# RETHINKING OPEN-WORLD VIDEO ANOMALY DETECTION: DIAGNOSING DEFINITION BLINDNESS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Open-world video anomaly detection (OWVAD) aims to localize anomalous events in videos according to a user-specified definition. An OWVAD detector should therefore use the supplied definition to decide what is anomalous. We call this behavior **definition following**, and its failure **definition blindness**. However, existing temporal evaluation protocols may obscure this distinction, making strong generic anomaly detection appear to reflect definition following. For example, Drift@5 (Liu et al., 2026b) and the customizable VAD protocol (Ahn et al., 2026) treat both normal frames and anomalies outside the current definition as negatives. Across the three benchmarks used in our study, normal frames account for 78.6%–96.4% of these negatives. As a result, the reported score can be driven largely by target-versus-normal separation rather than target-versus-other-anomaly separation. Consistent with this diagnosis, a LaGoVAD (Liu et al., 2026b) branch that never receives the anomaly definition nearly matches the full model under Drift@5. We therefore introduce three complementary definition-conditioned evaluations that separate definition following from generic anomaly detection. Applying them to VAD systems and general vision–language models on UCF-Crime, XD-Violence, and MSAD reveals that strong performance under existing evaluation can coexist with weak definition following. Motivated by these results, we introduce Definition-Contrastive Scoring (DeCoS), which reduces anomaly evidence shared across competing definitions and improves definition following across the three benchmarks. Our results show that generic anomaly detection and definition following are distinct capabilities and should be evaluated separately in OWVAD.

## 1 INTRODUCTION

Video anomaly detection (VAD) aims to localize anomalous events in videos (Wu et al., 2024b; Song et al., 2025). In conventional VAD, which events are considered anomalous is predefined in the benchmark (Sultani et al., 2018; Wu et al., 2020; Zhu et al., 2024). Open-world video anomaly detection (OWVAD), by contrast, allows users to specify which events should be considered anomalous (Liu et al., 2026b; Ahn et al., 2026). The detector takes this user-specified anomaly definition as input. OWVAD therefore requires more than detecting anomalous events. For the same video, changing the anomaly definition should change which events receive high scores. Figure 1 illustrates this distinction with separate shooting and explosion intervals. In the top row (a), changing the definition from shooting to explosion shifts the highest scores from the shooting interval to the explosion interval. We call this behavior **definition following**. In the bottom row (b), the two definitions produce nearly identical score curves, with both anomalous intervals receiving high scores. We call this behavior **definition blindness**.

However, existing VAD evaluation may not clearly distinguish definition following from definition blindness. Standard VAD metrics operate under a fixed benchmark definition and therefore do not compare anomaly scores across different anomaly definitions. Recently introduced OWVAD evaluation protocols, such as Drift@5 (Liu et al., 2026b) and customizable VAD (C-VAD) evaluation (Ahn et al., 2026), specify which anomaly classes are treated as targets. Yet both compare target anomalies against a negative set that contains normal frames and anomalies excluded by the current definition. A high score can therefore come from separating target anomalies from normal frames, even if the model does not distinguish target anomalies from other anomalies present in the video.



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.

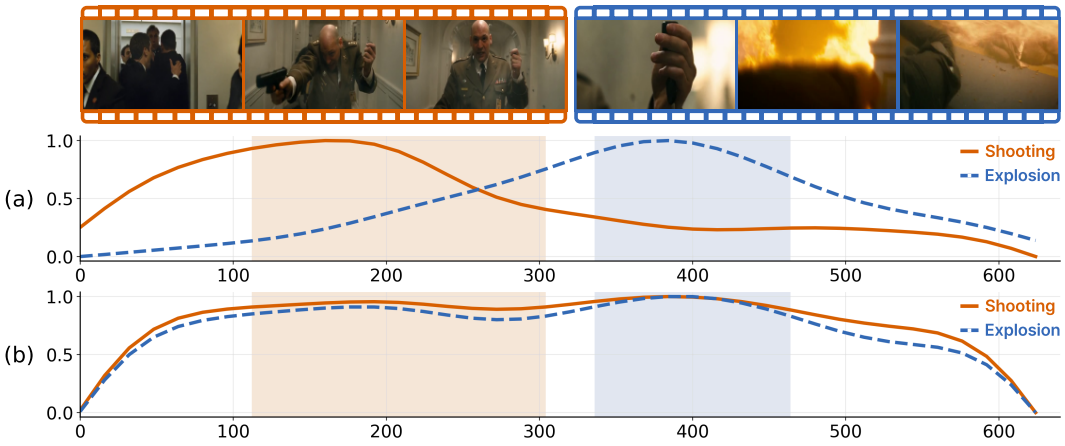


Figure 1: **Definition following and blindness in OWVAD.** (a) DeCoS (Ours) assigns higher frame-level scores to the event specified by the definition. (b) LaGoVAD (Liu et al., 2026b) produces nearly identical score curves for the two definitions.

We find that this imbalance is substantial on current VAD benchmarks. Across UCF-Crime, XD-Violence, and MSAD, normal frames make up 87.9%–96.4% of the negative set under Drift@5 and 78.6%–93.0% under class-wise C-VAD evaluation. To test whether this imbalance can hide definition blindness in practice, we evaluate a branch of LaGoVAD (Liu et al., 2026b) that never receives the anomaly definition. Although this branch produces the same scores for every anomaly definition, it nearly matches the full model under Drift@5, with gaps of only 0.6–1.2 AUROC points and 0.3–0.5 AP points on XD-Violence and MSAD. These results reveal that even evaluation protocols with varying anomaly definitions may fail to distinguish definition following from definition blindness. We therefore introduce three complementary definition-conditioned metrics to evaluate definition following: DC-Disc, $\Delta$DC-Det, and $\Delta$DC-Sel. They test whether a model distinguishes the specified anomaly from other anomalies, detects the specified anomaly better with the matching definition than with non-matching definitions, and selects the matching event when multiple anomalies occur. These metrics separate definition following from generic anomaly detection.

We apply these metrics to VAD systems spanning conventional, open-vocabulary, and open-world settings, together with general vision–language models, on UCF-Crime, XD-Violence, and MSAD. Across our experiments, we find that models that perform well under existing VAD evaluation often struggle to follow the supplied anomaly definition. Motivated by this gap, we introduce Definition-Contrastive Scoring (DeCoS), which scores the supplied definition relative to competing definitions and reduces anomaly evidence shared across them. DeCoS improves all three definition-conditioned metrics across the three benchmarks.

This study provides new insights into the evaluation of OWVAD by showing that strong generic anomaly detection performance can mask definition blindness, even when anomaly definitions vary during evaluation. Our core contribution is to make the distinction between anomaly detection and definition following explicit and measurable, through dedicated evaluation metrics, empirical analysis, and Definition-Contrastive Scoring. Together, these contributions highlight that progress in OWVAD should be measured not only by how well models detect anomalies, but also by whether their predictions follow the supplied anomaly definition.

## 2 Related Work

**Video anomaly detection.** Semi-supervised VAD learns from normal training videos (Hasan et al., 2016; Liu et al., 2018; Gong et al., 2019; Song & Lee, 2026a), while weakly supervised VAD uses video-level labels to localize anomalous events (Sultani et al., 2018; Tian et al., 2021; Feng et al., 2021; Song et al., 2025). These settings are typically evaluated under a fixed benchmark definition of anomaly (Sultani et al., 2018; Wu et al., 2020; Zhu et al., 2024). Open-set and open-vocabulary methods extend detection to unseen anomaly classes (Zhu et al., 2022; Wu et al., 2024a;

Xu et al., 2025; Li et al., 2025), while cross-domain and multi-domain methods study generalization across datasets and anomaly labels (Aich et al., 2023; Cho et al., 2024). These lines of work ask whether anomaly detection generalizes to new classes or domains. We instead study whether temporal anomaly scores follow a supplied definition.

**Language-based anomaly detection.** CLIP-based VAD uses image–text alignment to localize and recognize anomalous events (Radford et al., 2021; Wu et al., 2024b; Zhu et al., 2026). Multimodal language-model methods use captions and reasoning to detect and explain anomalies (Zanella et al., 2024; Ye et al., 2025; Zhang et al., 2025; Song et al., 2024; 2026). These approaches establish language as a useful source of supervision or conditioning for anomaly detection. Our method uses language for a different purpose: it constructs temporal anomaly scores relative to competing definitions and trains the readout to distinguish the matching definition from non-matching definitions.

**User-defined anomaly detection and evaluation.** Recent work extends VAD to user-specified anomaly definitions (Liu et al., 2026b; Ahn et al., 2026; Liu et al., 2026a). In this setting, Drift@5 (Liu et al., 2026b) and the C-VAD protocol (Ahn et al., 2026) assess model performance as the user-provided anomaly definition changes. These protocols, however, do not isolate how the supplied definition changes the ranking of anomaly scores, because their evaluations reward generic anomaly detection more heavily than definition following. Our work separates these effects and evaluates definition following directly.

## 3 Diagnosing Definition Blindness

Given a video $V$ and an anomaly definition $z$, an OWVAD detector assigns an anomaly score to each frame:

$$S_\tau(V, z) \in \mathbb{R}, \qquad \tau = 1, \ldots, |V|, \tag{1}$$

where $|V|$ is the number of frames in $V$. Following Liu et al. (2026b), we assume that abnormality is determined by the video and the supplied definition. We write the frame-level target as

$$y^z_\tau = F_\tau(V, z) \in \{0, 1\}, \tag{2}$$

where $y^z_\tau = 1$ means that the frame at time $\tau$ is anomalous under $z$, and $y^z_\tau = 0$ otherwise. Definition following requires the detector to assign higher anomaly scores to frames that are anomalous under the supplied definition than to frames that are not.

A clear case of definition blindness occurs when the score function does not depend on the supplied definition:

$$S_\tau(V, z) = a_\tau(V) \qquad \forall\tau,\ \forall z, \tag{3}$$

where $a_\tau(V)$ depends only on the video. Such a detector can localize anomalous events, but cannot adjust which events receive high scores to match the supplied definition.

### 3.1 Why existing evaluation may conceal definition blindness

In this section, we show why definition-blind detectors can perform well under existing VAD evaluation. We identify each evaluation frame by $i = (n, \tau)$, where $n$ indexes videos $V_n$ and $\tau$ indexes a frame within it. We write $s_i(z) = S_\tau(V_n, z)$ and use $s(z)$ to denote the collection of these scores. All frame sets below use these indices. Let $\mathcal{A}$ and $\mathcal{N}$ denote the index sets of frames labeled anomalous and normal by the benchmark, respectively. Let $m$ denote either the AUROC or AP evaluation function. We write $m(s(z); \mathcal{P}^+, \mathcal{P}^-)$ for its value with $\mathcal{P}^+$ as the positive set and $\mathcal{P}^-$ as the negative set, using their scores in $s(z)$.

Standard VAD evaluation treats all anomalous frames as positives ($\mathcal{P}^+ = \mathcal{A}$) and normal frames as negatives ($\mathcal{P}^- = \mathcal{N}$). It measures generic anomaly detection under a fixed benchmark definition. Consider a detector that reuses the scores obtained under this definition for every definition, as in Equation (3). Its standard VAD AUROC and AP remain unchanged despite its definition blindness. High performance under this evaluation therefore does not establish definition following.

Drift@5 (Liu et al., 2026b) and the C-VAD protocol (Ahn et al., 2026) instead vary which anomaly classes are treated as targets. Drift@5 selects subsets of classes, whereas C-VAD evaluates one class

at a time. Let $\mathcal{C}$ denote the benchmark's anomaly classes. For each $c \in \mathcal{C}$, let $\mathcal{A}_c$ be the set of indices $i = (n, \tau) \in \mathcal{A}$ for frames labeled as class $c$. Under the single-label evaluation protocol considered here, these sets partition $\mathcal{A}$. For a target class set $\Gamma \subseteq \mathcal{C}$, let $z_\Gamma$ specify the classes in $\Gamma$ as anomalous. The target and excluded anomaly frame sets are

$$\mathcal{A}_\Gamma = \bigcup_{c \in \Gamma} \mathcal{A}_c, \qquad \mathcal{A}_{\neg\Gamma} = \mathcal{A} \setminus \mathcal{A}_\Gamma. \tag{4}$$

Under this class-based labeling, $F_\tau(V_n, z_\Gamma) = 1$ exactly when $i = (n, \tau) \in \mathcal{A}_\Gamma$ (Equation (2)). Both protocols treat $\mathcal{A}_\Gamma$ as positives and $\mathcal{N} \cup \mathcal{A}_{\neg\Gamma}$ as negatives. We analyze both AUROC and AP with these labels and denote the target-versus-rest metric by

$$m^{\text{rest}}(\Gamma) = m\big(s(z_\Gamma); \mathcal{A}_\Gamma, \mathcal{N} \cup \mathcal{A}_{\neg\Gamma}\big). \tag{5}$$

This evaluation combines separation of target anomalies from normal frames with separation from anomalies excluded by the definition.

For AUROC, the overall value is an exact weighted average of the two comparisons. Assume that $\mathcal{A}_\Gamma$, $\mathcal{N}$, and $\mathcal{A}_{\neg\Gamma}$ are nonempty. Keeping the target frames and scores $s(z_\Gamma)$ unchanged, we evaluate the two negative sets separately:

$$\begin{aligned} m^{\text{normal}}(\Gamma) &= m\big(s(z_\Gamma); \mathcal{A}_\Gamma, \mathcal{N}\big), \\ m^{\text{excluded}}(\Gamma) &= m\big(s(z_\Gamma); \mathcal{A}_\Gamma, \mathcal{A}_{\neg\Gamma}\big). \end{aligned} \tag{6}$$

Each target frame with index $i \in \mathcal{A}_\Gamma$ is compared with every negative frame in $\mathcal{N} \cup \mathcal{A}_{\neg\Gamma}$. AUROC averages these comparisons, assigning a value of one when the target score is higher, one half for a tie, and zero otherwise (Song & Lee, 2026b). The fraction of comparisons involving a normal negative frame is

$$\omega_\Gamma = \frac{|\mathcal{N}|}{|\mathcal{N}| + |\mathcal{A}_{\neg\Gamma}|}. \tag{7}$$

Grouping the comparisons by the negative frame's set gives

$$\text{AUROC}^{\text{rest}}(\Gamma) = \omega_\Gamma \text{AUROC}^{\text{normal}}(\Gamma) + (1 - \omega_\Gamma)\text{AUROC}^{\text{excluded}}(\Gamma). \tag{8}$$

When $\omega_\Gamma$ is large, a definition-blind detector can obtain high overall AUROC by separating anomalies from normal frames, even if it poorly distinguishes target anomalies from excluded anomalies.

AP has the same negative-pool problem, although it does not decompose into a weighted average as AUROC does. Normal frames and excluded anomalies both contribute false positives to the same precision calculation. As a result, lowering normal-frame scores can increase AP even when the scores of all anomalous frames remain unchanged. High AP under target-versus-rest evaluation therefore does not by itself establish improved definition following. We provide the exact expression in Appendix A.1.

Normal frames dominate the negative pool under both Drift@5 and C-VAD-style evaluation on current VAD benchmarks. For Drift@5, normal frames account for $96.4\%$, $87.9\%$, and $88.5\%$ of the negative pool on UCF-Crime, XD-Violence, and MSAD, respectively. For AUROC, this means that comparisons with excluded anomalies receive only $3.6\%$–$12.1\%$ of the total weight. Under C-VAD evaluation, the corresponding normal-frame fractions are $93.0\%$, $80.2\%$, and $78.6\%$. Thus, in both protocols, target-versus-normal discrimination can contribute substantially to the reported score even when the detector does not distinguish the target anomaly from other anomalies.

We next test whether a definition-blind detector can retain strong performance under Drift@5. We compare the full LaGoVAD model (Liu et al., 2026b) with its pre-fusion branch, which never receives an anomaly definition and therefore produces the same frame scores for every definition. As shown in Table 1, this branch nearly matches the full model under both standard VAD evaluation and Drift@5. On XD-Violence, the Drift@5 gaps are only $1.2$ AUROC points and $0.5$ AP points. On MSAD, they shrink to $0.6$ AUROC points and $0.3$ AP points. Thus, strong Drift@5 performance can coexist with complete definition blindness.

Table 1: **A definition-blind branch nearly matches full LaGoVAD.** Definition-blind denotes the pre-fusion branch that never receives the queried definition.

| Dataset | Score | Standard VAD | | D@5 | |
|---|---|---|---|---|---|
| | | AUROC | AP | AUROC | AP |
| XD | LaGoVAD | 89.5 | 71.5 | 85.9 | 41.9 |
| | Definition-blind | 88.1 | 70.6 | 84.7 | 41.5 |
| | Δ | **1.4** | **0.8** | **1.2** | **0.5** |
| MSAD | LaGoVAD | 90.0 | 66.1 | 85.2 | 37.3 |
| | Definition-blind | 89.5 | 65.5 | 84.7 | 37.0 |
| | Δ | **0.6** | **0.5** | **0.6** | **0.3** |

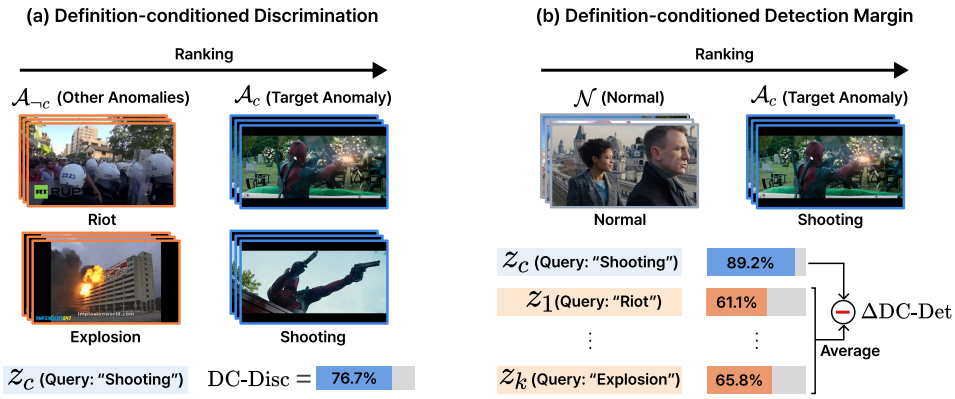


Figure 2: **Illustration of definition-conditioned evaluation for an example target class.** (a) DC-Disc measures whether the supplied definition ranks target anomaly frames above other anomalies, excluding normal frames. and (b) ΔDC-Det measures the performance difference of the matching definition over the non-matching average on the same anomaly and normal frames.

### 3.2 Definition-conditioned evaluation

The analysis above suggests two complementary ways to isolate definition following: remove normal frames from the comparison, or hold the evaluated frames fixed and change only the supplied definition. DC-Disc follows the first approach. ΔDC-Det and ΔDC-Sel follow the second, measuring the difference from the matching definition for detection and multi-event selection, respectively. We report all three using base metric $m \in \{\text{AUROC}, \text{AP}\}$.

**DC-Disc: distinguish target anomalies from excluded anomalies. (Figure 2-a)** For each class $c \in \mathcal{C}$, we use the definition $z_c = z_{\{c\}}$ that specifies only class $c$ as anomalous. We evaluate the scores $s(z_c)$ using $\mathcal{A}_c$ as positives and $\mathcal{A}_{\neg c} = \mathcal{A} \setminus \mathcal{A}_c$ as negatives. Averaging the metric values equally over classes gives

$$\text{DC-Disc}_m = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} m\big(s(z_c); \mathcal{A}_c, \mathcal{A}_{\neg c}\big). \tag{9}$$

Normal frames do not enter this comparison, so strong separation from normal frames cannot mask poor discrimination between target and excluded anomalies.

**ΔDC-Det: measure the matching-definition difference. (Figure 2-b)** For each target class $c \in \mathcal{C}$, we use $\mathcal{A}_c$ as positives and $\mathcal{N}$ as negatives. We evaluate this comparison under the matching definition $z_c$ and each non-matching definition $z_k$, where $k \in \mathcal{C} \setminus \{c\}$. We subtract the mean metric value under non-matching definitions from the value under $z_c$, then average equally over target classes:

$$\Delta\text{DC-Det}_m = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left[ m\big(s(z_c); \mathcal{A}_c, \mathcal{N}\big) - \frac{1}{|\mathcal{C}| - 1} \sum_{k \in \mathcal{C} \setminus \{c\}} m\big(s(z_k); \mathcal{A}_c, \mathcal{N}\big) \right]. \tag{10}$$

A detector satisfying Equation (3) receives zero, regardless of how well it separates anomalous frames from normal frames.

**ΔDC-Sel: select the matching event in a multi-event video. (Figure 1)** We measure whether the matching definition improves separation of the target event from other anomalous events in the same video. For each multi-event video $V_n$, let $\mathcal{C}_n \subseteq \mathcal{C}$ be the anomaly classes present, with $|\mathcal{C}_n| \geq 2$. Let $\mathcal{A}_{n,c}$ contain the frame indices $i = (n, \tau)$ labeled as class $c$ in $V_n$. The other anomalous frames in that video form

$$\mathcal{A}_{n,\neg c} = \bigcup_{k \in \mathcal{C}_n \setminus \{c\}} \mathcal{A}_{n,k}. \tag{11}$$

We use $\mathcal{A}_{n,c}$ as positives and $\mathcal{A}_{n,\neg c}$ as negatives under both the matching definition $z_c$ and each non-matching definition $z_k$, where $k \in \mathcal{C}_n \setminus \{c\}$. Let $\Omega_{\text{sel}}$ contain one pair $(n, c)$ for each evaluated

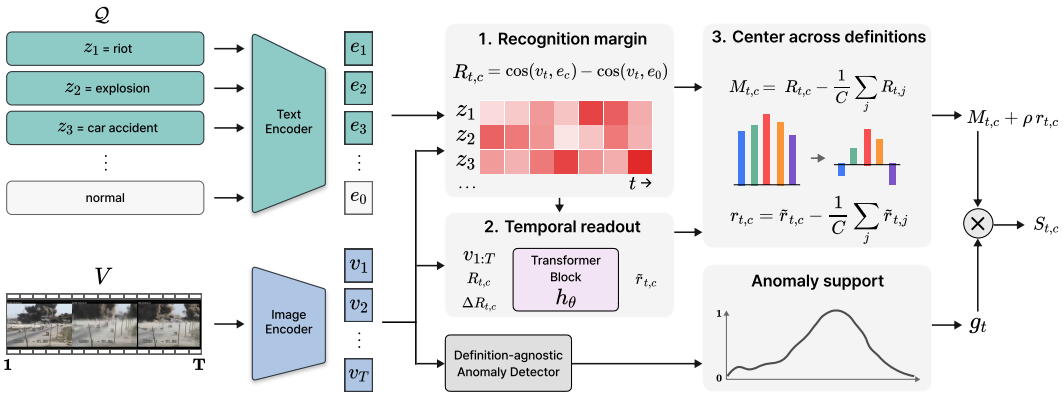


Figure 3: DeCoS converts anomaly support into definition-contrastive scores. A frozen definition-agnostic detector estimates anomaly support $g_t$. The readout produces a zero-sum contrast $M_{t,c} + \rho r_{t,c}$ across the definitions in $\mathcal{Z}$. Their product preserves anomaly support while preventing a definition-shared score offset.

video $V_n$ and each target $c \in \mathcal{C}_n$. We average the matching-definition difference equally over these pairs:

$$\Delta\text{DC-Sel}_m = \frac{1}{|\Omega_{\text{sel}}|} \sum_{(n,c)\in\Omega_{\text{sel}}} \left[ m\big(s(z_c); \mathcal{A}_{n,c}, \mathcal{A}_{n,\neg c}\big) - \frac{1}{|\mathcal{C}_n| - 1} \sum_{k \in \mathcal{C}_n \setminus \{c\}} m\big(s(z_k); \mathcal{A}_{n,c}, \mathcal{A}_{n,\neg c}\big) \right]. \tag{12}$$

Normal frames may remain in the video as context but are excluded from the metric. A detector satisfying Equation (3) receives zero even if it accurately localizes all anomalous events.

## 4 DeCoS: Definition-Contrastive Scoring

DeCoS is a definition-relative scorer. It evaluates a queried definition against a comparison set $\mathcal{Z}$. This set-conditioned interface is also used by existing language-based VAD models, which compare video features against multiple textual labels (Wu et al., 2024b; Liu et al., 2026b).

**Scoring.** We score the supplied definition $z_c$ within a comparison set $\mathcal{Z} = \{z_1, \ldots, z_C\}$, where $C$ is the number of comparison entries. For a video $V$, let $v_t$ denote its CLIP visual feature at feature time step $t$. Let $e_c$ and $e_0$ be the text embeddings of $z_c$ and the normal reference, respectively. We compute recognition margins relative to normality and remove their mean across definitions:

$$R_{t,c} = \cos(v_t, e_c) - \cos(v_t, e_0), \tag{13}$$

$$M_{t,c} = R_{t,c} - \frac{1}{C}\sum_{k=1}^{C} R_{t,k}. \tag{14}$$

A temporal readout $h_\theta$, shared across definitions, uses the visual sequence, recognition margins, and their first-order temporal differences to produce bounded residuals $\tilde{r}_{t,c} \in [-1, 1]$. Let $g_t \in [0, 1]$ be the frozen detector's anomaly support aligned to the feature sequence and shared across definitions. We center the residuals and combine the refined evidence with this support:

$$r_{t,c} = \tilde{r}_{t,c} - \frac{1}{C}\sum_{k=1}^{C} \tilde{r}_{t,k}, \tag{15}$$

$$S_{t,c} = g_t \left(M_{t,c} + \rho r_{t,c}\right), \tag{16}$$

where $\rho > 0$ is a fixed residual scale. Centering both evidence terms gives $\sum_{c=1}^{C} S_{t,c} = 0$, so the readout cannot raise all definition scores together.

Table 2: **Drift@5 can obscure differences in anomaly discrimination.** $\mathrm{D@5}^{\mathrm{N}}$ and $\mathrm{D@5}^{\mathrm{X}}$ average $\mathrm{AUROC}^{\mathrm{normal}}(\Gamma)$ and $\mathrm{AUROC}^{\mathrm{excluded}}(\Gamma)$, respectively, over the five definitions used for D@5.

| Method | $\mathrm{D@5}_{\mathrm{AUROC}}$ (↑) | | | $\mathrm{D@5}^{\mathrm{N}}_{\mathrm{AUROC}}$ (↑) | | | $\mathrm{D@5}^{\mathrm{X}}_{\mathrm{AUROC}}$ (↑) | | |
|---|---|---|---|---|---|---|---|---|---|
| | UCF | XD | MSAD | UCF | XD | MSAD | UCF | XD | MSAD |
| LaGoVAD | **79.7** | 85.9 | 85.2 | **80.8** | 90.3 | 89.9 | 49.9 | 55.4 | 49.4 |
| VadCLIP | 78.2 | 89.2 | **86.8** | 78.9 | **91.6** | **90.7** | 58.9 | 73.4 | 57.5 |
| **DeCoS** | 78.5±3.7 | **90.5±1.3** | 85.2±0.9 | 79.1±3.8 | 90.9±1.2 | 88.3±0.9 | **64.6±0.6** | **88.1±1.9** | **62.3±1.3** |

Table 3: **Definition-conditioned evaluations separate generic anomaly detection from definition following.** DC-Disc values are AUROC percentages; Δ values are AUROC point margins.

| Method | $\mathrm{DC\text{-}Disc}_{\mathrm{AUROC}}$ (↑) | | | $\Delta\mathrm{DC\text{-}Det}_{\mathrm{AUROC}}$ (↑) | | | $\Delta\mathrm{DC\text{-}Sel}_{\mathrm{AUROC}}$ (↑) | | |
|---|---|---|---|---|---|---|---|---|---|
| | UCF | XD | MSAD | UCF | XD | MSAD | UCF | XD | MSAD |
| VadCLIP | 52.0 | 60.8 | 59.4 | 3.2 | 5.0 | 2.3 | 8.7 | 21.9 | 14.2 |
| LaGoVAD | 45.2 | 41.0 | 49.0 | 0.0 | -0.1 | 0.0 | -0.2 | -0.7 | 0.0 |
| PLOVAD | 47.8 | 52.3 | 48.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| LLaVA-1.5 7B | 56.4 | 55.4 | 59.9 | 6.6 | 6.1 | 9.8 | 6.0 | 8.9 | 20.1 |
| InternVL3 8B | 64.9 | 71.1 | 71.0 | 14.5 | 20.9 | 18.8 | 26.7 | 36.2 | 35.3 |
| Qwen2.5-VL 7B | 62.5 | 68.4 | 68.5 | 13.2 | 19.2 | 17.5 | 21.8 | 34.2 | 36.1 |
| Qwen3-VL 8B | 61.3 | 66.5 | 67.4 | 11.7 | 17.7 | 16.7 | 20.0 | 27.1 | 31.4 |
| **DeCoS** | **76.1±0.8** | **87.1±2.1** | **78.3±1.8** | **30.0±0.1** | **49.2±1.7** | **37.1±2.8** | **39.1±2.1** | **53.5±3.6** | **49.6±1.7** |

**Training.** We train the readout on single-event and multi-event clips using PreVAD video-level labels (Liu et al., 2026b). At each supervised time step, cross-entropy over the comparison definitions uses $\alpha(M_{t,1:C} + \rho r_{t,1:C})$ as logits and the matching definition as its target. We optimize $\theta$ and the logit scale $\alpha$, averaging first over time steps within each supervised region and then over regions in the batch. Original clips inherit their video-level label at all valid time steps; in synthesized clips, only the inserted event regions enter the loss. The readout must distinguish matching from non-matching definitions rather than merely detect anomalous events. The readout architecture, training construction, full loss, and frame alignment are detailed in Appendix B.

## 5 Experimental Results

We now test whether the distinction between anomaly detection and definition following changes how existing VAD systems should be evaluated. We first examine how Drift@5 combines target-versus-normal and target-versus-other-anomaly discrimination. We then evaluate definition following directly across model families, test definition-conditioned selection on natural multi-event videos, and examine whether the results depend on exact event names. Finally, we ablate DeCoS to connect its components to the behaviors they are designed to support.

**Protocol.** We evaluate on UCF-Crime (UCF), XD-Violence (XD), and MSAD (Sultani et al., 2018; Wu et al., 2020; Zhu et al., 2024). The comparison includes weakly supervised VAD, open-vocabulary VAD, language-guided OWVAD, and generative vision–language models (Liu et al., 2024; Zhu et al., 2025; Bai et al., 2025b;a). For a fair zero-shot comparison, VadCLIP, LaGoVAD, PLOVAD, and DeCoS use the same PreVAD training source (Wu et al., 2024b; Liu et al., 2026b; Xu et al., 2025). Model-specific inputs and scoring rules are given in Appendix C. We evaluate all three metrics with both AUROC and AP. The main tables report AUROC for compactness, with the corresponding AP results in Appendix D.

### 5.1 High Drift@5 does not imply definition following

Table 2 shows that large improvements in separating target anomalies from excluded anomalies can produce much smaller gains in Drift@5. On XD-Violence, DeCoS exceeds LaGoVAD by 32.7 AUROC points on $\mathrm{D@5}^{\mathrm{X}}$, but by only 4.6 points on D@5. The difference on $\mathrm{D@5}^{\mathrm{N}}$ is just 0.6 points. On MSAD, DeCoS exceeds LaGoVAD by 12.8 points on $\mathrm{D@5}^{\mathrm{X}}$, yet both models obtain 85.2% on D@5. Thus, higher Drift@5 does not necessarily indicate better discrimination between target and excluded anomalies.

Table 4: **Definition-conditioned performance persists beyond event-name.** Left: evaluated event names are absent from the training vocabulary. Right: event names are removed and replaced by natural-language definitions. Each name-free entry averages the metrics from three descriptions.

(a) **Unseen event names**

| Method | DC-Disc$_{\text{AUROC}}$ UCF | XD | MSAD | $\Delta$DC-Det$_{\text{AUROC}}$ UCF | XD | MSAD |
|---|---|---|---|---|---|---|
| LaGoVAD | 44.3 | 15.2 | 45.4 | 0.0 | -0.2 | 0.0 |
| VadCLIP | <u>48.1</u> | <u>28.7</u> | <u>54.0</u> | <u>2.2</u> | <u>2.3</u> | <u>1.5</u> |
| **DeCoS** | **75.8** | **75.6** | **75.4** | **29.2** | **29.9** | **31.7** |

(b) **Name-free definitions**

| Method | DC-Disc$_{\text{AUROC}}$ UCF | XD | MSAD | $\Delta$DC-Det$_{\text{AUROC}}$ UCF | XD | MSAD |
|---|---|---|---|---|---|---|
| LaGoVAD | 45.2 | 40.7 | 49.0 | -0.1 | -0.2 | +0.0 |
| VadCLIP | <u>50.6</u> | <u>52.3</u> | <u>55.4</u> | <u>1.6</u> | <u>3.0</u> | <u>0.9</u> |
| **DeCoS** | **72.0** | **86.3** | **71.8** | **27.4** | **53.0** | **28.3** |

### 5.2 Definition-conditioned metrics reveal definition blindness

Table 3 reveals large differences in definition following that are not apparent from Drift@5 alone. On XD-Violence, VadCLIP obtains 89.2% AUROC under Drift@5, but only 60.8% on DC-Disc, with $\Delta$DC-Det and $\Delta$DC-Sel values of 5.0 and 21.9 points. LaGoVAD and PLOVAD show little or no matching-definition gain across the three benchmarks. In contrast, InternVL3 and Qwen2.5-VL show substantially larger gains, indicating that models with strong anomaly-detection performance can differ sharply in how their scores respond to the supplied definition.

DeCoS achieves the highest values on all three AUROC-based metrics across UCF-Crime, XD-Violence, and MSAD. Relative to the best baseline in each comparison, it improves DC-Disc by 7.3–16.0 AUROC points, $\Delta$DC-Det by 15.5–28.3 points, and $\Delta$DC-Sel by 12.4–17.3 points.

### 5.3 Definition-conditioned selection in natural multi-event videos

The controlled $\Delta$DC-Sel evaluation uses synthesized two-event clips. We next ask whether the differences in definition following observed in this controlled setting also appear in natural multi-event videos. We apply the same $\Delta$DC-Sel comparison to 43 XD-Violence videos containing at least two anomaly classes in separable intervals, yielding 89 target-label queries. DeCoS obtains 36.1 AUROC points, compared with 19.2 for VadCLIP. All other baselines obtain at most 1.2 points, while the generative vision–language models produce near-zero or negative margins (Table 5). These results show that the differences in definition following observed in the controlled evaluation persist in natural multi-event videos.

Table 5: Natural multi-event evaluation in AUROC points.

| Method | $\Delta$DC-Sel |
|---|---|
| VadCLIP | <u>19.2</u> |
| LaGoVAD | 1.2 |
| PLOVAD | 0.0 |
| LLaVA-1.5 7B | 0.6 |
| InternVL3 8B | -2.6 |
| Qwen2.5-VL 7B | -3.2 |
| Qwen3-VL 8B | -6.1 |
| **DeCoS** | **36.1±8.6** |

### 5.4 Definition following beyond fixed event names

A user-specified definition need not use an event name seen during training. We therefore evaluate two settings that break this correspondence. In the unseen-name setting, the evaluated event names do not appear in the PreVAD training labels (Table 4-a). In the name-free setting, the event name is removed entirely and replaced by natural-language visual descriptions (Table 4-b). DeCoS retains strong definition-conditioned performance in both settings, while LaGoVAD remains near zero on the $\Delta$ metrics and VadCLIP shows much smaller gains. These results show that DeCoS's definition following is not limited to exact event-name matching.

### 5.5 Ablation study

Table 6 ablates the components of DeCoS on XD-Violence. The −centering and −zero-sum variants remove the mean subtraction from the base margin and residual, respectively. Each variant uses a separately trained checkpoint with the same definition-agnostic anomaly detector. Removing the residual readout gives the largest drop

Table 6: Component ablation on XD-Violence.

| Variant | DC-Disc | $\Delta$DC-Det | $\Delta$DC-Sel |
|---|---|---|---|
| DeCoS | 87.1±2.1 | 49.2±1.7 | 53.5±3.6 |
| −residual | 76.7±0.0 | 40.9±0.0 | 48.8±0.0 |
| −centering | 84.1±2.3 | 45.5±0.8 | 50.3±1.2 |
| −zero-sum | 82.6±5.0 | 40.5±1.9 | 52.7±4.6 |
| −multi-event | 80.4±2.7 | 44.4±4.2 | 37.3±3.6 |

in DC-Disc. Removing residual centering reduces $\Delta$DC-Det by 8.7 points. Removing base centering reduces all three metrics on XD-Violence. Removing multi-event training examples reduces $\Delta$DC-Sel by 16.2 points.

## 6 Discussion and Limitations

**Definition following should be evaluated separately.** Changing the anomaly definition during evaluation is not, by itself, enough to evaluate definition following. In Drift@5 and C-VAD-style evaluation, target anomalies are compared against a negative pool containing both normal frames and anomalies excluded by the current definition. Because normal frames make up most of this pool on the benchmarks we study, a detector can score well by separating anomalies from normality even if its ranking changes little with the definition. OWVAD evaluation should therefore report definition following separately from conventional anomaly detection.

**Definition sensitivity is not definition following.** A model can respond to changes in the input text without ranking the event specified by that text more highly. Definition following therefore requires a directional comparison: on the same visual evidence, the matching definition should rank the matching event higher than non-matching definitions do. This is why our $\Delta$ metrics compare matching and non-matching definitions rather than measuring score variation alone.

**The evaluated definitions remain benchmark-grounded.** We evaluate definition following across the anomaly concepts in UCF-Crime, XD-Violence, and MSAD, including settings where the exact event name is unseen during training or removed from the test-time definition. These controls show that definition following need not depend on exposure to the exact event name during training or its inclusion in the test-time definition. They do not establish generalization to arbitrary user definitions or entirely new anomaly concepts, since some events with unseen names are conceptually close to those represented in PreVAD and our evaluation remains anchored to benchmark events. Evaluating definition following with more open-ended user definitions is an important direction for future benchmarks.

**Natural multi-event evaluation remains limited in scale.** Our controlled two-event clips provide a clean test of whether changing the definition changes the relative scores of co-occurring anomalous events. We also observe the same distinction on 43 naturally multi-event XD-Violence videos, but this subset is small and less balanced than the controlled evaluation. A larger benchmark with naturally co-occurring anomalies and query-specific temporal annotations would allow definition-conditioned selection to be evaluated in natural videos at a larger scale.

## 7 Conclusion

Open-world video anomaly detection should be evaluated by whether temporal scores follow the supplied anomaly definition, not only by whether anomalous frames are separated from normal ones. We show that existing temporal evaluations can obscure this distinction. Across UCF-Crime, XD-Violence, and MSAD, normal frames make up 87.9%–96.4% of the negative pool under Drift@5 and 78.6%–93.0% under C-VAD evaluation. As a result, strong target-versus-normal separation can contribute substantially to the reported score even when a model poorly distinguishes the target anomaly from other anomalies. Our definition-conditioned evaluation confirms this distinction across multiple model families: strong generic anomaly detection can coexist with weak definition following.

These findings change what should count as evidence of OWVAD behavior. Definition following should be evaluated explicitly alongside conventional anomaly detection, rather than inferred from text conditioning or strong target-versus-rest performance. DeCoS provides a constructive example that definition following can be improved with definition-relative scoring. More generally, testing how anomaly scores change when the anomaly definition changes provides a direct way to determine whether an OWVAD system follows the supplied anomaly definition.

### AI USE STATEMENT

Generative AI tools were used to assist with sentence editing, grammar checking, typographical checks, isolated utility functions, and generation of candidate descriptions used in the name-free evaluation. The authors reviewed every AI-assisted change, verified the mathematical definitions and all reported values against the evaluation outputs, and take responsibility for the final text, claims, code, and artifacts. No reported experimental result was generated or selected by a language model.

### ETHICS STATEMENT

This work uses existing public video anomaly detection datasets and pretrained models and does not collect new human-subject data. Video anomaly detection may be used in surveillance settings, where deployment can raise privacy and fairness concerns. Our experiments are limited to benchmark evaluation and do not study or make claims about deployment decisions involving individuals.

### REPRODUCIBILITY STATEMENT

The main paper provides the definitions of all proposed evaluation metrics and the experimental protocols. The appendix specifies the DeCoS architecture, training and inference procedure, baseline scoring rules, implementation and runtime settings, AP counterparts of the main results, and the definitions used in the semantic controls. These details are intended to support reproduction of both the evaluation and the reported experiments.

## REFERENCES

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

Abhishek Aich, Kuan-Chuan Peng, and Amit K Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2579–2591, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14009–14018, 2021.

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.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29203–29212, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 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, Jianqin Wu, and Linlin Yang. Esom: Efficiently understanding streaming video anomalies with open-world dynamic definitions. *arXiv preprint arXiv:2604.07772*, 2026a.

Zihao Liu, Xiaoyu Wu, Jianqin Wu, Xuxu Wang, and Linlin Yang. Language-guided open-world video anomaly detection under weak supervision. In *International Conference on Learning Representations*, volume 2026, pp. 153197–153224, 2026b.

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.

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

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*, 2026b.

Inpyo Song, Minjun Joo, Joonhyung Kwon, and Jangwon Lee. Video question answering for people with visual impairments using an egocentric 360-degree camera. *arXiv preprint arXiv:2405.19794*, 2024.

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

Inpyo Song, Minjun Joo, Joonhyung Kwon, Eunji Jeon, and Jangwon Lee. Instance-aligned captions for explainable video anomaly detection. *arXiv preprint arXiv:2601.08155*, 2026.

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

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.

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

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, 2024b.

Chenting Xu, Ke Xu, Xinghao Jiang, and Tanfeng Sun. Plovad: Prompting vision-language models for open vocabulary video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):5925–5938, 2025.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8679–8688, 2025.

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

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the computer vision and pattern recognition conference*, pp. 13843–13853, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024.

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-clip: Abnormality-aware latent-enhanced representation tuning of clip for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35545–35554, 2026.

Yuansheng Zhu, Wentao Bao, and Qi Yu. Towards open set video anomaly detection. In *European Conference on Computer Vision*, pp. 395–412. Springer, 2022.

# A Additional Evaluation Details

**Drift@5.** XD-Violence and MSAD use the five class subsets in LaGoVAD's Table D. UCF-Crime uses a local extension: five balanced seven-class subsets of its 13 anomaly classes, constructed with seed 7. For each method, the subset score is the maximum over the selected definition-conditioned score columns, using the same scoring rule as in the main evaluation. Anomaly spans inherit the video's single class label; the five subset metrics are averaged equally.

**Controlled selection.** For each ordered pair of distinct classes, we sample 20 pairs of test anomaly clips with seed 7. At least five annotated feature steps are required; extracted segments are capped at 32 steps. Each feature sequence is constructed as $[N_0, A, N_1, B, N_2]$, with normal segments sampled at lengths of 24–44 steps (or the entire normal clip when shorter). Each sequence is evaluated with both events as the target; this swaps the target and competitor masks, not the temporal sequence. The resulting $3120/600/2200$ sequences for UCF/XD/MSAD yield $6240/1200/4400$ directed evaluations, averaged equally. For DeCoS, the query occupies slot 0 and is replaced between the matching and competing evaluations. Four filler definitions, sampled from classes other than the two events, remain fixed. Normal context is excluded from the metric.

**Natural selection.** The XD-Violence evaluation uses label-specific intervals from the released multi-class annotations and contains 43 videos and 89 target-label queries (40 videos with two queries and three with three). Intervals are mapped to the feature grid by flooring the start and rounding the end upward at the 16-frame interval. For each target, positives are its annotated steps excluding overlap with any competing label; negatives are the union of competing-label steps excluding the target. The matching-definition metric minus the mean competing-definition metric is averaged equally over target-label queries. DeCoS uses the full XD anomaly vocabulary in this evaluation.

## A.1 AP under Target-vs-Rest Evaluation

The main text notes that AP uses the same mixed negative pool as target-vs-rest AUROC, but does not admit the same additive decomposition. Here we give the exact expression.

For a target frame $p \in \mathcal{A}_\Gamma$, let $T_\Gamma(p)$ denote the number of target frames with scores at least $s_p(z_\Gamma)$, and let $B^N_\Gamma(p)$ and $B^E_\Gamma(p)$ denote the corresponding numbers of normal frames and excluded-anomaly frames. Then

$$\mathrm{AP}^{\mathrm{rest}}(\Gamma) = \frac{1}{|\mathcal{A}_\Gamma|} \sum_{p \in \mathcal{A}_\Gamma} \frac{T_\Gamma(p)}{T_\Gamma(p) + B^N_\Gamma(p) + B^E_\Gamma(p)}. \tag{17}$$

Normal frames and excluded anomalies therefore enter the same precision denominator, even though their contributions cannot be separated into the weighted average used for AUROC.

# B DeCoS Implementation Details

## B.1 Comparison Set and Text Features

DeCoS evaluates the queried anomaly definition within a comparison set $\mathcal{Z} = \{z_1, \ldots, z_C\}$, where $C$ is the number of comparison entries. For the full-benchmark evaluation, $\mathcal{Z}$ contains the anomaly definitions of the corresponding dataset, without an additional normal column. Controlled selection instead uses the query and four fixed fillers. When only one anomaly definition is supplied, we additionally include a normal definition as a second comparison entry ($C = 2$), so that the queried definition is compared against normality.

We use frozen CLIP ViT-B/16 image and text encoders. Let $\mathbf{v}_t \in \mathbb{R}^{512}$ denote the normalized visual feature at feature time step $t$, and let $\mathbf{e}_c$ denote the normalized text embedding of definition $z_c$. Each anomaly definition is encoded using the template

"a video of {}".

The normal reference $\mathbf{e}_0$ is the normalized mean embedding of three fixed descriptions of normal video content. The single-query comparison entry uses the text "a video of normal". The recognition margin used in the main text is

$$R_{t,c} = \cos(\mathbf{v}_t, \mathbf{e}_c) - \cos(\mathbf{v}_t, \mathbf{e}_0). \tag{18}$$

### B.2 Residual Readout

We obtain the anomaly support $g_t$ from the frozen pre-fusion, definition-agnostic binary head of LaGoVAD. This branch receives only CLIP visual features and no anomaly definition. We apply its sigmoid output to obtain $g_t \in [0, 1]$.

The temporal readout uses the CLIP visual sequence together with the recognition margin and its temporal derivative. We first project each 512-dimensional visual feature to 256 dimensions and process the sequence with a two-layer Transformer encoder:

$$\mathbf{H}_{1:T} = E_{\text{vis}}(W_v \mathbf{v}_{1:T}), \qquad W_v : \mathbb{R}^{512} \to \mathbb{R}^{256}. \tag{19}$$

Both Transformer encoders used by DeCoS have hidden dimension 256, four attention heads, feed-forward dimension 512, and dropout 0.1.

For each definition $c$, the readout also uses the recognition margin $R_{t,c}$ and its temporal derivative

$$\Delta R_{t,c} = \begin{cases} R_{2,c} - R_{1,c}, & t = 1, \\ (R_{t+1,c} - R_{t-1,c})/2, & 1 < t < T, \\ R_{T,c} - R_{T-1,c}, & t = T. \end{cases} \tag{20}$$

These two scalars are concatenated with $\mathbf{H}_t$ and projected back to 256 dimensions:

$$\mathbf{u}_{t,c} = \text{GELU}\left(W_f[\mathbf{H}_t; R_{t,c}; \Delta R_{t,c}] + \mathbf{b}_f\right). \tag{21}$$

A second two-layer Transformer encoder, with parameters shared across definitions, processes each definition-specific temporal sequence:

$$\mathbf{h}^{\text{res}}_{1:T,c} = E_{\text{res}}(\mathbf{u}_{1:T,c}), \tag{22}$$

$$\tilde{r}_{t,c} = \tanh\left(\mathbf{w}_r^\top \mathbf{h}^{\text{res}}_{t,c} + b_r\right). \tag{23}$$

The residual head is initialized with zero weights and bias. There is no attention across definitions: the same $E_{\text{res}}$ is applied independently to each definition sequence.

We center both the recognition margin and the residual over the comparison set:

$$M_{t,c} = R_{t,c} - \frac{1}{C}\sum_{k=1}^{C} R_{t,k}, \tag{24}$$

$$r_{t,c} = \tilde{r}_{t,c} - \frac{1}{C}\sum_{k=1}^{C} \tilde{r}_{t,k}. \tag{25}$$

Let $g_t \in [0, 1]$ denote the anomaly support produced by the frozen definition-agnostic detector and aligned to the CLIP feature sequence. The final definition-conditioned score is

$$S_{t,c} = g_t\left(M_{t,c} + \rho r_{t,c}\right), \qquad \rho = 0.45. \tag{26}$$

### B.3 Training and Inference

We train only the definition-conditioned readout using PreVAD video-level annotations. The CLIP encoders and the definition-agnostic detector remain frozen, and no component is adapted on UCF-Crime, XD-Violence, or MSAD. Training examples include original single-event clips, one-event splices with normal context, and two-event splices containing distinct anomaly categories. For original anomaly clips, all valid feature time steps inherit the clip-level anomaly definition. For synthesized clips, the known insertion locations determine the supervised time steps and their matching

Table 7: Inference efficiency of feature-based methods on the median-length anomalous MSAD video. Parameter counts and GFLOPs exclude the shared frozen CLIP backbone. Timings average 20 repetitions at batch size one; DeCoS processes all dataset definitions in one pass and excludes temporal smoothing. End-to-end runtime is the post-CLIP time plus the shared 23.8 ms feature-extraction measurement.

| Method | Scorer params | After CLIP (ms) | GFLOPs | E2E (ms) |
|---|---|---|---|---|
| VadCLIP | 9.5M | 83 | 176 | 107 |
| LaGoVAD | 19.1M | 98 | 73 | 122 |
| PLOVAD | 4.0M | 10 | 66 | 34 |
| DeCoS | 8.8M | 10 | 6.7 | 34 |

definitions. Each example is paired with $C = 6$ anomaly definitions, including the matching definition and distractor definitions. Normal frames are retained as temporal context. Synthetic normal context is excluded from the selection loss; original clips retain weak video-level supervision even when some of their steps depict normal activity. No additional human frame-level labels are used.

Let $\mathcal{R}$ be the supervised regions in a batch: all valid steps of an original anomaly clip, or each inserted event region of a synthetic clip. For region $r$ in example $q(r)$, let $y_r \in \{1, \dots, C\}$ be its video-level event label. Training uses ungated readout logits with learned scale $\alpha$:

$$L_{t,c}^{(q)} = \alpha \left( M_{t,c}^{(q)} + \rho r_{t,c}^{(q)} \right). \tag{27}$$

The readout is trained with cross-entropy over the $C$ anomaly-definition columns:

$$\mathcal{L}_{\text{sel}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{1}{|r|} \sum_{t \in r} \text{CE} \left( L_{t,1:C}^{(q(r))}, y_r \right). \tag{28}$$

The logit scale $\alpha$ is initialized to 10 and optimized jointly with the readout.

At inference time, full-vocabulary evaluations retain the same comparison set; controlled selection retains the fillers while replacing the query. The visual encoder and the definition-agnostic detector are evaluated once per video, while the margin and residual branches are applied to all definitions using shared parameters. Each definition is processed independently along the temporal dimension, with no attention between definitions. For a queried definition $z_c$, DeCoS returns the corresponding temporal score sequence $S_{1:T,c}$.

The detector support $g_t$ and the DeCoS readout are evaluated on the same CLIP feature grid. To obtain frame-level scores, we repeat each feature-level score over the corresponding video frames. Let $\pi_d(\tau) \in \{1, \dots, T\}$ map frame $\tau$ in dataset $d$ to its corresponding feature block. Before smoothing, this gives

$$S_\tau(V, z_c) = S_{\pi_d(\tau),c}. \tag{29}$$

The feature interval is 8 frames for UCF-Crime and MSAD and 16 frames for XD-Violence. For all DeCoS evaluations, Gaussian smoothing is applied to the gated score after frame repetition, averaging standard deviations of 1–9 feature intervals.

### B.4 Implementation and Efficiency

The DeCoS readout contains 2.3M trainable parameters. Together with the 6.5M frozen detector, the deployed scorer contains 8.8M parameters, excluding CLIP. We optimize it with AdamW using a learning rate of $2 \times 10^{-4}$, weight decay of $10^{-4}$, and gradient clipping at 1.0. The residual scale is fixed to $\rho = 0.45$, and the learned logit scale $\alpha$ is initialized to 10. The main benchmark tables report mean and standard deviation over three training seeds.

The evaluation server has four NVIDIA RTX 6000 Ada GPUs (48 GB each) and two AMD EPYC 9124 CPUs.

Table 8: Inference cost of the generative VLM baselines on the same MSAD video. GFLOPs and runtime cover full-model inference.

| Method | Model size | GFLOPs | E2E (ms) |
|---|---|---|---|
| LLaVA-1.5 | 7B | 588,148 | 8,829 |
| InternVL3 | 8B | 340,253 | 9,923 |
| Qwen2.5-VL | 7B | 1,870,142 | 36,287 |
| Qwen3-VL | 8B | 1,259,435 | 21,421 |

### B.5 Effect of Query-Class Count

DeCoS is evaluated in the main paper with the full dataset vocabulary as the comparison set. At deployment, however, a user may provide only a small number of anomaly definitions. We therefore vary the number of queried anomaly classes $c$ at test time and keep the trained checkpoint fixed.

The only degenerate case is a single anomaly class without any competing column: centering would make the definition-relative score identically zero. For the operational one-query interface, we therefore include a normal column,

$$\mathcal{Q} = \{z_0, z_1\}, \qquad z_0 = \text{normal},$$

so the single supplied anomaly still competes against normality. Before the residual correction,

$$M_{t,1} = \frac{1}{2}(R_{t,1} - R_{t,0}), \qquad M_{t,0} = -M_{t,1}.$$

All rows with $c \geq 2$ use anomaly-class comparison sets of size $c$, sampled as in the candidate-count sweep. We report DC-Disc and DC-Det$\Delta$ only; DC-Sel$\Delta$ uses the splice-specific comparison-set construction and is not the same cardinality axis.

Table 9: Effect of the number of queried anomaly classes. The $c = 1$ row adds a normal comparison column. Intermediate sizes average five candidate-set draws per seed; the one-query and full-vocabulary rows use one set. Values are means and standard deviations across three seed-level means, in AUROC percentages or AUROC-point margins.

| Dataset | Query classes $c$ | DC-Disc AUROC | DC-Det$\Delta$ AUROC |
|---|---|---|---|
| UCF | 1 + normal | 69.3±1.8 | 22.0±1.3 |
| | 2 | 67.3±0.4 | 21.1±0.7 |
| | 3 | 71.7±0.2 | 26.6±0.7 |
| | 5 | 74.4±0.4 | 29.3±0.9 |
| | 8 | 76.8±0.2 | 31.2±0.5 |
| | 13 (full) | 76.1±0.8 | 30.0±0.1 |
| XD | 1 + normal | 79.7±0.3 | 32.3±0.7 |
| | 2 | 80.2±1.2 | 39.3±1.2 |
| | 3 | 84.2±1.4 | 45.3±1.1 |
| | 5 | 87.1±1.9 | 48.6±1.6 |
| | 6 (full) | 87.1±2.1 | 49.2±1.7 |
| MSAD | 1 + normal | 72.5±2.7 | 26.2±0.9 |
| | 2 | 70.8±1.5 | 27.3±1.7 |
| | 3 | 73.7±1.0 | 31.1±1.6 |
| | 5 | 76.5±1.2 | 34.3±1.9 |
| | 8 | 77.5±1.9 | 36.2±2.9 |
| | 11 (full) | 78.3±1.8 | 37.1±2.8 |

The trend is the expected one: more competing anomaly definitions usually sharpen definition following, while the smallest operational setting already remains strong. In particular, the $c = 1$ row is above the trained VAD/OWVAD baselines in Table 3, and its DC-Det$\Delta$ margin is larger than every non-DeCoS method there. Thus DeCoS does not require a large user-provided candidate set to improve definition-conditioned scoring.

Table 10: AP counterpart of Table 2. Values are AP percentages. DeCoS results are mean and standard deviation over three seeds.

| Method | $D@5_{AP}$ (↑) | | | $D@5^{N}_{AP}$ (↑) | | | $D@5^{X}_{AP}$ (↑) | | |
|---|---|---|---|---|---|---|---|---|---|
| | UCF | XD | MSAD | UCF | XD | MSAD | UCF | XD | MSAD |
| LaGoVAD | 12.8 | 41.9 | 37.3 | 14.5 | 57.7 | 52.3 | 53.8 | 56.9 | 56.5 |
| VadCLIP | 15.1 | 55.1 | 44.7 | 16.3 | 59.6 | 56.3 | 65.1 | 77.6 | 65.4 |
| DeCoS | **19.8±2.8** | **64.4±1.5** | **45.1±1.6** | **20.8±3.1** | **67.9±1.3** | **56.7±1.1** | **69.3±0.6** | **84.5±1.1** | **66.2±1.3** |

# C Baseline Scoring Details

All baselines are evaluated with the same anomaly definitions used in the definition-conditioned evaluation. For each definition $z_c$, we obtain a temporal score sequence $S_{1:T,c}$ as follows.

**PLOVAD.** We use the official PLOVAD architecture and losses, retrained on PreVAD. Its temporal anomaly score $a_t$ does not depend on the supplied definition. We therefore use

$$S_{t,c} = a_t \qquad \text{for all } c. \tag{30}$$

**VadCLIP.** We use the official VadCLIP architecture and losses, retrained on PreVAD. For definition-conditioned evaluation, we use the frame-level alignment scores produced by its A-branch as $S_{t,c}$.

**LaGoVAD.** We use the LaGoVAD model trained on PreVAD released by author Liu et al. (2026b). Its binary detection head produces a single temporal anomaly logit for the entire supplied anomaly-definition set, rather than a separate score for each definition. It therefore does not provide a per-definition anomaly score when multiple anomaly definitions are supplied. For each queried definition, we evaluate the model separately with $[\texttt{normal}, z_c]$ and use the sigmoid of the binary anomaly logit as $S_{t,c}$. Prompt sampling is disabled during evaluation.

**Generative vision–language models.** We evaluate LLaVA-1.5, InternVL3, Qwen2.5-VL, and Qwen3-VL as frozen frame-level scorers. For each sampled frame and definition $z_c$, we use the same numeric-rating instruction:

```
Is the following happening in this video: {z}? Reply with
ONLY a single number between 0.00 and 1.00 (e.g. 0.80).
```

The first generated number in $[0, 1]$ is used as $S_{t,c}$. Unparseable responses receive score 0. Model-specific chat templates wrap the same instruction.

# D AP Counterparts to Main Results

The main tables report AUROC for compactness. Here we report the corresponding AP results using the same evaluation protocols and score sequences.

# E Semantic-Control Details

## E.1 Unseen Event Names

We define an event name as unseen when the corresponding benchmark event name does not appear in the PreVAD training labels. Table 13 lists the event names used in this evaluation.

## E.2 Name-Free Definitions

For the name-free evaluation, we replace each benchmark event name with three short natural-language visual definitions. The definition bank combines the project's existing visual definitions

Table 11: AP counterpart of the DC-Disc and ΔDC-Det results in Table 3. DC-Disc values are AP percentages, and ΔDC-Det values are AP-point margins.

| Method | $\text{DC-Disc}_{\text{AP}}$ (↑) | | | $\Delta\text{DC-Det}_{\text{AP}}$ (↑) | | |
|---|---|---|---|---|---|---|
| | UCF | XD | MSAD | UCF | XD | MSAD |
| VadCLIP | 13.5 | 44.8 | 20.2 | 2.6 | 16.6 | 8.6 |
| LaGoVAD | 8.6 | 19.2 | 11.4 | 0.0 | 0.0 | 0.0 |
| PLOVAD | 10.3 | 21.2 | 11.6 | 0.0 | 0.0 | 0.0 |
| LLaVA-1.5 7B | 10.2 | 18.9 | 14.5 | 0.9 | 3.0 | 8.6 |
| InternVL3 8B | 20.7 | 36.4 | 29.7 | 8.7 | 17.6 | **29.8** |
| Qwen2.5-VL 7B | 15.2 | 31.1 | 24.4 | 2.4 | 14.6 | 22.7 |
| Qwen3-VL 8B | 15.1 | 31.5 | 25.7 | 2.9 | 14.4 | 22.6 |
| DeCoS | **27.2±1.8** | **55.7±1.1** | **33.3±2.4** | **10.1±0.8** | **36.5±0.6** | 25.2±2.5 |

Table 12: AP counterpart of the ΔDC-Sel results in Table 3 and the natural-video evaluation in Table 5. Values are AP-point margins.

| Method | $\Delta\text{DC-Sel}_{\text{AP}}$ (↑) | | | Natural XD (↑) |
|---|---|---|---|---|
| | UCF | XD | MSAD | XD |
| VadCLIP | 6.0 | 14.5 | 9.8 | 13.5 |
| LaGoVAD | -0.1 | -0.3 | 0.0 | 0.9 |
| PLOVAD | 0.0 | 0.0 | 0.0 | 0.0 |
| LLaVA-1.5 7B | 2.7 | 3.9 | 9.7 | -0.4 |
| InternVL3 8B | 12.8 | 16.3 | 16.6 | -1.2 |
| Qwen2.5-VL 7B | 10.4 | 15.5 | 17.3 | -2.4 |
| Qwen3-VL 8B | 9.3 | 11.9 | 14.3 | -2.7 |
| DeCoS | **26.5±1.4** | **36.0±2.4** | **33.7±1.2** | **23.1±4.6** |

with positive descriptions from GPT-4o-generated discriminative anchors. The three selected descriptions per class are fixed for evaluation.

The benchmark event name is used only to generate these definitions and is not included in the text provided to the evaluated model. We evaluate each of the three definition sets independently and average the resulting metrics, rather than averaging temporal scores before evaluation. Scoring and temporal processing are unchanged from the class-name evaluation. The complete definitions are listed in Appendix E.3.

### E.3 Complete Name-Free Definitions

The class names below are shown only as identifiers and are not included in the text provided to the evaluated models. The three definition sets are evaluated independently and their metrics are averaged.

**UCF-Crime.**

**Abuse**
1. someone mistreating or harming another person
2. A person is striking another person with visible force and aggression.
3. An individual is restraining another person against their will with clear physical dominance.

**Arrest**
1. police officers detaining and restraining a person
2. Police officers place handcuffs on an individual's wrists while they are standing against a wall.
3. A person is escorted by officers with hands restrained behind their back.

**Arson**
1. a person deliberately setting property on fire
2. A person is seen igniting a flammable liquid on a structure with a lighter.
3. An individual throws a lit object into a building, causing flames to erupt.

**Assault**
1. a violent physical attack on a person

Table 13: Event names used in the unseen-name evaluation. The count for each dataset gives the number of unseen event names over the total number of anomaly categories.

| Dataset | Unseen event names |
|---|---|
| UCF-Crime $(9/13)$ | Abuse, Arrest, Arson, Burglary, Fighting, RoadAccidents, Shooting, Shoplifting, Stealing |
| XD-Violence $(3/6)$ | Abuse, Fighting, Shooting |
| MSAD $(6/11)$ | Fighting, Object_falling, People_falling, Shooting, Traffic_accident, Water_incident |

2. One person repeatedly strikes another with fists in a public area.
3. A group surrounds an individual, pushing and hitting them aggressively.

**Burglary** 1. someone breaking into a building to take valuables
2. A person is seen breaking a window to enter a building at night.
3. An individual is carrying valuables out of a house through a forced-open door.

**Explosion** 1. a sudden violent burst of flames smoke and flying debris
2. A sudden burst of flames and smoke rapidly expands outward from a central point.
3. Debris and objects are violently propelled outward in all directions from an epicenter.

**Fighting** 1. people physically attacking each other
2. Two individuals exchange rapid punches and kicks in close proximity.
3. A person aggressively grabs another, pulling them forcefully towards the ground.

**RoadAccidents** 1. vehicles crashing or colliding on a road
2. A vehicle collides with another vehicle, causing visible damage or sudden stop.
3. A pedestrian is struck by a moving vehicle, resulting in immediate physical impact.

**Robbery** 1. someone taking belongings by force or threat
2. A person forcefully grabs a bag from another individual and runs away.
3. An individual points a weapon at a cashier while demanding money from the register.

**Shooting** 1. a person firing a gun
2. A person is seen discharging a firearm with visible muzzle flash and recoil.
3. An individual aims a gun and a projectile visibly impacts a target or surface.

**Shoplifting** 1. a customer secretly hiding store goods to avoid paying
2. An individual conceals merchandise under clothing while glancing around for observers.
3. A person quickly exits the store with items in hand, bypassing the checkout area.

**Stealing** 1. a person taking belongings that are not theirs
2. A person discreetly places an item into their bag without paying.
3. An individual quickly exits a store with merchandise in hand, bypassing the checkout.

**Vandalism** 1. a person deliberately damaging or defacing property
2. A person is spray-painting graffiti on a public wall.
3. An individual is smashing a car window with a blunt object.

**XD-Violence.**

**Fighting** 1. people physically attacking each other
2. Two individuals are exchanging rapid punches and kicks in close proximity.
3. One person is forcefully pushing another against a wall while raising a fist.

**Shooting** 1. a person firing a gun
2. A person is seen aiming a firearm directly at another individual or group.
3. A muzzle flash is visible as a firearm is discharged in a public setting.

**Riot** 1. a violent crowd disturbance
2. Crowds of people are aggressively clashing with law enforcement, throwing objects and pushing barriers.
3. Multiple individuals are seen vandalizing property and setting fires in a chaotic street scene.

**Abuse** 1. someone mistreating or harming another person
2. A person is forcefully grabbing another's arm while the latter visibly struggles to break free.
3. One individual is repeatedly striking another who is visibly trying to shield themselves.

**Car accident** 1. vehicles crashing or colliding on a road
2. Two vehicles collide, causing visible damage and debris on the road.
3. A car abruptly stops after hitting another vehicle, triggering airbags.

**Explosion** 1. a sudden violent burst of flames smoke and flying debris
2. A sudden burst of flames and smoke rapidly expands outward from a central point.
3. Debris is violently propelled in all directions following a bright flash.

## MSAD.

**Assault** 1. a violent physical attack on a person
2. One person repeatedly strikes another with fists in a public setting.
3. A group surrounds an individual, pushing and hitting them aggressively.

**Explosion** 1. a sudden violent burst of flames smoke and flying debris
2. A sudden burst of flames and smoke rapidly expands outward from a central point.
3. Debris is violently ejected in all directions following a bright flash of light.

**Fighting** 1. people physically attacking each other
2. Two individuals are exchanging rapid punches and kicks in close proximity.
3. One person is aggressively grappling another, attempting to throw them to the ground.

**Fire** 1. flames and smoke spreading through a scene
2. Flames visibly flicker and emit smoke from a concentrated area.
3. Bright orange and red hues rapidly spread across a surface.

**Object_falling** 1. an object dropping or toppling to the ground
2. An object visibly detaches from a surface and descends rapidly to the ground.
3. A stationary object suddenly moves downward without external force visible.

**People_falling** 1. a person losing balance and falling down
2. A person suddenly collapses to the ground without any external force.
3. An individual loses balance and falls from a standing position to the floor.

**Robbery** 1. someone taking belongings by force or threat
2. A person forcibly takes a bag from another individual who visibly resists.
3. An individual brandishes a weapon while demanding items from a store clerk.

**Shooting** 1. a person firing a gun
2. A person is seen aiming a firearm and discharging it with visible recoil.
3. Muzzle flash is observed from a gun held by an individual.

**Traffic_accident** 1. vehicles crashing or colliding on a road
2. A vehicle collides with another, causing visible damage or abrupt stops.
3. Pedestrians or cyclists are struck by a moving vehicle, resulting in sudden falls or injuries.

**Vandalism** 1. a person deliberately damaging or defacing property
2. A person is spray-painting graffiti on a public wall in broad daylight.
3. An individual is smashing car windows with a blunt object in a parking lot.

**Water_incident** 1. a person struggling or submerged in water
2. Water rapidly flooding an indoor space, visibly rising above floor level.
3. A burst pipe spraying water forcefully across a room or corridor.