0.388	80.28	79.66	80.36	61.75	57.85
VADTree	0.388	84.74	82.10	84.78	74.17	63.70
VERA	0.388	85.58	80.82	85.62	73.46	70.27
XD-Violence						
LAVAD	0.071	85.40	85.32	85.42	64.78	59.94
VADTree	0.071	90.49	88.62	90.50	80.80	73.94
VERA	0.071	85.50	87.39	85.51	71.20	64.28
ShanghaiTech (107-video test split)						
AnomalyRuler-base	0.814	84.53	63.86	84.53	84.10	84.80
AnomalyRuler-Full	0.814	84.69	62.96	84.67	87.30	87.38
Avenue						
AnomalyRuler-base	5.079	70.85	50.12	70.53	76.68	78.07
AnomalyRuler-Full	5.079	61.96	44.45	61.82	64.54	72.12
UCSD Ped2						
AnomalyRuler-base	6.802	92.54	71.39	92.42	94.24	94.61
AnomalyRuler-Full	6.802	94.61	75.34	94.61	94.70	95.10

On XD-Violence, LAVAD changes from 85.40% to 85.32% Micro-AUROC after replacement, and VADTree changes from 90.49% to 88.62%. Thus, high Micro-AUROC after assigning one anomaly score to every frame of a video also occurs in training-free results. The effect is not uniform across methods and test splits. For example, AnomalyRuler-Full on Avenue decreases from 61.96% to 44.45%. The same decomposition applies throughout, but it does not predict a common retention level.

F.3 Evaluation on common test videos

The weakly-supervised and semi-supervised ShanghaiTech protocols use different test sets. Table 15 compares predictions on a shared set of 44 videos and 17,472 frames, using the same frame annotations. All 44 videos contain an anomaly, and within-video pairs account for 1.891% of anomalous–normal comparisons. This fixes the evaluation frames, but does not hold the models or their training procedures constant.

Table 15: AUROC (%) on the same 44 ShanghaiTech videos and 17,472 frames. All rows have $w = 1.891\%$. Mean Micro is Micro-AUROC after video-mean replacement. † denotes the mean over three runs.

Method	Features	Micro	Mean Micro	Cross	Within	Macro
Weakly-supervised						
BN-WVAD†	CLIP	54.09	49.55	53.95	61.03	61.42
BN-WVAD†	I3D	56.84	52.31	56.58	70.30	71.35
CLIP-TSA†	CLIP	68.58	45.22	68.32	82.13	81.37
CLIP-TSA†	I3D	61.07	44.68	60.76	76.96	74.24
RTFM†	CLIP	46.85	50.05	46.88	45.11	44.16
RTFM†	I3D	56.96	41.34	56.61	74.97	74.45
Sultani†	CLIP	63.31	57.62	63.26	65.84	66.03
Sultani†	I3D	69.00	52.62	68.83	77.90	75.96
S3R	I3D	68.22	44.70	67.96	81.80	81.98
Semi-supervised						
TrajVAD-P	Trajectory + pose	89.66	63.16	89.59	93.17	93.35
TrajVAD-T	Trajectory	85.47	65.82	85.48	84.92	87.84
MoCoDAD	Pose	79.50	59.71	79.41	83.80	83.01
SPARTA	Pose	84.24	58.13	84.17	87.60	88.34
STG-NF	Pose	85.07	56.36	85.00	88.38	89.01
UniVAD	CLIP + pose	89.25	62.65	89.20	91.63	90.43
Training-free						
AnomalyRuler-base	Descriptions	84.16	65.65	84.12	85.92	85.83
AnomalyRuler-Full	Descriptions + Mistral	84.81	64.80	84.75	87.77	86.97

On the shared test set, CLIP-TSA with CLIP features has Micro-AUROC of 68.58% and Cross-AUROC of 68.32%. Its Within-AUROC is 82.13%. TrajVAD-P has Micro-AUROC of 89.66%, Cross-AUROC of 89.59%, and Within-AUROC of 93.17%. Video-mean replacement reduces their Micro-AUROC to 45.22% and 63.16%, respectively. The dominance of cross-video comparisons persists on the shared frames, while the change under replacement depends on the predictions being evaluated. The shared test subset is a separate evaluation and is not included in the 120-run WSVAD retention summary.

G Within-video weighting and anomaly-only evaluation

Equation 2 requires pair weights $a_i n_i$. Let $\alpha_i = a_i n_i / W$ for $i \in \mathcal{M}$ and define

$$N_{\text{eff}} = \frac{1}{\sum_{i \in \mathcal{M}} \alpha_i^2}. \tag{9}$$

Table 16: Within-pair concentration on protocol GT. V_{mix} counts mixed-label videos and $N_{\text{eff}} = (\sum_i a_i n_i)^2 / \sum_i (a_i n_i)^2$ is a concentration statistic, not an independent sample size. MSAD uses protocol GT (97 mixed videos); its detector support has 96 mixed videos and $N_{\text{eff}} = 8.17$.

Benchmark	V_{mix}	N_{eff}	$N_{\text{eff}}/V_{\text{mix}}$
ShanghaiTech	43	28.3	65.9%
XD-Violence	475	43.7	9.2%
UCF-Crime	140	18.7	13.3%
MSAD	97	8.3	8.6%

Table 17: Anomaly-only pooling under benchmark protocol GT. Cross mass kept is the fraction of original cross-video pairs surviving removal of all-normal videos. The final column gives the maximum AUC_A with one anomaly score per video. MSAD uses the 146,410-frame protocol axis, not the truncated detector support.

Benchmark	Videos kept	Within-video fraction: Micro → AUC_A	Cross mass kept	Constant optimum
ShanghaiTech	44/199	0.14% → 1.90%	7.2%	77.68
XD-Violence	500/800	0.07% → 0.24%	30.1%	84.92
UCF-Crime	140/290	0.39% → 1.05%	36.6%	80.87
MSAD	120/240	0.21% → 0.98%	20.8%	85.20

This inverse-concentration count is the number of equally contributing videos that would produce the same pair-mass concentration. It is not a count of statistically independent videos and does not estimate the variance of Within-AUROC.

Macro-AUROC gives equal weight to each video containing anomalous and normal frames. Within-AUROC instead weights each video by $a_i n_i$, as required by Equation (2). Across the 120 detector training runs, the absolute difference between the two metrics averages 3.76 percentage points and reaches 12.73 percentage points. Table 9 reports both metrics for each method and feature combination.

G.1 Anomaly-only pooling

Removing all-normal videos changes the set of frame comparisons but does not restrict the remaining pairs to the same video. Table 17 reports the resulting pair fractions and the maximum AUC_A when every video receives one anomaly score.

H Temporal Interval Evaluation

Temporal detection evaluates the overlap between predicted anomaly intervals and annotated events. We report binary anomaly-event AP for scalar detector scores and video-level classifiers, and class-wise mAP for VadCLIP and DSANet. The primary evaluation includes all test videos. We also report evaluations restricted to anomaly-containing videos. At each temporal intersection-over-union (tIoU) threshold, binary AP evaluates one anomaly class, whereas class-wise mAP averages the AP values of the retained classes. Average AP and average mAP refer to the mean over tIoU thresholds $0.1, 0.2, 0.3, 0.4$, and 0.5. Metrics are calculated for each run before averaging across runs.

Evaluation population and class-wise reproduction. VadCLIP describes its fine-grained evaluation as using only abnormal videos (Wu et al., 2024c). The released-code evaluation that includes all test videos instead reproduces its reported average mAP of 6.68% on UCF-Crime and 24.70% on XD-Violence. Our corresponding values are 6.6806% and 24.6953% (Table 21). Restricting the evaluation to anomaly-containing videos gives 6.9226% and 18.1577%, respectively. We therefore use all test videos for the primary binary evaluation and keep the restricted population as a separate comparison.

Table 18: **Interval extraction and whole-video predictions give different results.** Binary anomaly-event AP (%) on all test videos. The first two rows in each block use interval extraction. Whole-video rows submit one interval per video. Detector results use released checkpoints, and mean-pool results average three runs with the indicated features. Average AP uses tIoU thresholds $0.1, 0.2, 0.3, 0.4$, and 0.5.

Prediction	Anomaly-event AP	
	Average	tIoU 0.5
MSAD, RTFM, I3D features		
Original scores	46.64	21.82
Video-mean replacement	0.00	0.00
Whole video, detector mean	61.12	35.69
Whole video, mean-pool	63.04	36.33
XD-Violence, VadCLIP, CLIP features		
Original scores	19.19	8.85
Video-mean replacement	0.00	0.00
Whole video, detector mean	14.73	6.69
Whole video, mean-pool	12.49	5.49

Original scores and video-mean replacement. The binary comparison uses a scalar adaptation of VadCLIP's released interval evaluation. We evaluate the intervals obtained from the original frame scores and repeat interval extraction after replacing each video's scores by their mean. The predictions and their labels remain aligned to the same recorded video boundaries in both conditions. All 120 WSVAD training runs and 13 released WSVAD results yield zero intervals and zero AP after replacement.

The released VadCLIP proposal rule retains consecutive frames whose scores strictly exceed a threshold derived from their within-video range. For a constant score sequence, the threshold equals that constant, so no frame is selected. The zero-proposal outcomes in the binary runs agree with this behavior. Class-wise replacement independently averages each output class over time, and also yields zero proposals in all VadCLIP and DSANet runs.

Explicit whole-video predictions. We separately submit one interval $[0, T_i)$ for every evaluated video, where T_i is its evaluated length. For a detector, the interval confidence is its mean anomaly score $\bar{s}_i$. For a video-level classifier, the confidence is its single output. These intervals are supplied directly to the event evaluator. They are not proposals extracted after video-mean replacement. Within each comparison, both interval constructions use the same evaluated frames, labels, and video boundaries. Normal videos have no anomaly event, so anomaly predictions on these videos are false positives.

Representative comparisons. Table 18 separates interval extraction from explicitly submitted whole-video predictions. Video-mean replacement yields no proposals and zero AP. On MSAD, however, whole-video intervals scored by RTFM's video mean achieve 61.12% average AP, compared with 46.64% for intervals extracted from its original scores. The mean-pool classifier achieves 63.04% with the same whole-video interval construction. On XD-Violence, VadCLIP's whole-video intervals instead achieve 14.73%, below 19.19% for extracted intervals. The zero AP after replacement and the positive AP of directly submitted whole-video intervals reflect different interval constructions.

WSVAD results. Table 19 gives the 40 WSVAD method, feature, and benchmark configurations, averaged over three runs, and the 13 released WSVAD results. LAVAD and VADTree are reported in a separate panel. After video-mean replacement, both reported AP quantities are zero for every row, so the table shows the original-score and explicit whole-video conditions. Whole-video intervals have higher average AP in 29 of 30 MSAD training runs, compared with 7 of 30 on ShanghaiTech, 6 of 30 on UCF-Crime, and 8 of 30 on XD-Violence. The comparison therefore depends on the detector and benchmark.

Table 19: Binary anomaly-event AP (%) on all test videos. Original denotes interval extraction from the original scalar scores. Whole video denotes one interval per video scored by the detector mean. Video-mean replacement followed by extraction gives zero AP throughout. † marks three-run means.

			Average AP		AP@0.5	
Benchmark	Method	Features	Original	Whole video	Original	Whole video
Three-run detector means						
ShanghaiTech	Sultani†	CLIP	31.85	28.04	8.96	4.85
ShanghaiTech	Sultani†	I3D	21.41	23.31	8.07	5.51
ShanghaiTech	RTFM†	CLIP	23.70	27.89	7.51	6.71
ShanghaiTech	RTFM†	I3D	27.79	23.21	8.45	4.26
ShanghaiTech	CLIP-TSA†	CLIP	34.69	24.94	12.88	4.53
ShanghaiTech	CLIP-TSA†	I3D	26.77	22.16	9.90	5.13
ShanghaiTech	BN-WVAD†	CLIP	25.70	26.43	5.98	5.00
ShanghaiTech	BN-WVAD†	I3D	28.72	24.84	10.26	6.27
ShanghaiTech	S3R†	CLIP	38.10	23.34	17.91	4.41
ShanghaiTech	S3R†	I3D	31.18	23.87	12.27	4.08
UCF-Crime	Sultani†	CLIP	11.88	10.35	2.04	0.50
UCF-Crime	Sultani†	I3D	7.92	9.44	1.24	0.45
UCF-Crime	RTFM†	CLIP	13.83	9.79	0.85	0.69
UCF-Crime	RTFM†	I3D	13.99	9.36	2.06	0.71
UCF-Crime	CLIP-TSA†	CLIP	14.88	9.88	3.40	0.71
UCF-Crime	CLIP-TSA†	I3D	8.13	9.60	1.28	0.41
UCF-Crime	BN-WVAD†	CLIP	14.08	10.43	3.24	0.63
UCF-Crime	BN-WVAD†	I3D	9.28	9.81	2.12	0.58
UCF-Crime	S3R†	CLIP	11.15	9.55	2.23	0.54
UCF-Crime	S3R†	I3D	12.75	8.84	1.14	0.67
XD-Violence	Sultani†	CLIP	9.32	13.49	2.37	6.27
XD-Violence	Sultani†	I3D	6.67	13.84	1.19	5.95
XD-Violence	RTFM†	CLIP	19.45	12.66	8.08	5.41
XD-Violence	RTFM†	I3D	24.57	12.79	9.29	5.20
XD-Violence	CLIP-TSA†	CLIP	19.86	11.57	6.89	4.38
XD-Violence	CLIP-TSA†	I3D	18.48	12.38	6.77	4.90
XD-Violence	BN-WVAD†	CLIP	16.80	12.00	6.44	4.87
XD-Violence	BN-WVAD†	I3D	15.55	13.23	5.39	5.15
XD-Violence	S3R†	CLIP	21.70	12.28	7.89	5.10
XD-Violence	S3R†	I3D	12.19	11.94	3.91	4.07
MSAD	Sultani†	CLIP	32.72	67.99	17.96	38.20
MSAD	Sultani†	I3D	24.30	63.85	10.80	36.87
MSAD	RTFM†	CLIP	46.85	62.13	28.22	36.46
MSAD	RTFM†	I3D	42.67	60.53	20.29	35.42
MSAD	CLIP-TSA†	CLIP	40.58	65.92	18.74	36.10
MSAD	CLIP-TSA†	I3D	39.47	60.04	18.76	32.27
MSAD	BN-WVAD†	CLIP	46.10	65.92	24.86	38.00
MSAD	BN-WVAD†	I3D	53.17	62.63	30.24	35.93
MSAD	S3R†	CLIP	54.69	59.86	31.24	33.99
MSAD	S3R†	I3D	44.73	58.48	17.76	33.24
Released WSVAD checkpoints and scores						
ShanghaiTech	S3R	I3D	34.39	25.24	10.85	5.22
UCF-Crime	DSANet	CLIP	19.08	10.98	6.04	0.82
UCF-Crime	MGFN	I3D	13.06	9.77	1.92	1.04
UCF-Crime	S3R	I3D	14.57	9.56	2.48	0.83
UCF-Crime	UR-DMU	I3D	18.72	10.86	6.31	0.79
UCF-Crime	VadCLIP	CLIP	14.55	11.71	2.84	0.73

Continued on the next page

Table 19 (continued)

			Average AP		AP@0.5	
Benchmark	Method	Features	Original	Whole video	Original	Whole video
XD-Violence	BN-WVAD	I3D	17.36	15.42	8.84	6.86
XD-Violence	DSANet	CLIP	21.53	15.24	10.29	7.26
XD-Violence	MGFN	I3D	22.82	14.17	10.58	6.85
XD-Violence	UR-DMU	I3D	26.00	14.13	11.07	6.48
XD-Violence	VadCLIP	CLIP	19.19	14.73	8.85	6.69
MSAD	MGFN	I3D	61.75	56.25	39.99	33.51
MSAD	RTFM	I3D	46.64	61.12	21.82	35.69
Additional training-free results						
UCF-Crime	LAVAD	VLM	3.83	9.84	0.75	1.01
UCF-Crime	VADTree	VLM	18.13	13.43	4.85	1.50
XD-Violence	LAVAD	VLM	4.66	13.25	0.87	4.93
XD-Violence	VADTree	VLM	16.48	13.22	6.68	4.69

Video-level classifiers. Table 20 reports all four classifier architectures with both feature representations. Each value is a mean of three separately evaluated training runs. Every prediction is a whole-video interval, so the classifiers do not predict event boundaries within a video. The table reports average AP and AP@0.5 on all test videos.

Table 20: Binary anomaly-event AP (%) for whole-video intervals scored by video-level classifiers. Each entry averages three runs. The interval construction is identical for every classifier.

Classifier	Features	ShanghaiTech	UCF-Crime	XD-Violence	MSAD
All test videos, Average AP					
Mean-pool	CLIP	28.11	9.99	12.49	66.44
Sequence-linear	CLIP	27.13	11.11	14.17	67.04
Bi-GRU	CLIP	29.17	10.92	13.31	66.40
Transformer	CLIP	29.51	11.08	11.28	66.37
Mean-pool	I3D	23.28	9.43	13.89	63.04
Sequence-linear	I3D	23.37	9.90	13.65	63.87
Bi-GRU	I3D	22.51	9.35	13.16	61.73
Transformer	I3D	21.56	10.29	11.54	61.12
All test videos, AP@0.5					
Mean-pool	CLIP	5.80	0.48	5.49	36.01
Sequence-linear	CLIP	4.85	0.66	6.61	36.77
Bi-GRU	CLIP	5.60	0.62	5.87	37.17
Transformer	CLIP	6.40	0.45	3.86	36.08
Mean-pool	I3D	5.56	0.57	5.90	36.33
Sequence-linear	I3D	4.62	0.64	5.43	37.08
Bi-GRU	I3D	7.21	0.62	4.97	35.74
Transformer	I3D	3.65	0.59	3.52	32.21

Anomaly-containing videos. Restricting the binary evaluation to anomaly-containing videos preserves the contrast between the RTFM and VadCLIP examples. On MSAD, RTFM's whole-video intervals achieve 69.71% average AP, compared with 51.49% for extracted intervals. On XD-Violence, VadCLIP's whole-video intervals achieve 14.78%, below 19.27% for extracted intervals. Video-mean replacement still yields zero AP for every detector. The separate class-wise comparisons are reported in Table 21.

Stricter temporal overlap. The binary runs additionally report AP at tIoU 0.7 and 0.9. For released RTFM on MSAD, the original-score results are 11.71% and 1.99%, compared with 20.82% and 8.68% for whole-video intervals. A whole-video interval has tIoU equal to the length of an

Table 21: Class-wise mAP (%) from released VadCLIP and DSANet checkpoints. Each population retains the classes recorded by the released evaluation, 14 on UCF-Crime and 7 on XD-Violence. Every class-wise video-mean-replacement run produces zero proposals and zero mAP at all reported thresholds.

		All test videos		Anomaly-containing videos	
Benchmark	Method	Average	mAP@0.5	Average	mAP@0.5
UCF-Crime	VadCLIP	6.68	2.93	6.92	3.04
UCF-Crime	DSANet	13.01	8.00	13.17	8.06
XD-Violence	VadCLIP	24.70	14.34	18.16	9.66
XD-Violence	DSANet	28.87	17.89	18.88	9.45

individual annotated event divided by the evaluated video length. It can match at most one event, so short or multiple events constrain this baseline independently of its confidence ordering.

I METRIC REPORTING ACROSS VAD TRAINING SETTINGS

We surveyed 113 peer-reviewed VAD papers published from 2018 through 2026 that report a numerical result for their own method on at least one benchmark in Table 6. The 59 weakly supervised methods use target-domain video-level anomaly labels. This group includes methods with category-level video labels, few-shot supervision, or continual updates. The 45 semi-supervised methods learn normality from normal target videos. They may use pretrained feature extractors or synthetic perturbations of normal data, but do not train on real labeled anomaly clips. The nine training-free methods do not fit VAD detector weights to the target benchmark. Some use normal references, update an inference memory, or select score-combination parameters using benchmark results. These differences in available information matter when comparing performance across groups.

We count a metric when a paper gives a numerical result for its proposed method, including in ablations or analysis. An unspecified frame-level AUC counts as frame-level reporting without assuming that the authors pooled all frames to compute Micro-AUROC. AUC_A requires one score over all test videos containing anomalies, rather than separate scores for anomaly categories. Macro-AUROC requires an explicit average of per-video AUROCs. Event-level mAP requires temporal overlap between a predicted interval and a ground-truth event. For FAR, we count numerical values reported by the authors, while recording differences in evaluation conditions below. Table 22 summarizes these reports.

Table 22: **Frame-level AUROC or AP dominates VAD evaluation.** Values show the percentage of surveyed papers reporting each metric, with paper counts in parentheses. Only papers reporting their own numerical results on a benchmark in Table 6 are counted. AUC_A pools frames from all videos containing anomalies. FAR follows the definition reported in each paper.

Metric	WSVAD ($n = 59$)	Semi-supervised ($n = 45$)	Training-free ($n = 9$)	Overall ($n = 113$)
Frame-level AUROC or AP	98.3% (58)	100% (45)	100% (9)	**99.1**% (112)
AUC_A	18.6% (11)	0% (0)	0% (0)	9.7% (11)
Macro-AUROC	0% (0)	22.2% (10)	0% (0)	8.8% (10)
Event-level mAP at temporal IoU	13.6% (8)	0% (0)	0% (0)	7.1% (8)
Reported FAR	13.6% (8)	0% (0)	0% (0)	7.1% (8)

The weakly supervised papers published from 2018 through 2022 are (Sultani et al., 2018; Zhong et al., 2019; Wu et al., 2020; Zaheer et al., 2020; Tian et al., 2021; Lv et al., 2021; Feng et al., 2021; Li et al., 2022; Wu et al., 2022b;c; Yu et al., 2022; Panariello et al., 2022). The 2023 and 2024 papers are (Zhou et al., 2023; Chen et al., 2022; Lv et al., 2023; Zhang et al., 2023; Joo et al., 2023; Pang et al., 2023; Zhou et al., 2024; Wu et al., 2024c; Chen et al., 2024; Karim et al., 2024; Tan et al., 2024; Ghadiya et al., 2024; Wu et al., 2024b; Jain et al., 2024; Cho et al., 2024; Leng et al.,

2024; Sun et al., 2024; Sun & Gong, 2024; Lee et al., 2023; Yang et al., 2024b; Pu et al., 2024; Wu et al., 2024a). The 2025 and 2026 papers are (Song et al., 2025; Amicantonio et al., 2025; Ye et al., 2025; Qiu et al., 2025; Jin et al., 2025; Li et al., 2025; Meng et al., 2025; Chu et al., 2025; Dalvi et al., 2025; Xu et al., 2025; Qiu & Hou, 2025; He et al., 2025; Xu et al., 2026; Acharya et al., 2026; Zhao et al., 2026; Sun et al., 2026a; Zhang et al., 2026; Yin et al., 2026; Chu et al., 2026; Sun et al., 2026b; Yang et al., 2026; Wang & Zhao, 2026; Senapati et al., 2026; Zheng et al., 2026; Hashimoto et al., 2026). Open-vocabulary, few-shot, and continual evaluations appear in this group (Wu et al., 2024a; Zheng et al., 2026; Hashimoto et al., 2026). They retain video-level supervision but use protocols that differ from conventional binary WSVAD. Song et al. (2025) also use direction labels for an auxiliary task on VIEW360.

The semi-supervised papers published from 2018 through 2022 are (Liu et al., 2018; Morais et al., 2019; Gong et al., 2019; Ionescu et al., 2019; Nguyen & Meunier, 2019; Park et al., 2020; Yu et al., 2020; Lu et al., 2020; Georgescu et al., 2021; Astrid et al., 2021b;a; Luo et al., 2021; Szymanowicz et al., 2021; Wu et al., 2022a; Wang et al., 2022; Ristea et al., 2022; Park et al., 2022; Cho et al., 2022). The 2023 and 2024 papers are (Barbalau et al., 2023; Cao et al., 2023; Yang et al., 2023; Astrid et al., 2023; Yan et al., 2023; Shi et al., 2023; Baradaran & Bergevin, 2023; Wang et al., 2023; Zhang et al., 2024; Micorek et al., 2024; Zeng et al., 2024; Cheng et al., 2024; Rajapaksha et al., 2024; Liu et al., 2024; Jiang & Mao, 2024; Lappas et al., 2024; Rai et al., 2024; Zhu et al., 2024). The 2025 and 2026 papers are (Jiang, 2025; Zhang et al., 2025; Reiss & Hoshen, 2022; Zhou et al., 2025; Hu et al., 2025; Yu et al., 2026; Li et al., 2026b; Mumcu et al., 2025; Song & Lee, 2026a). Macro-AUROC is reported by (Wang et al., 2022; Ristea et al., 2022; Barbalau et al., 2023; Zhang et al., 2024; Micorek et al., 2024; Zeng et al., 2024; Zhu et al., 2024; Jiang, 2025; Reiss & Hoshen, 2022; Zhang et al., 2025). The MSAD study reports Macro-AUROC under cross-view and cross-scenario tests, and Mumcu et al. (2025) report an Avenue result for their method combined with T-EVAL. Several other semi-supervised studies use externally pretrained object or pose models, normal support sets, or synthetic anomalies. Their frame-level AUROC reports still concern normal-only VAD training, but their data and adaptation protocols are not identical.

The training-free papers are (Zanella et al., 2024; Yang et al., 2024a; Ahn et al., 2024; Shao et al., 2025; Li et al., 2026a; Ahn et al., 2026; Feng et al., 2026; Lin et al., 2026; Mo et al., 2026). Normal references are used by Yang et al. (2024a); Ahn et al. (2024). Feng et al. (2026); Mo et al. (2026) update their inference context as a video stream is processed. Ahn et al. (2026); Lin et al. (2026) select score parameters using benchmark results. Thus, training-free here describes detector-weight fitting, not the absence of reference data or score calibration.

Among weakly supervised papers, $\mathrm{AUC_A}$ is reported by (Lv et al., 2021; Zhou et al., 2023; Lv et al., 2023; Zhou et al., 2024; Wu et al., 2024c; Amicantonio et al., 2025; Qiu et al., 2025; Xu et al., 2025; 2026; Acharya et al., 2026; Chu et al., 2026). Temporal event-level mAP appears in (Wu et al., 2022c; 2024c; Li et al., 2025; Zhang et al., 2026; Yin et al., 2026; Chu et al., 2026; Wang & Zhao, 2026; Senapati et al., 2026). Numerical FAR results appear in (Sultani et al., 2018; Zhong et al., 2019; Feng et al., 2021; Wu et al., 2022c; Zhou et al., 2023; Pang et al., 2023; Pu et al., 2024; Xu et al., 2026). Zhou et al. (2023) do not specify the score threshold for their FAR results. Feng et al. (2021) also omit the threshold and describe their FAR as measured on anomaly videos. These reporting differences limit direct comparisons of FAR values.

The metric names need a few qualifications. Zhang et al. (2026) report temporal event mAP but no frame-level AUROC or AP. Shao et al. (2025) report frame-level AUROC and AP, but not temporal event-level mAP on the included benchmarks. The category-specific anomaly-only AUCs of Pang et al. (2023) are not one pooled $\mathrm{AUC_A}$. Most semi-supervised test splits in Table 6 contain no all-normal videos. On those splits, overall pooled AUROC already uses the same test videos as $\mathrm{AUC_A}$, even when the paper does not name it separately.

Frame-level AUROC or AP is reported in 112 of the 113 surveyed papers across weakly supervised, semi-supervised, and training-free VAD. Alternative evaluations are substantially less common. Macro-AUROC appears in 10 papers, all in the semi-supervised setting, while temporal event-level mAP appears in 8 weakly supervised papers. These reporting patterns motivate separating the different questions that frame-level pooled ranking, within-video ranking, and interval-based evaluation answer.