JUST IMAGINE REAL-WORLD ANOMALIES! CLASS-SPECIALIZED VISION-FREE TRAINING FOR VIDEO ANOMALY DETECTION AND UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world anomalies are difficult to collect: they are rare, unsafe, and often occur in privacy-sensitive environments, yet most video anomaly detection (VAD) methods require training videos. We ask: can a detector recognize anomalies it has only imagined through language? Building on evidence that text-trained detectors can transfer to video, we introduce NOVAD, a video-free framework for class-specialized anomaly detection and understanding. An LLM imagines normal and anomalous events as narratives ordered into beginning, following, climax, and resolution, with controlled variation over locations, objects, and participant counts. Encoded with a frozen vision–language text encoder, these narratives train a gated mixture of category-specialized experts, and their known temporal structure provides pseudo-labels for localization. At inference, the paired image encoder projects video frames into the same space. The predicted scores then guide multimodal event segmentation, and a video-language model explains only the resulting segments. Experiments on UCF-Crime, XD-Violence, and MSAD show that language-imagined events enable real-world anomaly localization without domain videos during detector training. On Vad-R1, detection-guided segments improve zero-shot reasoning over the untrimmed video, and a model trained with supervised fine-tuning only surpasses the full Vad-R1 model when fed with these segments.

1 INTRODUCTION

Video anomaly detection (VAD) aims to localize unusual, unsafe, or harmful events in untrimmed videos by assigning an anomaly likelihood to each frame. It supports applications such as public safety, traffic monitoring, industrial inspection, and surveillance. Despite substantial progress, most VAD methods still require domain-specific training videos. This dependence limits practical deployment because anomalous events are rare, expensive to collect, difficult to annotate temporally, and often captured in privacy-sensitive environments.

Existing methods reduce annotation requirements through different supervision paradigms. Weakly supervised approaches (Sultani et al., 2018; Tian et al., 2021; Wu et al., 2024) learn from video-level labels but still require large collections of normal and anomalous footage. One-class methods (Hasan et al., 2016; Lu et al., 2013; Wang & Cherian, 2019) train only on normal videos, yet often confuse unseen benign activities with anomalies and provide limited category-level discrimination. Unsupervised methods (Thakare et al., 2023b;a) remove manual labels but continue to assume access to domain videos. Consequently, all these paradigms remain difficult to apply when collecting real training footage is impractical, unsafe, or restricted.

TD-VAD (Zhang et al., 2026) introduced natural language as an alternative source of anomaly knowledge. Textual descriptions can be generated at scale, assigned category labels automatically, and adapted to diverse environments and interactions. Moreover, pretrained vision–language models (Radford et al., 2021; Li et al., 2026) project text and images into an aligned representation space, creating the possibility of learning anomaly patterns from language and transferring them to videos at inference.

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.

However, learning VAD models from imagined textual events introduces three challenges. First, **contextual coverage**: the same anomaly can occur across different locations, participant configurations, objects, and viewpoints, whereas descriptions generated from category names alone may be narrow or stereotypical. Second, **category and temporal ambiguity**: related anomalies, such as fighting and assault or robbery and shoplifting, share similar semantics, while short textual descriptions provide limited supervision for frame-level localization. Third, **limited interpretability**: anomaly scores indicate when an event occurs but not what happened or which visual evidence supports the decision. Directly applying a large video-language model to every frame or clip can provide richer semantics, but is computationally expensive for long videos. Moreover, anomalous events typically occupy a small fraction of untrimmed videos, so a video-language model processing the full video must reason over a context dominated by normal activity. These challenges raise a central question: *Can we learn category-specific and temporally localized video anomalies entirely from imagined textual events, while providing explanations without using large models as the primary detector?*

Motivated by this question, we introduce **NOVAD**, a video-free framework for video anomaly detection and understanding. Its anomaly detector is trained exclusively on synthetic textual narratives, and it never observes training videos. The video-language model used for understanding is applied frozen or after task-specific fine-tuning. Rather than describing an event using a single static sentence, we prompt a large language model to imagine its temporal evolution through four ordered stages: *beginning*, *following*, *climax*, and *resolution*. These stages approximate how an event unfolds in a real video, providing explicit semantic and temporal structure to the generated supervision.

Generating many linguistic variations alone does not guarantee sufficient visual diversity. We therefore introduce controlled contextual imagination through three factors: scene location, participant count, and visible objects. Each factor is sampled from a category-specific vocabulary and inserted into the event-generation prompt. For example, a fighting event may occur near a nightclub, inside a parking area, or in a public park; it may involve two individuals or a larger group; and it may contain objects that alter how the interaction is visually expressed. This controlled process produces multiple observable realizations of an anomaly while preserving its underlying category semantics. The generated narratives are encoded phrase-by-phrase using the frozen text encoder of a vision-language embedding model.

To prevent semantically related anomaly categories from collapsing into a coarse shared representation, NOVAD employs a gated mixture of category-specialized experts. Each expert is trained on one anomaly category together with shared normal events, allowing it to learn the distinctive semantic and temporal characteristics of that category. Since the anomaly category is unknown during inference, a lightweight gating network analyzes the predictions of all experts and combines them into a unified frame-level anomaly score sequence. Beyond category specialization, NOVAD turns the structure of imagined events into temporal supervision. The beginning–following–climax–resolution organization provides an approximate progression of anomaly intensity, from event onset to its most discriminative stage and subsequent resolution. We use this organization to derive temporal pseudo-labels for training the experts, thereby encouraging localized anomaly predictions without requiring manually annotated video boundaries.

During inference, video frames are projected into the same representation space using the corresponding frozen image encoder and processed directly by the text-trained experts. Thus, NOVAD produces frame-level anomaly scores through frozen feature encoding and lightweight expert inference, without relying on a large video-language model for primary detection.

To complement localization with semantic understanding, we introduce a detector-guided understanding stage. Predicted anomaly scores modulate optical-flow, depth, human-pose, and scene-segmentation features, which are fused with RGB features and partitioned into temporally coherent event segments. A video-language model is invoked only on these detected segments to generate factual event descriptions and explain the visible evidence supporting each anomaly decision. This separates efficient temporal detection from computationally expensive semantic reasoning: the text-trained experts determine *where to look*, while the video-language model explains *what happened*. Beyond reducing cost, restricting the video-language model to these segments removes normal context that would otherwise dominate its input, which we show improves anomaly reasoning on Vad-R1.

Overall, NOVAD learns anomalies from controlled textual narratives, detects them in real videos with category-specialized experts, and explains them with a video-language model restricted to the detected segments. We evaluate detection on UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), and MSAD (Zhu et al., 2024), showing that a detector trained only on imagined events can localize real anomalies. We evaluate understanding on Vad-R1 (Huang et al., 2026), showing that detection-guided segments improve the reasoning of a video-language model.

Our contributions are summarized as follows:

- We introduce **NOVAD**, a video-free framework that learns category-specific anomaly patterns from LLM-generated textual narratives and transfers them to real videos through an aligned vision–language representation space.
- We propose a **controlled event-imagination strategy** that represents each event through its beginning, following, climax, and resolution while varying locations, objects, and participant counts.
- We introduce a **gated mixture of category-specialized experts** with structure-derived temporal pseudo-supervision for fine-grained and temporally localized anomaly prediction.
- We propose a **detector-guided multimodal event segmentation pipeline** that detects anomalous events, improving the anomaly reasoning capabilities of video-language models while reducing computational cost.
- We evaluate NOVAD on UCF-Crime, XD-Violence, and MSAD for real-world video anomaly detection and on the Vad-R1 benchmark for video anomaly understanding.

2 Related Work

Video Anomaly Detection: VAD research predominantly utilizes training strategies that minimize manual labeling, such as weakly supervised, unsupervised, and zero-shot (training-free) models. Weakly supervised methods typically rely on multiple-instance learning (MIL), where only video-level binary labels are available (Sultani et al., 2018; Tian et al., 2021; Wu et al., 2020). Sultani et al. (2018) introduced the UCF-Crime benchmark and a ranking-based MIL objective; subsequent methods improve temporal localization through feature magnitude learning (Tian et al., 2021), graph-based label denoising (Zhong et al., 2019), or richer temporal modeling with transformers (Zhou et al., 2023; Chen et al., 2023). More recently, CLIP-based methods (Wu et al., 2024) have leveraged vision-language pretraining to enhance cross-modal feature alignment under weak supervision. Shifting to a target-video-free paradigm, TD-VAD (Zhang et al., 2026) trained a model exclusively on LLM-generated textual descriptions, leveraging CLIP's multimodal space to perform anomaly detection using videos only at inference. Unsupervised approaches remove the labeling requirement entirely, relying on reconstruction (Hasan et al., 2016), diffusion models (Tur et al., 2023a;b), or scene-dynamics self-training (Thakare et al., 2023b), at the cost of lower detection accuracy. Training-free methods (Zanella et al., 2024; Shao et al., 2025) avoid any training by querying large VLMs at test time, but incur high inference costs even for detection. In video understanding, MoE has been applied to action recognition (Jain et al., 2024) and, more recently, to VAD. GS-MoE (Amicantonio et al., 2025) introduces class-specific expert models for weakly-supervised VAD, where each expert is trained on a single anomaly category and a shared gating network integrates their predictions. This design addresses a fundamental limitation of shared-model MIL approaches: the inability to capture fine-grained, category-specific cues for subtle anomalies such as shoplifting or burglary.

LLMs and VLMs for Video Anomaly Explanation and Understanding: The integration of large language models and vision-language models into VAD has opened a new direction that goes beyond binary anomaly scoring toward interpretable natural language descriptions of observed events. Kim et al. (2023) first demonstrated that LLM-generated text descriptions can serve as semantic anchors for unsupervised anomaly scoring. LAVAD (Zanella et al., 2024) operationalized this at scale by captioning individual frames with a VLM and scoring each caption with an LLM for anomaly relevance, achieving competitive detection performance without any training. Building on this, Holmes-VAU (Zhang et al., 2025) introduced a hierarchical understanding framework operating simultaneously at clip, event, and video levels, paired with an anomaly-focused temporal sampler

that concentrates VLM attention on anomaly-dense regions, demonstrating that multi-granularity reasoning substantially improves both detection accuracy and understanding capacity. EventVAD (Shao et al., 2025) similarly showed that organizing video into temporally coherent segments before VLM querying yields sharper temporal localization than frame-by-frame captioning. Ex-VAD (Huang et al., 2025) combines VLM-generated anomaly explanations with visual features for explainable fine-grained detection. Vad-R1 (Huang et al., 2026) frames anomaly understanding as a perception-to-cognition chain of thought, fine-tunes a video-language model with supervised and reinforcement learning, and introduces a benchmark for evaluating anomaly reasoning. Despite their expressivity, these methods either rely on the VLM as the primary detection mechanism or feed it the untrimmed video, requiring large model calls at test time and exposing the VLM to long stretches of normal context. In NOVAD, detection is handled by the lightweight text-trained model, and the VLM is invoked only on detector-guided segments. Detection reduces reasoning costs and improves quality by filtering out normal context.

3 Proposed NOVAD Method

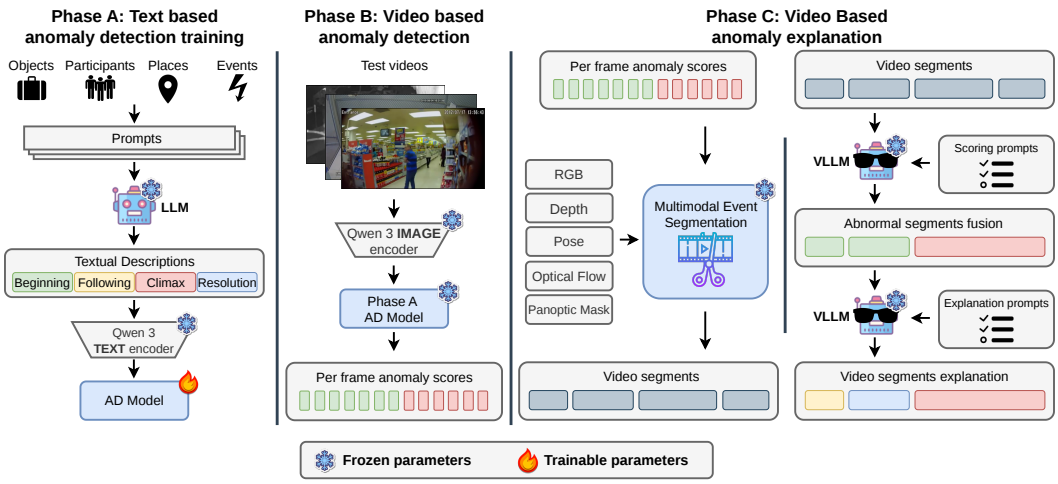


Figure 1: **Overview of NOVAD. Phase A: Text-Based Training.** Structured anomaly descriptions train a gated mixture of category-specialized experts in a shared vision–language space. **Phase B: Video Detection.** Frozen image features are processed by the text-trained model to produce frame-level anomaly scores. **Phase C: Video Explanation.** The scores guide multimodal event segmentation; a VLLM scores the segments, abnormal ones are fused, and a second VLLM pass explains them.

3.1 Problem Formulation and Overview

Let an untrimmed test video be $\mathcal{V} = \{f_t\}_{t=1}^{T_v}$. Video anomaly detection aims to estimate a temporally dense anomaly sequence: $\mathcal{F}_\theta(\mathcal{V}) = \mathbf{s} = [s_1, \ldots, s_{T_v}] \in [0,1]^{T_v}$, where s_t denotes the anomaly likelihood at temporal position t. Unlike conventional VAD methods, NOVAD does not use target-domain videos to train or fine-tune its anomaly detector. Its only domain-specific inputs are an event ontology $\mathcal{C} = \{c_k\}_{k=1}^{K}$ of K anomaly categories plus a normal category c_0, and small category-specific context vocabularies written manually (Sec. 3.2). The setting is thus ontology-supervised and closed-set.

As shown in Fig. 1, NOVAD contains three phases. Phase A constructs temporally structured textual descriptions and trains a category-specialized AD model. Phase B transfers the text-trained detector to video through a shared vision–language embedding space. Phase C transforms the predicted anomaly scores into temporally grounded natural-language explanations.

3.2 Phase A: Text-Based Anomaly Detection Training

Controlled textual description generation. For every event category c_k, we define category-specific sets of plausible places $\mathcal{L}_k$, participant counts $\mathcal{N}_k$, and visible objects $\mathcal{O}_k$. For the i-th

sample, a context $\boldsymbol{\xi}_i = (\ell_i, n_i, \mathcal{O}_i)$ is drawn from these sets and provided to an LLM together with the event category. The LLM generates:

$$\mathbf{c}_i = \mathcal{G}(y_i, \boldsymbol{\xi}_i) = [c_i^{\text{beg}}, c_i^{\text{fol}}, c_i^{\text{cli}}, c_i^{\text{res}}], \tag{1}$$

where $y_i \in \mathcal{C} \cup \{c_0\}$ is the category of the i-th sample and the four phrases describe the *Beginning*, *Following*, *Climax*, and *Resolution* of the event.

This structure introduces an event-evolution prior rather than representing an anomaly through a single static sentence. Controlled variation over places, participants, and objects additionally produces multiple observable manifestations of the same event. For normal samples, places are drawn from the anomaly-place vocabulary while the prompt enforces peaceful activity. This discourages the model from using scene context alone as an anomaly shortcut.

Text encoding and temporal construction. Each phrase is independently encoded using the frozen Qwen3-VL-Embedding text encoder ϕ_t:

$$\mathbf{z}_i^q = \phi_t(c_i^q), \qquad q \in \{\text{beg}, \text{fol}, \text{cli}, \text{res}\}. \tag{2}$$

Following Matryoshka representation learning (Kusupati et al., 2022), the native embeddings are truncated to $d = 512$ dimensions and subsequently ℓ_2-normalized; image embeddings undergo the same truncation. The four embeddings form $\mathbf{Z}_i \in \mathbb{R}^{4\times d}$.

To construct a temporally dense input, we follow TD-VAD (Zhang et al., 2026) and repeat each phrase embedding four times: $\widehat{\mathbf{Z}}_i = \text{Repeat}(\mathbf{Z}_i, 4) \in \mathbb{R}^{16\times d}$. Repetition introduces no additional parameters and preserves the discrete phrase-level structure, providing a known stage identity for every temporal position.

Category-specialized AD model: Mixture of Experts. Anomalies such as fighting and assault or robbery and shoplifting share related semantics and may be difficult to separate using a single coarse anomaly representation. NOVAD therefore employs K category-specialized experts $\mathbb{E} = \{E_k\}_{k=1}^K$. Expert E_k is trained using textual descriptions from anomaly category c_k together with normal descriptions shared across all experts.

Our AD model builds on the expert backbone of GS-MoE (Amicantonio et al., 2025), a weakly-supervised video anomaly detection framework that combines class-specific expert models with a Temporal Gaussian Splatting (TGS) training objective. We adapt this architecture to the vision-free setting by replacing the original I3D visual features with 512-dimensional Qwen3-VL text embeddings produced by our generation pipeline, and by replacing the standard weak video-level labels with synthetic point annotations derived directly from the structure of our generated sequences.

Expert Architecture: We instantiate K independent expert models, one per anomaly category: 13 for UCF-Crime, 11 for MSAD, 6 for XD-Violence, and 37 for Vad-R1. Each expert is implemented as a UR-DMU module (Zhou et al., 2023) operating in the shared text-video embedding space. Each expert proceeds in three stages. First, a temporal convolution layer projects the input sequence of $T = 16$ text embeddings into a 512-dimensional latent space, capturing local temporal dynamics across neighboring timesteps. Second, a two-layer transformer refines this representation, each layer combining standard scaled dot-product attention, which models global temporal dependencies, with a position-only, distance-weighted decay term that penalizes long-range interactions and keeps the model sensitive to local sequential structure. Third, two learnable memory units, one for anomalous patterns and one for normal patterns, augment the sequence representation through attention-weighted retrieval over a bank of stored prototype vectors, allowing each expert to contrast the current input against a learned set of class-specific anomaly and normality anchors. The augmented representation is concatenated with the transformer output and passed through a two-layer MLP with sigmoid activation to produce the final frame-level anomaly score sequence $\hat{s} \in [0, 1]^T$.

Each expert is trained exclusively on sequences from its assigned anomaly class paired with normal sequences, with the normal sequences shared across all experts at each training step. This encourages each expert to develop a compact and discriminative representation of its own anomaly type rather than collapsing into a shared coarse-level anomaly representation.

Gating Network: Since the true anomaly class is unknown at test time, the K expert score sequences $\{\hat{s}_k\}_{k=1}^K$ must be fused into a single prediction without class-oracle routing. A lightweight gating

network g_ψ applies single-head self-attention across the T timesteps over the K-dimensional per-frame expert-score vector, letting the fusion at each position be informed by the expert-agreement pattern elsewhere in the sequence, before mapping the resulting context vector to a fused score $\hat{s} \in [0,1]^T$ through a four-layer MLP with sigmoid activation.

Temporal Gaussian Splatting Loss: A key challenge of training on synthetic text sequences is the absence of frame-level temporal annotations. We address this by adopting the TGS objective of GS-MoE (Amicantonio et al., 2025). For each anomalous training sequence, a synthetic point label is placed at $t^* = \lfloor T/2 \rfloor + 1$, the first timestep of the Climax stage, simulating a glance annotation that localizes the anomaly at its peak. Writing $u(t) = \frac{2(t-1)}{T-1} - 1$ for the position of timestep t rescaled to $[-1, 1]$, a peak-normalized Gaussian kernel is rendered around the anchor:

$$\hat{y}_t = \exp\left(-\frac{(u(t) - u(t^*))^2}{2\sigma^2}\right), \quad t \in \{1, \ldots, T\} \tag{3}$$

When training the gating network, the anchor is first expanded by propagating outward from t^* while the fused score stays above a fraction α of its value at t^*, and the kernel above is rendered over this widened point set, letting the model's own confidence refine the extent of the pseudo-label. For normal sequences, the pseudo-label is set to zero at all timesteps.

Training Objective: Each expert is optimized with a composite loss combining six terms. An MIL loss applies binary cross-entropy to the mean of the top-k scores, pushing it towards 1 for anomalous sequences and towards 0 for normal ones. A memory-consistency loss $\mathcal{L}_{\text{mem}}$, composed of four binary cross-entropy terms, enforces that the anomalous memory attends strongly to anomalous inputs and weakly to normal inputs, and conversely for the normal memory. A triplet margin loss encourages normal-memory representations to lie closer to normal features than to anomalous ones. A margin loss on the ℓ_2 norm of the memory-augmented representations, following Tian et al. (2021), further separates anomalous from normal magnitudes. A KL divergence term regularizes the variational reparameterization of the normal memory augmentation. Finally, the TGS loss minimizes the binary cross-entropy between the predicted score sequence and the rendered pseudo-label: $\mathcal{L} = \mathcal{L}_{\text{MIL}} + \lambda_1(\mathcal{L}_{\text{mem}} + \mathcal{L}_{\text{triplet}}) + \lambda_2\mathcal{L}_{\text{KL}} + \lambda_3\mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{TGS}}$ with $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, $\lambda_3 = 0.0001$. The gating network is optimized in a separate stage following expert training, with expert parameters frozen, using $\mathcal{L}_{\text{MIL}}$ and $\mathcal{L}_{\text{TGS}}$ applied directly to the fused output $\hat{s}$.

3.3 Phase B: Video-Based Anomaly Detection

Given a test video, the frozen Qwen3-VL-Embedding image encoder ϕ_v corresponding to ϕ_t projects sampled frames into the same embedding space: $\mathbf{X}_\mathcal{V} = [\phi_v(f_1), \ldots, \phi_v(f_{T_v})] \in \mathbb{R}^{T_v \times d}$. The Phase-A AD model is applied without modification: $\mathbf{s}^{\text{det}} = g_\psi\left(E_1(\mathbf{X}_\mathcal{V}), \ldots, E_K(\mathbf{X}_\mathcal{V})\right) \in [0,1]^{T_v}$. Thus, only the frozen encoder branch changes from text to image; the category-specialized experts and gate remain unchanged.

3.4 Phase C: Video-Based Anomaly Explanation and Understanding

Phase C converts the per-frame predictions into temporally grounded explanations. For each frame, we extract RGB appearance features $\mathbf{r}_t$ with a frozen CLIP ViT-B/16 image encoder (Radford et al., 2021), together with four auxiliary features $\mathbf{f}_t^m$: optical flow, depth, human pose, and panoptic mask, obtained using RAFT (Teed & Deng, 2020), Depth Anything V2 (Yang et al., 2024), YOLOv7-Pose (Maji et al., 2022), and SAM (Kirillov et al., 2023), respectively.

The Phase-B anomaly score modulates the auxiliary features, which are then fused with the appearance features:

$$\mathbf{h}_t = \left[\, \alpha_{\text{clip}}\, \mathbf{r}_t \,\middle\|\, \beta\, s_t^{\text{det}}\, \mathbf{f}_t^{\text{flow}} \,\middle\|\, \beta\, s_t^{\text{det}}\, \mathbf{f}_t^{\text{depth}} \,\middle\|\, \beta\, s_t^{\text{det}}\, \mathbf{f}_t^{\text{pose}} \,\middle\|\, \beta\, s_t^{\text{det}}\, \mathbf{f}_t^{\text{mask}} \,\right], \tag{4}$$

with $\beta = (1 - \alpha_{\text{clip}})/M$, where M is the number of available auxiliary streams. In low-score regions, the auxiliary features shrink towards zero, so event boundaries there are driven by appearance changes only; where the detector fires, motion, geometric, pose, and object-level changes also contribute to boundary detection.

The fused sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{T_v}]$ is partitioned using the event segmentation procedure of EventVAD (Shao et al., 2025): $\mathcal{S} = \Psi_{\text{MES}}(\mathbf{H}) = \{[\tau_j^s, \tau_j^e]\}_{j=1}^J$. Each segment is uniformly sampled

and scored by a video-language model $\mathcal{G}_{\text{VLLM}}$ (Qwen3.5 (Team, 2026) by default, Qwen2.5-VL-7B (Bai et al., 2025) on Vad-R1), which returns an anomaly confidence $q_j \in [0, 1]$; propagating q_j to frames yields $\mathbf{s}^{\text{exp}}$. Each run of contiguous segments with $q_j > 0.5$ is then merged into a single abnormal segment, while the other segments are kept unchanged, and a second prompt asks $\mathcal{G}_{\text{VLLM}}$ to describe each resulting segment and explain the visible evidence supporting its decision, restricted to observable people, objects, actions, and interactions. Explanations thus justify the VLLM's segment-level decision, while $\mathbf{s}^{\text{det}}$ determines where segments are formed. For Vad-R1, which expects one answer per video, the untrimmed video is kept if no segment satisfies $q_j > 0.5$; otherwise, the highest-scoring abnormal segment, merged with its contiguous abnormal neighbors and padded with one segment on each side, replaces the untrimmed video under the original Vad-R1 prompt.

4 Experiments

Evaluation protocol. We evaluate the NOVAD detector (Phases A and B) on the official test splits of UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), and MSAD (Zhu et al., 2024); the detector never accesses target-domain training videos. For each benchmark, Qwen3.5 (Team, 2026) generates 2,000 descriptions per anomaly category and 10,000 normal descriptions. We report frame-level AUC on UCF-Crime and AUC and AP on XD-Violence and MSAD, plus AUC_A and AP_A, computed on abnormal videos only to measure temporal localization without the dominant contribution of normal videos. Anomaly understanding is evaluated on Vad-R1 (Huang et al., 2026) with its reasoning metrics, BLEU-2 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-2 (Lin, 2004), using Qwen2.5-VL-7B with the original Vad-R1 prompt and decoding parameters. Implementation details are given in Appendix A.1.

4.1 Comparison with the State of the Art

4.1.1 Video Anomaly Detection

We compare NOVAD with representative weakly supervised, unsupervised, training-free, and text-supervised VAD methods in Table 1. Since NOVAD does not observe target-domain videos during detector training, the most direct comparison is with text-supervised approaches. The other supervision settings are included to contextualize its performance against methods trained on or directly reasoning over real videos.

On UCF-Crime, NOVAD achieves 82.78% AUC and 65.69% AUC_A. Compared with our TD-VAD reimplementation (Appendix F), NOVAD improves these metrics by 8.59 and 7.86 percentage points, respectively. It also surpasses the originally reported TD-VAD AUC by 1.96 points and outperforms VideoLLaMA2, LAVAD, and EventVAD. NOVAD remains below the strongest video-dependent baselines, indicating that a residual text–video representation gap still exists.

On XD-Violence, NOVAD obtains 74.03% AP, 91.32% AUC, 75.88% AP_A, and 76.88% AUC_A. Relative to TD-VAD†, this corresponds to gains of 4.78, 3.05, 3.63, and 4.54 points, respectively. NOVAD also improves the original TD-VAD AUC by 1.82 points, although TD-VAD retains a 1.80-point advantage in AP. Compared with the strongest training-free baseline, NOVAD attains nearly identical AUC (91.32% versus 91.34%) while improving AP by 5.96 points.

On MSAD, NOVAD obtains the highest AUC among all compared methods. It improves TD-VAD† by 3.24, 1.83, and 0.19 points in AUC, AUC_A, and AP, respectively, while remaining 0.71 points lower in AP_A. NOVAD also exceeds the weakly supervised GS-MoE baseline in AUC, AUC_A, and AP. The consistent gains across three benchmarks indicate that the detector does not merely memorize benchmark category names. Instead, controlled variation over places, participants, and visible objects supplies multiple observable manifestations of each anomaly, while the category-conditioned experts provide complementary views of each anomaly (Appendix E).

4.1.2 Video Anomaly Understanding

Table 2 evaluates the reasoning generated on the Vad-R1 benchmark, measured by the overlap between generated and reference reasoning texts. In both settings, NOVAD's detection-guided segments lead to higher scores on all three metrics. In the zero-shot setting, NOVAD segments improve

Table 1: State-of-the-art comparison on UCF-Crime, XD-Violence and MSAD. **Bold** denotes the best result within each supervision setting. † denotes our reimplementation (Appendix F).

Model	UCF-Crime		XD-Violence				MSAD			
	AUC	AUC_A	AP	AUC	AP_A	AUC_A	AUC	AP	AUC_A	AP_A
Weakly-Supervised VAD										
MIL (Sultani et al., 2018)	75.41	54.25	73.20	–	–	–	–	–	–	–
RTFM (Tian et al., 2021)	84.30	62.96	77.81	–	78.57	–	86.70	66.30	–	–
MGFN (Chen et al., 2023)	86.98	–	80.11	–	–	–	85.00	63.50	–	–
UR-DMU (Zhou et al., 2023)	86.97	70.81	81.66	–	83.94	–	85.78	67.35	67.95	75.30
GS-MoE (Amicantonio et al., 2025)	**91.58**	**83.86**	**82.29**	–	**85.74**	–	**87.72**	**68.26**	**69.54**	**76.68**
Training-Free VAD										
VideoLLaMA2 (Cheng et al., 2024)	74.42	–	53.57	80.21	–	–	–	–	–	–
LAVAD (Zanella et al., 2024)	80.28	–	62.01	85.36	–	–	–	–	–	–
EventVAD (Shao et al., 2025)	82.03	–	64.04	87.51	–	–	–	–	–	–
United Frame VAA (Lin et al., 2026)	**84.28**	–	**68.07**	**91.34**	–	–	**86.00**	**76.40**	–	–
Unsupervised VAD										
BODS (Wang & Cherian, 2019)	68.26	–	57.32	–	–	–	–	–	–	–
GODS (Wang & Cherian, 2019)	70.20	–	61.56	–	–	–	–	–	–	–
DyAnNet (Thakare et al., 2023b)	**84.50**	–	–	–	–	–	–	–	–	–
RareAnom (Thakare et al., 2023a)	–	–	**68.33**	–	–	–	–	–	–	–
Text-Supervised VAD										
TD-VAD (Zhang et al., 2026)	80.82	–	**75.83**	89.50	–	–	–	–	–	–
TD-VAD† (Zhang et al., 2026)	74.19	57.83	69.25	88.27	72.25	72.34	85.65	70.83	69.47	**77.13**
NOVAD (ours)	**82.78**	**65.69**	74.03	**91.32**	**75.88**	**76.88**	**88.89**	**71.02**	**71.30**	76.42

Table 2: Anomaly reasoning on Vad-R1. Zero-shot: Qwen2.5-VL-7B on untrimmed videos or NOVAD segments. Fine-tuned: Qwen2.5-VL-7B trained with SFT only *vs.* the full Vad-R1 (SFT + RL). **Bold** denotes the best results.

Model	Input	BLEU-2	METEOR	ROUGE-2
Zero-shot				
Qwen2.5-VL-7B	Untrimmed	0.113	0.264	0.116
Qwen2.5-VL-7B	NOVAD	**0.129**	**0.283**	**0.146**
Fine-tuned				
Vad-R1 (SFT+RL)	Untrimmed	0.233	0.406	0.194
Ours (SFT only)	NOVAD	**0.241**	**0.425**	**0.208**

Table 3: Cross-dataset generalization (test AUC). Rows: source ontology.

Source	UCF	XD	MSAD
TD-VAD (Zhang et al., 2026)			
UCF	80.82	59.53	–
XD	77.08	89.50	–
NOVAD (Ours)			
UCF	**82.78**	82.79	88.54
XD	73.53	**91.32**	81.49
MSAD	74.59	86.33	**88.89**

Qwen2.5-VL by 0.016 BLEU-2, 0.019 METEOR, and 0.030 ROUGE-2, corresponding to relative gains of 14.2%, 7.2%, and 25.9%. The largest gain on ROUGE-2 suggests that the generated descriptions recover more of the event-specific content found in the references. This is consistent with prior work showing that focusing VLM inputs on anomaly-dense (Zhang et al., 2025) or temporally coherent (Shao et al., 2025) regions benefits anomaly understanding and localization, compared with processing the untrimmed video, where normal context dominates. In the fine-tuned setting, our model trained with SFT only and fed with NOVAD segments surpasses the full Vad-R1 model, trained with both SFT and RL on untrimmed videos (+0.008 BLEU-2, +0.019 METEOR, +0.014 ROUGE-2). Since the RL stage further improves reasoning in Vad-R1 (Huang et al., 2026), this comparison is conservative: detection-guided segments offset the gain that Vad-R1 obtains from this additional stage. Across all three reasoning metrics, Phase C thus provides a cleaner context for anomaly understanding, which justifies its use as an interpretation layer on top of the Phase-B detector.

4.2 Cross-dataset generalization

Table 3 evaluates whether anomaly knowledge learned from one source ontology transfers to videos from unseen benchmarks. Here, the source denotes only the anomaly categories and their generated textual descriptions; no source or target-domain training videos are used.

The largest improvement occurs for UCF-Crime→XD-Violence, where NOVAD obtains 82.79% AUC compared with 59.53% for TD-VAD. Its transfer performance is nearly identical to its UCF-

specific result, suggesting that the learned anomaly representation is not strongly tied to UCF-Crime visual contexts. Generalization remains asymmetric: for XD-Violence→UCF-Crime, NOVAD achieves 73.53%, below TD-VAD's 77.08%. This direction is particularly challenging because the smaller XD-Violence ontology must transfer to the more diverse and fine-grained anomaly categories of UCF-Crime.

We attribute this generalization to two components. Controlled textual description generation varies places, participant counts, and visible objects, expanding the contextual support of each category and reducing reliance on benchmark-specific scenes. The category-conditioned experts also transfer beyond their own category: removing the expert of a category changes its AUC by only -0.34 points on UCF-Crime (Appendix E).

4.3 Ablation studies

Table 4: Controlled text generation

No. of Participants	Place Diversity	Object Diversity	UCF AUC
–	–	–	76.59
✓	–	–	78.49
✓	✓	–	80.39
✓	✓	✓	**82.78**

Table 5: Multimodal segmentation

Optical Flow	Depth	Pose	Panoptic Mask	UCF AUC	UCF AUC_A
✓	–	–	–	79.16	64.99
✓	✓	–	–	79.96	66.51
✓	✓	✓	–	80.12	66.65
✓	✓	✓	✓	**80.39**	**66.95**

Effect of controlled textual description generation. Table 4 evaluates the three contextual factors used in Phase A. Introducing participant-count variation improves AUC from 76.59% to 78.49%, while adding place diversity further raises it to 80.39%. The complete formulation, including visible objects, reaches 82.78%, providing an overall gain of 6.19 points over uncontrolled generation. Notably, objects provide the largest marginal improvement of 2.39 points. This is expected because objects often define category-specific human–object relations, such as taking merchandise in shoplifting, carrying a bag during robbery, or interacting with a weapon during assault.

Effect of complementary modalities for Phase C. Table 5 evaluates multimodal event segmentation. Starting from optical flow, depth increases AUC and AUC_A by 0.80 and 1.52 points, showing that geometric changes help distinguish coherent interactions from background motion. Pose provides additional cues for human-centric anomalies, while panoptic masks introduce object- and scene-level structure. Their combination reaches 80.39% AUC and 66.95% AUC_A.

Compared with the Phase-B detector, Phase C improves AUC_A from 65.69% to 66.95%, but reduces global AUC from 82.78% to 80.39%. This reveals an important trade-off: multimodal segmentation and VLLM reasoning sharpen localization inside abnormal videos, but can introduce false positives when normal segments contain visually unusual context. We therefore retain Phase B as the primary anomaly detector. Together with Table 2, this defines a clear division of labor: Phase C lowers frame-level detection AUC but improves the reasoning of video-language models.

5 Conclusion

We introduced **NOVAD**, a framework whose anomaly detector is trained only on events imagined through language and whose video-language model explains the detected segments. Experiments on UCF-Crime, XD-Violence, and MSAD show that structured textual supervision yields effective category-specific anomaly representations without target-domain training videos. On Vad-R1, detection-guided segments improve zero-shot reasoning and allow a model with supervised fine-tuning only to surpass the full Vad-R1 model. We hope this work encourages language-imagined supervision where rare, sensitive, or unsafe footage is impractical to collect.

LLM USAGE STATEMENT

Generative AI tools were used in this work in two distinct roles: as components of the proposed method, and as research and writing assistants.

LLMs as part of the method. NOVAD relies on generative models by design, and all such uses are described in the paper. We used generative AI tools to generate synthetic datasets: Qwen3.5-27B produces the textual training corpus, which is the only training data of the anomaly detector. Frozen Qwen3-VL-Embedding encoders map text and video frames into a shared space. Qwen3.5-27B, and Qwen2.5-VL-7B on Vad-R1, score and explain the detected video segments; on Vad-R1, Qwen2.5-VL-7B is additionally fine-tuned with the Vad-R1 supervised fine-tuning stage.

LLMs as research and writing assistants. We used an LLM to translate the authors' notes and draft text from French into English, to produce first drafts of parts of the manuscript, and to improve readability. We also used an LLM to obtain feedback on a draft of the paper, including its methodology and experimental design, and to identify typographical errors. All resulting changes to the methodology and experiments were decided and implemented by the authors. All drafted and translated text was revised by the authors, and every technical claim, reported number and citation was checked against our experiments and the cited sources.

Undisclosed tasks. We did not use generative AI tools to develop the conceptual framework, propose or refine hypotheses, implement the method, clean or reformat datasets, support qualitative analysis, or interpret results. Formulating mathematical claims and writing proofs are not applicable to this work, which contains no theoretical results.

We take full responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

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

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, 2025. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 387–395, 2023.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

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

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *Forty-second International Conference on Machine Learning*, 2025.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026.

Gagan Jain, Nidhi Hegde, Aditya Kusupati, Arsha Nagrani, Shyamal Buch, Prateek Jain, Anurag Arnab, and Sujoy Paul. Mixture of nested experts: Adaptive processing of visual tokens. *Advances in Neural Information Processing Systems*, 37:58480–58497, 2024.

Jaehyun Kim, Seongwook Yoon, Taehyeon Choi, and Sanghoon Sull. Unsupervised video anomaly detection based on similarity with predefined text descriptions. *Sensors*, 23(14):6256, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2637–2646, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

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.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140:109567, 2023a.

Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pp. 5541–5550, 2023b.

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

Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection. In *2023 IEEE international conference on image processing (ICIP)*, pp. 2540–2544. IEEE, 2023a.

Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In *International Conference on Image Analysis and Processing*, pp. 49–62. Springer, 2023b.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8201–8211, 2019.

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 (ECCV)*, 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.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

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

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

Shuangqing Zhang, Lei-Lei Ma, Zhao Wang, Wen Dong, Xinyi Xu, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. Td-vad: Breaking visual dependence in video anomaly detection with text-driven learning. In *Forty-third International Conference on Machine Learning*, 2026.

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

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

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.

A Implementation Details

All experiments are run on a single NVIDIA H100 GPU.

A.1 Evaluation Setting

Detection benchmarks. UCF-Crime (Sultani et al., 2018) covers 13 anomaly categories, while XD-Violence (Wu et al., 2020) and MSAD (Zhu et al., 2024) cover 6 and 11, respectively. Only the official test splits are used, and the detector never accesses training videos. One expert is trained per anomaly category on the generated corpus (Appendix A.2). Text descriptions and test frames are encoded with the paired frozen text and image encoders, so the experts and the gating network are trained only on textual features. On these three benchmarks, Phase C uses Qwen3.5-27B (Team, 2026) for segment-level scoring and explanation (Appendix A.5).

Anomaly understanding on Vad-R1. Vad-R1 (Huang et al., 2026) covers 37 anomaly categories. Following the same text-only protocol, we generate 2,000 descriptions per anomaly category and 10,000 normal descriptions, train 37 experts, and run Phases B and C on the Vad-R1 test videos. Since Vad-R1 builds on Qwen2.5-VL, we use Qwen2.5-VL-7B (Bai et al., 2025) both for segment scoring in Phase C and as the evaluated video-language model, with the exact prompt and decoding parameters of Vad-R1. In the fine-tuned setting, Qwen2.5-VL-7B is trained on the Vad-R1 training set with the Vad-R1 supervised fine-tuning (SFT) stage only, without its reinforcement learning (RL) stage, and compared with the full Vad-R1 model trained with both SFT and RL, as reported by Huang et al. (2026). Reasoning quality is measured with the metrics used by Vad-R1, BLEU-2 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-2 (Lin, 2004), to ensure direct comparability.

A.2 Text Corpus Generation

Textual descriptions are generated with Qwen3.5-27B (Team, 2026), sampled with temperature 0.7, top-p 0.9, and up to $1{,}024$ new tokens per call, in batches of 64 prompts. We generate $2{,}000$ descriptions for each anomaly category and $10{,}000$ for the normal category (UCF-Crime (Sultani et al., 2018): 13 anomaly categories; the same protocol applies with 6 and 11 categories for XD-Violence (Wu et al., 2020) and MSAD (Zhu et al., 2024), respectively). For UCF-Crime and MSAD, which consist of surveillance footage, each description follows the template below, adapted from the UCF-Crime prompt of TD-VAD (Zhang et al., 2026) with our controlled context slots:

```
As a surveillance camera generate 4 phrases describing a
[CLASS] event involving [N] participants in a [LOCATION]
scene.  The description of the scene must include the
following visible objects:  [OBJECT]. [NORM] The phrases
need to be between 20 and 35 words and from a security
camera perspective.  Ensure that your phrases create logical
progression:  beginning→following→climax→resolution.
Focus on visible behavior only - no internal thoughts or
off-camera events.  Use surveillance terminology:  CCTV,
security footage, monitoring system, camera view.
```

XD-Violence is mostly collected from movies and online videos rather than surveillance cameras, so its prompt drops the surveillance framing and asks for richer descriptions of objects and surroundings, also adapted from TD-VAD:

```
Generate 4 sequential sentences describing a [CLASS] event
involving [N] participants in a [LOCATION] scene.  The
description must include the following visible [OBJECT].
[NORM] The phrases need to be between 20 and 35 words.  Be
sure to follow a logical and natural progression with a
start→development→climax→resolution.  Describe as much as
possible the objects involved and the surroundings.  Focus
only on visible things and behavior no internal thoughts.
```

In both templates, `[NORM]` is only appended for the normal category ("For the normal event the description should depict normal and peaceful events, nothing illegal or that might attract attention.").

Three contextual factors are sampled independently per description. *Participants*: an integer count drawn uniformly in $[n_{\min}, 10]$, with $n_{\min} = 2$ for *Fighting*, *Assault*, *Abuse*, *Arrest*, *Robbery*, and *Shooting*, and $n_{\min} = 1$ otherwise. *Places*: one location drawn uniformly from a category-specific vocabulary of 3–4 candidate locations (e.g., *Shooting*: school campus, shopping mall, street corner, parking garage); normal-category locations are drawn from the pooled vocabulary of all anomaly locations, while the prompt enforces peaceful, non-suspicious activity, preventing the model from using scene context alone as a shortcut. *Objects*: for every anomaly category except *Road Accidents* and *Explosion*, between 1 and 3 objects are sampled without replacement from a category-specific pool of 10 candidate objects and inserted into the prompt. Each 4-phrase description is split and encoded phrase-by-phrase with Qwen3-VL-Embedding-8B (Li et al., 2026). Embeddings are truncated to the first 512 dimensions and ℓ_2-renormalized. Each phrase embedding is repeated 4 times to form the $T = 16$ sequence of Sec. 3.2.

A.3 Mixture-of-Experts

Expert Architecture: Each of the K category-specialized UR-DMU experts receives a sequence of $T = 16$ text (training) or image (inference) embeddings of dimension $d = 512$. A temporal convolution (kernel size 3, stride 1, padding 1, followed by ReLU) first projects the sequence into a 512-dimensional latent space. A 2-layer transformer then refines the representation; each layer combines a 4-head, 128-dim-per-head self-attention branch with a position-only temporal-decay attention branch, concatenated before the output projection, with dropout 0.5. Two memory units of 60 prototype slots each (one anomalous, one normal) augment the representation via attention-weighted retrieval; both augmented representations are projected through a shared 512×512 linear layer, summed, and concatenated with the transformer output. A 2-layer MLP ($1024 \to 128 \to 1$, ReLU, sigmoid output) maps the resulting vector to the per-timestep anomaly score $\hat{s} \in [0,1]^T$.

At inference, frame-level embeddings $X_V \in \mathbb{R}^{T_v \times d}$ (Sec. 3.3) are mean-pooled into non-overlapping 16-frame segments and ℓ_2-renormalized, giving $N = \lfloor T_v/16 \rfloor$ segment embeddings. A sliding window of $T = 16$ segments (stride $T/2 = 8$) is fed to the experts and gate; scores from overlapping windows are averaged per segment, following the fixed-segment MIL convention of weakly-supervised VAD (Sultani et al., 2018). Each segment score is repeated 16 times and truncated/tail-padded to length T_v, yielding the frame-level sequence $\hat{s} \in [0,1]^{T_v}$.

Gating Network: The K per-timestep expert scores are stacked into a $T \times K$ tensor and processed by a single-head self-attention layer (`embed_dim` $= K$), followed by a 4-layer MLP ($K \to 64 \to 128 \to 64 \to 1$, LeakyReLU, sigmoid output) producing the fused score $\hat{s} \in [0,1]^T$.

Temporal Gaussian Splatting: The Gaussian kernel radius is set to $\sigma = 0.1$ and the kernel-mining expansion ratio to $\alpha = 0.7$. The experts are trained with the point pseudo-label directly; the confidence-driven expansion is only applied when training the gating network (steps 1,501–3,000), once the experts have converged.

Training Schedule: Each expert is optimized with AdamW (weight decay 5×10^{-4}, gradient norm clipped to 1.0), using batches of 16 normal and 16 category-matched anomalous sequences. Training proceeds in two stages over 3,000 steps in total, summarized in Table 6. *(1) Expert training* (steps 1–1,500): all K experts are optimized jointly (losses summed across experts) at learning rate 10^{-4}, halved at steps 500, 750, and 1,250; the gating network is untouched. *(2) Gate training* (steps 1,501–3,000): expert parameters are frozen (evaluation mode, no gradient); the gating network alone

is optimized at a constant learning rate of 10^{-4} using $\mathcal{L}_{\text{MIL}}$ ($k = \lfloor T/16 \rfloor + 1$) plus $0.1\,\mathcal{L}_{\text{TGS}}$ on the fused score.

Table 6: Two-stage optimization schedule.

Stage	Steps	Trained params.	LR
Expert training	1–1,500	Experts ($\times K$)	10^{-4} (decayed)
Gate training	1,501–3,000	Gate only	10^{-4} (const.)

A.4 Multimodal Event Segmentation

Given a test video, frame-level RGB features are extracted once with the frozen CLIP ViT-B/16 (Radford et al., 2021) image encoder. Four auxiliary per-frame feature streams are extracted: optical flow (from pre-computed RAFT (Teed & Deng, 2020) flow fields, projected to 128 dimensions via a fixed random orthogonal projection), monocular depth (Depth Anything V2 (Yang et al., 2024)), 2D human pose (YOLOv7-Pose (Maji et al., 2022); the first two detected people are encoded as two 51-dimensional keypoint blocks packed into a 128-dimensional vector, zero-padded when fewer than two people are detected), and panoptic segmentation masks (SAM (Kirillov et al., 2023), projected to 128 dimensions via a fixed random orthogonal projection).

The auxiliary streams are multiplied by the Phase-B anomaly score and concatenated with the RGB features, which are not score-weighted, with weight $\alpha_{\text{clip}} = 0.8$ for RGB and $(1 - \alpha_{\text{clip}})/M$ for each of the M available auxiliary streams (Eq. 4). Following Shao et al. (2025), a single similarity graph is built over the fused sequence: each node is connected to its approximate k-nearest neighbors under a combined score, 0.8 times the cosine similarity of the (renormalized) RGB parts plus 0.2 times $\exp(-d)$, where d is the Euclidean distance between auxiliary parts, divided by a temporal penalty $1 + \gamma_{\text{time}}|i - j|$ with $\gamma_{\text{time}} = 0.05$; k starts at 5 and decreases to a floor of 3 over the video. Graph construction is performed in blocks of 200 frames to bound memory. Event boundaries are detected by combining, at each timestep, the squared Euclidean distance and the cosine dissimilarity between consecutive fused features, smoothing with a Savitzky–Golay filter (polynomial order 2, window $\lfloor 2.0 \times \text{fps} \rfloor$), and flagging peaks exceeding median $+ 3.0 \times$ MAD of the smoothed-to-local-average ratio; peaks closer than 2 seconds are merged. The resulting boundaries define a single partition of the video.

A.5 Segment-Level Prompting and Explanation Generation

Phase C queries Qwen3.5-27B (Team, 2026) in two passes (Qwen2.5-VL-7B (Bai et al., 2025) on Vad-R1). In both passes, 16 frames are uniformly sampled from the segment, converted to RGB, and resized so that their longer side does not exceed 448 px, preserving aspect ratio.

Scoring pass. Each segment of the partition is scored with the following prompt, whose output is the segment confidence q_j (Sec. 3.4):

```
``If you were a law enforcement officer monitoring this
video scene, how suspicious does this situation appear on a
scale from 0 to 1, where 0 means a completely normal scene
and 1 means highly suspicious activity?  Carefully observe
the actions, interactions, and context in the scene.  Return
only one number between 0 and 1 (example:  0.73).  Do not
include any explanation.''
```

Explanation pass. After contiguous abnormal segments are merged, each resulting segment is described and explained with the following prompt, where `[CATEGORIES]` is the category vocabulary of the target dataset:

```
``You are a forensic video analyst reviewing a short clip
extracted from a CCTV surveillance video.  Carefully observe
```

```
the actions, interactions and context shown across the
frames.  Think step by step about what is happening before
answering.  After your reasoning, respond with EXACTLY the
following four fields, each starting on its own line with
the exact tag shown below (no markdown, no extra commentary
after EXPLANATION):
DESCRIPTION: <a factual description of the clip in exactly 4
sentences>
SCORE: <a single number between 0 and 1, where 0 is a
completely normal scene and 1 is a highly suspicious or
anomalous scene>
TYPE: <exactly one label chosen from this list:
[CATEGORIES]>
EXPLANATION: <2 to 3 sentences:  (a) which specific visual
or behavioral cues make this scene normal or anomalous, and
(b) the likely severity or consequence of the situation>''
```

The SCORE and TYPE fields make the model commit to an explicit decision that the explanation must justify; the TYPE field provides the category labels shown in Fig. 2. On Vad-R1, the explanation pass is replaced by the original Vad-R1 prompt applied to the selected clip.

B QUALITATIVE ANOMALY EXPLANATIONS

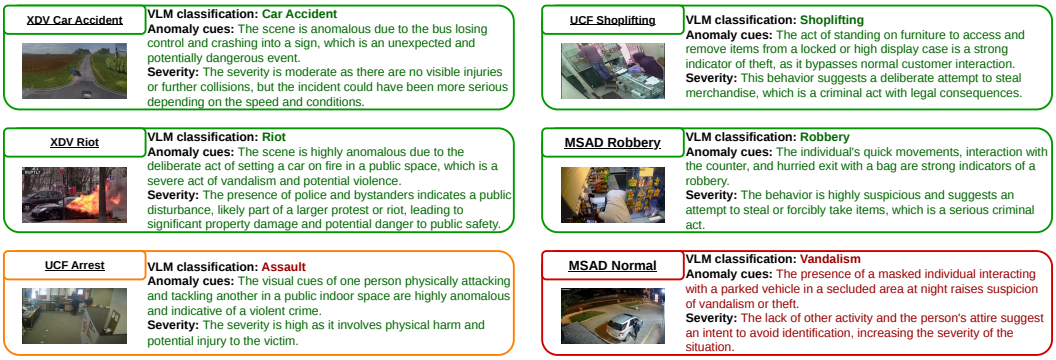


Figure 2: **Qualitative video-segment explanations produced by NOVAD.** Green boxes show correctly classified segments with explanations supported by observable actions, objects, and scene changes. The orange example illustrates fine-grained category ambiguity: an *Arrest* event is classified as *Assault* because both contain similar local physical interactions. The red example shows an unsupported contextual inference, where a normal MSAD segment is classified as *Vandalism* without an observable destructive action. These examples demonstrate both the interpretability of Phase C and its remaining sensitivity to category overlap and contextual bias.

Figure 2 shows both successful predictions and representative failures. NOVAD correctly identifies car accident, riot, shoplifting, and robbery by relying on complementary cues such as collision dynamics, crowd disruption, and human–object interactions. These examples illustrate the interpretability of detector-guided segments.

The failure cases reveal the remaining limitations. An arrest is misclassified as assault because both share similar local physical interactions, while a normal MSAD scene is predicted as vandalism based on contextual appearance rather than an observed destructive action. These examples show that NOVAD's explanations are useful not only for interpreting correct predictions, but also for diagnosing errors caused by category overlap, insufficient temporal context, and unsupported contextual inference.

C CLASS-WISE RESULTS

To provide further analytical insight into complex anomaly detection, Tables 7–9 compare performance across specific anomaly categories between NOVAD and the baseline method, our reimple-

mentation of TD-VAD. These results are computed on the UCF-Crime test set, the XD-Violence test set and the MSAD test set.

Table 7: UCF-Crime Class-Wise AUC. **Bold** denotes the best result. † denotes our reimplementation.

Class	TD-VAD†	NOVAD
Abuse	43.90	**44.84**
Arrest	66.90	**68.53**
Arson	**69.42**	63.15
Assault	84.49	**89.71**
Burglary	61.06	**70.74**
Explosion	**77.39**	69.22
Fighting	**60.89**	59.43
Road accident	52.96	**57.59**
Robbery	61.82	**85.91**
Shooting	56.83	**67.81**
Shoplifting	36.75	**49.24**
Stealing	59.22	**72.21**
Vandalism	56.16	**67.86**

Table 8: XD-Violence Class-Wise AP and AUC. **Bold** denotes the best result. † denotes our reimplementation.

	TD-VAD†		NOVAD	
Class	AP	AUC	AP	AUC
Abuse	**59.88**	**83.64**	52.77	79.45
Car accident	**46.46**	63.20	43.60	**64.70**
Explosion	**70.66**	**83.37**	45.11	69.64
Fighting	70.46	77.97	**83.63**	**85.58**
Riot	93.31	81.58	**95.19**	**87.08**
Shooting	47.55	68.83	**58.45**	**74.37**

Table 9: MSAD Class-Wise AUC and AP. **Bold** denotes the best result. † denotes our reimplementation.

	TD-VAD†		NOVAD	
Class	AUC	AP	AUC	AP
Assault	47.12	47.40	**65.58**	**69.15**
Explosion	**88.39**	88.70	86.58	**90.19**
Fighting	86.57	92.47	**88.85**	**93.68**
Fire	83.39	87.19	**86.04**	**93.59**
Object falling	**82.25**	**85.22**	72.10	78.02
People falling	**71.73**	**65.27**	59.92	53.71
Robbery	50.04	78.47	**57.18**	**83.74**
Shooting	**67.45**	**81.21**	52.98	72.37
Traffic accident	62.10	49.37	**67.83**	**51.80**
Vandalism	**59.16**	61.36	58.84	**78.39**
Water incident	96.05	99.17	**98.55**	**99.68**

Across the three benchmarks, NOVAD outperforms TD-VAD† on the majority of individual classes (10/13 on UCF-Crime; 3/6 in AP and 4/6 in AUC on XD-Violence; 6/11 in AUC and 8/11 in AP on MSAD). The most consistent exception involves *Explosion*, where TD-VAD retains an edge on both UCF-Crime and XD-Violence, and *Arson* on UCF-Crime. These categories are characterized by short, high-magnitude visual discontinuities (flash, smoke), which can yield embeddings already well separated from the normal class regardless of the semantic richness of the training text. Feature-magnitude cues in these classes are strong and largely scene-invariant indicators for high-intensity anomalies, reducing the relative benefit of category-specialized experts for this specific subset. The *Abuse* class on XD-Violence, where TD-VAD leads clearly in both AP and AUC,

coincides with a known imbalance in the dataset's category composition, *Abuse* being markedly under-represented relative to the other five categories, the resulting small test-set size for this class likely inflates the variance of both AP and AUC estimates, making the comparison less reliable than for better-populated categories. On MSAD, TD-VAD retains an advantage on *Object falling* and *People falling*, two categories with comparatively subtle and scene-dependent motion signatures; given that MSAD spans fourteen heterogeneous capture scenarios, this is broadly consistent with the sensitivity of magnitude-based anomaly cues to scene variation. The *Shooting* class shows a less consistent pattern, with NOVAD ahead on both UCF-Crime and XD-Violence but behind on MSAD despite an identical training protocol across the three benchmarks; we do not currently have a structural explanation for this discrepancy and leave a more detailed investigation of this class to future work.

D TEXT–VIDEO ALIGNMENT

Table 10 and Figs. 3–5 test whether generated descriptions and real videos share a usable embedding space. Text prototypes average the generated descriptions of each category; video prototypes average the annotated abnormal segments of the test videos of each category (all segments for normal videos). Although a modality gap remains ($\cos(\mu_t, \mu_v)$ between 0.60 and 0.83 (Liang et al., 2022)), every retrieval metric is far above chance: the correct category is ranked first for 28.6–87.6% of abnormal test videos (chance: 7.7–16.7%). Without any training, scoring segments by their similarity to the text prototypes, $\max_k \cos(x, t_k) - \cos(x, t_0)$, already yields 74.93, 88.00, and 75.43 AUC. Training the detector adds 7.9, 3.3, and 13.5 points, and helps most where alignment is weakest (MSAD).

Table 10: Text–video alignment in the Qwen3-VL embedding space. Chance level in parentheses.

Dataset	$\cos(\mu_t, \mu_v)$	V→T top-1	V→T top-3	T→V mAP	Zero-shot AUC	NOVAD AUC
UCF-Crime	0.749	33.6 (7.7)	62.9 (23.1)	22.8 (7.7)	74.93	82.78
XD-Violence	0.830	87.6 (16.7)	98.9 (50.0)	66.5 (16.7)	88.00	91.32
MSAD	0.601	28.6 (9.1)	63.0 (27.3)	24.6 (9.1)	75.43	88.89

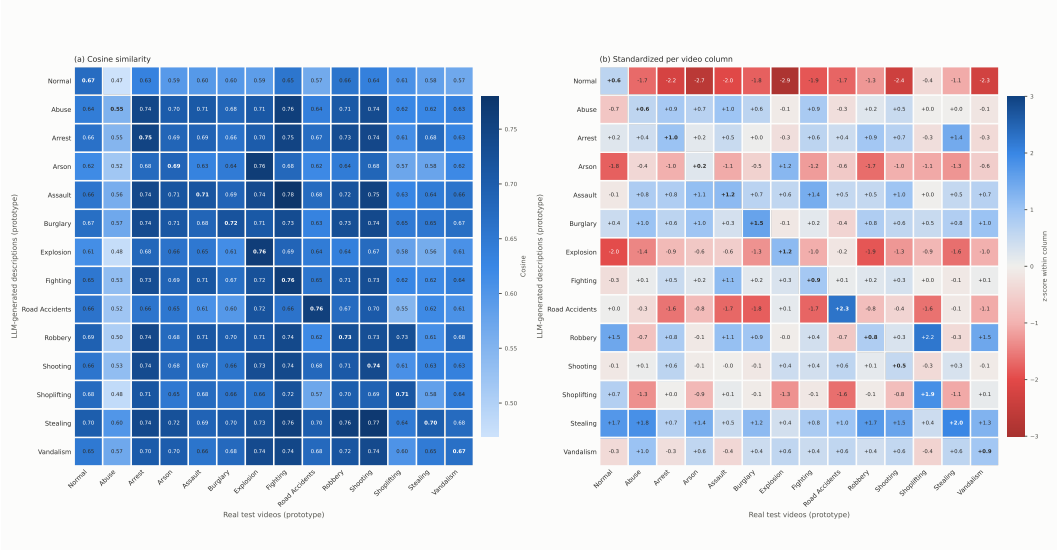


Figure 3: Text–video alignment on UCF-Crime in the Qwen3-VL embedding space. Rows: text prototypes (averaged generated descriptions of each category). Columns: video prototypes (averaged annotated abnormal segments of the test videos of each category; all segments for normal videos). (a) Raw cosine similarity. (b) Similarity standardized within each video column, which preserves the ranking used for retrieval while removing the constant offset caused by the modality gap. Boxed cells mark matching categories.

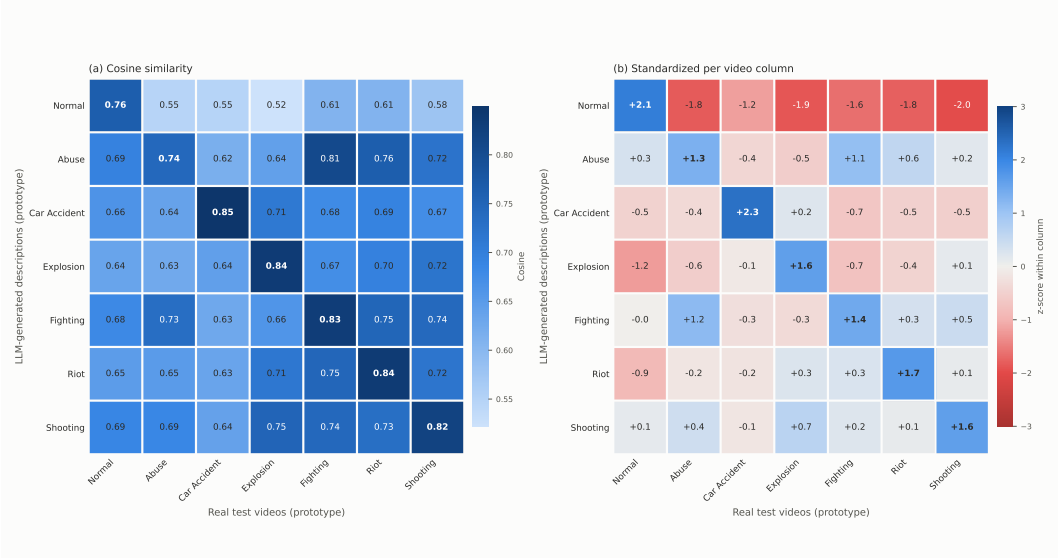


Figure 4: Text–video alignment on XD-Violence, same layout as Fig. 3. Videos with several labels are excluded.

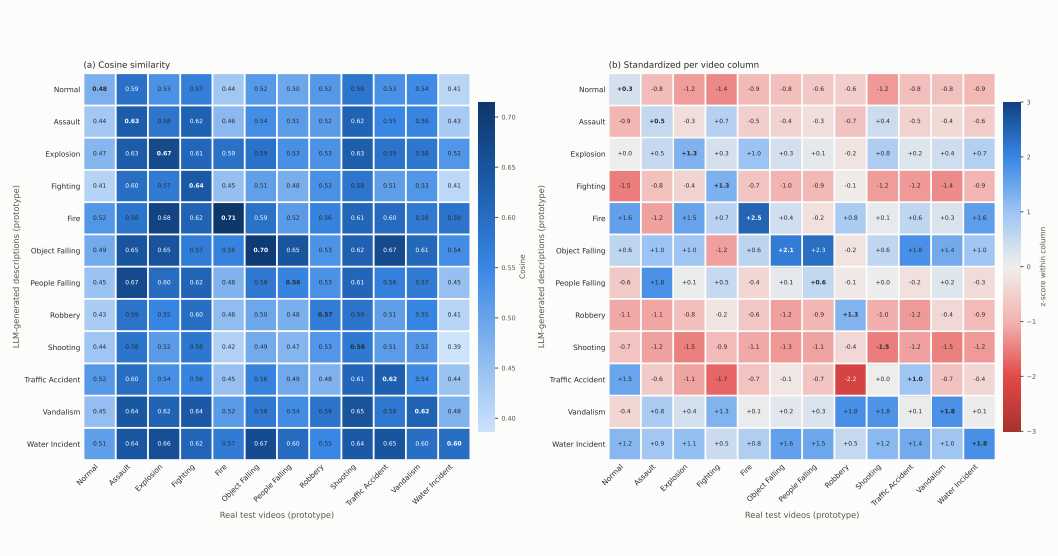


Figure 5: Text–video alignment on MSAD, same layout as Fig. 3.

E Unseen Categories and Expert Specialization

Each expert only sees its own category and normal descriptions, so removing expert c yields a detector that has never read a description of category c. Since the learned gate expects all K experts, this analysis fuses the remaining experts by averaging. On UCF-Crime, this changes the AUC on category c (its videos plus all normal videos) by only -0.34 points on average (worst case -1.88, Burglary). The learned anomaly representation therefore extends to categories absent from the textual supervision. Specialization is real but partial: the expert of the true category ranks first for 31% of the categories (chance: 7.7%), with a mean AUC of 89.38 versus 85.39 for the other experts (Fig. 6), and averaging all experts (91.82) outperforms the category's own expert. The experts thus act as complementary category-conditioned detectors rather than as a hard routing mechanism.

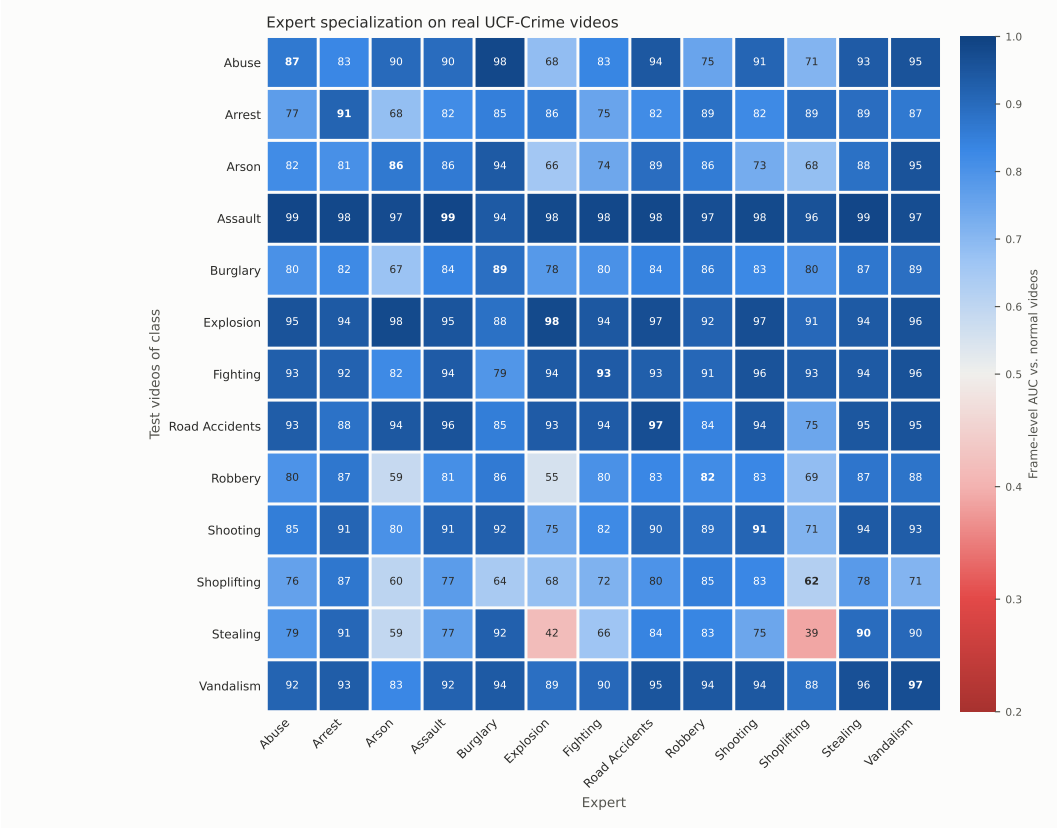


Figure 6: Expert specialization on UCF-Crime. Each cell is the frame-level AUC of expert k (columns) for separating the annotated abnormal frames of the test videos of category c (rows) from the frames of all normal test videos; 50 corresponds to chance. Boxed cells mark the expert trained on the true category. This metric is independent of the calibration of each expert.

F TD-VAD Reproduction

The code and training corpus of TD-VAD (Zhang et al., 2026) are not publicly available, so we reimplemented it from the paper. The reimplementation follows the described architecture and objectives: event-focus causal attention over non-overlapping windows of 16 steps, event-context causal attention with a learnable distance-dependent bias, summation of both representations, a binary and a multi-class head trained with top-k MIL losses ($k = 16$) sharing the same selected positions, and an anomaly score of $1 - p_{\text{normal}}$ from the multi-class head. Optimization follows the reported settings for UCF-Crime: AdamW, learning rate 2×10^{-4}, batch size 64, 10 epochs, and either the per-class text distribution of TD-VAD (3,253 descriptions) or our full corpus. Unspecified details were set as follows: a linear causal bias with a learnable slope per head; 8 attention heads and two-layer prediction heads with 128 hidden units. Since TD-VAD's corpus is unavailable, the model is trained on our generated descriptions. Test videos are processed with exactly the protocol used for NOVAD (16-frame segments, sliding windows of 16 segments with stride 8).

Table 11 disentangles the effects of the encoder and of the corpus size. With the original encoder and class distribution (CLIP ViT-B/16, 3,253 descriptions), our reimplementation reaches 67.07 AUC on UCF-Crime, below the reported 80.82. We attribute this gap to the unavailable training corpus and to unspecified implementation details. Qwen3-VL-Embedding improves TD-VAD with both corpus sizes (+4.24 and +14.97 AUC), so all text-supervised comparisons in the main paper use this encoder, which favors the baseline. The effect of the corpus size depends on the encoder: enlarging the corpus lowers the AUC with CLIP (-7.85) but raises it with Qwen3-VL ($+2.88$), suggesting that additional textual supervision only transfers to video when the text and image embeddings are

sufficiently aligned. TD-VAD† in Table 1 therefore uses its most favorable setting (Qwen3-VL, full corpus).

Table 11: Effect of encoder and corpus size for our TD-VAD reimplementation on UCF-Crime. We report frame-level AUC. The originally reported result is 80.82 AUC.

Encoder	3,253 descriptions	Full corpus
CLIP ViT-B/16	67.07	59.22
Qwen3-VL-Embedding	71.31	74.19

G Computational cost

We provide a computational cost analysis in Table 12 on the UCF-Crime benchmark.

Table 12: Computational cost for UCF-Crime (13 experts).

	1 expert	13 experts	Gate func.	NOVAD
GFLOPs	0.1859	2.4162	0.0006	2.4168
Params (M)	5.7065	74.1845	0.0183	74.2027

Scaling from 1 to $K = 13$ experts multiplies both GFLOPs and parameters by almost exactly 13, confirming that experts are architecturally identical and that cost grows linearly with the number of categories; the gating network adds a negligible 0.0006 GFLOPs / 0.018M parameters on top.

H Visual representations of textual features

Figures 7–9 show the t-SNE (Van der Maaten & Hinton, 2008) plot of the embedded text descriptions generated for the three benchmarks.

Unlike XD-Violence's six broad and visually distinctive violence categories, UCF-Crime and MSAD include several classes that share very similar physical actions and scene context (e.g., Fighting, Assault and Abuse in UCF-Crime are largely distinguished by intent and legal framing rather than a visually or semantically distinctive pattern), which naturally places their embeddings closer together in the shared representation space. This effect is further amplified on MSAD by its greater scene diversity across fourteen heterogeneous capture scenarios, as scene-level variation is known to interfere with the discriminability of anomaly-specific visual cues.

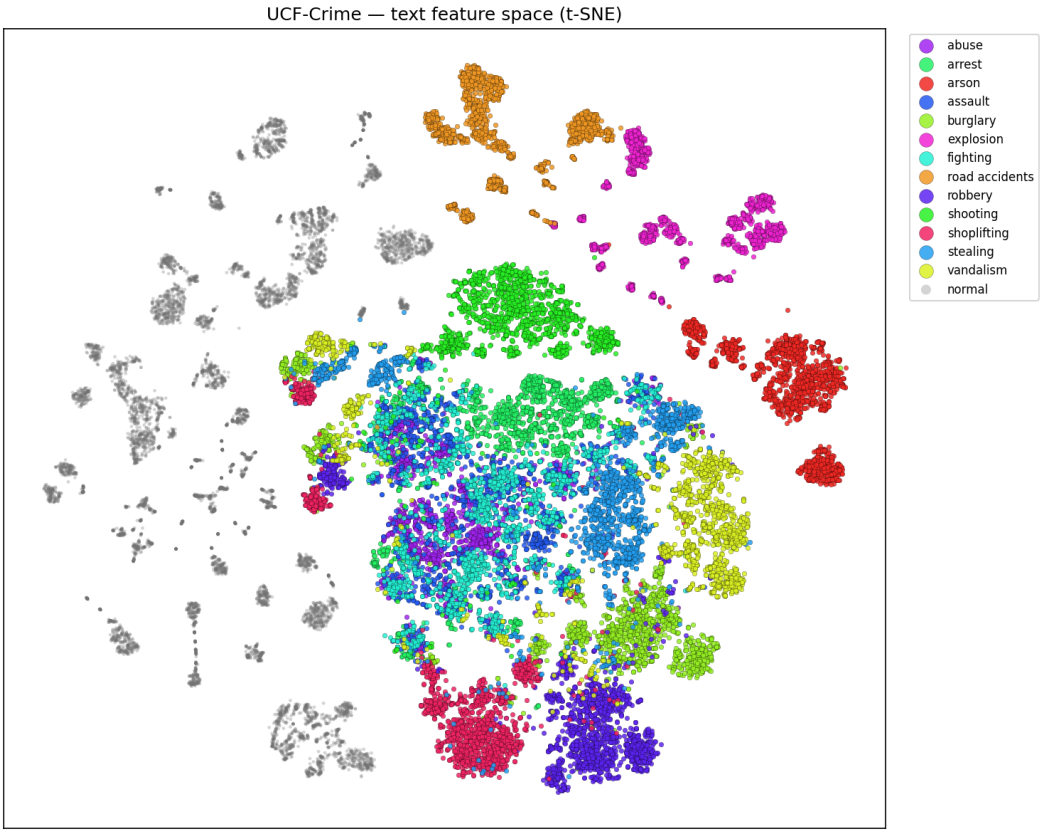


Figure 7: UCF-Crime textual features

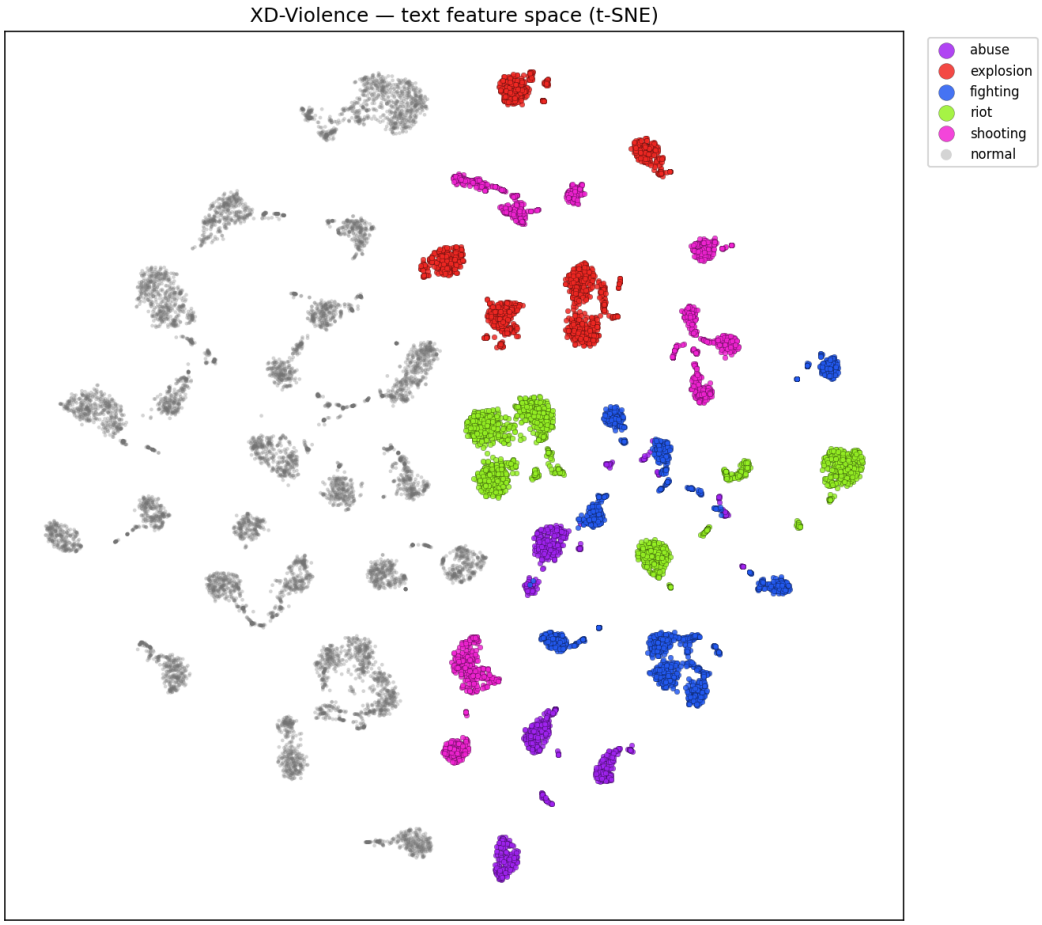


Figure 8: XD-Violence textual features

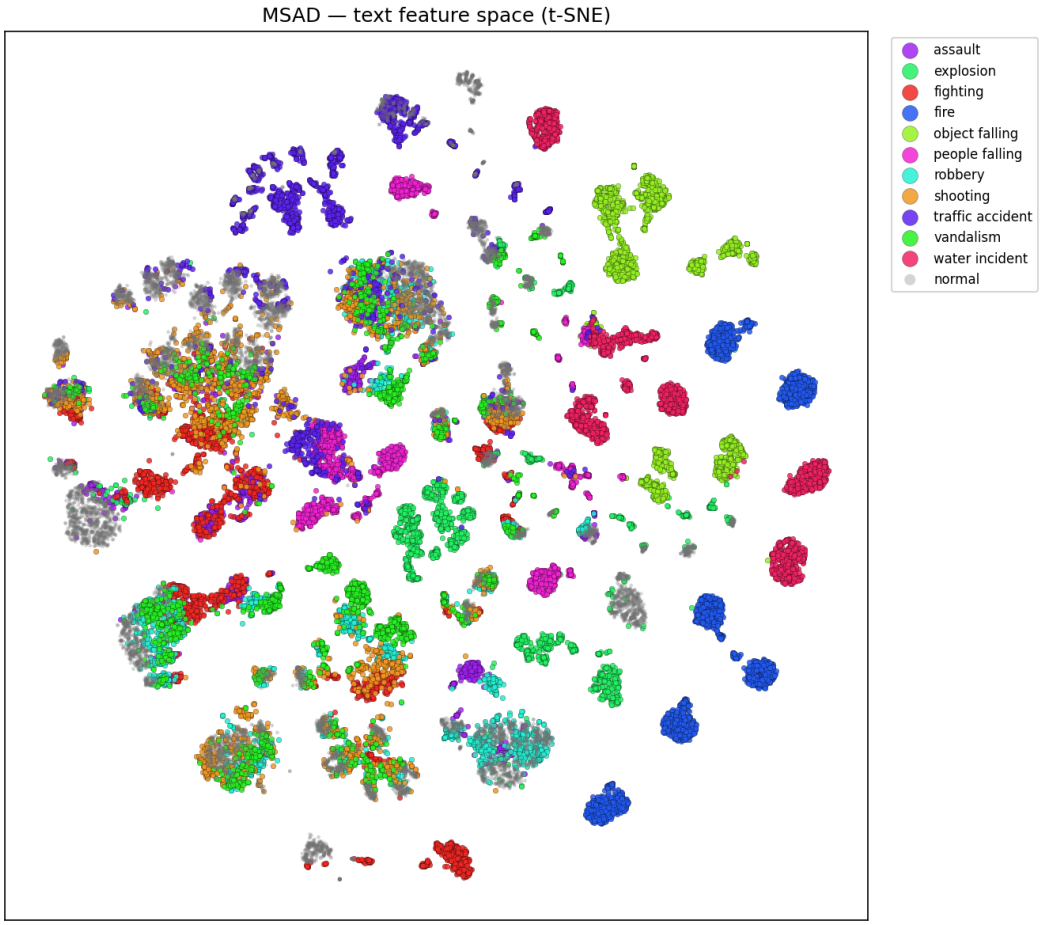


Figure 9: MSAD textual features