WHAT DRIVES ANOMALY SCORES IN VLM-BASED VIDEO ANOMALY DETECTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly used for video anomaly detection (VAD), yet it remains unclear how much of a VAD system's anomaly-scoring behavior is inherited from its pretrained VLM and how much is introduced or transformed by downstream VAD processing. For example, if a system assigns higher anomaly scores to segments containing running people than to otherwise comparable segments without running, is this preference already present in the pretrained VLM, or does it emerge after the VLM is incorporated into the VAD system? Final detection accuracy alone cannot distinguish these possibilities. We introduce a systematic framework for tracing anomaly-scoring behavior across the pretrained-to-VAD transition. We characterize scoring behavior through *cue-score associations*: the relative ordering of anomaly scores between matched cue-present and cue-absent segments, while matching anomaly status and observed context. This score-scale-independent comparison allows pretrained VLMs and their corresponding VAD systems to be analyzed on the same video content despite differences in architectures and scoring functions, while preserving each VAD method's native inference procedure. Across 11 open-weight VLMs, 7 representative VAD methods, and 3 video-anomaly benchmarks, we find that VAD processing *selectively reshapes* pretrained scoring behavior: cue-score associations can persist, attenuate, disappear, reverse, or emerge after downstream processing, with these transformations depending jointly on the VLM and VAD method. Intermediate analysis identifies *where* these transformations arise within VAD pipelines, while temporal analysis shows *when* downstream processing changes anomaly evidence before, during, and after annotated events. These findings reveal that anomaly scoring in VLM-based VAD reflects both behavior inherited from the pretrained model and systematic modifications introduced by downstream detection mechanisms, providing a complementary view of VAD systems beyond final detection accuracy.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly used for video anomaly detection (VAD), providing visual and semantic knowledge through semantic reasoning, representation readout, evidence selection, or parameter adaptation (Ye et al., 2025; Dai et al., 2026; Cai et al., 2025; Shao et al., 2025; Wu et al., 2024; Zanella et al., 2024; Zhu et al., 2026; Xu et al., 2025; Kim et al., 2026). Yet an important question remains largely unexplored: *how much of a VAD system's anomaly-scoring behavior is inherited from its pretrained VLM, and how much is introduced or transformed by downstream VAD processing?* Standard metrics such as AUC and average precision evaluate the final detector, but do not show how its scoring behavior develops from the pretrained model.

Consider a VAD system that assigns higher anomaly scores to segments containing running people than to otherwise comparable segments without running people. We call this difference a *cue-score association*. The same association may already be present in the pretrained VLM and persist after VAD processing, or downstream processing may strengthen, weaken, eliminate, reverse, or newly introduce it (Gu et al., 2026; Senapati et al., 2026). These cases can yield similar final detection performance, but imply fundamentally different relationships between the pretrained model and the resulting detector. Final detection accuracy alone cannot distinguish these possibilities.

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.

We therefore ask: *which cue-score associations are already present in pretrained VLMs, and how are they transformed when these models are incorporated into VAD systems?* We compare pretrained VLMs and their corresponding VAD systems on the *same video segments*. For each visual or contextual cue, such as people, vehicles, or high motion, we compare naturally occurring matched pairs of cue-present and cue-absent segments that are otherwise comparable while matching anomaly status and observed context. This reveals whether a cue is associated with higher or lower anomaly scores before and after VAD processing. Because the comparison relies on within-pair score ordering rather than absolute score values, it does not require a common score scale across heterogeneous models or methods, while each VAD system retains its native inference procedure.

We refer to the collection of cue-score associations within a scoring function as its *anomaly-scoring behavior*, and to their differential change from the pretrained model to the VAD system as *selective reshaping* across the pretrained-to-VAD transition. Across 11 open-weight VLMs, 7 representative VAD methods, and 3 video-anomaly benchmarks, we find that downstream VAD processing neither uniformly preserves nor replaces pretrained scoring behavior. Instead, individual cue-score associations can persist, attenuate, disappear, reverse, or emerge after downstream processing, with these transformations depending jointly on the VLM and VAD method. Thus, the anomaly-scoring behavior of a VLM-based VAD system reflects both substantial behavior inherited from its pretrained model and systematic modifications introduced by the downstream detection mechanism.

We further trace *where* and *when* this selective reshaping occurs. Analysis of intermediate outputs identifies stages at which cue-score associations change within VAD pipelines, while temporal analysis of anomaly scores around annotated events reveals how downstream processing changes anomaly evidence before, during, and after anomalous events. These analyses move beyond whether a system detects anomalies to characterize how its anomaly-scoring behavior evolves from the pretrained VLM through the VAD pipeline. Our **main contributions** are:

i. We introduce a behavioral tracing framework for comparing pretrained VLMs with the VAD systems built on them. Using matched cue-present and cue-absent video segments, the framework quantifies cue-score associations on a common behavioral basis without requiring shared architectures or score scales, while preserving each VAD method's native inference procedure.

ii. Across 11 open-weight VLMs, 7 representative VAD methods, and 3 video-anomaly benchmarks, we identify *selective reshaping*: pretrained cue-score associations can persist, attenuate, disappear, reverse, or emerge after VAD processing, with the resulting transformations varying jointly across VLM and VAD combinations depending on the downstream method.

iii. We trace this reshaping through intermediate VAD outputs and across time, identifying *where* cue-score associations change within VAD pipelines and *when* anomaly evidence is assigned before, during, and after annotated anomalous events over the same video content.

Related work is reviewed in Appendix A; below, we present our framework and analysis.

2 METHOD

We compare each pretrained VLM with VAD systems built on it to characterize how anomaly-scoring behavior changes across the pretrained-to-VAD transition. Specifically, we examine *what* changes through cue-score associations, *where* changes arise across VAD stages, and *when* anomaly evidence is assigned around anomalous events. We use the same video content and matched comparisons throughout, while preserving each VAD method's native inference procedure. Here, *VAD processing* denotes all method-specific operations that transform the backbone's inputs, representations, parameters, or outputs into anomaly scores. Let x denote an evaluation segment, and let b, m, d, and c index backbones, VAD methods, datasets, and cues, respectively. We study eight contextual cues (person, vehicle, running, night, indoor, street, dense people, and high motion) and two event attributes (early event and long event). See Appendix C for more details.

2.1 CUE-LEVEL COMPARISON

Pretrained reference score. For each backbone b, we query its pretrained checkpoint on segment x with a fixed probe, p_0: "Does this video contain an abnormal event? Answer Yes or No.", and use

the Yes/No log-odds as its anomaly score:

$$s_b^{\mathrm{ref}}(x) = \log \frac{P_b(\mathrm{Yes} \mid x, p_0)}{P_b(\mathrm{No} \mid x, p_0)}, \tag{1}$$

where the two probabilities are those of the corresponding answer tokens at the first answer position, clipped to $[10^{-6}, 1 - 10^{-6}]$. Larger values indicate a stronger preference for an abnormal-event answer. The reference always uses the original pretrained checkpoint and serves as the baseline against which downstream changes are measured.

VAD score. Each VAD method m produces anomaly scores using its native inference procedure, potentially at a different temporal resolution. We therefore align its scores to evaluation segment x only after inference. If the method assigns score s_k to temporal unit I_k, its segment-level score is

$$s_{b,m}^{\mathrm{VAD}}(x) = \frac{\sum_k |I_k \cap x| \, s_k}{\sum_k |I_k \cap x|}, \tag{2}$$

where $|I_k \cap x|$ denotes the overlap duration. This projection changes only the evaluation representation, not the method's native sampling, evidence selection, inference, or scoring. For adapted-backbone methods, the reference remains the original pretrained checkpoint, whereas the VAD score comes from the adapted model and its downstream pipeline. Because anomaly scores can have different scales and semantics across backbones and VAD methods, we do not compare their absolute values. Instead, all subsequent comparisons use within-pair score ordering.

Matched cue comparison. For each cue c, we construct matched pairs (x_i^+, x_i^-), where x_i^+ contains the cue and x_i^- does not. Pairs are drawn from different videos and matched on anomaly status and prespecified observed contextual factors, including people category, scene, and motion category, with the factor corresponding to the cue under study excluded from the matching variables. Appendix C specifies the categories and pair-selection procedure. The same pairs are evaluated using the pretrained reference, final VAD scores, and available intermediate outputs, ensuring that changes reflect the scoring system rather than changes in the compared video content.

For any scoring function s, we define the *cue-score association* as

$$D_c(s) = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathrm{sign}\left(s(x_i^+) - s(x_i^-)\right), \tag{3}$$

where N_c is the number of valid pairs. Thus, $D_c(s) \in [-1, 1]$ is positive when cue-present segments more often receive higher scores, negative when they more often receive lower scores, and zero when neither ordering dominates. Because D_c depends only on within-pair score ordering, it is invariant to strictly increasing transformations of the scores. The measure therefore captures *behavioral ordering* rather than score magnitude, enabling comparisons across heterogeneous scoring systems without score calibration. For dataset d, we compute the pretrained and VAD cue-score associations as

$$D_{b,d,c}^{\mathrm{ref}} = D_c\big(s_b^{\mathrm{ref}}\big), \qquad D_{b,m,d,c}^{\mathrm{VAD}} = D_c\big(s_{b,m}^{\mathrm{VAD}}\big), \tag{4}$$

and define the downstream change

$$\Delta D_{b,m,d,c} = D_{b,m,d,c}^{\mathrm{VAD}} - D_{b,d,c}^{\mathrm{ref}}. \tag{5}$$

Hence, ΔD measures how a cue-score association is reshaped from the pretrained model to VAD: changes in magnitude indicate attenuation or amplification, while a sign change indicates reversal.

Statistical characterization. An association or change is considered *detectable* at adjusted $q < 0.05$. Inference accounts for dependence among segments from the same source video using a source-video cluster bootstrap with 2,000 replicates, percentile 95% confidence intervals, and two-sided bootstrap tests; multiple comparisons are controlled using Benjamini-Hochberg correction within each backbone-cue family Appendix C.3. We classify detectable transformations as *preserved*, *attenuated*, *amplified*, *suppressed*, *introduced*, or *reversed*, according to the direction and detectability of the pretrained and VAD associations. Cases in which neither association is detectable are classified as *no difference*. The complete classification criteria are given in Appendix C.

Across-cue summary. For each configuration, let $\boldsymbol{d}_{b,d}^{\mathrm{ref}}$ and $\boldsymbol{d}_{b,m,d}^{\mathrm{VAD}}$ collect the cue-level associations over the C eligible cues. We summarize their relationship using

$$\rho_{b,m,d} = \mathrm{Spearman}\left(\boldsymbol{d}_{b,d}^{\mathrm{ref}}, \boldsymbol{d}_{b,m,d}^{\mathrm{VAD}}\right), \tag{6}$$

which measures how well the relative ordering of cue associations is preserved, and

$$R_{b,m,d} = \left[\frac{1}{C} \sum_{c=1}^{C} \left(\Delta D_{b,m,d,c} \right)^2 \right]^{1/2}, \tag{7}$$

which measures the overall magnitude of behavioral reshaping across cues.

2.2 Tracing Changes Across Stages and Time

Changes across VAD stages. To identify *where* cue-score associations change within a VAD pipeline, we evaluate D_c at each stage that produces an anomaly score. Let $s_{b,m}^{(k)}$ denote the score after stage k, with $s_{b,m}^{(0)} = s_b^{\mathrm{ref}}$. After mapping each stage output to the common evaluation segments using Eq. 2, we compute

$$D_c^{(k)} = D_c\left(s_{b,m}^{(k)}\right), \qquad k = 0, \ldots, K. \tag{8}$$

We measure both cumulative change from the pretrained reference $D_c^{(0)}$,

$$\Delta D_{c,\mathrm{ref}}^{(k)} = D_c^{(k)} - D_c^{(0)}, \tag{9}$$

and the change introduced between adjacent stages,

$$\Delta D_c^{(k-1)\to k} = D_c^{(k)} - D_c^{(k-1)}. \tag{10}$$

The first identifies how far each stage has moved from pretrained behavior, whereas the second localizes changes to individual transitions. Across-cue summaries and detectable-change counts are reported in Appendix F. Because a stage may simultaneously alter visual evidence, prompts, representations, temporal processing, or model parameters, these analyses identify *where* scoring behavior changes but do not establish component-level causality.

Temporal allocation of anomaly evidence. To examine *when* anomaly evidence is assigned, we compare pretrained and VAD score timelines around the same annotated events. For each video, let F^{cov} denote frames for which both scores are available. We convert the pretrained and VAD scores separately to within-video percentile ranks $r(t) \in [0, 1]$ over these shared frames, removing score-scale differences while preserving temporal score ordering.

For each event, let E denote its covered frames, and let F^- and F^+ denote the covered frames in the 10 s intervals immediately before and after the event, excluding other events and their surrounding intervals. Let F^{base} contain the remaining covered background frames, excluding all annotated events and these surrounding intervals. We quantify relative anomaly evidence before, during, and after the event as

$$\begin{aligned} L^{\mathrm{pre}} &= \mathbb{E}[r(t) \mid t \in F^-] - \mathbb{E}[r(t) \mid t \in F^{\mathrm{base}}], \\ L^{\mathrm{ev}} &= \mathbb{E}[r(t) \mid t \in E] - \mathbb{E}[r(t) \mid t \in F^{\mathrm{base}}], \\ L^{\mathrm{post}} &= \mathbb{E}[r(t) \mid t \in F^+] - \mathbb{E}[r(t) \mid t \in F^{\mathrm{base}}]. \end{aligned} \tag{11}$$

$\mathbb{E}[\cdot]$ denotes the average over the specified frames. Positive values indicate higher relative anomaly scores than background. We also measure how strongly scores concentrate within the event:

$$C^{\mathrm{ev}} = \frac{\sum_{t \in E} r(t) / \sum_{t \in F^{\mathrm{cov}}} r(t)}{|E|/|F^{\mathrm{cov}}|}. \tag{12}$$

The numerator is the fraction of total relative score assigned to event frames, while the denominator is the fraction of covered frames belonging to the event. Thus, $C^{\mathrm{ev}} = 1$ corresponds to score allocation proportional to event duration, whereas $C^{\mathrm{ev}} > 1$ indicates greater concentration within the event. We first average these quantities across events within each video and then across videos. For each temporal quantity T, we define

$$\Delta T = T^{\mathrm{VAD}} - T^{\mathrm{ref}},$$

so that ΔT directly measures how downstream VAD processing changes the temporal allocation of anomaly evidence. We additionally compare matched early-versus-late and long-versus-short events using the sign-based estimator in Eq. 3, yielding D_{early} and D_{long}. Implementation details are provided in Appendix B.3. Below we present our experiments and analysis.

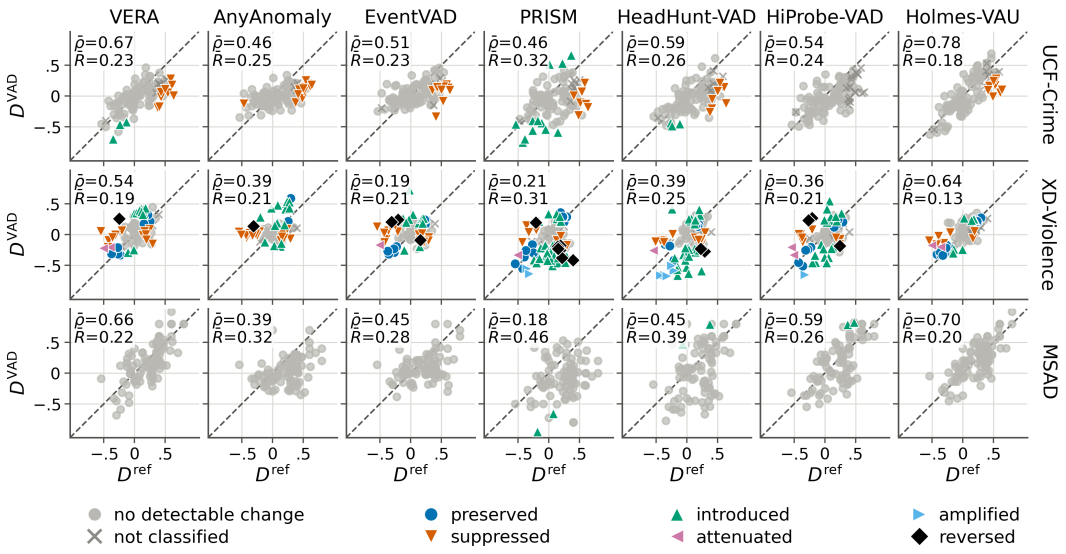


Figure 1: Pretrained-reference versus final VAD cue-score associations. Rows show datasets and columns show VAD methods. Each point represents one backbone-cue cell, with $x = D^{\text{ref}}_{b,d,c}$ and $y = D^{\text{VAD}}_{b,m,d,c}$, colored by its change category. The dashed line ($y = x$) denotes unchanged cue associations. Insets report the mean $\bar{\rho}$ and $\bar{R}$ over backbones, where larger $\bar{\rho}$ and smaller $\bar{R}$ indicate greater preservation of pretrained cue ordering. All $\bar{\rho}$ values are positive, showing that VAD systems generally retain pretrained cue ordering, but to a method-dependent degree. Holmes-VAU shows the strongest overall preservation ($\bar{\rho} \geq 0.64$, $\bar{R} \leq 0.20$), whereas PRISM shows the largest overall reshaping ($\bar{\rho} \leq 0.46$, $\bar{R} \geq 0.31$). Detectable changes are selective: UCF-Crime is dominated by suppressions of pretrained preferences, XD-Violence shows a broader mixture of change types and generally lower $\bar{\rho}$, while MSAD exhibits no detectable ΔD despite cue-level variation.

3 Experiment

3.1 Experimental Setup

Datasets, VLMs, and VAD methods. We evaluate the proposed framework on the test splits of UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), and MSAD (Zhu et al., 2024), covering surveillance, web and movie, and multi-scenario anomaly videos. We study eleven open-weight VLMs spanning multiple model families and scales, including Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), InternVL3 (Zhu et al., 2025), SmolVLM2 (Marafioti et al., 2025), GLM-4.1V (Hong et al., 2025), VideoLLaMA3 (Zhang et al., 2025a), VideoChat3 (Li et al., 2026), and JoyAI-VL (Video Understanding Team of JoyAI-VL @ Joy Future Academy, JD, 2026). Open-weight models are required because our framework accesses model likelihoods, hidden representations, attention heads, and adapted backbones. We evaluate seven representative VAD methods covering four processing families: reasoning (VERA (Ye et al., 2025)), visual evidence selection (AnyAnomaly (Ahn et al., 2026), EventVAD (Shao et al., 2025)), representation readout (PRISM (CHEN et al., 2026), HeadHunt-VAD (Cai et al., 2026), HiProbe-VAD (Cai et al., 2025)), and backbone adaptation (Holmes-VAU (Zhang et al., 2025b)). Each method is implemented using compatible backbones while preserving its native inference procedure. Details of models, adaptations, reconstructed components, and detection performance are provided in Appendix B.

Evaluation protocol. Videos are divided into non-overlapping 10 s evaluation segments, each represented by eight uniformly sampled frames. We evaluate eight visual and contextual cues together with two event attributes using the matched-pair behavioral framework described in Sec. 2.1. Cue labels were retained only when all three independent VLM annotators, DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), Ling-3.0-flash-VL[1], and Gemma-4-26B-A4B-IT (Team et al., 2026), agreed. None of these models was included among the evaluated backbones. Details of the annota-

[1]Access from `https://huggingface.co/inclusionAI/Ling-3.0-flash-VL`

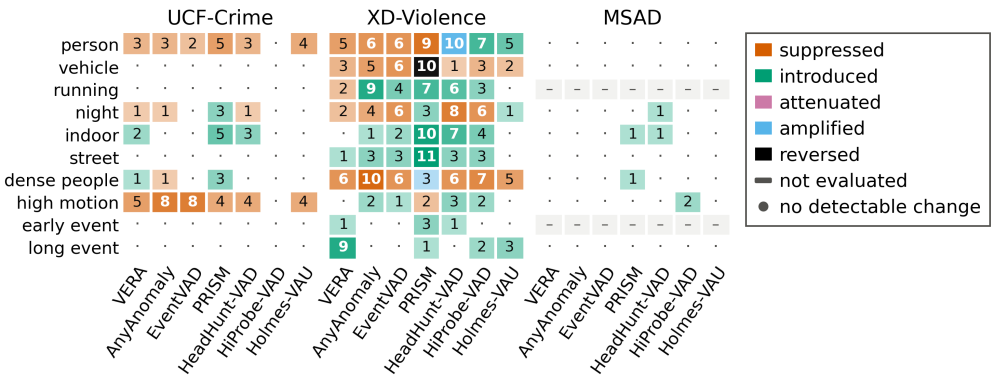


Figure 2: Change categories across cues, VAD methods, and datasets. Each cell is colored by the dominant change category among the 11 backbones; the number gives the number of backbones assigned a change category, and darker shading indicates more backbones. Dots indicate that no backbone is assigned a change, while dashes indicate that the cue was not evaluated. On UCF-Crime, detectable changes are concentrated on *person* and *high motion*, with predominantly *suppressed* preferences, while *dense people* shows both *suppressed* and *introduced* changes across methods. On XD-Violence, *suppressed* changes for *person*, *vehicle*, and *dense people* coexist with widespread *introduced* changes for *running*, *indoor*, *street*, and *high motion*, showing that VAD processing both removes inherited cue-score associations and introduces new ones. MSAD exhibits no detectable ΔD, consistent with its smaller number of matched pairs per cue.

tion procedure are provided in Appendix C. Each cue-present segment is paired with a cue-absent segment from a different video with the same anomaly status and, where applicable, the same people count, scene, and motion level. For each backbone, method, and dataset, $\rho_{b,m,d}$ and $R_{b,m,d}$ summarize all eligible cues and event attributes. Detection performance is reported as frame-level AUC on UCF-Crime and MSAD and average precision on XD-Violence. Complete details of cue annotation, matched-pair construction, implementation, and statistical analysis are provided in Appendix B.

3.2 Behavioral Reshaping After VAD Processing

Behavioral retention. Despite downstream processing, VAD systems retain a considerable portion of the pretrained VLM's cue-level scoring behavior. Across all methods and datasets, the cue-level association between pretrained and final VAD behavior remains predominantly positive (Fig. 1), indicating that cues receiving relatively higher anomaly scores in the pretrained model often continue to do so after VAD processing. However, the degree of retention varies substantially across methods, backbones, and datasets. On UCF-Crime, average cue-order correlations range from 0.46 (AnyAnomaly and PRISM) to 0.78 (Holmes-VAU), with individual backbone-method combinations spanning an even wider range. Lower correlations are mainly observed for weaker 2B backbones, whereas stronger VLMs generally preserve pretrained behavioral ordering more consistently.

Selective behavioral reshaping. Figs. 2 and 3 show that VAD processing selectively reshapes individual cue-score associations rather than uniformly preserving or replacing them. On UCF-Crime, pretrained preferences for *high motion* and *person* are suppressed by most VAD methods, while *dense people* is both suppressed and introduced. In contrast, XD-Violence exhibits a broader mix of transformations: pretrained preferences for *person*, *vehicle*, and *dense people* are often suppressed, while new associations emerge for cues such as *running*, *indoor*, *street*, and *high motion*. Across cues and datasets, suppression, attenuation, amplification, introduction, and reversal coexist, showing that downstream processing modifies specific behavioral components rather than globally shifting anomaly scores. Moreover, the same cue can be transformed in different directions by different methods, as illustrated by *person* on XD-Violence.

Dependence on VLM-VAD combinations. Behavioral reshaping varies with the combination of pretrained VLM and downstream VAD method. Holding the backbone fixed, different VAD methods can produce markedly different transformations of the same cue (Fig. 3; Tables 11 and 12 in Appendix E.2). For example, on UCF-Crime, for GLM-4.1V-9B, the pretrained preference for *high motion* is suppressed by AnyAnomaly, EventVAD, PRISM, and Holmes-VAU, whereas no detectable

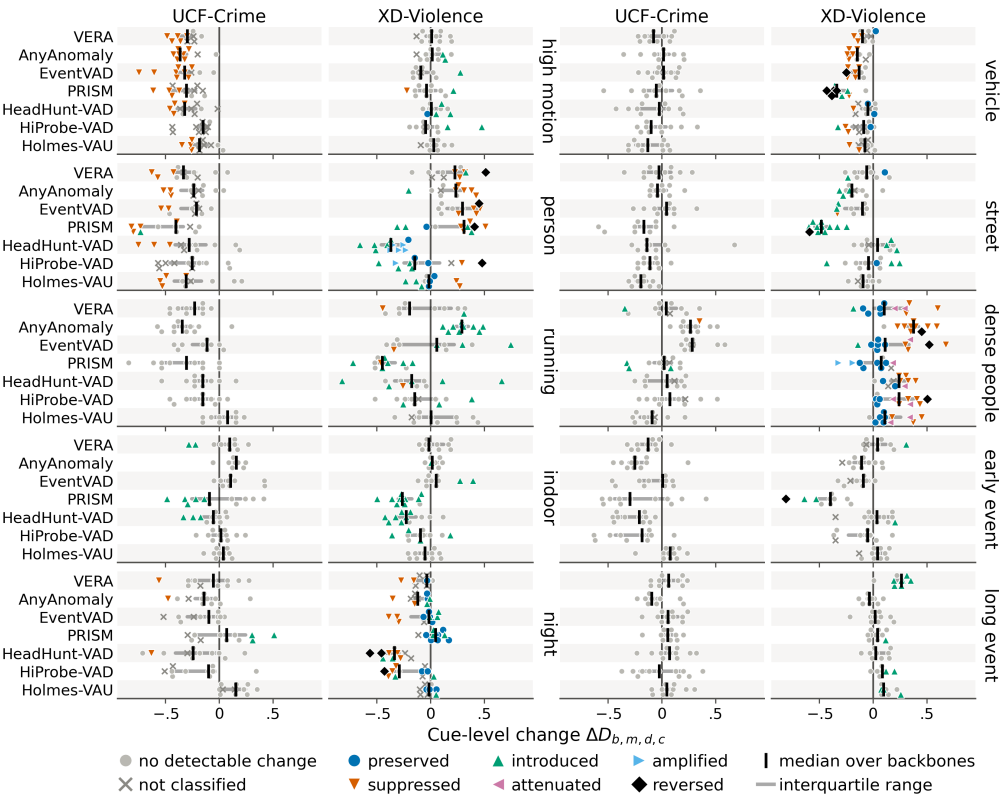


Figure 3: Cue-level changes $\Delta D_{b,m,d,c}$ across backbones, with points coloured by change category; gray circles indicate no detectable change and gray crosses unclassified cases. The black tick and gray bar show the median and interquartile range across backbones for each method. Columns correspond to UCF-Crime and XD-Violence, with eight cues and two attributes in the panel. The same cue can change in opposite directions across methods, as shown by *person* on XD-Violence, highlighting the selective and method-dependent reshaping of cue-score associations.

change is found under VERA, HeadHunt-VAD, and HiProbe-VAD. Conversely, for Qwen3-VL-4B, PRISM uniquely introduces a negative association with *indoor* and a positive association with *night*. More broadly, even with the same backbone, cue-order correlations vary substantially across VAD methods, ranging from 0.10 to 0.85 for Qwen2.5-VL-3B on UCF-Crime. These results show that downstream VAD processing does not impose a uniform behavioral transformation: the resulting anomaly-scoring behavior varies with both the pretrained VLM and the VAD pipeline.

Beyond detection performance. Behavioral reshaping captures aspects of VAD systems that conventional detection metrics do not reveal. Across all 231 configurations, cue-order retention $\rho_{b,m,d}$ shows only a moderate association with the within-dataset detection-performance rank (Spearman .51; Fig. 4), and this association weakens among stronger detectors (.29 for configurations above 0.75 AUC or .50 AP). Within individual methods, it is absent for VERA ($-.17$) and Holmes-VAU (.00) and weak for PRISM (.14), while it remains moderate for the other evidence and representation methods (.45 to .48). Moreover, similar cue-order retention can accompany very different amounts of behavioral modification: on UCF-Crime, HiProbe-VAD preserves cue ordering (mean $\rho_{b,m,d} = .54$) with no detectable cue-level change, whereas PRISM and AnyAnomaly show comparable retention (.46) but 16 detectable changes each. Thus, detection performance does not by itself determine whether a VAD system preserves, suppresses, or introduces pretrained scoring behavior.

3.3 Tracing Behavioral Reshaping

Where reshaping occurs. Behavioral reshaping is concentrated at the earliest VAD transformation rather than accumulated gradually across the pipeline (Fig. 5; Tables 19 and 20 in Appendix F). For all four methods with intermediate scores, the first transition has the lowest agreement with

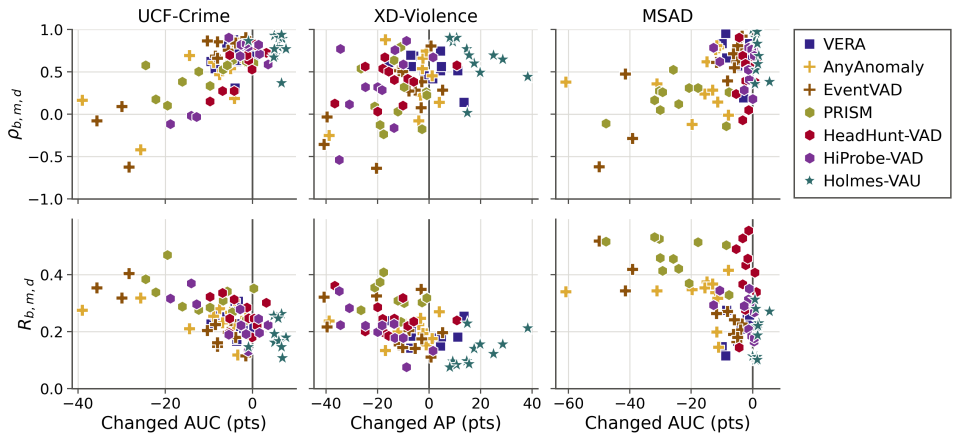


Figure 4: Detection-performance change relative to the pretrained reference score of the same backbone (AUC; AP on XD-Violence) versus cue-level behavioral retention. Each point represents one configuration (b, m, d); columns show datasets, rows show $\rho_{b,m,d}$ (top) and $R_{b,m,d}$ (bottom), with colour denoting VAD method and marker shape its processing family. Lower ρ and higher R indicate greater departure from pretrained cue-scoring behavior. Larger performance losses generally coincide with greater behavioral departure, but the relationship remains loose and method-dependent (e.g., Holmes-VAU improves detection on every dataset while its ρ varies across backbones).

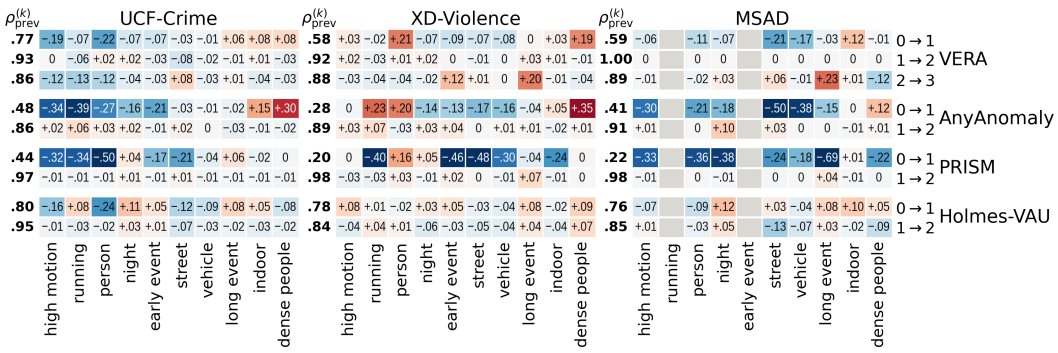


Figure 5: Stage-wise changes in cue-score associations for the four VAD methods exposing intermediate scores (VERA, AnyAnomaly, PRISM, and Holmes-VAU). Rows show transitions $k{-}1 \to k$, with stage 0 denoting the pretrained reference; columns show cues for each dataset. Cells report the mean $\Delta D_c^{(k-1)\to k}$ over backbones (Eq. 10), with blue/red indicating decreases/increases and grey denoting fewer than ten valid pairs. The left values give the mean $\rho^{(k)}_{\text{prev}}$. For VERA, AnyAnomaly, and PRISM, most cue-level reshaping occurs at the first transition and later stages produce smaller changes; for Holmes-VAU the two transitions are of similar size on XD-Violence and MSAD.

the preceding stage (mean $\rho^{(1)}_{\text{prev}}$ from .20 to .80), whereas subsequent transitions largely preserve the established ordering ($\rho^{(k)}_{\text{prev}} \geq .84$). Later stages are not negligible, however: VERA's temporal smoothing adds further changes, and for Holmes-VAU the two transitions are of similar size on XD-Violence and MSAD. The operation responsible for the main departure differs across pipelines-guided reflection, key-frame question answering, semantic readout, or parameter adaptation-so behavioral modification is localized at each method's first method-specific step rather than uniformly distributed across the pipeline.

When reshaping occurs. VAD processing also reshapes *when* anomaly evidence is assigned (Fig. 6; Table 21 in Appendix G). On UCF-Crime, all six frozen-backbone methods weaken the pretrained rise in anomaly evidence during annotated events, whereas Holmes-VAU largely preserves the pretrained temporal profile. On XD-Violence the pattern is method-dependent: the frozen-backbone methods except PRISM again weaken the during-event evidence, while Holmes-VAU raises it. Together with the cue-level results, these findings show that behavioral reshaping has distinct temporal and cue-level dimensions: on UCF-Crime, Holmes-VAU alters which cues receive high anomaly

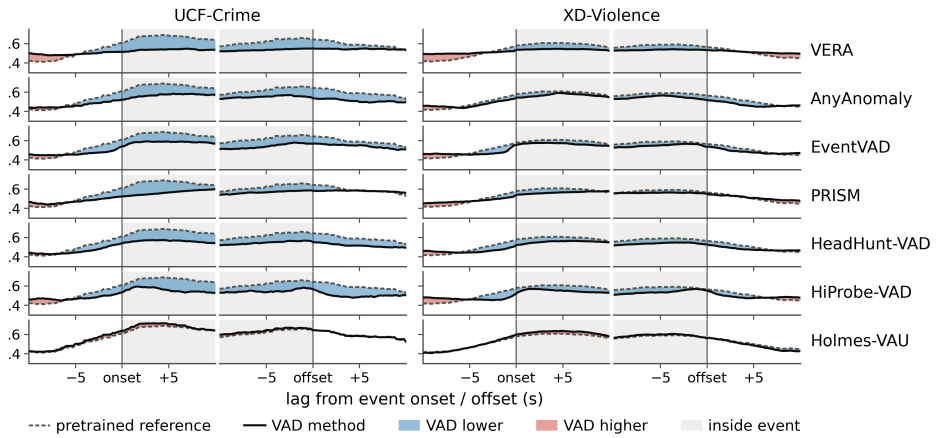


Figure 6: Event-aligned anomaly evidence.Each panel shows within-video percentile anomaly scores around event onset (left) and offset (right), averaged over events, videos, and backbones. The dashed curve is the pretrained reference and the solid curve is the final VAD score; blue/red shading indicates lower/higher VAD scores than the reference, and grey marks the event interval. On UCF-Crime, all six frozen-backbone methods weaken the pretrained rise in anomaly evidence during the event, whereas the adapted Holmes-VAU closely follows it. XD-Violence shows the same weakening for all frozen-backbone methods except PRISM, which follows the reference, while Holmes-VAU raises the evidence during the event. MSAD is shown in Fig. 10 (Appendix G).

scores (14 detectable cue-level changes) without detectably changing when evidence is concentrated, whereas HiProbe-VAD changes the temporal profile without any detectable cue-level change.

4 Discussion

Behavioral transformation depends on the VLM-VAD combination. The same backbone can exhibit markedly different cue transformations under different VAD methods, while the same method can behave differently across backbones. This indicates that reshaping emerges from the interaction between the pretrained VLM and the VAD pipeline rather than from either component alone. Moreover, the transformations are selective: a method can suppress one pretrained preference while introducing another, showing that downstream processing changes specific behavioral components rather than globally shifting anomaly scores.

Reshaping is localized and multidimensional. Most cue-level changes occur at the first method-specific transition, while later stages generally preserve the established ordering. This suggests that major behavioral changes are introduced early rather than accumulated uniformly through the pipeline. At the same time, cue-level and temporal behavior are not interchangeable. Holmes-VAU changes multiple cue-level associations while largely preserving the pretrained temporal profile, whereas HiProbe-VAD changes temporal allocation without detectable cue-level changes. VAD processing therefore affects both *what* receives anomaly evidence and *when* that evidence is assigned, along partly independent dimensions.

Implications for evaluation. These findings show why detection performance alone is insufficient to characterize VLM-based VAD. The only moderate relationship between cue-order retention and detection performance, which weakens among stronger detectors, shows that similar detection performance can accompany substantially different behavioral transformations. Reporting pretrained-to-VAD behavioral changes alongside conventional detection metrics can therefore reveal what a system inherits from its backbone and what its downstream design changes.

Limitations and future directions. Our study covers eleven open-weight VLMs, seven VAD methods, three benchmarks, and a predefined set of cues and event attributes. Matched pairs control observed context but do not establish causal cue effects, and the limited number of pairs on MSAD reduces the ability to detect cue-level changes. Moreover, evaluated methods are adapted to a common protocol rather than reproduced exactly in every implementation. Future work could extend the framework to richer concepts, broader anomaly domains, and interventional analyses that directly test how downstream mechanisms transform pretrained behavior.

REFERENCES

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.

Muaz Al Radi and Sajid Javed. Hierarchical vision-language model with comprehensive language description for video anomaly detection. *Knowledge-Based Systems*, pp. 115466, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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.

Kyle Buettner and Adriana Kovashka. Investigating the role of attribute context in vision-language models for object recognition and detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5462–5472. IEEE, 2024.

Zhaolin Cai, Fan Li, Ziwei Zheng, and Yanjun Qin. Hiprobe-vad: Video anomaly detection via hidden states probing in tuning-free multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 592–601, 2025.

Zhaolin Cai, Fan Li, Ziwei Zheng, Haixia Bi, and Lijun He. Headhunt-vad: Hunting robust anomaly-sensitive heads in mllm for tuning-free video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19835–19843, 2026.

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5549–5559. IEEE, 2023.

YUANTONG CHEN, Zhengyan Ding, and YanFeng Shang. PRISM: Training-free video anomaly detection via intrinsic statistical modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0ZhgSSbA6i.

Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Reza Esfandiarpoor, Cristina Menghini, and Stephen Bach. If clip could talk: Understanding vision-language model representations through their preferred concept descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9797–9819, 2024.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *International Conference on Learning Representations*, volume 2025, pp. 61715–61746, 2025.

Jihao Gu, Kun Li, He Wang, and Kaan Akşit. Text-guided fine-grained video anomaly understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1251–1261, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Hamza Karim and Yasin Yilmaz. Semvad: Fusing semantic and vision features for weakly supervised video anomaly detection. *Transactions on Machine Learning Research*, 2026.

Dongjun Kim, Changjae Oh, Andrea Cavallaro, and Jeonghoon Mo. Context-structured video anomaly detection with large vision-language models. *arXiv preprint arXiv:2607.19077*, 2026.

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, 2025a.

Shuo Li, Fang Liu, Licheng Jiao, Zehua Hao, Jiahao Wang, Long Sun, Lingling Li, Xu Liu, and Puhua Chen. Promptvad: Abnormal prompt via vision-language model. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Hyeongmuk Lim and Youngbum Hur. Corevad: A contextual reasoning framework for training-free video anomaly detection. In *International Conference on Pattern Recognition*, pp. 418–432. Springer, 2026.

Zhiming Liu, Yujie Wei, Lei Feng, Xiu Su, Xiaobo Xia, Weili Guan, Zeke Xie, and Shuo Yang. Do all individual layers help? an empirical study of task-interfering layers in vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9597–9607, June 2026a.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026b.

Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24500–24510. IEEE, 2023.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of lvlms by contrasting chain-of-embedding. In *International Conference on Learning Representations*, volume 2026, pp. 94723–94754, 2026.

Yiming Ma, Hongkun Yang, Lionel Z. Wang, Bin Chen, Weizhi Xian, and Jianzhi Teng. Dear: Fine-grained vlm adaptation by decomposing attention head roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31514–31523, June 2026.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

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.

Yanting Pei, Minhao Hao, Yaping Huang, and Qi Zou. Dr-vad: Definition-guided reasoning for training-free video anomaly detection. *Neurocomputing*, pp. 132836, 2026.

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

Maan Qraitem, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Web artifact attacks disrupt vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1048–1057. IEEE, 2025.

Debi Prasad Senapati, Santosh Kumar Pani, Santos Kumar Baliarsingh, Prabhu Prasad Dev, and Hrudaya Kumar Tripathy. Ca-vad: Caption aware video anomaly detection in surveillance videos. *Journal of Visual Communication and Image Representation*, 111:104521, 2025.

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.

Shengjie Shen, Ziteng Guo, Yahui Li, Liejun Wang, and Zhiqing Guo. Trclip-vad: Weak supervised video anomaly detection by improving clip training with text rewriting. *Neural Networks*, pp. 108951, 2026.

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.

Chenchen Tao, Xiaohao Peng, Chong Wang, Jiafei Wu, Puning Zhao, Jun Wang, and Jiangbo Qian. Learning suspected anomalies from event prompts for video anomaly detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(5):1–20, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Video Understanding Team of JoyAI-VL @ Joy Future Academy, JD. Joyai-vl-interaction: Real-time vision-language interaction intelligence. Technical report, Joy Future Academy, JD, June 2026.

Binghui Wang, Chunjuan Yan, Da Song, Lanxiao Li, and Chuanxu Wang. Prompt-guided fine-grained semantic alignment for weakly supervised video anomaly detection. *Knowledge-Based Systems*, pp. 115944, 2026.

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

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, 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, 2024.

Chen Xu, Zheng Wang, Xing Xu, Jingkuan Song, Zhe Sun, and Andrzej Cichocki. A2vad: Attribute-augmented prompt learning for weakly supervised video anomaly detection. *Pattern Recognition*, pp. 113987, 2026a.

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.

Guanghui Xu, Sifan Long, Hongwei Ge, Enxuan Gu, Zidi Li, and Zhaoqin Wang. Semantic consistency learning across temporal scales for weakly supervised video anomaly detection. *Knowledge-Based Systems*, pp. 115903, 2026b.

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, 2024.

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.

Jun Yi, Xu Sun, Ye Yu, Wei Liu, Wei Jia, and Jun Yu. Mrtp: Multiscale video anomaly detection with representative text prompt. *Image and Vision Computing*, pp. 105959, 2026.

Chenyu You, Yifei Mint, Weicheng Dai, Jasjeet S Sekhon, Lawrence Staib, and James S Duncan. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26140–26150. IEEE, 2024.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853. IEEE, 2025b.

Yi Zhang, Rui Zhu, Channi Li, Xiaoxu Li, Zhanyu Ma, and Jing-Hao Xue. Coda: Co-adaptive dual-path alignment for vision-language models. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 2100–2108. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/234. URL https://doi.org/10.24963/ijcai.2026/234. Main Track.

Lingru Zhou, Peng Wu, Manqing Zhang, Qingsheng Wang, Guansong Pang, and Peng Wang. Targetvau: Multimodal anomaly-aware reasoning for target behavior understanding in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13710–13718, 2026.

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.

Shu Zou, Xinyu Tian, Lukas Wesemann, Fabian Waschkowski, Zhaoyuan Yang, and Jing Zhang. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4223–4233. IEEE, 2026.

APPENDIX CONTENTS

A Related Work **15**

B Experimental Setup **16**

B.1 Datasets 16

B.2 Models 16

B.3 Evaluation Protocol 17

C Framework Details **18**

C.1 Cue Definitions 18

C.2 Matched Pair Construction 18

C.3 Statistical Testing 19

C.4 Transformation Categories 20

D Implementation Details **20**

D.1 Backbone-specific Details 20

D.2 Intermediate Outputs 22

D.3 Hyperparameters 22

D.4 Compute 22

E Additional Behavioral Results **22**

E.1 Reference versus VAD Comparisons 22

E.2 Cue-Level Changes per Dataset 23

E.3 Cross-Dataset Comparisons 28

E.4 Robustness and Sensitivity Analyses 29

E.5 Layered Analysis 30

E.6 Sources of Variation in Cue-Level Changes 33

E.7 Fine-Tuning versus Frozen-Backbone Methods 33

F Stage-wise Behavioral Analysis **34**

G Temporal Behavioral Analysis **35**

H Detection Performance and Correlation Analysis **37**

H.1 Detection Performance of the Pretrained Score 37

H.2 Cue-Related Differences and Detection Performance 37

I LLM Usage Declaration **38**

J Ethics Statement **38**

K Reproducibility Statement **38**

A Related Work

VLMs for VAD. Recent work has increasingly incorporated VLMs into VAD, motivated by the broad semantic knowledge acquired through large-scale multimodal pretraining. Early approaches such as VadCLIP (Wu et al., 2024) transfer CLIP's visual-language representations to weakly supervised VAD, using fine-grained image-text associations to improve anomaly localization and recognition. Other methods exploit language-guided semantic information, prompt learning, or textual descriptions to connect video content with anomalous events (Yang et al., 2024; Xu et al., 2026a; Shen et al., 2026). More recent approaches use larger VLMs for anomaly reasoning, fine-grained prompting, visual evidence selection, or temporal reasoning (Zhang et al., 2025b; Zou et al., 2026). For example, Holmes-VAU (Zhang et al., 2025b) combines anomaly-focused temporal sampling with hierarchical anomaly understanding across clip, event, and video levels, while other recent methods integrate semantic reasoning with spatio-temporal interaction modeling or definition-guided temporal reasoning (Zhou et al., 2026; Pei et al., 2026). These works demonstrate that pretrained multimodal knowledge can provide richer semantic information for anomaly detection than conventional visual representations (Karim & Yilmaz, 2026; Al Radi & Javed, 2026).

Our work addresses a different question from these VLM-based VAD methods. Existing approaches primarily ask *how to use pretrained VLMs to build better anomaly detectors*, with success measured mainly by final detection performance. We instead ask *what anomaly-scoring behavior is already present in a pretrained VLM and how that behavior differs after the VLM is incorporated into a VAD pipeline*. Rather than proposing another detector, we use existing VAD systems as downstream settings in which to study the behavior of their pretrained backbones. We therefore use their architectures and inference mechanisms as downstream settings for examining how pretrained behavior is preserved or modified.

Semantic cues and context in VAD. A related line of work improves VAD by explicitly incorporating semantic information about objects, actions, scene context, or anomaly categories. CLIP-based methods use textual prompts or semantic prototypes to associate visual observations with anomaly descriptions (Li et al., 2025b; Qiu et al., 2025; Li et al., 2025a; Yi et al., 2026; Wang et al., 2026). Other approaches generate or use textual descriptions of video content to identify suspicious events and improve semantic discrimination (Chen et al., 2023; Liu et al., 2023; Tao et al., 2026; Al Radi & Javed, 2026; Senapati et al., 2025). More recent methods explicitly model environmental, object-level, or temporal context to make anomaly detection more sensitive to the surrounding scene (Wang & Zhao, 2026; Dai et al., 2026; Lim & Hur, 2026; Mo et al., 2026; Xu et al., 2026b). These studies demonstrate that anomaly scores can depend on semantic and contextual information beyond low-level appearance or motion.

Our work uses such visual and contextual cues for a different purpose. Rather than introducing semantic cues as additional information for improving detection, we use naturally occurring cues as *probes of anomaly-scoring behavior*. For a cue such as people, vehicles, running, or high motion, we compare matched video segments with and without the cue and examine whether the cue is associated with different anomaly scores. Crucially, we perform this comparison both for the pretrained VLM and for VAD systems built on that VLM. This allows us to assess whether a cue-related score difference is already visible in the pretrained score or differs after downstream VAD processing. Thus, semantic cues serve as a common basis for comparing model behavior across the VAD pipeline rather than as components of a new detection mechanism.

Understanding pretrained multimodal model behavior. Beyond VAD, recent studies have shown that pretrained VLMs can respond differently to inputs containing particular semantic concepts, contextual information, or visual attributes (Gavrikov et al., 2025; Esfandiarpoor et al., 2024; Buettner & Kovashka, 2024; You et al., 2024; Qraitem et al., 2025). Such work suggests that pretrained multimodal models can respond differently to particular visual or semantic information before task-specific adaptation. Other studies examine how model representations and predictions vary across layers, components, prompts, or downstream adaptation, providing tools for understanding how pretrained behavior is expressed in downstream tasks (Long et al., 2026; Liu et al., 2026b; Zhang et al., 2026; Liu et al., 2026a; Ma et al., 2026).

These studies motivate our use of the pretrained VLM as an explicit reference, but they do not address the VAD setting considered here. In particular, they generally study general-purpose vision-language tasks or individual model components rather than comparing a pretrained VLM with multi-

ple VAD systems built on the same backbone. Our work brings this perspective into VAD and makes the pretrained-versus-downstream comparison explicit. For each backbone, we retain the pretrained model as a reference and evaluate the corresponding downstream VAD system on the same video content. We then examine not only whether cue-related score differences are preserved or modified, but also where these differences emerge within the downstream pipeline and whether their timing changes around anomalous events.

Positioning of our work. The existing literature therefore provides two important foundations. VLM-based VAD shows how pretrained multimodal models can be incorporated into anomaly detection, while work on multimodal model behavior shows that pretrained models can exhibit systematic responses to semantic and contextual information. What remains largely unexplored is the connection between these two perspectives: *how does anomaly-scoring behavior present in a pretrained VLM relate to the behavior of the VAD system built on it?*

We address this gap through a reference-based analysis across heterogeneous VLMs and VAD pipelines. We compare each downstream system with its corresponding pretrained VLM on the same video segments, use matched cue-present and cue-absent comparisons to obtain a common measure of cue-related score differences, and preserve each VAD method's algorithmic recipe. We further examine intermediate pipeline outputs and temporal scores to identify where and when observed differences emerge. This shifts the focus from asking only whether VLM-based VAD detects anomalies accurately to understanding *what behavior is already present in the pretrained model, what differs after downstream processing, where those differences arise, and when anomaly evidence is assigned*.

B Experimental Setup

This appendix describes the datasets, models, and evaluation protocol. The components of the framework are detailed in Appendix C and the implementation in Appendix D.

B.1 Datasets

Datasets. We use the test splits of two primary benchmarks and one external-validation benchmark. *UCF-Crime* (Sultani et al., 2018) contains long untrimmed surveillance videos; its test split has 290 videos (140 anomalous, 150 normal, 13 anomaly categories) with temporal annotations, yielding 3,709 evaluation segments. *XD-Violence* (Wu et al., 2020) is our primary cross-dataset replication; its 800 test videos (500 violent, 300 non-violent) are collected from movies, sports, surveillance, news, and games, and have a content distribution different from UCF-Crime; we use the visual stream only, yielding 9,779 evaluation segments. *MSAD* (Zhu et al., 2024) spans 14 scenarios with varying lighting and weather and both human- and non-human-related anomalies; we use the test split of its official weakly supervised protocol (Protocol ii; 240 videos, 120 anomalous and 120 normal) to test whether the cue-related difference pattern found on the crime-centric benchmarks persists in a more heterogeneous multi-scenario setting. It yields 514 evaluation segments from 239 videos: one anomalous video (`Object_falling_19`, 2.9 s) is shorter than the minimum segment length of $L/2 = 5$ s (*Evaluation and scoring settings* below) and contributes no segment. Cue-related score differences are estimated on test videos only; training splits are used only where a method's recipe requires them (Appendix D.1). Cue coverage (numbers of cue-present, cue-absent, and unresolved segments and videos under both consensus rules) is reported per dataset in Appendix C.2.

B.2 Models

Vision-language models. We evaluate eleven open-weight VLM backbones: Qwen2.5-VL-3B/7B (Bai et al., 2025b), Qwen3-VL-4B/8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), SmolVLM2-2.2B (Marafioti et al., 2025), GLM-4.1V-9B-Thinking (Hong et al., 2025), VideoLLaMA3-2B/7B (Zhang et al., 2025a), VideoChat3-4B (Li et al., 2026), and JoyAI-VL-Interaction-8B (Video Understanding Team of JoyAI-VL @ Joy Future Academy, JD, 2026). The set contains controlled variations in *scale* (Qwen2.5-VL 3B vs. 7B, Qwen3-VL 4B vs. 8B, VideoLLaMA3 2B vs. 7B) and *generation* (Qwen2.5-VL vs. Qwen3-VL), as well as independent families, image-centric and video-native designs, an efficiency-oriented small model, and a reasoning-

Table 1: VLM-based VAD methods, grouped by main processing step. A method may use more than one type of processing. Data: method-specific labelled data (fixed across backbones). "Scores during inference" lists the intermediate scores $s_{b,m}^{(k)}$ compared in Appendix F.

Family	Method	Main processing step	θ_b / data	Scores during inference
Reasoning	VERA	guided reflection with learned guiding questions	frozen / –	reflection → scene context → temporal smoothing
Evidence	AnyAnomaly	text-guided key frames with position and temporal context	frozen / –	key-frame VQA → context scores
	EventVAD	event-aware segmentation and event-level scoring	frozen / –	–
Representation	PRISM	projection onto text-derived semantic anomaly axes	frozen / –	semantic readout → temporal smoothing
	HeadHunt-VAD	robust anomaly-sensitive attention heads with a linear scorer	frozen / ≈1% calib.	–
	HiProbe-VAD	salient-layer hidden-state probing with a linear scorer	frozen / ≈1% calib.	–
Parameter	Holmes-VAU	LoRA adaptation on hierarchical VAU instructions	updated / HIVAU train	adapted VLM score → final VAD score

oriented model. Open weights are an inclusion criterion rather than a convenience: the framework requires next-token likelihoods (Eq. 1), hidden states and attention-head outputs for representation-based methods, trainable weights for parameter adaptation, and deterministic, reproducible inference. All backbones use their released instruction-tuned checkpoints and are decoded greedily.

Table 1 summarizes the evaluated methods.

VLM-based VAD methods. We use seven methods that differ in how they process anomaly scores (Table 1): prompting and reasoning (VERA (Ye et al., 2025)), visual evidence selection (AnyAnomaly (Ahn et al., 2026), EventVAD (Shao et al., 2025)), representation readout (PRISM (CHEN et al., 2026), HeadHunt-VAD (Cai et al., 2026), HiProbe-VAD (Cai et al., 2025)), and backbone adaptation (Holmes-VAU (Zhang et al., 2025b)). Each method uses a pretrained VLM to produce continuous anomaly scores and has a public implementation or sufficient reproduction details. We consider backbone combinations that support the method's required inputs and outputs.

B.3 Evaluation Protocol

Evaluation and scoring settings. Let x be an evaluation segment of length L. Segments are consecutive and non-overlapping; a final shorter segment is retained if it lasts at least $L/2$. A segment is labeled anomalous, $y(x) = 1$, when the fraction of its frames inside annotated anomalous intervals reaches a threshold τ. Our experimental choices are $L = 10\,\text{s}$ and $\tau = 0.5$. For content annotation and pretrained or adapted VLM scoring, we sample eight evenly spaced frames per segment, with the long image side capped at 640 pixels. Uniform sampling covers the segment while limiting inference cost across backbones; sensitivity to frame count is examined in Sec. E.4. These choices define the evaluation setting, not a required input format for downstream detectors. The implementation clips the two-answer probability at 10^{-6} and $1 - 10^{-6}$ before taking log-odds, which can create ties at the extremes. Motion is the mean absolute grayscale difference between frames five apart, sampled every fifth frame within the segment.

Detection performance. Detection performance provides context for interpreting cue-related score differences. We therefore report frame-level performance for the same configurations: AUC on UCF-Crime and MSAD and average precision (AP) on XD-Violence, following each benchmark's standard protocol. Segment-level scores are expanded to frames for this purpose.

Event-analysis settings. Temporal comparisons use each VAD method's score sequence at its output resolution, before segment averaging; the reference score is held constant over each evaluation segment. Percentiles use midranks over frames covered by both timelines. Pre/post intervals are clipped at video boundaries. Event concentration is undefined if the total rank mass is zero. For segment-based event attributes, we use the first annotated event of each anomalous video. The early event attribute contrasts onset before the first third of the video with onset after the second third; the long event attribute contrasts the top and bottom dataset-specific thirds of event duration divided

by video duration. Intermediate groups are excluded. These video-level attributes are attached to all segments of the source video and use all four matching variables. For the event-aligned analysis, pre/post intervals use $h = 10\,\mathrm{s}$, converted to frames at each video's frame rate; frames in other events or their pre/post intervals are excluded. We average available event quantities within each video and then across videos. The separate event-level D_{early} comparison uses the same onset boundaries, whereas D_{long} contrasts event durations in seconds above versus below the dataset median, excluding equality. Events are matched one-to-one without replacement on anomaly category across different source videos, using a fixed random order (seed 0) and the same pairs for the compared scores; estimates require at least ten valid pairs.

C Framework Details

C.1 Cue Definitions

Cues and annotation. We evaluate eight visual or contextual cues: person, vehicle, running, night, indoor, street, dense people, and high motion. The seven semantic cues are annotated by VLMs; high motion is computed from frame differences. The three annotators are DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), Ling-3.0-flash-VL[2], and Gemma-4-26B-A4B-IT (Team et al., 2026), accessed with a pinned inference provider, temperature 0, fixed seed, and cached raw responses. None of them is an evaluated backbone. Each annotator receives the same eight sampled frames in temporal order and a content-only prompt across all three datasets, without anomaly labels, event boundaries, file names, or detector outputs. Using three annotators balances cross-model agreement checks against annotation cost. The main analysis retains only non-uncertain labels on which all three agree; a two-of-three rule is examined in Sec. E.4. Disagreeing or uncertain labels are unresolved and excluded only from the affected cue comparison. Unanimous annotations resolve between 50% (night) and 96% (vehicle) of the UCF-Crime segments; the rarest analysable cue is running, with 26 unanimous cue-present segments from 21 videos.

Table 2 summarizes the ten evaluated cues of Sec. 2.1 and their positive and negative group definitions. The person cue contrasts segments with a person visible in at least one sampled frame against those with none; the vehicle cue similarly contrasts the presence versus absence of a visible car, truck, bus, motorcycle, or bicycle; and the running cue contrasts a clearly running person with no running person. The night cue contrasts darkness or predominantly artificial lighting with daylight, the indoor cue contrasts indoor with outdoor camera views, and the street cue contrasts the presence versus absence of a visible road, street, or parking lane. The dense people cue contrasts at least six people in some frame with at most two people in every frame, excluding segments with three to five people, while the high motion cue contrasts the top and bottom dataset-specific thirds of the frame-difference motion measure, excluding the middle third.

C.2 Matched Pair Construction

Pairing settings. Within each dataset, pairs match exactly on segment anomaly status, people-count category (0, 1–2, 3–5, or $6+$), scene category (indoor or outdoor), and motion category (lower, middle, or upper third of within-dataset motion ranks). People count and scene use majority annotation, retaining unresolved answers as a separate category. We omit people count when studying person or dense people, scene when studying indoor, and motion when studying high motion. All other visual cues use every matching variable. An unresolved label matches only another unresolved label and does not establish that the underlying contexts are identical. Within each matching stratum, cue-present and cue-absent segments are shuffled with a fixed seed (0). Cue-present segments are visited in that order; each takes the first unused eligible cue-absent segment from a different source video. Segments without an eligible partner are dropped.

Pair coverage. Table 3 reports, per dataset and cue, how many segments support each estimate: cue-present candidates, candidates with an eligible control, label-based matched pairs, the number of source videos, and the common valid pairs after requiring finite reference and VAD scores on both segments of a pair. Every method scores every segment, so the valid pairs equal the matched

[2]Access from `https://huggingface.co/inclusionAI/Ling-3.0-flash-VL`

Table 2: Evaluated cues. Each cue compares segments where it is present with segments where it is absent; segments that fit neither description (disagreeing or `uncertain` annotators, 3–5 people for dense people, middle third of motion, onset, or duration) are left out of that cue's analysis only. The event-timing attributes use ground-truth anomaly annotations and exist only for anomalous videos.

Cue	Present means	Compared with	How the label is obtained
Visual attributes of a segment			
person	a person is visible in at least one frame	no person visible	agreement of 3 VLMs
vehicle	a car, truck, bus, motorcycle, or bicycle is visible	no vehicle visible	agreement of 3 VLMs
running	a person is clearly running (not walking)	no running person	agreement of 3 VLMs
night	dark scene or artificial lighting dominates	daylight	agreement of 3 VLMs
indoor	the camera views an indoor scene	outdoor scene	agreement of 3 VLMs
street	a road, street, or parking lane is visible	none visible	agreement of 3 VLMs
dense people	≥ 6 people in some frame	at most 2 people in every frame	agreed people count of 3 VLMs
Motion of a segment			
high motion	top third of frame-difference motion	bottom third	computed from pixels
Event timing of an anomalous video (attached to all its segments)			
early event	event starts in the first third of the video	event starts in the last third	ground-truth interval
long event	event duration (share of the video) in the top third	bottom third	ground-truth interval

Table 3: Pair coverage per dataset and cue (unanimous annotations). Cand.: cue-present segments; ctrl.: candidates with at least one eligible control; pairs: label-based matched pairs; videos: source videos on either side; valid: common valid pairs used by every method.

Dataset	Cue	Cand.	Ctrl.	Pairs	Videos	Valid
UCF-Crime	person	2869	2828	424	188	424
	vehicle	1089	1017	224	122	224
	running	26	26	26	40	26
	night	552	543	448	178	448
	indoor	2248	2247	1045	259	1045
	street	1034	893	71	76	71
	dense people	431	431	299	154	299
	high motion	1237	1181	555	234	555
	early event	577	396	161	58	161
	long event	403	281	230	92	230
XD-Violence	person	8173	7705	1218	468	1218
	vehicle	3084	3083	1962	633	1962
	running	458	458	458	284	458
	night	1621	1599	1306	595	1306
	indoor	3623	3622	2952	723	2952
	street	3752	3742	1093	527	1093
	dense people	4338	4338	1839	718	1839
	high motion	3260	3258	1265	650	1265
	early event	3854	2999	182	154	182
	long event	1581	1570	617	293	617
MSAD	person	312	289	98	123	98
	vehicle	261	223	76	108	76
	running	6	5	5	10	5
	night	84	81	77	93	77
	indoor	143	143	123	145	123
	street	252	234	48	67	48
	dense people	66	66	44	64	44
	high motion	172	136	57	86	57
	early event	127	14	4	8	4
	long event	68	29	10	17	10

pairs for all methods; only the EventVAD variant that drops unparsable answers (Appendix E.4) uses fewer.

C.3 Statistical Testing

Statistical protocol. The main setting uses unanimous content annotations, prompt p_0, the cue-level estimator D_c (Eq. 3), and the same valid pairs for each before/after comparison. The correlation

$\rho_{b,m,d}$ and root-mean-square change $R_{b,m,d}$ use eligible attributes from the eight visual or contextual cues and two event attributes, with a separate summary for visual cues only.

Uncertainty and detectability. An estimated cue-level difference can be nonzero without being statistically detectable. Segments from the same video are dependent, and a video can contribute to either side of multiple pairs. We therefore resample source videos from both sides: in each bootstrap replicate, independently drawn video multiplicities give pair i weight $w_i = n_{v_i^+} n_{v_i^-}$. We recompute weighted cue-level values and their changes using identical weights for the scores being compared. Replicates with no valid pair weight are redrawn. The analysis uses $2{,}000$ replicates, percentile 95% intervals, and two-sided tests of zero, with Benjamini-Hochberg correction across backbone-cue comparisons separately for each dataset and score or change comparison. A difference or change is called *detectable* at adjusted $q < 0.05$; not detectable does not mean zero.

Score extraction and coverage. Reference answer likelihoods are read at the first answer position, with probability clipping as specified in Sec. B.3. Segments without the required score coverage are excluded, and the same valid pairs are retained across compared scores; score ties contribute zero to Eq. 3.

C.4 Transformation Categories

Table 4 gives the exact rules of the categories of change of Sec. 2.1.

Design of the rules. The rules are checked in the order of the table, so the categories are mutually exclusive and cover every cell. They differ in what they require of ΔD. Introduced and reversed are decided from D^{ref} and D^{VAD} alone: a difference that is detectable after the method but not before, or detectable before and after with opposite signs, already shows that the method changed the cue-related difference, and a reversal implies a change at least as large as either magnitude. Losing detectability is weaker evidence, because D^{VAD} can fall just below the threshold after a small shift. Suppressed therefore also requires a detectable ΔD, and a reference difference that becomes undetectable without a detectable change is left as not classified instead of being counted as suppressed. Likewise, preserved means that no change is detectable, not that the difference is unchanged.

Reading the counts. Because every rule depends on detectability, the counts depend on the number of matched pairs and understate how often each kind of change occurs; on MSAD, with the fewest pairs, no cell has a detectable change (Table 15). Not classified is the residue of the conservative suppression rule. On UCF-Crime it outnumbers suppressed cells, so the suppression count there is a lower estimate of how often reference differences weaken. Preserved, amplified, and attenuated require both differences to be detectable, and the latter two also a detectable change; they occur only on XD-Violence, the dataset with the most pairs. Finally, a category describes how the detectability and magnitude of a difference change, not its direction: whether cue-present segments score higher or lower is given by the sign of D.

D Implementation Details

D.1 Backbone-specific Details

Each port keeps the method's algorithmic recipe (evidence construction, scoring rule, and temporal processing), but its input units, frame sampling, and in some cases auxiliary components are adapted to the common evaluation or reconstructed where the official code is incomplete; these adaptations are listed below, and the method's scores are then aligned to evaluation segments. The ported configurations are therefore not reproductions of the published systems, and their detection performance is not comparable to the published numbers. *VERA*: the log-odds of class label `1` versus `0` after reflection with the released dataset-specific guiding questions (stage 1), followed by scene-context weighting over the most similar segments of the same video (stage 2) and temporal smoothing (stage 3, the final VERA score). The questions are fixed across backbones, and the UCF-Crime questions are used on MSAD. Reflection is run on the evaluation segments themselves (eight frames per $10\,\mathrm{s}$), so on short videos the official top-15% neighbourhood of the scene-context step reduces to the segment itself. *AnyAnomaly*: anomaly text guides key-frame selection, and position and temporal context provide additional evidence for scoring. Each dataset anomaly category serves

Table 4: Categories of cue-level change (Sec. 2.1) as implemented. "Detectable" means BH-adjusted $q < 0.05$ for D^{ref}, D^{VAD}, or ΔD; not detectable does not mean zero. Each backbone-method-dataset-cue cell receives the first category whose rule it satisfies. Counts of categories differ from N_{chg}, the number of cells with a detectable ΔD: introduced and reversed do not require a detectable ΔD, and a cell with no detectable difference can still have one.

Category	Meaning	Rule (checked in this order)
No detectable difference	neither score shows a detectable cue-related difference	D^{ref} and D^{VAD} not detectable
Introduced (+)	a detectable difference appears only after the method	D^{ref} not detectable; D^{VAD} detectable
Reversed (±)	detectable before and after, in opposite directions	both detectable; opposite signs
Preserved	detectable before and after in the same direction; no detectable change	both detectable; same sign; ΔD not detectable
Suppressed (∅)	detectable before; detectably changed and no longer detectable	D^{ref} detectable; ΔD detectable; D^{VAD} not detectable
Not classified	detectable before; neither the VAD difference nor the change is detectable	D^{ref} detectable; D^{VAD} and ΔD not detectable
Amplified (↑)	same direction, detectably larger magnitude	both detectable; ΔD detectable; $\lvert D^{\mathrm{VAD}}\rvert > \lvert D^{\mathrm{ref}}\rvert$
Attenuated (↓)	same direction, detectably smaller magnitude	both detectable; ΔD detectable; $\lvert D^{\mathrm{VAD}}\rvert < \lvert D^{\mathrm{ref}}\rvert$

as an anomaly text. The original input unit is a clip of 24 consecutive frames decoded from the video (the official clip length): CLIP-based key-frame selection, the WinCLIP position map, and the 2×2 temporal grid are built inside the clip, and the three VQA scores per anomaly text are fused with the official weights $(0.6/0.3/0.1)$ after taking the maximum over anomaly texts. One 24-frame input clip is scored at the centre of each 10 s evaluation segment instead of the official back-to-back clips, which would multiply the number of VQA calls by the frame rate; the official frame-level Gaussian is an identity at this clip spacing and is omitted. Stage 1 is the key-frame VQA score and stage 2 the fused score. *EventVAD*: events are obtained once per video from CLIP and optical-flow features using dynamic-graph propagation and boundary detection; the backbone scores each event from 16 uniformly sampled frames with a description-then-score prompt. Following the official evaluation, an answer without a parsable score is assigned -1; the fraction of such events is recorded per evaluation segment and a variant that excludes them is reported in Sec. E.4. *PRISM*: clip embeddings every 16 frames and text embeddings of the released normal and anomaly descriptions are taken from the backbone's last hidden layer. Clip scores use whitened semantic anomaly axes and the official softmax weighting (stage 1, semantic readout), followed by temporal smoothing (stage 2, final score). *HeadHunt-VAD*: attention heads are ranked by anomaly saliency across five prompts on a calibration subset. A logistic-regression scorer uses the top-5 heads; test videos are scored every 48 frames and smoothed. Frames given to a backbone are capped at 640 px on the long side in every method. For HeadHunt-VAD on VideoLLaMA3-7B the vision encoder's attention over sixteen 640 px frames does not fit a 32 GB GPU because the implementation materialises the full attention matrix; since that attention is block-diagonal per frame, we run the encoder over four frames at a time, which is mathematically identical to a single call. *HiProbe-VAD*: the layer with the strongest anomaly signal is selected on the same calibration subset. A logistic-regression probe uses its hidden states; test videos are scored every 24 frames and smoothed. *Holmes-VAU*: the language model receives LoRA adapters trained with the official hyperparameters (rank 64, one epoch) on HIVAU-70k instructions from UCF-Crime and XD-Violence *training* videos; the vision encoder and projector stay frozen. The adapted VLM score is the stage-1 score $s^{(1)}_{b,m}$ of Holmes-VAU: the same frames, question, and readout as the pretrained score, evaluated with the adapted weights; it measures adaptation separately from the final Holmes score $s^{(2)}_{b,m}$. The final score queries the adapted model through its own trained interface: HIVAU's video-level judgement question ("Are there any anomaly events detected in the video?") on 12 uniformly decoded frames of the segment, read out as the log-odds of the anomaly-answer openings ("The/An anomaly …") against the no-anomaly openings ("No anomaly …", "There is no …") at the first answer token. Holmes's separate supervised anomaly scorer used for dense frame sampling is trained on task-specific video features and does not depend on the backbone, so it is not part of this estimand. We report the adapted VLM score and final pipeline separately to distinguish the change after adaptation from the change after the remaining VAD processing. The calibration subset for HeadHunt-VAD and HiProbe-VAD contains about 1% of training videos, includes normal and anomalous videos, and is shared across the two methods

and backbones. Each Holmes adapter is trained once and evaluated on all three test sets, so MSAD also tests transfer to a different dataset; adapters are trained for all eleven backbones, the six backbones above 4B with four-GPU data-parallel training that keeps the 512-sample optimisation steps and a shared LoRA initialisation across ranks. This differs from benchmark-specific training in the original paper.

Scope of reproduction. We did not reproduce the published systems with their official backbones and full protocols. Three ports rest on reconstructed components: the public EventVAD code lacks the graph-propagation step and the scoring prompt, HiProbe-VAD has no public code and is implemented from the paper, and the PRISM feature-extraction code is not released; for these methods the ported configurations can differ from the published systems beyond the backbone. As a single anchor for the gap between the ported configurations and the pretrained score, VERA is run with its official backbone InternVL2-8B under the same common evaluation setting on UCF-Crime (10 s segments, eight frames, backbone scene features): it reaches 77.3 frame-level AUC (79.0 after reflection, 80.3 after scene-context weighting; pretrained score of the same backbone 82.1), against 86.55 reported for the original protocol and a mean of 75.7 over the eleven evaluated backbones (reference mean 80.1). The official backbone under our common evaluation setting thus exhibits a similar performance reduction, suggesting that the gap is substantially associated with the adapted evaluation and inference setting rather than with backbone substitution alone.

D.2 Intermediate Outputs

For the staged methods, the intermediate scores $s_{b,m}^{(k)}$ recorded during inference are listed in Table 1 and defined with each method in Appendix D.1; stage 0 is the pretrained score.

Stage summaries. Each stage is further summarized by its agreement with the reference and with the preceding stage,

$$\rho_{\mathrm{ref}}^{(k)} = \mathrm{Spearman}\left(\boldsymbol{D}^{(0)}, \boldsymbol{D}^{(k)}\right), \qquad \rho_{\mathrm{prev}}^{(k)} = \mathrm{Spearman}\left(\boldsymbol{D}^{(k-1)}, \boldsymbol{D}^{(k)}\right), \tag{13}$$

where $\boldsymbol{D}^{(k)}$ is the vector of cue-level differences at stage k, by its cumulative magnitude of change

$$R^{(k)} = \left[\frac{1}{C}\sum_{c=1}^{C}\left(D_c^{(k)} - D_c^{(0)}\right)^2\right]^{1/2}, \tag{14}$$

over the C eligible cues, and by $N_{\mathrm{chg}}^{(k)}$, the number of backbone-cue cells whose change from the reference is detectable.

D.3 Hyperparameters

Table 5 collects the fixed settings of the main analysis; each is defined where it is used (Appendices B.3, C, and D.1).

D.4 Compute

All inference and adaptation are run in PyTorch with Hugging Face Transformers and PEFT on NVIDIA RTX 5090 GPUs and A100 80GB. Scoring is sharded by video and resumable, and methods with temporal state are always run over a video in chronological order, so that sharding does not change any method's temporal context. The accompanying materials include scores, cue labels, matched pairs, and bootstrap replicates.

E Additional Behavioral Results

E.1 Reference versus VAD Comparisons

Table 6 summarizes the comparison per dataset and method (the insets of Figure 1), and Table 7 gives the complete backbone-level results and the frame-level detection performance of each configuration. Every method is evaluated on all eleven backbones and three datasets, so both tables are complete (231 configurations).

Table 5: Fixed settings of the main analysis.

Component	Setting	Value
Evaluation grid	segment length L; anomaly threshold τ	10 s; 0.5
Reference score	frames per segment; image size; decoding	8 evenly spaced; long side $\leq$ 640 px; greedy
	probability clipping	$[10^{-6}, 1-10^{-6}]$ before log-odds
Motion	frame difference	mean absolute grayscale difference, frames five apart, every fifth frame
Pairing	shuffle seed; minimum valid pairs	0; 10
Statistics	bootstrap; interval; correction	2,000 video-resampling replicates; percentile 95%; BH, $q < 0.05$
Event analysis	pre/post window h	10 s
AnyAnomaly	clip length; fusion weights	24 frames; $0.6/0.3/0.1$
EventVAD	frames per event	16
PRISM	clip stride	16 frames
HeadHunt-VAD	heads; scoring stride	top-5; 48 frames
HiProbe-VAD	scoring stride	24 frames
HeadHunt / HiProbe	calibration subset	$\approx 1\%$ of training videos
Holmes-VAU	LoRA	rank 64, one epoch, 512-sample steps
	final-score input	12 uniformly decoded frames

Table 6: Overview of the pretrained-reference versus VAD comparison. For each dataset and method over the eleven backbones: the range $[\min, \max]$ of $\rho_{b,m,d}$, the mean $\rho_{b,m,d}$, the mean $R_{b,m,d}$, and N_{chg}, the number of backbone-cue cells with a detectable $\Delta D_{b,m,d,c}$. Ranges are not confidence intervals. Appendix E.1 lists every backbone-level value with its interval indicator and detection performance.

Dataset	Method	$\rho_{b,m,d}$ range	Mean ρ	Mean R	N_{chg}
UCF-Crime	VERA	[.31, .88]	.67	.23	14
	AnyAnomaly	[−.42, .85]	.46	.25	16
	EventVAD	[−.62, .90]	.51	.23	14
	PRISM	[.10, .64]	.46	.32	16
	HeadHunt-VAD	[.15, .90]	.59	.26	10
	HiProbe-VAD	[−.12, .90]	.54	.24	0
	Holmes-VAU	[.37, .94]	.78	.18	14
XD-Violence	VERA	[.14, .75]	.54	.19	40
	AnyAnomaly	[−.25, .88]	.39	.21	36
	EventVAD	[−.64, .81]	.19	.21	41
	PRISM	[−.24, .79]	.21	.31	54
	HeadHunt-VAD	[.03, .67]	.39	.25	38
	HiProbe-VAD	[−.54, .87]	.36	.21	30
	Holmes-VAU	[.02, .90]	.64	.13	19
MSAD	VERA	[.19, .95]	.66	.22	0
	AnyAnomaly	[−.12, .81]	.39	.32	0
	EventVAD	[−.62, .92]	.45	.28	0
	PRISM	[−.14, .52]	.18	.46	0
	HeadHunt-VAD	[−.07, .90]	.45	.39	0
	HiProbe-VAD	[.18, .79]	.59	.26	0
	Holmes-VAU	[.36, 1.0]	.70	.20	0

E.2 Cue-Level Changes per Dataset

Table 8 gives the cue-level changes behind Figures 3 and 2 on UCF-Crime, and Tables 9 and 10 repeat it for the other two datasets. Cues without at least ten valid pairs are omitted for that dataset. Tables 11 and 12 resolve the UCF-Crime counts into the individual backbone-method cells, giving D^{ref}, D^{VAD}, the category of change, and $\rho_{b,m,d}$ of every configuration.

Table 7: Backbone × method × dataset comparisons. Rows: dataset and method (Rsn. = reasoning, Evid. = evidence construction, Repr. = representation readout, Par. = parameter adaptation); columns: backbones. **(a)** Correlation $\rho_{b,m,d} = \mathrm{Spearman}(\boldsymbol{d}^{\mathrm{ref}}_{b,d}, \boldsymbol{d}^{\mathrm{VAD}}_{b,m,d})$ over eligible cues (Eq. 6; unanimous annotations); †: 95% source-video bootstrap interval includes 0. Summary columns give the mean ρ, the mean root-mean-square change $R_{b,m,d}$ (Eq. 7), and N_{chg}, the number of backbone-cue cells with a detectable $\Delta D_{b,m,d,c}$; they are reported only for rows with every supported backbone. **(b)** Frame-level detection performance of the same configurations (×100): AUC on UCF-Crime and MSAD, AP on XD-Violence. AnyA. = AnyAnomaly, Event = EventVAD, HeadH. = HeadHunt-VAD, HiPr. = HiProbe-VAD, Holmes = Holmes-VAU (final pipeline score).

		Backbone											Summary		
Dataset	Method	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen3-VL-4B	Qwen3-VL-8B	InternVL3-8B	SmolVLM2-2B	GLM-4.1V-9B	VideoLLaMA3-2B	VideoLLaMA3-7B	VideoChat3-4B	JoyAI-VL-8B	Mean ρ	Mean R	N_{chg}
(a) Correlation $\rho_{b,m,d}$ with the pretrained reference															
UCF-Crime	VERA	.62	.64	.81	.72	.81	.62†	.78	.54†	.31	.68	.88	.67	.23	14
	AnyA.	.19†	.84	.85	.47†	.61†	.16†	.54	−.42†	.52†	.69	.62	.46	.25	16
	Event	.85	.90	.62	.68	.87	−.08†	.09†	−.62†	.76	.66	.84	.51	.23	14
	PRISM	.10†	.50	.53	.60†	.18†	.58	.62	.33†	.38†	.64†	.58†	.46	.32	16
	HeadH.	.64	.53	.83	.70	.90	.15†	.27†	.27†	.70†	.81	.72	.59	.26	10
	HiPr.	.59	.79	.82	.68	.90	−.12†	.71	−.02†	−.03†	.83	.82	.54	.24	0
	Holmes	.37	.70	.67	.79	.93	.87	.77	.89	.77	.90	.94	.78	.18	14
XD-Violence	VERA	.14†	.52†	.42†	.56	.75	.68†	.56	.59	.52	.47†	.70	.54	.19	40
	AnyA.	.54†	.66	.75	.88	−.08†	−.25†	.45	.22†	.14†	.45†	.53†	.39	.21	36
	Event	.28†	.38	.28	.52	.81	−.03†	−.36	−.64†	.08†	.27†	.50	.19	.21	41
	PRISM	.32	.79	−.18†	.03†	.09†	.54	.64	.26†	.22†	−.24†	−.13†	.21	.31	54
	HeadH.	.67	.13†	.42	.41†	.54	.10†	.03†	.58	.38	.50	.56	.39	.25	38
	HiPr.	.32†	.12†	.67	.87	.77	.32†	−.16†	.76	.59	.28†	−.54†	.36	.21	30
	Holmes	.70	.60	.77	.85	.49	.90	.02†	.44	.65	.70	.90	.64	.13	19
MSAD	VERA	.19†	.83	.57†	.74†	.65†	.95	.64†	.67†	.83†	.40†	.76†	.66	.22	0
	AnyA.	−.01†	.79†	.62†	.76†	.14†	.38†	.24†	.36†	.29†	−.12†	.81†	.39	.32	0
	Event	.92†	.62†	.79†	.55†	.40†	−.62†	.48†	−.29†	.61†	.71†	.81†	.45	.28	0
	PRISM	.28†	.26†	.12†	.31†	−.11†	.17†	.52†	.31†	.26†	−.14†	.05†	.18	.46	0
	HeadH.	−.07†	.90	.62†	.60†	.40†	.69†	.38†	.60†	.24†	.05†	.52†	.45	.39	0
	HiPr.	.18†	.79	.79	.71†	.41†	.62†	.76†	.69†	.52†	.29†	.79†	.59	.26	0
	Holmes	.38†	.90	.60	.52†	.83†	1.0	.36†	.69†	.98	.60†	.83	.70	.20	0
(b) Detection performance (UCF-Crime, MSAD: AUC; XD-Violence: AP)															
UCF-Crime	VERA	73.6	78.1	76.0	77.3	81.9	72.8	75.0	69.3	76.0	80.0	73.0	75.7		
	AnyA.	73.1	78.4	77.5	73.0	75.1	43.3	73.5	52.3	73.1	65.2	75.8	69.1		
	Event	69.1	80.2	76.9	72.1	73.7	46.6	48.6	49.6	74.4	71.7	73.4	67.0		
	PRISM	57.7	69.4	75.1	80.9	61.9	57.9	77.2	68.2	63.9	75.1	71.5	69.0		
	HeadH.	76.3	81.6	82.2	79.3	78.7	72.6	71.7	73.7	74.7	80.7	80.4	77.5		
	HiPr.	80.7	80.7	79.3	77.8	78.6	63.5	80.2	63.8	67.2	78.2	78.7	75.3		
	Holmes	83.8	86.9	86.7	86.5	89.1	81.4	83.6	84.8	87.6	85.7	84.0	85.5		
XD-Violence	VERA	68.6	68.2	65.6	67.2	69.6	56.7	69.6	31.1	66.0	64.0	65.1	62.9		
	AnyA.	52.4	61.4	63.1	57.2	61.3	23.9	66.0	43.5	58.8	56.5	60.1	54.9		
	Event	60.2	60.9	58.9	62.4	65.9	23.0	24.1	26.4	51.7	58.1	62.0	50.3		
	PRISM	45.4	52.8	61.5	62.0	44.2	36.2	51.2	45.7	53.9	48.3	53.2	50.4		
	HeadH.	37.9	27.2	56.0	60.6	47.3	52.5	44.9	57.5	48.8	50.3	47.4	48.2		
	HiPr.	35.1	32.9	65.8	65.3	30.8	38.1	51.3	33.6	44.8	47.7	37.4	43.9		
	Holmes	80.0	81.2	78.9	83.3	85.1	73.4	79.7	85.0	83.4	81.4	80.3	81.1		
MSAD	VERA	84.0	82.6	87.4	88.8	91.7	82.1	90.3	83.1	92.0	91.0	90.7	87.6		
	AnyA.	79.1	81.4	77.4	80.5	81.2	30.0	76.6	60.4	79.9	72.5	79.8	72.6		
	Event	83.6	87.3	88.3	86.6	84.5	40.9	50.7	52.5	81.2	85.9	86.6	75.3		
	PRISM	56.5	63.2	68.2	71.9	44.9	59.0	66.2	73.8	85.3	83.6	61.4	66.7		
	HeadH.	83.8	88.1	93.5	93.0	90.4	87.6	92.9	89.3	87.6	90.8	90.0	89.7		
	HiPr.	87.0	91.2	92.7	93.2	91.5	80.1	91.8	92.1	91.6	89.4	78.7	89.0		
	Holmes	92.4	93.4	93.8	93.7	93.9	90.9	92.9	93.6	94.4	93.8	93.0	93.3		

Table 8: Cue-level changes on UCF-Crime. N: label-based matched pairs for the cue. Ref.: number of backbones (of 11) with a detectable $D^{\mathrm{ref}}_{b,d,c}$ and its mean over backbones, both evaluated on the valid pairs of the VERA comparison. Method columns: number of backbones whose cell is classified as amplified (↑), attenuated (↓), suppressed (∅), introduced (+), or reversed (±) (Table 4), followed by the most frequent categories; introduced and reversed do not require a detectable $\Delta D_{b,m,d,c}$, so these counts need not match N_{chg} (Table 6); a dash marks methods without complete coverage. XD-Violence and MSAD are given in Appendix E.2.

		Ref.		# backbones changed						
Cue	N	det.	mean	VERA	AnyA.	Event	PRISM	HeadH.	HiPr.	Holmes
person	424	5	+.35	3 ∅	3 ∅	2 ∅	5 ∅	3 ∅	0	4 ∅
vehicle	224	0	−.06	0	0	0	0	0	0	0
running	26	0	+.37	0	0	0	0	0	0	0
night	448	2	+.27	1 ∅	1 ∅	0	3 +	1 ∅	0	0
indoor	1045	0	−.20	2 +	0	0	5 +	3 +	0	0
street	71	0	+.00	0	0	0	0	0	0	0
dense people	299	1	−.38	1 +	1 ∅	0	3 +	0	0	0
high motion	555	9	+.44	5 ∅	8 ∅	8 ∅	4 ∅	4 ∅	0	4 ∅
early event	161	0	+.22	0	0	0	0	0	0	0
long event	230	0	−.08	0	0	0	0	0	0	0

Table 9: Cue-level changes on XD-Violence (layout as in Table 8).

		Ref.		# backbones changed						
Cue	N	det.	mean	VERA	AnyA.	Event	PRISM	HeadH.	HiPr.	Holmes
person	1218	6	−.20	5 ∅	6 ∅	6 ∅	9 ∅+	10 ↑ +	7 +	5 +
vehicle	1962	6	+.13	3 ∅	5 ∅	6 ∅	10 ±	1 ∅	3 ∅	2 ∅
running	458	1	+.11	2 ∅+	9 +	4 +	7 +	6 +	3 +	0
night	1306	8	+.22	2 ∅	4 ∅+	6 ∅+	3 +	8 ∅	6 ∅	1 +
indoor	2952	0	+.01	0	1 +	2 +	10 +	7 +	4 +	0
street	1093	1	+.09	1 +	3 +	3 +	11 +	3 +	3 +	0
dense people	1839	10	−.35	6 ∅	10 ∅	6 ∅	3 ↑	6 ∅	7 ∅	5 ∅
high motion	1265	1	+.06	0	2 +	1 +	2 ∅+	3 +	2 +	0
early event	182	1	+.22	1 +	0	0	3 +	1 +	0	0
long event	617	0	+.10	9 +	0	0	1 +	0	2 +	3 +

Table 10: Cue-level changes on MSAD (layout as in Table 8).

		Ref.		# backbones changed						
Cue	N	det.	mean	VERA	AnyA.	Event	PRISM	HeadH.	HiPr.	Holmes
person	98	0	+.16	0	0	0	0	0	0	0
vehicle	76	0	+.28	0	0	0	0	0	0	0
night	77	0	+.26	0	0	0	0	1 +	0	0
indoor	123	0	−.06	0	0	0	1 +	1 +	0	0
street	48	0	+.30	0	0	0	0	0	0	0
dense people	44	0	−.22	0	0	0	1 +	0	0	0
high motion	57	0	+.46	0	0	0	0	0	2 +	0
long event	10	0	+.51	0	0	0	0	0	0	0

Table 11: Backbone-level cue-related score differences on UCF-Crime (Qwen2.5-VL, Qwen3-VL, InternVL3, SmolVLM2). Ref.: $D^{\mathrm{ref}}_{b,d,c}$ of the pretrained reference; method rows: $D^{\mathrm{VAD}}_{b,m,d,c}$ of the final score, with the category of change as superscript (Table 4): amplified ($\uparrow$), attenuated ($\downarrow$), suppressed ($\varnothing$), introduced ($+$), reversed ($\pm$), preserved ($=$), not classified ($\circ$); no superscript: no detectable difference. Bold: detectable (BH-adjusted $q < 0.05$). ρ: $\rho_{b,m,d}$ (Eq. 6). Every method scores every segment, so the valid pairs, and hence D^{ref}, are the same for all methods. Table 12 gives the remaining backbones.

Backbone	Score	person	vehicle	running	night	indoor	street	dense	h. motion	early	long	ρ
Qwen2.5-VL-3B	Ref.	$\mathbf{.49}$	.12	.23	.01	−.13	.07	−.42	$\mathbf{.51}$	−.09	−.30	
	VERA	$.15^{\circ}$	.08	.08	.29	$\mathbf{-.42}^{+}$	.15	−.08	$.27^{\circ}$	−.14	−.11	.62
	AnyA.	$-.04^{\varnothing}$	−.23	−.08	.30	.03	−.04	−.18	$.08^{\varnothing}$	.15	−.10	.19
	Event	$.30^{\circ}$	.04	.15	.06	−.00	.18	−.21	$.25^{\circ}$	−.22	−.20	.85
	PRISM	$-.32^{\varnothing}$	−.42	.08	$\mathbf{.52}^{+}$	.02	−.46	$\mathbf{-.75}^{+}$	$-.11^{\varnothing}$	.32	−.23	.10
	HeadH.	$.16^{\circ}$	−.30	.15	.06	−.20	−.21	−.38	$.08^{\varnothing}$	−.25	−.17	.64
	HiPr.	$.07^{\circ}$	−.11	.08	.35	−.07	−.10	−.49	$.39^{\circ}$	.02	−.05	.59
	Holmes	$-.01^{\varnothing}$	−.13	.08	.36	−.13	−.10	−.59	$.35^{\circ}$	.14	.00	.37
Qwen2.5-VL-7B	Ref.	.25	−.18	.23	.25	−.05	−.01	−.54	.27	.09	−.19	
	VERA	−.15	−.34	.15	.32	.04	−.18	−.45	.11	.01	−.13	.64
	AnyA.	.22	−.16	.35	.25	.04	−.11	−.34	.24	−.14	−.31	.84
	Event	.17	−.05	.15	.23	−.10	.01	−.25	.22	.20	−.15	.90
	PRISM	−.32	−.02	.08	.43	$\mathbf{-.54}^{+}$	−.07	$\mathbf{-.46}^{+}$	.01	.28	−.10	.50
	HeadH.	.03	−.21	.08	.48	−.11	−.13	−.33	.10	.03	.13	.53
	HiPr.	−.04	−.10	.31	.18	−.05	−.04	−.46	.15	−.09	−.06	.79
	Holmes	−.07	−.12	.38	.40	−.20	−.10	−.63	.26	.23	.03	.70
Qwen3-VL-4B	Ref.	$\mathbf{.61}$	.09	.38	.22	−.26	−.04	−.34	$\mathbf{.54}$	.24	−.06	
	VERA	$.18^{\varnothing}$	−.13	.23	.17	.01	−.15	−.30	$.05^{\varnothing}$	.01	−.09	.81
	AnyA.	$.16^{\varnothing}$	−.11	.19	.19	−.13	−.08	−.16	$.25^{\varnothing}$	−.01	−.22	.85
	Event	$.13^{\varnothing}$	−.13	.42	.21	−.12	−.08	−.04	$.15^{\varnothing}$	.27	.13	.62
	PRISM	$-.13^{\varnothing}$	.06	.38	$\mathbf{.53}^{+}$	$\mathbf{-.56}^{+}$	−.10	−.32	$.07^{\varnothing}$	.33	−.01	.53
	HeadH.	$.15^{\varnothing}$	−.22	.15	−.03	−.30	−.10	−.48	$.21^{\varnothing}$	.07	.08	.83
	HiPr.	$.05^{\circ}$	−.11	.23	.12	−.19	−.27	−.36	$.35^{\circ}$	.14	−.09	.82
	Holmes	$.06^{\varnothing}$	−.22	.38	.49	−.18	−.32	−.43	$.28^{\varnothing}$	.35	.04	.67
Qwen3-VL-8B	Ref.	$\mathbf{.64}$	−.03	.46	.25	−.18	−.04	−.29	$\mathbf{.43}$	.25	−.03	
	VERA	$.01^{\varnothing}$	−.23	.00	.33	−.01	−.21	−.35	$-.01^{\varnothing}$	.12	−.13	.72
	AnyA.	$.18^{\varnothing}$	−.06	.04	.14	−.03	.03	.02	$.10^{\varnothing}$	−.12	−.11	.47
	Event	$.09^{\varnothing}$	−.07	.35	.26	−.09	.00	−.02	$.11^{\varnothing}$	.35	−.03	.68
	PRISM	$-.18^{\varnothing}$	−.31	.00	.32	$\mathbf{-.40}^{+}$	−.30	−.41	$.12^{\circ}$	.30	.17	.60
	HeadH.	$-.12^{\varnothing}$	−.21	.15	.25	−.18	−.18	−.45	$.41^{\circ}$	.14	−.26	.70
	HiPr.	$.06^{\circ}$	−.03	.31	.19	−.13	−.08	−.43	$.28^{\circ}$	−.23	−.40	.68
	Holmes	$.10^{\varnothing}$	−.25	.62	.40	−.14	−.30	−.57	$.16^{\varnothing}$	.33	.09	.79
InternVL3-8B	Ref.	.30	−.26	.46	.17	−.22	−.01	−.47	$\mathbf{.35}$	.40	.00	
	VERA	−.01	−.24	.23	.35	−.07	−.37	−.50	$.05^{\circ}$	.38	.19	.81
	AnyA.	.14	−.06	.08	.10	−.06	−.10	.04	$.15^{\circ}$	.12	−.10	.61
	Event	.09	−.11	.46	.16	−.10	.06	.05	$.10^{\varnothing}$	.41	.00	.87
	PRISM	−.10	.10	−.15	.31	−.31	−.13	−.37	$-.08^{\circ}$	−.08	.01	.18
	HeadH.	.01	−.46	.62	.10	−.21	−.15	−.25	$.14^{\circ}$	.06	.04	.90
	HiPr.	.04	−.41	.42	.08	−.28	−.18	−.26	$.14^{\circ}$	.07	−.18	.90
	Holmes	.03	−.24	.69	.40	−.25	−.21	−.54	$.16^{\circ}$	.42	.09	.93
SmolVLM2-2B	Ref.	.15	−.01	.46	.36	−.26	.18	−.39	.27	.40	−.01	
	VERA	.22	.06	.08	.07	−.31	.46	−.16	.03	.12	−.01	.62
	AnyA.	−.00	.00	−.12	−.03	−.01	.18	−.31	−.10	.20	−.10	.16
	Event	−.11	−.14	.04	−.02	.16	−.08	.19	.09	−.03	−.20	−.08
	PRISM	$\mathbf{-.59}^{+}$	−.18	.15	$\mathbf{.67}^{+}$	.05	−.41	$\mathbf{-.70}^{+}$	.14	.01	.01	.58
	HeadH.	.30	−.04	.23	−.35	−.20	.04	−.21	−.07	−.06	.25	.15
	HiPr.	−.04	−.03	.15	−.08	−.02	−.04	−.03	.18	−.23	.15	−.12
	Holmes	.28	−.14	.38	.37	−.14	−.04	−.18	.30	.16	−.01	.87

Table 12: Backbone-level cue-related score differences on UCF-Crime (GLM-4.1V, VideoLLaMA3, VideoChat3, JoyAI-VL; layout as in Table 11).

Backbone	Score	person	vehicle	running	night	indoor	street	dense	h. motion	early	long	ρ
GLM-4.1V-9B	Ref.	.32	−.02	.31	.36	−.11	−.15	**−.47**	**.52**	.16	−.12	
	VERA	.08	−.12	.15	.07	−.02	.01	−.40°	.28°	.09	.03	.78
	AnyA.	.06	−.02	−.23	.20	−.17	.06	−.12∅	.14∅	.01	−.14	.54
	Event	−.14	.00	−.04	−.00	−.10	−.06	−.19°	−.08∅	−.01	.12	.09
	PRISM	−.08	−.16	.08	.37	**−.44**+	−.04	−.40°	.08∅	.09	.02	.62
	HeadH.	.28	.18	.08	−.16	**−.45**+	.52	−.36°	.16°	−.06	−.22	.27
	HiPr.	.20	−.16	.31	−.08	−.33	−.07	−.25°	.37°	.03	−.24	.71
	Holmes	−.02	−.33	.31	.36	−.16	−.18	−.54°	.17∅	.32	−.11	.77
VideoLLaMA3-2B	Ref.	.07	−.13	.38	**.38**	−.24	.01	−.34	**.41**	.34	−.13	
	VERA	.03	−.04	.08	−.19∅	**−.46**+	.10	−.16	.18°	.12	.10	.54
	AnyA.	.11	.07	.04	−.10∅	−.06	.08	.11	.03∅	−.11	.04	−.42
	Event	−.09	.03	.00	−.15°	.18	.34	−.08	−.34∅	−.12	.06	−.62
	PRISM	−.15	−.10	−.15	.20°	.00	−.15	−.36	.26°	−.22	.00	.33
	HeadH.	.26	.04	.23	−.26∅	**−.48**+	.07	−.20	−.05∅	−.08	.01	.27
	HiPr.	.15	.20	−.08	−.13°	−.08	−.04	−.18	−.03°	−.28	.25	−.02
	Holmes	.28	−.13	.54	.41°	−.17	−.15	−.28	.33°	.40	−.10	.89
VideoLLaMA3-7B	Ref.	**.41**	−.12	.23	.24	−.19	.07	−.35	**.50**	.27	−.05	
	VERA	−.17∅	−.01	−.08	.23	−.04	.01	**−.70**+	.11∅	−.06	−.16	.31
	AnyA.	.21°	−.01	−.04	.08	−.07	−.04	−.02	.12∅	−.06	−.10	.52
	Event	.19°	−.11	.12	.14	−.11	−.10	−.06	.10∅	.06	.00	.76
	PRISM	.01∅	−.18	.23	−.02	.23	−.15	−.18	.26°	−.28	−.06	.38
	HeadH.	.05°	−.06	.08	.06	−.31	−.07	−.30	.08∅	−.18	.02	.70
	HiPr.	.15°	.05	.15	−.08	−.25	−.10	.16	.07°	−.12	.15	−.03
	Holmes	.09∅	−.05	.31	.50	−.21	−.15	−.49	.24∅	.37	−.01	.77
VideoChat3-4B	Ref.	.23	.01	.38	**.39**	−.26	.04	−.26	**.43**	.14	.07	
	VERA	−.16	−.07	.23	.21°	−.25	.01	−.58	.07∅	.23	.13	.68
	AnyA.	−.01	−.06	.15	.10°	−.08	.06	.00	.11∅	.00	−.06	.69
	Event	.04	−.02	.31	.15°	−.16	.15	−.04	.14∅	.18	.03	.66
	PRISM	.04	−.06	−.15	.09°	−.15	−.21	−.25	.23°	−.24	−.11	.64
	HeadH.	.02	−.22	.38	.09°	−.19	−.18	−.42	.20°	−.27	.06	.81
	HiPr.	.33	−.11	.15	−.04°	−.36	−.07	−.26	.32°	.01	−.16	.83
	Holmes	.17	−.18	.54	.53°	−.17	−.27	−.52	.29°	.16	−.03	.90
JoyAI-VL-8B	Ref.	**.41**	−.17	.54	.36	−.26	−.07	−.25	**.59**	.22	−.01	
	VERA	.20°	−.29	.08	.13	−.11	−.04	−.28	.28∅	.08	−.03	.88
	AnyA.	.16°	−.16	.12	.22	−.04	−.13	−.05	.19∅	−.04	−.24	.62
	Event	.18°	−.08	.38	.18	−.23	−.14	−.07	.18∅	.24	.09	.84
	PRISM	.13°	.02	−.31	.13	**−.40**+	−.13	−.08	.21∅	−.08	.04	.58
	HeadH.	−.20∅	−.14	.00	.12	**−.44**+	−.04	−.51	.32°	.07	.00	.72
	HiPr.	−.09°	−.27	.12	.31	−.25	−.18	−.46	.46°	.10	−.28	.82
	Holmes	.14°	−.16	.38	.38	−.19	−.18	−.49	.40°	.25	−.10	.94

E.3 Cross-Dataset Comparisons

Table 13 compares the vectors of cue-level values across datasets, and Figure 7 shows the spread over backbones behind each mean. Cross-dataset entries are means over the backbones available in both datasets.

Do cue-level patterns and their changes recur across datasets? We use the same cue definitions, annotation procedure, pairing criteria, and estimator on each dataset, then correlate the vectors of cue-level values over their shared eligible cues. Dataset-specific motion and duration thresholds still define relative groups within each dataset. Table 13 compares UCF-Crime, XD-Violence, and MSAD.

UCF-Crime and XD-Violence. The pretrained score vectors $\boldsymbol{d}^{\mathrm{ref}}_{b,d}$ of the two crime-centric datasets correlate at .33 on average over the eleven backbones. The VAD vectors agree across the two datasets for Holmes-VAU (.63) and PRISM (.52), less for VERA (.29), AnyAnomaly (.23), EventVAD (.13), and HeadHunt-VAD (.10), and not for HiProbe-VAD ($-.03$). The sign of a detectable $\Delta D_{b,m,d,c}$ can be compared only for cells detectable in both datasets: it agrees for all six such cells of HeadHunt-VAD, six of nine for PRISM, three of five for AnyAnomaly, two of four for Holmes-VAU, two of five for EventVAD, and two of six for VERA, and HiProbe-VAD has no such cell. Because no $\Delta D_{b,m,d,c}$ is detectable on MSAD, no direction comparison involving MSAD is defined.

MSAD. The reference vectors of MSAD correlate with those of UCF-Crime at .50 over eleven backbones and with XD-Violence at .55 over eleven, more than the two crime datasets with each other; the VAD vectors of VERA (.51 and .62), EventVAD (.53), HiProbe-VAD (.49), and Holmes-VAU (.67 and .82) agree to a similar degree, PRISM (.29) and AnyAnomaly (.13) less. Within MSAD the mean $\rho_{b,m,d}$ ranges from .18 (PRISM) to .70 (Holmes-VAU), but at most two backbones per frozen-backbone method (five of eleven for Holmes-VAU) have an interval above zero, no $\Delta D_{b,m,d,c}$ is detectable (six cells are classified as introduced, none with a detectable $\Delta D_{b,m,d,c}$), and detection performance is high for the reasoning, representation, and adaptation methods (87.6 to 93.3 AUC) and lower for the evidence methods (72.6 to 75.3) and PRISM (66.7). The multi-scenario dataset therefore reproduces the direction of the cue-level pattern found on the crime-centric benchmarks without the statistical power to resolve individual changes.

Table 13: Cue-level agreement across datasets. Left: Spearman correlation between the vectors of cue-level values $\boldsymbol{d}^{\mathrm{VAD}}_{b,m,d}$ ("pretrained score": $\boldsymbol{d}^{\mathrm{ref}}_{b,d}$) of two datasets over their shared eligible cues; dir.: UCF-Crime–XD-Violence backbone-cue cells whose $\Delta D_{b,m,d,c}$ is detectable in both datasets and has the same sign, out of all cells detectable in both (–: no such cell). No $\Delta D_{b,m,d,c}$ is detectable on MSAD (Table 6), so a direction agreement involving MSAD is undefined and not shown. Right: MSAD, mean $\rho_{b,m,d}$, number of backbones with an interval above 0, and mean AUC ($\times 100$). All means are over the eleven backbones; a bracketed count marks an entry based on fewer.

	Spearman($\boldsymbol{d}_d, \boldsymbol{d}_{d'}$)			dir.	MSAD		
Method	UCF–XD	UCF–MSAD	XD–MSAD	UCF–XD	ρ	CI>0	AUC
Pretrained reference	.33	.50	.55	–	–	–	91.7
VERA	.29	.51	.62	2/6	.66	2	87.6
AnyAnomaly	.23	.13	.68	3/5	.39	0	72.6
EventVAD	.13	.53	.20	2/5	.45	0	75.3
PRISM	.52	.29	.04	6/9	.18	0	66.7
HeadHunt-VAD	.10	.45	.16	6/6	.45	1	89.7
HiProbe-VAD	$-.03$	.49	.20	–	.59	2	89.0
Holmes-VAU	.63	.67	.82	2/4	.70	5	93.3

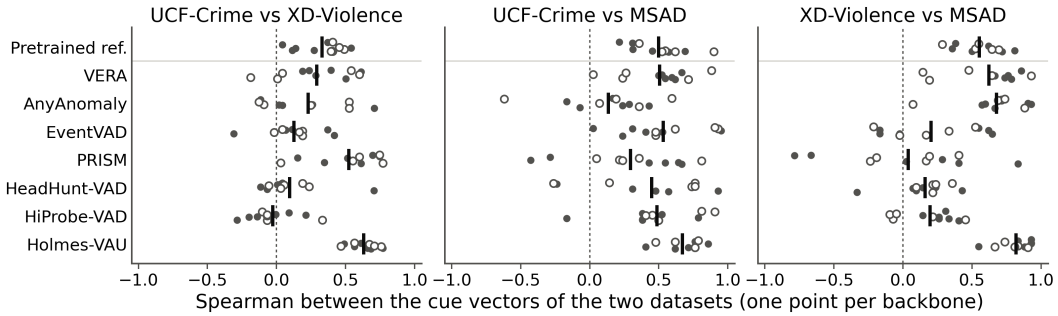


Figure 7: Cross-dataset agreement per backbone. Each point is one backbone: the Spearman correlation between its cue vectors on two datasets over their shared eligible cues ($\boldsymbol{d}^{\mathrm{ref}}_{b,d}$ for the pretrained reference, $\boldsymbol{d}^{\mathrm{VAD}}_{b,m,d}$ for the methods); hollow: backbones of at most 4B; black tick: mean over backbones, the value in Table 13.

E.4 Robustness and Sensitivity Analyses

Do the comparisons depend on the annotation rule, prompt, or summary measure? Table 14 changes one choice at a time.

Frame-count sensitivity. Eight uniformly sampled frames are a fixed compromise between temporal coverage and inference cost. On UCF-Crime we rescore the pretrained score of all eleven backbones with 4 and 16 frames, holding the 10 s segments, content annotations, anomaly question, and valid pairs of each comparison fixed; the VAD scores keep each method's own inference settings. The ten-cue vector of $D^{\mathrm{ref}}_{b,d,c}$ changes little: its Spearman correlation with the eight-frame vector is .95 on average over backbones with four frames (range .90 to .99) and .94 with sixteen (.87 to .99), and the four- and sixteen-frame vectors correlate at .93 (.81 to .99), comparable to the prompt paraphrases below. The recurring reference differences keep their direction and frequency: high motion is detectable on nine backbones with four frames and on ten with sixteen (nine with eight), person on five and six (five), and dense people remains negative ($-.40$ and $-.36$ against $-.38$). Comparing the VAD scores with the four- or sixteen-frame reference gives a VERA mean $\rho_{b,m,d}$ of .69 and .68 (eight and nine of eleven intervals above zero; main setting .67, nine) and a mean over all methods of .59 in both cases (49 and 51 of 77 cells above zero; main setting .57, 52). The frame count therefore shifts individual reference differences slightly but leaves the cue-related ordering against which the VAD scores are compared, and the retention measured by $\rho_{b,m,d}$, essentially unchanged. The check varies only the reference; the VAD methods are not rerun with other frame counts.

Prompt paraphrases. The pretrained score pattern is stable under rewording: the Spearman correlation between the ten-cue vectors of $D^{\mathrm{ref}}_{b,d,c}$ obtained with p_0 and with p_1 is .93 on average over

Table 14: Robustness of the reference-VAD correlation $\rho_{b,m,d}$ on UCF-Crime. Left: VERA, mean over the eleven backbones, range, and number of backbones whose interval lies above 0. Right: all methods with complete coverage on UCF-Crime, mean ρ and the number of backbone-method cells whose interval lies above 0. Each variant changes one choice of the main setting (unanimous annotations, prompt p_0, D_c, Spearman).

	VERA (UCF-Crime)			All methods	
Variant	mean ρ	range	CI>0	mean ρ	CI>0
Main: unanimous, p_0, D_c, Spearman	.67	[.31, .88]	9	.57	52
Majority ($\geq 2/3$) annotations	.76	[.56, .92]	9	.63	57
Pearson instead of Spearman	.71	[.50, .90]	9	.61	51
Matched Cohen's d instead of D_c	.68	[.52, .90]	9	.59	54
Pretrained reference with p_1	.63	[.28, .88]	9	.58	51
Pretrained reference with p_2	.66	[.25, .92]	8	.58	49
EventVAD: failed answers excluded	–	–	–	.53	8

backbones (range .79 to 1.00), and .94 with p_2 (range .85 to .99). Using p_1 or p_2 as the reference in the VERA comparison gives mean $\rho_{b,m,d}$ of .63 and .66 instead of .67 (Table 14).

Annotation rule, estimator, and correlation. Majority annotations increase coverage and give a slightly higher correlation than unanimous annotations (VERA mean $\rho_{b,m,d}$ .76 vs. .67, with intervals above zero for nine of eleven backbones in both settings; all methods .63 vs. .57, 57 vs. 52 of 77 cells above zero). Pearson correlation gives .71 for VERA and .61 over all methods, and the matched Cohen's d estimator .68 and .59 (Table 14). Excluding EventVAD's unparsable answers instead of scoring them -1 changes its mean correlation from .51 to .53 and leaves eight of eleven intervals above zero.

Unmatched versus matched estimates. Without matching, running receives detectably higher reference scores in all eleven backbones and street in seven. After matching on anomaly status and the specified pairing criteria, neither is detectable in any backbone, whereas high motion ($10 \rightarrow 9$ backbones) and person ($6 \rightarrow 5$) survive matching. Unmatched differences should therefore not be read as the matched cue-related difference; the comparison does not establish that the remaining difference is zero.

Cue-related differences and detection performance. The association between $\rho_{b,m,d}$ and detection performance is analysed in Appendix H.2.

E.5 Layered Analysis

This section breaks the comparison down one factor at a time: by dataset, by dataset and method (Table 15), by method and backbone (Table 16, Figure 8, and Table 7), and by backbone and dataset (Table 17). Appendix E.6 models all three factors together, and Appendix E.7 contrasts fine-tuning with the frozen-backbone methods. Unless stated otherwise, a configuration is one backbone–method–dataset combination ($11 \times 7 \times 3 = 231$) and a cell is one configuration and cue.

Dataset. UCF-Crime has the most configurations with an interval above zero (52 of 77), XD-Violence the most detectable changes (258 of 770 cells), and MSAD none (0 of 616), although its

Table 15: Dataset and dataset×method summaries. ρ, R: means of $\rho_{b,m,d}$ and $R_{b,m,d}$ over backbones; CI>0: configurations whose $\rho_{b,m,d}$ interval lies above zero; N_{chg}: cells with a detectable $\Delta D_{b,m,d,c}$ out of all cells; Pres., Amp., Att., Sup., Int., Rev.: preserved, amplified, attenuated, suppressed, introduced, and reversed cells; N.c.: not classified; Perf.: mean frame-level AUC (UCF-Crime, MSAD) or AP (XD-Violence), $\times 100$.

	ρ	CI>0	R	N_{chg}	Pres.	Amp.	Att.	Sup.	Int.	Rev.	N.c.	Perf.
UCF-Crime												
All methods	.57	52/77	.25	84/770	0	0	0	56	18	0	63	74.1
VERA	.67	9/11	.23	14/110	0	0	0	9	3	0	8	75.7
AnyAnomaly	.46	5/11	.25	16/110	0	0	0	13	0	0	4	69.1
EventVAD	.51	8/11	.23	14/110	0	0	0	10	0	0	7	67.0
PRISM	.46	4/11	.32	16/110	0	0	0	8	12	0	9	69.0
HeadHunt-VAD	.59	7/11	.26	10/110	0	0	0	8	3	0	9	77.5
HiProbe-VAD	.54	8/11	.24	0/110	0	0	0	0	0	0	17	75.3
Holmes-VAU	.78	11/11	.18	14/110	0	0	0	8	0	0	9	85.5
XD-Violence												
All methods	.39	41/77	.22	258/770	62	8	9	86	139	18	55	56.0
VERA	.54	6/11	.19	40/110	8	0	2	12	14	1	11	62.9
AnyAnomaly	.39	4/11	.21	36/110	3	0	0	21	18	1	9	54.9
EventVAD	.19	5/11	.21	41/110	9	0	1	18	12	3	3	50.3
PRISM	.21	4/11	.31	54/110	14	2	1	7	41	8	2	50.4
HeadHunt-VAD	.39	7/11	.25	38/110	6	5	1	11	26	2	9	48.2
HiProbe-VAD	.36	5/11	.21	30/110	9	1	2	10	21	3	9	43.9
Holmes-VAU	.64	10/11	.13	19/110	13	0	2	7	7	0	12	81.1
MSAD												
All methods	.49	10/77	.30	0/616	0	0	0	0	6	0	0	82.0
VERA	.66	2/11	.22	0/88	0	0	0	0	0	0	0	87.6
AnyAnomaly	.39	0/11	.32	0/88	0	0	0	0	0	0	0	72.6
EventVAD	.45	0/11	.28	0/88	0	0	0	0	0	0	0	75.3
PRISM	.18	0/11	.46	0/88	0	0	0	0	2	0	0	66.7
HeadHunt-VAD	.45	1/11	.39	0/88	0	0	0	0	2	0	0	89.7
HiProbe-VAD	.59	2/11	.26	0/88	0	0	0	0	2	0	0	89.0
Holmes-VAU	.70	5/11	.20	0/88	0	0	0	0	0	0	0	93.3

Table 16: Agreement between cue vectors (mean Spearman correlation over the ten cues, eight on MSAD). Reference: $\boldsymbol{d}^{\mathrm{ref}}_{b,d}$ of two backbones. Methods, same backbone: vectors of two methods on the same backbone, averaged over backbones and method pairs. Backbones, same method: vectors of two backbones under the same method, averaged over methods and backbone pairs.

	Reference	$\boldsymbol{d}^{\mathrm{VAD}}_{b,m,d}$		$\Delta D_{b,m,d,c}$	
Dataset	backbones	methods, same backbone	backbones, same method	methods, same backbone	backbones, same method
UCF-Crime	.82	.47	.48	.47	.40
XD-Violence	.65	.31	.57	.35	.50
MSAD	.70	.48	.61	.38	.36

mean R is the largest (.30). With 44–123 pairs per analysed visual cue on MSAD (Appendix C.2), the absence of detectable change there reflects few pairs rather than an absence of change; six MSAD cells are nevertheless classified as introduced because their VAD difference is detectable.

Dataset×method. On UCF-Crime, HiProbe-VAD reaches a mean ρ of .54 without a single detectable cell change: its low correlations on the 2B backbones come from many small, individually undetectable shifts, whereas PRISM and AnyAnomaly combine a similar mean correlation with the most detectable changes. On XD-Violence, where visual cues have 458–2,952 pairs, every frozen-backbone method has many detectable changes (PRISM 54 of 110 cells, EventVAD 41, VERA 40, HeadHunt-VAD 38, AnyAnomaly 36, HiProbe-VAD 30) and Holmes-VAU fewer (19). All categories occur there: amplified (HeadHunt-VAD five cells, PRISM two, HiProbe-VAD one), attenuated (VERA, HiProbe-VAD, and Holmes-VAU two each, EventVAD, PRISM, and HeadHunt-VAD one), reversed (PRISM eight, EventVAD and HiProbe-VAD three, HeadHunt-VAD two, VERA and AnyAnomaly one), and preserved (three to fourteen cells per frozen-backbone method, thirteen for Holmes-VAU). Suppression concentrates on the cues whose reference difference is detectable on most backbones: *dense people* (ten backbones; suppressed by AnyAnomaly on all ten and by the other frozen-backbone methods on six or seven, while PRISM amplifies it on three) and *person* (six backbones; suppressed by VERA, AnyAnomaly, and EventVAD on five or six, whereas HeadHunt-VAD amplifies or introduces it on ten). Introduced differences dominate where the reference shows none: PRISM introduces *street* on all eleven backbones and *indoor* on ten, AnyAnomaly introduces *running* on nine, and VERA introduces *long event* on nine. Across datasets, the method order by ρ is similar (Holmes-VAU and VERA highest; EventVAD and PRISM lowest on XD-Violence, PRISM lowest on MSAD), while the number of detectable changes depends mainly on the dataset.

Method×backbone. Table 16 compares how similar the cue vectors are across backbones and across methods. The reference vectors of two backbones agree closely (.65–.82); after VAD processing, two backbones under the same method still agree at .48–.61, but two methods on the same backbone agree only at .31–.48. For the changes $\Delta D_{b,m,d,c}$ the two comparisons are close (.35–.47 against .36–.50), because methods on the same backbone often suppress the same strong reference differences. Holding the backbone fixed shows how differently methods act on the same reference: on GLM-4.1V-9B the reference *high motion* difference is suppressed by AnyAnomaly, EventVAD, PRISM, and Holmes-VAU but not detectably changed by VERA, HeadHunt-VAD, or HiProbe-VAD; on Qwen3-VL-4B it is suppressed by five of the six frozen-backbone methods and by Holmes-VAU, while PRISM additionally introduces *indoor* and *night*. The same backbone yields $\rho_{b,m,d}$ from .10 (PRISM) to .85 (EventVAD) on Qwen2.5-VL-3B and from .18 (PRISM) to .93 (Holmes-VAU) on InternVL3-8B. Holding the method fixed, VERA is positive on all eleven backbones (.31–.88), highest on JoyAI-VL-8B (.88), InternVL3-8B, and Qwen3-VL-4B (.81 each) and lowest on VideoLLaMA3-7B (.31); ρ and R can disagree within a method (VERA on Qwen3-VL-8B: $\rho = .72$ with the largest change, $R = .31$; on GLM-4.1V-9B and InternVL3-8B: $\rho \geq .78$ with no detectable cell change). Every negative $\rho_{b,m,d}$ on UCF-Crime belongs to a 2B backbone or to VideoLLaMA3-7B (EventVAD on VideoLLaMA3-2B, $-.62$, and SmolVLM2-2B, $-.08$; HiProbe-VAD on SmolVLM2-2B, VideoLLaMA3-2B, and VideoLLaMA3-7B, $-.12$ to $-.02$; AnyAnomaly on VideoLLaMA3-2B, $-.42$), and these configurations also detect poorly (AUC 46.6–67.2), although some weak configurations keep a positive ρ (AnyAnomaly on SmolVLM2-2B, 43.3 AUC; EventVAD on GLM-4.1V-9B, 48.6). On XD-Violence the negative correlations are mostly, but not only, weak configurations (EventVAD on SmolVLM2-2B, GLM-4.1V-9B, and VideoLLaMA3-2B at 23–26 AP, the GLM-4.1V-9B value of $-.36$ with an interval below zero; AnyAnomaly on SmolVLM2-2B at 23.9 AP; HiProbe-VAD on JoyAI-VL-8B at 37.4 AP; but also PRISM on Qwen3-

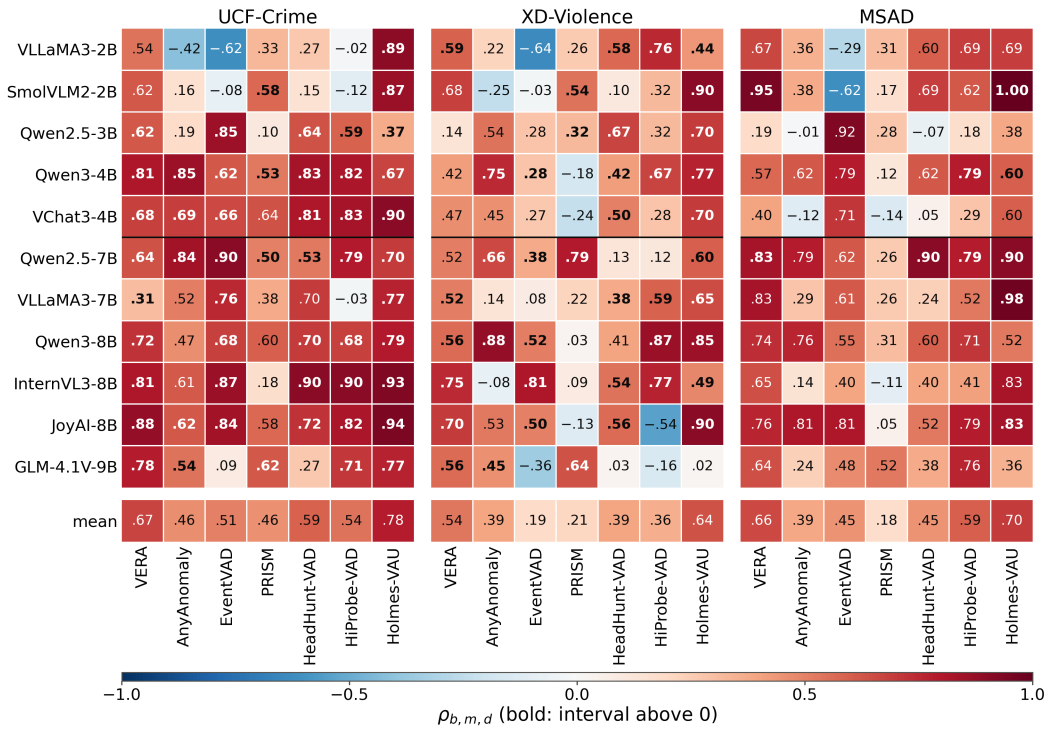


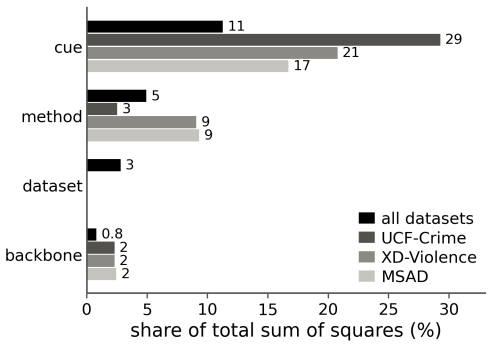
Figure 8: Reference–VAD cue-order correlation $\rho_{b,m,d}$ of every configuration. Rows: backbones ordered by parameter count (the line separates the five backbones of at most 4B) and the mean over backbones; columns: methods. Bold: bootstrap interval above zero. The values are those of Table 7.

Table 17: Backbone×dataset summaries. Ref. det.: cues with a detectable reference difference (VERA pairs); Ref. perf.: detection performance of the pretrained reference (×100); ρ: mean $\rho_{b,m,d}$ over the seven methods [min, max over methods]; N_{chg}: cells with a detectable $\Delta D_{b,m,d,c}$ over the seven methods; VAD perf.: mean detection performance of the seven methods.

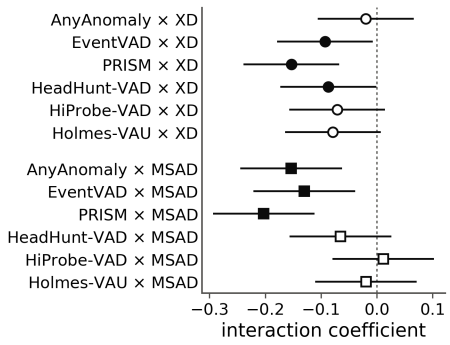
	UCF-Crime					XD-Violence					MSAD				
Backbone	Ref. det.	Ref. perf.	ρ	N_{chg}	VAD perf.	Ref. det.	Ref. perf.	ρ	N_{chg}	VAD perf.	Ref. det.	Ref. perf.	ρ	N_{chg}	VAD perf.
Qwen2.5-VL-3B	2	77.2	.48 [.10, .85]	10	73.5	4	55.0	.43 [.14, .70]	25	54.2	0	87.0	.27 [−.07, .92]	0	80.9
Qwen2.5-VL-7B	0	81.8	.70 [.50, .90]	5	79.3	2	63.8	.46 [.12, .79]	24	55.0	0	92.5	.73 [.26, .90]	0	83.9
Qwen3-VL-4B	2	81.9	.73 [.53, .85]	16	79.1	2	64.3	.45 [−.18, .77]	18	64.3	0	92.3	.59 [.12, .79]	0	85.9
Qwen3-VL-8B	2	80.6	.66 [.47, .79]	11	78.1	1	74.1	.59 [.03, .88]	18	65.4	0	92.6	.60 [.31, .76]	0	86.8
InternVL3-8B	1	84.1	.74 [.18, .93]	3	77.0	3	65.3	.48 [−.08, .81]	20	57.8	0	92.6	.39 [−.11, .83]	0	82.6
SmolVLM2-2B	0	82.3	.31 [−.12, .87]	3	62.6	2	62.7	.32 [−.25, .90]	22	43.4	0	90.9	.46 [−.62, 1.0]	0	67.2
GLM-4.1V-9B	2	78.6	.54 [.09, .78]	9	72.8	4	64.9	.17 [−.36, .64]	28	55.3	0	92.2	.48 [.24, .76]	0	80.2
VideoLLaMA3-2B	2	77.9	.14 [−.62, .89]	8	66.0	5	46.8	.32 [−.64, .76]	24	46.1	0	91.6	.43 [−.29, .69]	0	77.8
VideoLLaMA3-7B	2	80.0	.49 [−.03, .77]	10	73.9	4	54.9	.37 [.08, .65]	30	58.2	0	93.0	.53 [.24, .98]	0	87.4
VideoChat3-4B	2	79.7	.74 [.64, .90]	4	76.7	4	65.9	.35 [−.24, .70]	26	58.0	0	92.2	.26 [−.14, .71]	0	86.7
JoyAI-VL-8B	2	77.3	.77 [.58, .94]	5	76.7	3	72.2	.36 [−.54, .90]	23	57.9	0	91.5	.65 [.05, .83]	0	82.9

VL-4B, VideoChat3-4B, and JoyAI-VL-8B, −.24 to −.13 at 48.3–61.5 AP, HiProbe-VAD on GLM-4.1V-9B, −.16 at 51.3 AP, and AnyAnomaly on InternVL3-8B, −.08 at 61.3 AP, all with intervals including zero). Figure 8 shows every $\rho_{b,m,d}$ at once, and the full backbone-level values are in Table 7.

Backbone×dataset. The reference detects anomalies similarly well on all backbones (UCF-Crime AUC 77–84), yet the share of its cue ordering that survives VAD processing differs widely: the mean ρ over methods on UCF-Crime ranges from .14 (VideoLLaMA3-2B) and .31 (SmolVLM2-2B) to .77 (JoyAI-VL-8B). The reference performance of a backbone hardly predicts this mean ρ (Spearman over backbones .14, .35, and .27 on the three datasets), whereas the mean VAD performance does on UCF-Crime and XD-Violence (.68 and .50). Backbones with more detectable reference differences also have more detectable changes (Spearman .66 on UCF-Crime, .76 on XD-Violence), because most changes remove a reference difference. The ranking of backbones by mean ρ is not stable across datasets (for example, Qwen2.5-VL-3B is below the dataset mean on UCF-Crime and MSAD but slightly above it on XD-Violence). The backbone×dataset terms of the model in Ap-



(a) Type-II share of the total sum of squares.



(b) Method×dataset terms (95% intervals).

Figure 9: Sources of variation in $\Delta D_{b,m,d,c}$. (a) Share of each factor in the main-effects decomposition, over all datasets and within each dataset (where the dataset factor drops out). (b) The twelve method×dataset terms of the full model with Wald 95% intervals, relative to VERA on UCF-Crime; filled markers: $p < .05$ (uncorrected).

pendix E.6, however, reach $p < .05$ about as often as expected by chance, so this instability is descriptive rather than an established interaction.

E.6 Sources of Variation in Cue-Level Changes

This exploratory analysis (Figure 9) asks whether cue-level changes relative to the pretrained score vary systematically with backbone, method, and dataset. It models the signed changes $\Delta D_{b,m,d,c}$ of all 2,156 backbone-method-dataset-cue cells (eleven backbones, seven methods, three datasets; running and early event have too few pairs on MSAD) with

$$\Delta D_{b,m,d,c} = \beta_0 + \beta_b^{(B)} + \beta_m^{(M)} + \beta_d^{(D)} + \beta_{b,m}^{(BM)} + \beta_{m,d}^{(MD)} + \beta_{b,d}^{(BD)} + u_c + \epsilon, \qquad (15)$$

where the fixed terms describe backbone, method, dataset, and pairwise interactions, and u_c is a cue random intercept. This model concerns signed changes, not their absolute magnitude or causal contributions.

Because every method is evaluated on all eleven backbones and three datasets, the fixed-effects design has full column rank (111 of 111 columns), and the REML fit (statsmodels MixedLM) converges. The cue random intercept has variance .008 against a residual variance of .053, *i.e.,* cues account for 13% of the variance left after the fixed terms. Among the pairwise interactions, method × dataset carries the clearest structure: six of its twelve terms have $p < .05$, all negative relative to VERA on UCF-Crime, namely PRISM on MSAD ($-.20$) and XD-Violence ($-.15$), AnyAnomaly on MSAD ($-.15$), EventVAD on MSAD ($-.13$) and XD-Violence ($-.09$), and HeadHunt-VAD on XD-Violence ($-.09$). Backbone × method and backbone × dataset terms reach $p < .05$ in three of sixty and three of twenty cases, close to the number expected by chance without correction. A main-effects decomposition (OLS with Type-II sums of squares) gives the same ordering: the cue accounts for 11% of the total sum of squares, the method for 5%, the dataset for 3%, and the backbone for 1% ($R^2 = .19$); within a dataset the cue share is 29%, 21%, and 17% and the method share 3%, 9%, and 9% on UCF-Crime, XD-Violence, and MSAD. The variation of $\Delta D_{b,m,d,c}$ is therefore organised mainly by the cue and by the method-dataset combination, with little systematic backbone contribution once method and dataset are fixed. The p-values are not corrected for multiple testing, and the model treats cue estimates from shared videos as independent, so the analysis describes the variation rather than identifying its sources.

E.7 Fine-Tuning versus Frozen-Backbone Methods

Holmes-VAU is the only method that updates the backbone. Its first stage queries the LoRA-adapted backbone with the reference frames, anomaly question, and readout, so it isolates the effect of fine-tuning, whereas the frozen-backbone methods change the evidence or the readout. Table 18 com-

Table 18: Fine-tuning versus frozen-backbone methods. SD ratio: standard deviation of D_c across the eleven backbones after the method divided by that before it, averaged over cues, with a 95% bootstrap interval over backbones (< 1: backbones become more alike); *: the difference from Holmes-VAU stage 1 has an interval above zero. $|\Delta D|$: mean absolute change over the cue×backbone cells.

	UCF-Crime		XD-Violence		MSAD	
Method	SD ratio	$\lvert\Delta D\rvert$	SD ratio	$\lvert\Delta D\rvert$	SD ratio	$\lvert\Delta D\rvert$
Holmes-VAU, stage 1 (adapted backbone)	0.82 [0.61, 1.15]	.14	0.62 [0.54, 0.76]	.08	0.84 [0.65, 1.16]	.13
Holmes-VAU, final	0.84 [0.72, 1.07]	.15	0.70 [0.53, 0.98]	.10	0.99 [0.77, 1.49]	.17
VERA	1.44 [1.19, 2.00]*	.19	1.42 [1.00, 1.93]*	.15	1.27 [1.11, 1.71]*	.19
AnyAnomaly	0.99 [0.81, 1.30]	.21	0.90 [0.66, 1.23]*	.16	0.96 [0.76, 1.32]	.26
EventVAD	1.29 [0.97, 1.89]*	.19	1.36 [0.95, 1.91]*	.17	1.18 [0.82, 1.79]*	.23
PRISM	1.74 [1.45, 2.36]*	.26	1.33 [1.00, 1.92]*	.26	2.02 [1.61, 2.83]*	.38
HeadHunt-VAD	1.50 [1.13, 2.04]*	.21	1.30 [1.11, 1.69]*	.19	1.30 [0.93, 1.89]*	.31
HiProbe-VAD	1.38 [1.06, 2.07]*	.20	2.04 [1.56, 2.94]*	.17	1.38 [1.05, 2.05]*	.22

pares two properties of $\Delta D_{b,m,d,c}$ over the cue×backbone cells (unanimous annotations, reference p_0).

Magnitude. Fine-tuning changes the cue-level differences least. The mean $|\Delta D_{b,m,d,c}|$ of the adapted backbone is .14, .08, and .13 on UCF-Crime, XD-Violence, and MSAD, against .15–.38 for the frozen-backbone methods (paired Wilcoxon test over the same cells, $p < .01$ in all eighteen comparisons); the full Holmes-VAU pipeline gives .15, .10, and .17. Part of this gap may reflect that the first stage keeps the reference readout, whereas most frozen-backbone methods replace it.

Convergence across backbones. For each cue we divide the standard deviation of D_c across the eleven backbones after the method by the standard deviation before it, and average the ratio over cues. For the adapted backbone the ratio is below one on all three datasets (.82, .62, .84; the bootstrap interval over backbones excludes one on XD-Violence only). For the frozen-backbone methods it lies between .90 and 2.04 and is below one only for AnyAnomaly. The difference from the adapted backbone is detectable for sixteen of the eighteen frozen method–dataset pairs, the exceptions being AnyAnomaly on UCF-Crime and MSAD. Methods that train or reason on frozen features thus preserve or amplify the backbone-specific differences of the reference, whereas updating the backbone on shared VAD data pulls the backbones toward a common pattern. The cues it changes are those the frozen-backbone methods change: on UCF-Crime the adapted backbone lowers *person* ($-.24$ on average, detectable on six backbones) and *high motion* ($-.16$, four), which the frozen-backbone methods lower by $-.31$ and $-.30$ on average.

Backbone families. Over the twenty cue–dataset cells of UCF-Crime and XD-Violence, the ΔD profiles of the adapted backbones correlate more within a family (mean Spearman .56 over five pairs) than across families (.34 over fifty pairs; permutation $p = .03$ over family labels, uncorrected). The clearest case is the Qwen3-VL family (Qwen3-VL-4B, Qwen3-VL-8B, and JoyAI-VL-8B, which uses the Qwen3-VL architecture; pairwise .51–.71): all three lower the *person* difference on both datasets ($-.33$ to $-.53$ on UCF-Crime, $-.01$ to $-.15$ on XD-Violence; four detectable cells). These three backbones also have large reference *person* differences on UCF-Crime (.41–.64; outside the family only Qwen2.5-VL-3B, .49, and VideoLLaMA3-7B, .41, fall in this range), so the family pattern cannot be separated from the size of the reference difference. VideoLLaMA3-2B is the only backbone whose adapted version detectably raises the *person* difference, on both datasets. With two or three backbones per family and no reliable pattern on MSAD, where no change is detectable for any method, these family-level observations remain exploratory.

F Stage-wise Behavioral Analysis

Table 19 summarizes the stage-wise changes on UCF-Crime for the four staged methods. The full stage table lists every configuration on the three datasets; all rows are means over the eleven backbones. On MSAD every stage moves less than on UCF-Crime and no ΔD is detectable at any stage.

Table 19: Stage-wise scoring changes on UCF-Crime (means over the eleven backbones). Stage 0 is the pretrained score. $\rho^{(k)}_{\mathrm{ref}}$ and $\rho^{(k)}_{\mathrm{prev}}$ (Eq. 13) compare stage k with the reference and with the preceding stage; $R^{(k)}$ (Eq. 14) is the root-mean-square change from the reference; $N^{(k)}_{\mathrm{chg}}$ counts backbone-cue cells with a detectable $\Delta D^{(k)}_{c,\mathrm{ref}}$; AUC is frame-level ($\times 100$). The complete table for all datasets is in Appendix F.

Method	Stage k	$\rho^{(k)}_{\mathrm{ref}}$	$\rho^{(k)}_{\mathrm{prev}}$	$R^{(k)}$	$N^{(k)}_{\mathrm{chg}}$	AUC
VERA	0: pretrained reference	1	–	0	0	80.1
	1: guided reflection	.77	.77	.18	10	74.6
	2: scene-context weighting	.77	.93	.19	9	76.2
	3: temporal smoothing	.67	.86	.23	14	75.7
AnyAnomaly	0: pretrained reference	1	–	0	0	80.1
	1: key-frame VQA	.48	.48	.26	16	67.0
	2: context fusion	.46	.86	.25	16	69.1
PRISM	0: pretrained reference	1	–	0	0	80.1
	1: semantic readout	.44	.44	.32	11	68.9
	2: temporal smoothing	.46	.97	.32	16	69.0
Holmes-VAU	0: pretrained reference	1	–	0	0	80.1
	1: adapted VLM score	.80	.80	.17	11	84.4
	2: final Holmes score	.78	.95	.18	14	85.5

Table 20: Stage-wise scoring changes on UCF-Crime, XD-Violence, and MSAD (means over the eleven backbones). Stage 0 is the pretrained score on the common evaluation segments. $\rho^{(k)}_{\mathrm{ref}}$: correlation with stage 0; $\rho^{(k)}_{\mathrm{prev}}$: with the preceding stage; $R^{(k)}$: root-mean-square change from stage 0; $N^{(k)}_{\mathrm{chg}}$: backbone-cue cells with a detectable $\Delta D^{(k)}_{c,\mathrm{ref}}$; Perf.: AUC (UCF-Crime, MSAD) or AP (XD-Violence), $\times 100$.

		UCF-Crime					XD-Violence				MSAD			
Method	Stage	$\rho^{(k)}_{\mathrm{ref}}$	$\rho^{(k)}_{\mathrm{prev}}$	$R^{(k)}$	$N^{(k)}_{\mathrm{chg}}$	Perf.	$\rho^{(k)}_{\mathrm{ref}}$	$R^{(k)}$	$N^{(k)}_{\mathrm{chg}}$	Perf.	$\rho^{(k)}_{\mathrm{ref}}$	$R^{(k)}$	$N^{(k)}_{\mathrm{chg}}$	Perf.
VERA	0: pretrained reference	1	–	0	0	80.1	1	0	0	62.7	1	0	0	91.7
	1: guided reflection	.77	.77	.18	10	74.6	.58	.15	24	60.2	.59	.25	0	86.3
	2: scene-context weighting	.77	.93	.19	9	76.2	.50	.16	25	62.5	.59	.25	0	86.3
	3: temporal smoothing	.67	.86	.23	14	75.7	.54	.19	40	62.9	.66	.22	0	87.6
AnyAnomaly	0: pretrained reference	1	–	0	0	80.1	1	0	0	62.7	1	0	0	91.7
	1: key-frame VQA	.48	.48	.26	16	67.0	.28	.20	43	49.8	.41	.33	0	72.3
	2: context fusion	.46	.86	.25	16	69.1	.39	.21	36	54.9	.39	.32	0	72.6
PRISM	0: pretrained reference	1	–	0	0	80.1	1	0	0	62.7	1	0	0	91.7
	1: semantic readout	.44	.44	.32	11	68.9	.20	.31	61	51.6	.22	.46	0	67.6
	2: temporal smoothing	.46	.97	.32	16	69.0	.21	.31	54	50.4	.18	.46	0	66.7
Holmes-VAU	0: pretrained reference	1	–	0	0	80.1	1	0	0	62.7	1	0	0	91.7
	1: adapted VLM score	.80	.80	.17	11	84.4	.78	.10	17	78.9	.76	.16	0	93.2
	2: final Holmes score	.78	.95	.18	14	85.5	.64	.13	19	81.1	.70	.20	0	93.3

G Temporal Behavioral Analysis

Event-timing comparisons. As an additional analysis, D_{early} and D_{long} apply the matched sign comparison of Eq. 3 to the during-event evidence L^{ev} (Eq. 11) of matched early/late and long/short events. These comparisons use one response per event, whereas the segment-based event-attribute comparison in Sec. 2.1 uses all eligible segments of the source videos. The event-level grouping and matching settings are specified in Sec. B.3.

Table 21 gives the full set of temporal quantities on the three datasets, including the event-level comparisons D_{early} and D_{long}. These are estimated on UCF-Crime (17 early/late and 32 long/short event pairs) and XD-Violence (183 and 216) but not on MSAD: its test videos are short and mostly contain a single event, only 13 events fall into the early or late onset group and 22 into the long or short duration group, and requiring a partner of the same anomaly category from a different video leaves one and three pairs, below the minimum of ten.

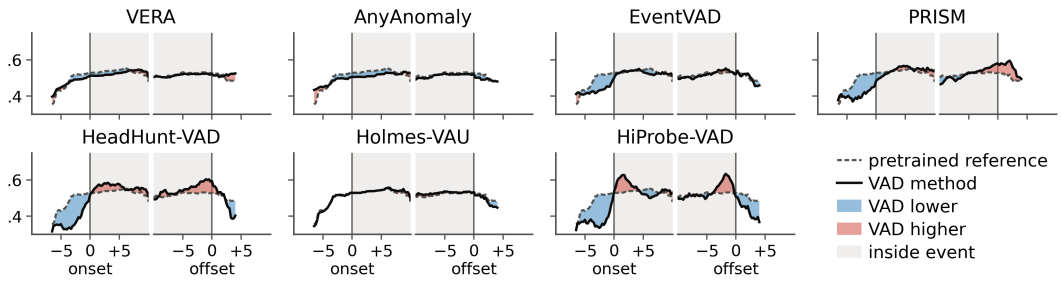


Figure 10: Event-aligned anomaly evidence on MSAD, drawn as in Figure 6. Lags with fewer than 20 contributing videos are omitted, so the curves end early on these short videos. HiProbe-VAD and HeadHunt-VAD lower the evidence before onset and raise it just after onset and before offset, which increases their event concentration C^{ev} over the reference.

Table 21: Event-aligned anomaly evidence on all three datasets, one panel per dataset (means over backbones; within-video percentile scores; $h = 10\,\mathrm{s}$). L^{pre}, L^{ev}, L^{post}: excess evidence before, during, and after the annotated event relative to background (Eq. 11); C^{ev}: event concentration (Eq. 12); D_{early}, D_{long}: event-level matched sign averages of L^{ev} for early versus late and long versus short events; –: fewer than ten matched event pairs. On MSAD, matching events of the same anomaly category from different videos yields one early/late pair (of 13 events with a defined onset group) and three long/short pairs (of 22 events), so both comparisons are omitted there. Each VAD entry is followed, in small type, by the mean change from the pretrained score on the same events and covered frames; * marks changes whose bootstrap interval excludes 0 in more than half of the backbones.

Method	L^{pre}	L^{ev}	L^{post}	C^{ev}	D_{early}	D_{long}
UCF-Crime						
Pretrained reference	.09	.24	.17	1.31	−.09	−.13
VERA	.04 (−.05)	.07 (−.18*)	.06 (−.11*)	1.08 (−.23*)	+.03 (+.12)	−.02 (+.11)
AnyAnomaly	.02 (−.07*)	.10 (−.14*)	.04 (−.13*)	1.13 (−.18*)	−.01 (+.09)	+.22 (+.35)
EventVAD	.00 (−.09*)	.11 (−.13*)	.07 (−.10*)	1.15 (−.16*)	−.19 (−.10)	+.07 (+.20)
PRISM	.03 (−.06)	.13 (−.12*)	.13 (−.05)	1.14 (−.17*)	−.21 (−.12)	+.14 (+.27)
HeadHunt-VAD	−.01 (−.09*)	.10 (−.15*)	.04 (−.14*)	1.14 (−.16*)	−.10 (−.01)	−.07 (+.06)
HiProbe-VAD	.00 (−.09*)	.09 (−.15*)	.02 (−.16*)	1.15 (−.16*)	−.09 (+.00)	−.07 (+.06)
Holmes-VAU	.12 (+.03)	.28 (+.04)	.18 (+.01)	1.35 (+.04)	−.21 (−.12)	−.10 (+.03)
XD-Violence						
Pretrained reference	.13	.31	.18	1.21	+.00	+.14
VERA	.10 (−.03*)	.17 (−.15*)	.13 (−.05*)	1.09 (−.12*)	−.04 (−.05)	+.10 (−.04)
AnyAnomaly	.08 (−.06*)	.20 (−.12*)	.09 (−.09*)	1.13 (−.08*)	+.03 (+.03)	+.19 (+.05)
EventVAD	.06 (−.07*)	.21 (−.10*)	.10 (−.08*)	1.16 (−.05)	+.03 (+.02)	+.11 (−.03)
PRISM	.15 (+.02)	.30 (−.02)	.20 (+.02)	1.16 (−.04*)	−.12 (−.12)	+.17 (+.03)
HeadHunt-VAD	.03 (−.11*)	.14 (−.17*)	.04 (−.14*)	1.12 (−.08*)	−.04 (−.04)	+.19 (+.05)
HiProbe-VAD	.01 (−.13*)	.14 (−.18*)	.04 (−.14*)	1.13 (−.07*)	−.04 (−.04)	+.08 (−.06)
Holmes-VAU	.18 (+.04*)	.39 (+.08*)	.21 (+.02)	1.25 (+.04*)	+.04 (+.03)	+.17 (+.03)
MSAD						
Pretrained reference	.13	.31	.18	1.06	–	–
VERA	.08 (−.05)	.21 (−.10)	.20 (+.01)	1.04 (−.02)	–	–
AnyAnomaly	.00 (−.12)	.14 (−.18)	.11 (−.08)	1.03 (−.03)	–	–
EventVAD	.02 (−.11)	.19 (−.12)	.08 (−.11)	1.09 (+.03*)	–	–
PRISM	.00 (−.12)	.17 (−.14)	.12 (−.06)	1.10 (+.04)	–	–
HeadHunt-VAD	.09 (−.04)	.35 (+.04)	.13 (−.05)	1.18 (+.12*)	–	–
HiProbe-VAD	.01 (−.11)	.32 (+.00)	.10 (−.08)	1.21 (+.15*)	–	–
Holmes-VAU	.12 (−.01)	.33 (+.02)	.18 (−.01)	1.07 (+.01)	–	–

H Detection Performance and Correlation Analysis

H.1 Detection Performance of the Pretrained Score

Table 22 reports the frame-level detection performance of the pretrained score itself, *i.e.,* the direct anomaly probe (Sec. 2.1) scored on the same 10 s evaluation grid with the segment score held constant over the segment's frames. It is the stage-0 configuration of every method in Table 6 and is reported separately because it is not a VAD method: it has no evidence construction, readout, or adaptation, and serves only as the reference against which retained and changed cue-related scoring behavior is measured.

H.2 Cue-Related Differences and Detection Performance

Table 7 lists the frame-level detection performance of every configuration next to its $\rho_{b,m,d}$.

Across the 231 backbone-method-dataset cells, $\rho_{b,m,d}$ and within-dataset detection performance are positively associated (Spearman .51 between $\rho_{b,m,d}$ and the within-dataset performance rank; .42 over 216 cells after excluding the fifteen configurations below 55 AUC or .30 AP), and the association weakens as the performance threshold is raised (.29 over the 157 cells above 75 AUC or .50 AP), so it is driven largely by the low-performing 2B configurations. Per dataset the association is .65 on UCF-Crime, .47 on MSAD, and .43 on XD-Violence. It is present within the evidence and representation methods (EventVAD .48, HiProbe-VAD .47, HeadHunt-VAD .46, AnyAnomaly .45, PRISM .14 with an interval including zero) and absent within VERA ($-.17$, 33 cells) and Holmes-VAU (.00, 33 cells). Retaining the reference ordering is therefore neither necessary nor sufficient for detection performance once the failing configurations are set aside.

Low-performing configurations. Table 23 repeats the reference–VAD comparison after removing the weakest configurations under two rules fixed in advance. Rule A removes five configurations per dataset, mostly AnyAnomaly and EventVAD on 2B backbones or GLM-4.1V-9B (plus HeadHunt-VAD on Qwen2.5-VL-7B on XD-Violence and PRISM on InternVL3-8B on MSAD); rule B removes the bottom quartile (19 per dataset). Removing them raises the mean $\rho_{b,m,d}$ (UCF-Crime .57 $\rightarrow$ .63 $\rightarrow$ .69, XD-Violence .39 $\rightarrow$ .43 $\rightarrow$ .45, MSAD .49 $\rightarrow$ .52 $\rightarrow$ .58) but leaves the qualitative pattern unchanged: on UCF-Crime the detectable changes remain suppressions of reference differences plus introduced context cues (52 and 18 cells under rule A), on XD-Violence introduced cells still outnumber suppressed ones (130 against 77 under rule A, 100 against 62 under rule B), and MSAD still has no detectable $\Delta D_{b,m,d,c}$. Holmes-VAU keeps the highest mean ρ on every dataset (tied with EventVAD on UCF-Crime under rule B), and PRISM stays the lowest on XD-Violence and MSAD. The method means most affected are those of EventVAD (.51 $\rightarrow$ .77 on UCF-Crime, .19 $\rightarrow$ .39 on XD-Violence) and AnyAnomaly on UCF-Crime (.46 $\rightarrow$ .59), whose lowest correlations come from the removed configurations; this reorders the middle of the method ranking on UCF-Crime (Spearman with the full ranking .61 and .50) but hardly on XD-Violence and MSAD (.86–.93). The low-performing configurations therefore lower the average retention without creating the cue-level findings.

Table 22: Frame-level detection performance of the pretrained direct anomaly score on the three benchmarks (annotated frames only).

	UCF-Crime		XD-Violence		MSAD	
Backbone	AUC	AP	AUC	AP	AUC	AP
Qwen2.5-VL-3B	77.2	22.9	85.2	55.0	87.0	56.0
Qwen2.5-VL-7B	81.8	27.6	89.0	63.8	92.5	73.3
Qwen3-VL-4B	81.9	35.8	89.2	64.3	92.3	73.2
Qwen3-VL-8B	80.6	31.0	91.9	74.1	92.6	74.9
InternVL3-8B	84.1	34.3	90.0	65.3	92.6	70.5
SmolVLM2-2B	82.3	30.5	87.9	62.7	90.9	73.6
GLM-4.1V-9B	78.6	27.2	89.2	64.9	92.2	74.2
VideoLLaMA3-2B	77.9	26.9	80.9	46.8	91.6	65.7
VideoLLaMA3-7B	80.0	29.2	85.3	54.9	93.0	74.8
VideoChat3-4B	79.7	31.9	89.5	65.9	92.2	74.1
JoyAI-VL-8B	77.3	25.5	91.1	72.2	91.5	70.8
Mean	80.1	29.3	88.1	62.7	91.7	71.0

Table 23: Sensitivity of the reference–VAD comparison to low-performing configurations. Excl. A drops configurations below 55 AUC (UCF-Crime, MSAD) or .30 AP (XD-Violence); Excl. B drops the bottom quartile of detection performance within each dataset. Both rules were fixed before the outcome was inspected. Conf.: configurations kept; ρ: mean $\rho_{b,m,d}$; CI>0: configurations whose $\rho_{b,m,d}$ interval lies above zero; N_{chg}: cells with a detectable $\Delta D_{b,m,d,c}$; Sup., Int., Oth.: suppressed, introduced, and other changed cells of the final score; Order: Spearman correlation of the seven method means of ρ with those of all configurations. Right: mean $\rho_{b,m,d}$ per method over the kept backbones. q-values are not recomputed, so a kept cell is detectable exactly when it is detectable with all configurations.

	Conf.	ρ	CI>0	N_{chg}	Sup.	Int.	Oth.	Order	VERA	AnyA.	Event	PRISM	HeadH.	HiPr.	Holmes
UCF-Crime															
All	77	.57	52	84	56	18	0	–	.67	.46	.51	.46	.59	.54	.78
Excl. A	72	.63	52	76	52	18	0	.61	.67	.59	.77	.46	.59	.54	.78
Excl. B	58	.69	47	64	45	9	0	.50	.69	.58	.78	.60	.59	.77	.78
XD-Violence															
All	77	.39	41	258	86	139	35	–	.54	.39	.19	.21	.39	.36	.64
Excl. A	72	.43	41	240	77	130	31	.86	.54	.45	.39	.21	.42	.36	.64
Excl. B	58	.45	34	192	62	100	21	.93	.53	.48	.39	.18	.44	.42	.64
MSAD															
All	77	.49	10	0	0	6	0	–	.66	.39	.45	.18	.45	.59	.70
Excl. A	72	.52	10	0	0	6	0	.89	.66	.39	.68	.21	.45	.59	.70
Excl. B	58	.58	10	0	0	4	0	.86	.66	.56	.68	.06	.45	.58	.70

I LLM USAGE DECLARATION

Large Language Models (LLMs) were used only as general-purpose assistive tools for tasks such as language editing and formatting. They were not used to generate scientific ideas, methodologies, analyses, experimental designs, or conclusions. All research contributions, results, references, and claims were independently developed and verified by the authors, who take full responsibility for the content of this paper.

J ETHICS STATEMENT

This research relies exclusively on publicly available datasets and does not involve human-subject recruitment, interaction, intervention, or the collection of private or personally identifiable information. The study does not investigate ethics-sensitive research questions or introduce additional risks related to privacy, security, discrimination, or harm beyond those associated with the original public data sources. All data are used solely for research purposes and in accordance with their stated access conditions and intended use.

K REPRODUCIBILITY STATEMENT

We provide detailed descriptions of our methodology, experimental setup, implementation details, and statistical analyses in the main paper and appendix. To further support reproducibility, we will release the code, annotations, configuration files, and other materials necessary to reproduce our experiments and results upon publication.