Training-Free Video Anomaly Detection via Uncertainty-Guided Hierarchical Retrieval with Vision-Language Models

Jiankang Zheng, Mengjingcheng Mo, Jiaxu Leng, Mingpi Tan, Zhanjie Wu, Ji Gan, Haosheng Chen and Xinbo Gao, *Fellow, IEEE*

***Abstract*—While many training-based video anomaly detection (VAD) methods achieve excellent results within their training domains, they often rely on target-domain annotations or adaptation, which may limit their deployment to unseen scenes. Conversely, training-free methods require no task-specific data and directly leverage the vast pre-training knowledge of vision–language models (VLMs) for strong generalization. However, in challenging scenarios with scarce visual cues (e.g., low light or occlusion), their performance deteriorates markedly due to insufficient auxiliary cues for reliable inference. Inspired by the human ability to recall high-confidence experiences in ambiguous situations, we propose Uncertainty-Guided Hierarchical Retrieval (UGHR), a training-free, plug-and-play framework built on VLMs, which effectively assesses scene uncertainty and performs hierarchical context retrieval to enhance performance under challenging scenarios. Accordingly, we present two elaborated modules: the uncertainty-aware retrieval gate (UG), which quantifies scene uncertainty to determine when additional context is needed to support accurate judgments; and the hierarchical context retriever (HR), which backtracks high-confidence examples to obtain the most relevant temporal, intra-scene, and semantically similar contexts, thereby reinforcing weakened visual cues. Extensive experiments on XD-Violence, UCF-Crime, and ShanghaiTech show that UGHR outperforms state-of-the-art training-free methods, even surpasses weakly supervised approaches in challenging scenarios, and provides both strong cross-dataset generalization and plug-and-play deployment. The code is available at https://github.com/lemonzjk/UGHR.**

***Index Terms*—Video anomaly detection, Vision–language models, Training-free inference**

I. Introduction

Video Anomaly Detection (VAD) automatically detects and localizes rare, unexpected, or hazardous events in video streams, enabling applications from public safety [1]–[6] to autonomous driving [7], [8]. While traditional training-based VAD pipelines [9] achieve high accuracy and robust feature representations on in-distribution data through supervised learning (Figure 1(a)(i)), they often depend on large target-domain annotation sets and require retraining or scene-specific adaptation, which can make cross-scene deployment more challenging in some settings (Figure 1(c)(ii)). Moreover, continuously gathering and labeling anomaly samples is resource-intensive, and is often hindered by data privacy and regulatory constraints. To reduce the reliance on extensive annotations, recent studies have explored instruction tuning of pretrained vision–language models (VLMs), injecting linguistic priors into visual features with only a small set of domain-specific image–text pairs [10]. While this lightweight tuning improves cross-domain transfer and interpretability, it still requires additional text annotations and nontrivial optimization cost.

By comparison, training-free methods [11], [12] require no task-specific data and build custom modules on top of a frozen VLM, allowing them to leverage broad pretraining knowledge and providing generalization potential across scenarios (Figure 1(a)(ii)). However, in severely degraded visual conditions such as low light, heavy occlusion, or camera motion, these approaches fail to acquire and exploit the most relevant contextual cues, leaving a large gap between standard and challenging scenarios (see Figure 1(c)(i)). To alleviate information poverty from degraded visual cues, a straightforward remedy is flat retrieval (FR): at inference it fetches from all previously observed samples most similar to the current query and supplies it as external context.

We adopt this as our baseline but identify two shortcomings: (1) retrieval ignores scene uncertainty, disrupting confident predictions and wasting compute; (2) global search drifts semantically, returning weakly related references that introduce noise and mask true anomalies.

Drawing inspiration from human strategies for resolving information scarcity, we note that human memory retrieval follows an efficient hierarchical process rather than exhaustive scanning [13], [14]. When confronted with explicit visual cues (e.g., a clearly visible weapon), decisions are made based on immediate perceptual input. In cases of ambiguous cues (e.g., a fleeting shadow), humans employ a structured retrieval approach: (1) first examining proximate temporal context; (2) then exploring broader scene context; and (3) finally accessing semantically related memories from distant experiences (Figure 1(b)). This hierarchical, relevance-driven search strategy efficiently balances information quality with processing demands, offering valuable insights for the design of VAD systems.

Guided by this insight, we propose Uncertainty-Guided

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123601; in part by the National Natural Science Foundation of China under Grants 62472060 and 62221005; in part by the Science and Technology Innovation Key R&D Program of Chongqing under Grant CSTB2023TIAD-STX0016; in part by the Natural Science Foundation of Chongqing under Grant CSTB2024NSCQ-QCXMX0060; in part by the China Postdoctoral Science Foundation under Grant 2025MD774186; in part by the Chongqing Special Postdoctoral Research Funding under Grant 2024CQBSHTB2002; in part by the Chongqing University of Posts and Telecommunications Ph.D. Innovative Talents Project under Grant BYJS202404; and in part by the Chongqing Graduate Research Innovation Project under Grant CYS260449.

J. Zheng, M. Mo, J. Leng, M. Tan, J. Wu, J. Gan, H. Chen and X. Gao are with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China (E-mail: s240201107@stu.cqupt.edu.cn, mo1031@live.com, lengjx@cqupt.edu.cn, tanmingp11@163.com, s230201119@stu.cqupt.edu.cn, ganji@cqupt.edu.cn, chenhs@cqupt.edu.cn, gaoxb@cqupt.edu.cn). They are also affiliated with the Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory, Chongqing, China. (Corresponding author: M. Mo.)

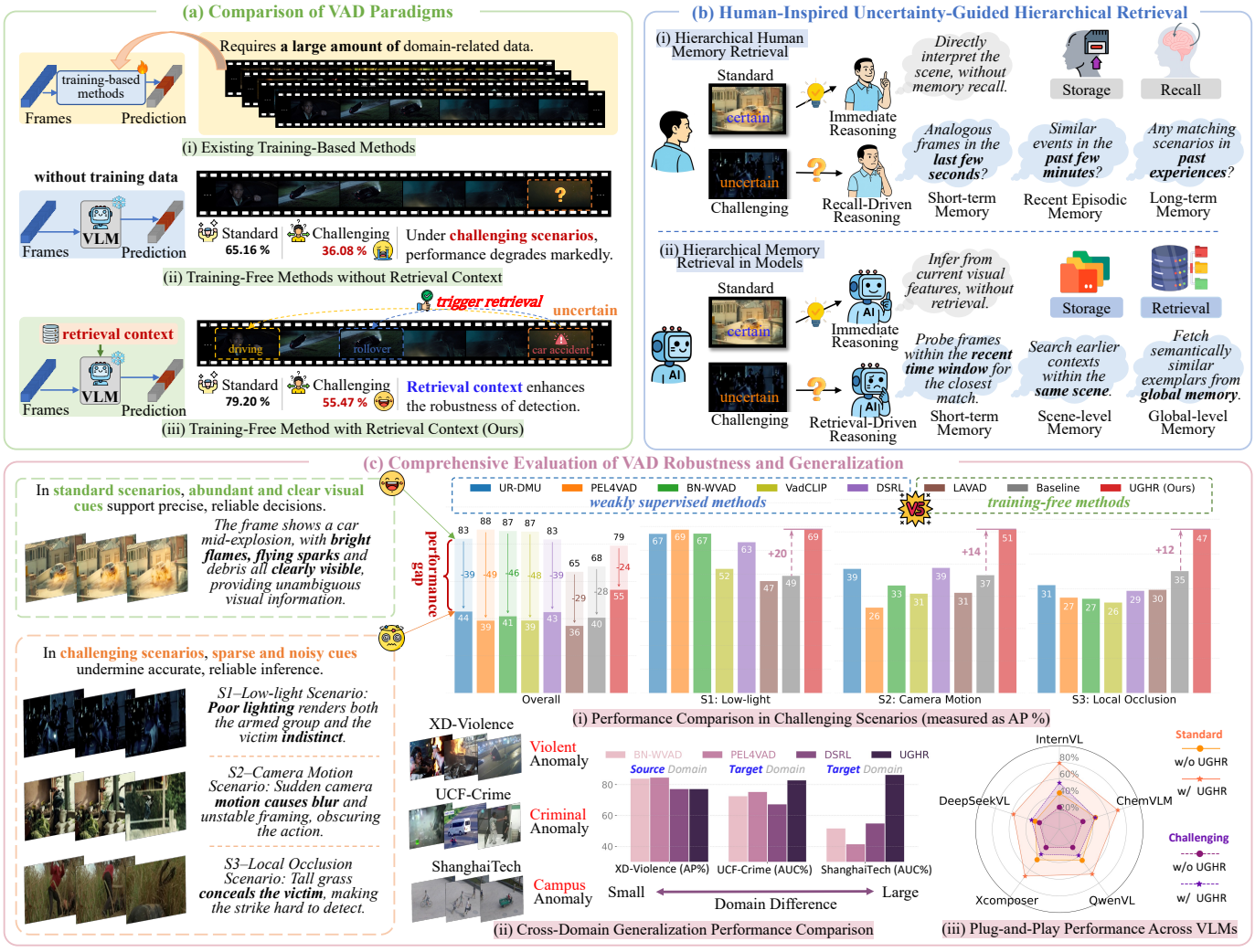


Fig. 1. Illustration of our research motivation. We integrate uncertainty-guided hierarchical retrieval into VAD, mimicking human strategies for resolving information scarcity. The design improves robustness under challenging scenarios, enables smooth domain transfer, and supports plug-and-play VLM integration.

Hierarchical Retrieval (UGHR), a training-free, plug-and-play framework built on VLMs, which effectively assesses scene uncertainty and performs hierarchical context retrieval to enhance performance in challenging scenarios (see Figure 1(a)(iii)). It is designed as a complementary training-free alternative, particularly for scenarios where target-domain training data or model adaptation is unavailable. Specifically, UGHR has two core modules plus a scoring stage: (1) **Uncertainty-aware Retrieval Gate (UG).** The UG estimates scene uncertainty using the Rényi entropy [15] to determine when additional context is necessary. By triggering hierarchical context retrieval only for high-uncertainty samples, this gate both exposes decision blind spots and avoids unnecessary interference with low-uncertainty ones. (2) **Hierarchical Context Retriever (HR).** When retrieval is required, the HR queries three levels of memory, each supplying supplementary evidence in a distinct form: (i) a short-term level that inspects the last few seconds, capturing the most temporally relevant cues; (ii) a scene level that scans earlier parts of the same video, supplying broader in-scene context; and (iii) a global level that queries accumulated high-confidence samples for semantically similar events, providing exemplars for novel or rare occurrences. Importantly, all memories are constructed from previously processed video segments, ensuring that no information is drawn from unseen test data. (3) **Vision–language Anomaly Scoring.** The current segment, its retrieved context, and our tailored prompt are fed into a frozen VLM for anomaly scoring and textual explanation. The scores are then refined by Evidence-aware Temporal Attention (EA) to suppress transient noise. Cooperatively, these modules determine both *when* and *where* to retrieve context, supplying the VLM with enriched visual cues to boost detection performance without additional training, thereby achieving robust, generalizable anomaly detection in information-sparse scenarios. Consequently, UGHR establishes a new training-free state of the art, achieving 77.19% AP on XD-Violence, 82.84% AUC on UCF-Crime, and 86.37% AUC on ShanghaiTech. It improves AP by 18.41% on challenging subsets and achieves gains of over 20% across multiple VLM backbones, underscoring its robustness, cross-domain generalization, and plug-and-play applicability.

In summary, the contributions of this paper are as follows:

- We propose UGHR, a novel training-free and plug-and-play framework for VAD that leverages frozen VLMs and effectively tackles challenging scenarios by adaptively reinforcing weak visual cues.
- We jointly leverage an uncertainty-aware retrieval gate (UG), which selectively triggers retrieval under high uncertainty, and a hierarchical context retriever (HR), which gathers high-confidence examples at short-term, scene, and global levels, to adaptively supplement contextual information for more accurate anomaly detection.
- We conduct extensive experiments on XD-Violence, UCF-Crime, and ShanghaiTech, and the results demonstrate that UGHR achieves new training-free state-of-the-art performance, highlighting its robustness, generalization, and plug-and-play applicability.

II. Related work

a) Video Anomaly Detection: Depending on how much annotation is assumed during training, research on training-based VAD methods is commonly divided into four branches: fully supervised, weakly supervised, one-class, and unsupervised approaches. **Fully supervised** approaches learn a frame-level normal-abnormal boundary on 3D-CNN or I3D backbones, but their reliance on dense frame-level annotations can limit scalability in practice [16], [17]. **Weakly supervised** methods rely only on video-level binary tags; since the multi-instance learning (MIL) paradigm of Sultani *et al.* [1], work combining 3D-CNNs, Transformers and self-attention has steadily improved clip selection while keeping labelling effort low [9], [18]–[24]. **One-class** VAD trains solely on normal videos, modelling normality via reconstruction error, adversarial generation or pseudo-anomalies, yet remains sensitive to data purity [25]–[30]. **Unsupervised** methods further relax these assumptions, allowing unlabelled anomalies to appear in the training split; representative techniques include GAN-based auto-encoders, diffusion denoising and pseudo-label recycling with One-Class SVM or iForest [31]–[34]. Recent unsupervised work, LANP-UVAD [35], introduces a normality prior and a normality propagation mechanism to inject prior knowledge into anomaly learning, thereby improving the modeling of less salient anomalies. Despite substantial progress, many training-based VAD methods still involve nontrivial annotation or data-curation cost, and their deployment to new domains often requires additional adaptation. Moreover, compared with recent VLM-based pipelines, their semantic interpretability is typically less explicit.

b) VLMs/LLMs in VAD: To break the cost–robustness barrier of conventional pipelines, recent studies tap into large vision–language or language models (VLMs/LLMs) along two complementary routes. **Instruction tuning with lightweight adapters.** HAWK [10], VAD-LLaMA [36] and Holmes-VAD/VAU [37], [38] leave all backbone weights frozen and adjust only a handful of LoRA layers or prompt tokens on thousands of video–description pairs. This brings human-readable explanations, but it still relies on curated instructional data, and its transferability may remain constrained by the tuning domain. **Weight-frozen inference.** In this line, the pre-trained model is queried as-is. Some methods remain *data-dependent*: VERA [39] learns textual queries from coarse labels, while AnomalyRuler [40] mines normal clips and derives symbolic rules with an LLM. SUVAD [41] further enriches this line by leveraging MLLMs to generate fine-grained semantic descriptions for LLM-based anomaly judgment, improving semantic understanding and anomaly explanation. Others are entirely *data-free*: LAVAD [11] captions every frame with a VLM and lets an LLM aggregate the captions temporally, and EventVAD [12] segments videos into semantically coherent events and hierarchically prompts an LLM for anomaly localization. AnyAnomaly [42] further broadens the data-free setting by enabling zero-shot customizable VAD through user-specified anomaly queries and context-aware visual question answering. However, both rely on uniform pipelines that fail to exploit the most relevant contextual cues, leading to a pronounced performance gap between standard and challenging scenarios. Unlike prior data-free, weight-frozen approaches, UGHR effectively assesses scene uncertainty and performs hierarchical context retrieval to enhance performance under challenging scenarios.

III. Method

A. Problem Definition

Let $\mathcal{V} = [I_1, I_2, \ldots, I_F]$ be a test video consisting of F frames, where each frame $I_f \in \mathcal{I}$ is either normal or anomalous. To keep the temporal units consistent throughout this section, we use f to index individual frames and partition $\mathcal{V}$ into $S = \lceil F/\eta \rceil$ non-overlapping segments $\{v_t\}_{t=1}^{S}$, where $v_t = [I_{(t-1)\eta+1}, \ldots, I_{\min(t\eta,F)}]$ and t indexes segments. Traditional VAD methods learn a mapping $g : \mathcal{I}^F \to [0,1]^F$ from a labeled dataset $\mathcal{D}$, such that $g(I_f) \to 1$ if I_f is anomalous and $g(I_f) \to 0$ otherwise. In our *training-free VAD* setting, no labeled dataset $\mathcal{D}$ or fine-tuning is available. Instead, using only a pre-trained VLM, we process the video segment by segment and compute an anomaly score $\alpha_t \in [0,1]$ with an explanation for each segment v_t at inference time.

For readability, we summarize the recurrent notation used below. K_S denotes the short-term temporal window used to construct the short-term memory level, and $\mathcal{N}_t$ denotes the segment-level temporal neighborhood centered at v_t for temporal anomaly-score refinement. Unless otherwise specified, s indexes historical segments, i indexes category prompts or retrieved references, ℓ indexes neighboring segments or retrieval stages, and k indexes memory entries; κ denotes the maximum number of references retrieved from each bank. We use U_t, α_t, and C_t for the uncertainty, anomaly, and confidence scores of segment v_t, respectively, while τ_u and τ_c denote the uncertainty and confidence thresholds. $\mathcal{M}_{\text{nor}}$ and $\mathcal{M}_{\text{ano}}$ denote the normal and anomaly memory banks, each of which is later organized into short-term, scene, and global levels.

B. Method Overview

To provide a unified view of the inference pipeline, we present a compact overview of UGHR in Figure 2. Given an input video segment, the Uncertainty-aware Retrieval Gate (UG) first estimates its uncertainty and decides whether additional

contextual retrieval is needed. If triggered, the Hierarchical Context Retriever (HR) retrieves supportive references from the memory banks for subsequent vision-language anomaly scoring. The resulting anomaly score is then refined by the Evidence-aware Temporal Attention (EA) module within a temporal neighborhood, while high-confidence segments are written back to the corresponding memory bank.

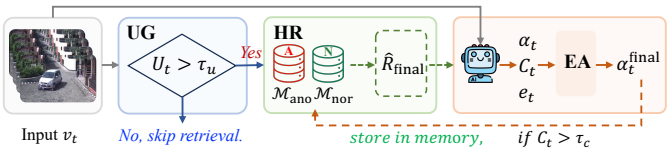


Fig. 2. Overall inference pipeline of the proposed UGHR framework.

The overall procedure is further summarized in Algorithm 1 in a concise sequential form. We next describe each module in detail.

Algorithm 1 Inference Pipeline of UGHR

Input: test video $\mathcal{V}$, frozen VLM Φ_{VLM}, category prompts $\{c_i\}_{i=1}^m$, segment length η, temporal radius W_R, thresholds τ_u and τ_c, cut-offs ε_{low} and $\varepsilon_{\text{high}}$, reference number κ.
Initialize: normal and anomalous banks $\mathcal{M}_{\text{nor}}, \mathcal{M}_{\text{ano}} \leftarrow \varnothing$.
partition $\mathcal{V}$ into $S = \lceil F/\eta \rceil$ non-overlapping segments $\{v_t\}_{t=1}^S$
for $t = 1$ to S **do**
  obtain the current segment v_t from $\mathcal{V}$
  # Step 1: Uncertainty-aware Retrieval Gate
  encode frames in v_t and all prompts in the shared embedding space
  compute frame-wise uncertainty $\{U_f\}$ and aggregate them to obtain U_t
  # Step 2: Hierarchical Context Retrieval
  if $U_t > \tau_u$ **then**
    retrieve $\hat{R}_{\text{final}}$ by progressively searching $\mathcal{M}_{\text{nor}}$ and $\mathcal{M}_{\text{ano}}$ across short-term, scene-level, and global memories
  else
    $\hat{R}_{\text{final}} \leftarrow \varnothing$
  end if
  # Step 3: Vision-language Anomaly Scoring
  obtain anomaly score α_t, confidence C_t, and explanation e_t by feeding v_t and $\hat{R}_{\text{final}}$ into Φ_{VLM}
  # Evidence-aware Temporal Attention
  obtain α_t^{final} by refining α_t over $\mathcal{N}_t = \{t - W_R, \ldots, t + W_R\}$
  # Online Memory Update
  form the current memory entry $b_t = (\mathbf{f}_t^v, U_t, C_t, \alpha_t)$
  if $C_t > \tau_c$ **then**
    insert b_t into $\mathcal{M}_{\text{nor}}$ if $\alpha_t \le \varepsilon_{\text{low}}$
    insert b_t into $\mathcal{M}_{\text{ano}}$ if $\alpha_t \ge \varepsilon_{\text{high}}$
  end if
end for
return final anomaly scores $\{\alpha_t^{\text{final}}\}_{t=1}^S$ and explanations $\{e_t\}_{t=1}^S$

C. Uncertainty-aware Retrieval Gate (UG)

As shown in Figure 3(a), we introduce an Uncertainty-aware Retrieval Gate at the front of the UGHR pipeline to expose decision blind spots for high-uncertainty samples and avoid unnecessary interference with low-uncertainty ones. This module uses the cosine distance between visual and textual features to measure the "hesitation" of the current sample across predefined category prompts, triggering retrieval only when uncertainty is classified as high. Concretely, the UG module operates on the segment sequence $\{v_t\}_{t=1}^S$, where each segment contains η frames. We extract the visual feature of every frame I_f, $f \in [1, F]$ via a frozen visual encoder E_v, yielding $\mathbf{f}_f^v = E_v(I_f)$. Meanwhile, a fixed set of m category prompts $\{c_i\}_{i=1}^m$ is encoded by a frozen text encoder E_t to obtain $\mathbf{f}_i^t = E_t(c_i)$. These prompts are predefined before inference rather than learned dynamically. Specifically, for each dataset, we construct one prompt for each official anomaly category, rewrite the category names into concise natural-language expressions, and include one additional normal prompt. Therefore, m denotes the total number of prompts used for inference. The complete prompt sets for all datasets are provided in Table I. Since E_v and E_t share the same embedding space, we directly compare the visual feature $\mathbf{f}_f^v$ with each prompt feature $\mathbf{f}_i^t$ and compute the cosine distance $d_{i,f} = 1 - \langle \mathbf{f}_f^v, \mathbf{f}_i^t \rangle / (\|\mathbf{f}_f^v\| \|\mathbf{f}_i^t\|)$, where $i = 1, \ldots, m$. A smaller $d_{i,f}$ indicates a closer match. We then apply the softmax function to the negated distance vector $\mathbf{d}_f = (d_{1,f}, \ldots, d_{m,f})$, obtaining the probability distribution $p_{i,f} = [\,\text{softmax}(-\mathbf{d}_f)\,]_i$. Thus $\{p_{i,f}\}$ represents the relative certainty that frame I_f matches each prompt. Because the retrieval gate needs to estimate whether a frame is strongly supported by a single dominant prompt or remains ambiguous among multiple competing prompts, it benefits from an uncertainty measure that captures the dispersion of the full prompt-matching distribution. We therefore quantify frame-wise uncertainty by the ρ-order Rényi entropy [15] of $\{p_{i,f}\}$:

$$U_f^{(\rho)} = \frac{1}{\log m} \frac{1}{1-\rho} \log\Big(\sum_{i=1}^{m} p_{i,f}^{\rho}\Big), \qquad \rho > 0,\ \rho \neq 1, \quad (1)$$

where $U_f^{(\rho)} \in [0, 1]$, ρ controls sensitivity to dominant vs. tail probabilities, and by choosing $\rho \approx 1$ we balance both—preventing moderate peaks from unduly lowering uncertainty or minor tails from unduly raising it.

We then obtain a segment-level uncertainty score via a Riemann integral over the continuous frame index ζ as follows: $U_t^{(\rho)} = \frac{1}{\eta} \int_{(t-1)\eta}^{t\eta} U^{(\rho)}(\zeta)\, d\zeta$. Here, the integral expresses temporal averaging over the segment; in practice, with discrete video frames, it is implemented by averaging the uncertainty scores of all frames in that segment. This aggregation converts frame-wise fluctuations into a segment-level uncertainty signal that aligns with the segment-level retrieval decision, suppresses isolated frame spikes, and yields a more temporally stable retrieval trigger. For brevity, we drop the superscript hereafter and simply write U_t. Comparing U_t with the preset threshold τ_u, if $U_t > \tau_u$ the segment is deemed *high-uncertainty* and triggers the hierarchical context retriever; otherwise it is *low-uncertainty* and skips retrieval, proceeding directly to vision–language anomaly scoring. This adaptive gating both ensures only ambiguous segments incur the extra cost of context lookup and prevents unnecessary interference with low-uncertainty segments.

D. Hierarchical Context Retriever (HR)

As shown in Figure 3(b), segments designated by the gate as high-uncertainty trigger the hierarchical context retriever to augment their visual context. Let $\mathbf{1}[\cdot]$ denote the indicator function that returns 1 if the enclosed condition holds (otherwise 0), and let $\Delta_{s,t} = t - s$. Here, t denotes the index of the current sample under evaluation, while s indexes a

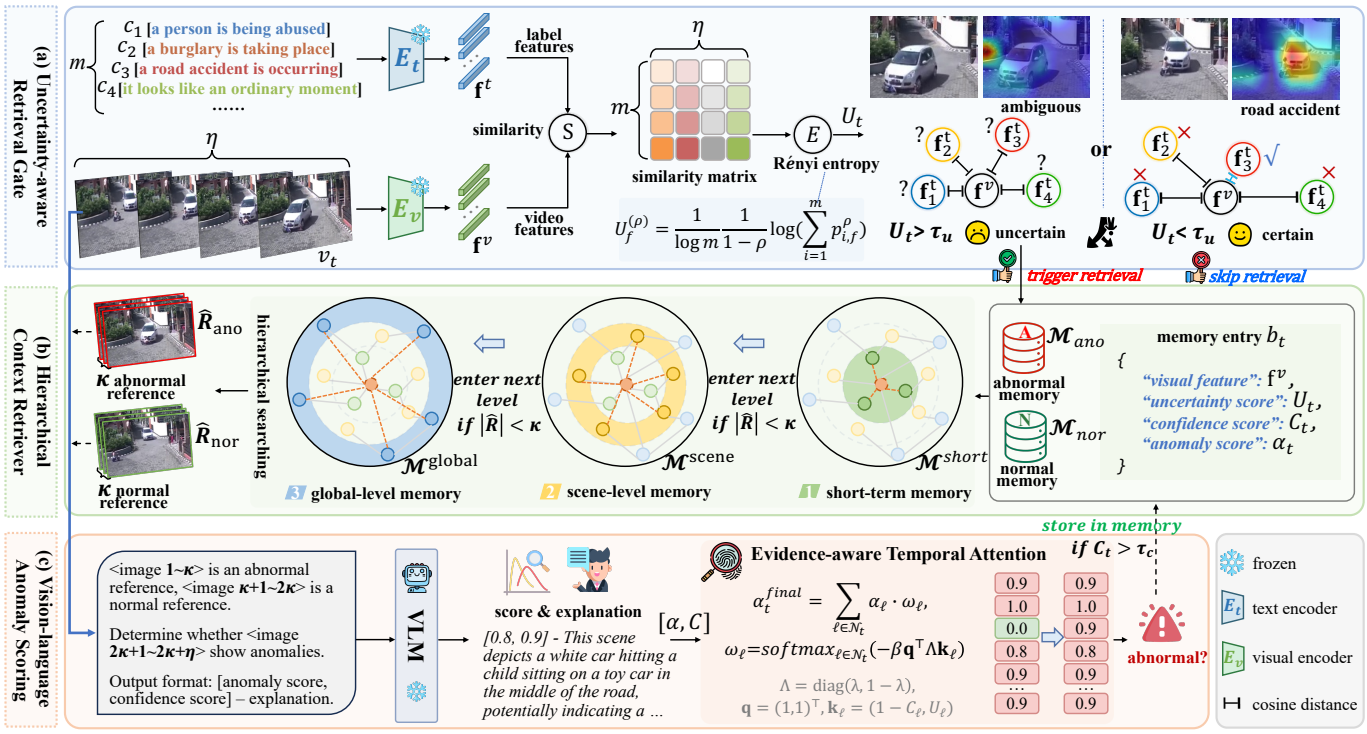


Fig. 3. Pipeline of our Uncertainty-guided Hierarchical Retrieval (UGHR) framework for training-free VAD. It uses an uncertainty-aware gate to trigger hierarchical context retrieval from short-term, scene, and global memories, dynamically updates those memories with high-confidence samples, and integrates the retrieved context into a frozen vision–language model for robust anomaly scoring and interpretation.

candidate sample from the set of previously processed ones. For clarity, the hierarchical memory notation uses a bank-type subscript and a context-level superscript. Specifically, the subscripts nor and ano denote the normal and anomalous memory banks, respectively, while the superscripts short, scene, and global denote the corresponding retrieval levels. For example, $\mathcal{M}_{\text{nor}}^{\text{short}}$ denotes the short-term normal memory pool. When the discussion applies to both bank types, we omit the bank-type subscript and write the level-specific pool generically as $\mathcal{M}^{\text{short}}$, $\mathcal{M}^{\text{scene}}$, or $\mathcal{M}^{\text{global}}$. We maintain two memory banks (normal bank $\mathcal{M}_{\text{nor}}$ and anomaly bank $\mathcal{M}_{\text{ano}}$), each organized into three hierarchical levels (short-term, scene, and global), defined as

$$\begin{aligned}\mathcal{M}^{\text{short}} &= \{v_s : \mathbf{1}[s<t]\,\mathbf{1}[C_s>\tau_c]\,\mathbf{1}[\Delta_{s,t}\le K_S]=1\},\\ \mathcal{M}^{\text{scene}} &= \{v_s : \mathbf{1}[s<t]\,\mathbf{1}[C_s>\tau_c]=1\},\\ \mathcal{M}^{\text{global}} &= \{v_s : \mathbf{1}[C_s>\tau_c]\,\mathbf{1}[v_s\in\mathcal{V}_{\text{hist}}]=1\},\end{aligned} \tag{2}$$

where C_s is the VLM's confidence for sample v_s, τ_c is the confidence threshold, $\mathcal{V}_{\text{hist}}$ contains all previously processed videos, and K_S bounds the short-term window.

Eq. (2) is not intended to define three isolated memory types. Instead, it defines hierarchical retrieval scopes from near to far according to the contextual range of evidence required for anomaly judgment. Specifically, short-term memory captures the nearest temporal cues, scene-level memory expands to earlier context within the same video, and global memory further extends to high-confidence samples from historical videos for broader cross-scene reference. Under this formulation, all three memory types share the same reliability filter $C_s > \tau_c$ but differ in contextual scope: $\mathcal{M}^{\text{short}} \subseteq \mathcal{M}^{\text{scene}}$, while $\mathcal{M}^{\text{global}}$ further extends retrieval to high-confidence samples from historical videos. This helps HR prioritize nearby and relevant evidence first, and to broaden the contextual range only when necessary.

We represent each segment v by the visual feature $\mathbf{f}^v$ of its key (middle) frame, and measure the similarity between query segment v_t and candidate v_s in memory bank as $\text{sim}(\mathbf{f}_t^v, \mathbf{f}_s^v) = \langle \mathbf{f}_t^v, \mathbf{f}_s^v \rangle / (\|\mathbf{f}_t^v\| \|\mathbf{f}_s^v\|)$. Each bank supplies at most κ references via an identical three-stage procedure. For the normal bank, let $R_{\text{nor}}^{(0)} = \varnothing$ and $L_1 = \text{short}$, $L_2 = \text{scene}$, $L_3 = \text{global}$. Then for $\ell = 1, 2, 3$:

$$R_{\text{nor}}^{(\ell)} = R_{\text{nor}}^{(\ell-1)} \cup \text{Top}_{\kappa - |R_{\text{nor}}^{(\ell-1)}|}\big(\mathcal{M}_{\text{nor}}^{L_\ell}, \text{sim}(\mathbf{f}_t^v, \cdot)\big), \tag{3}$$

where $\mathcal{M}_{\text{nor}}^{L}$ is the L-level normal pool, $R_{\text{nor}}^{(\ell)}$ denotes the references collected up to stage ℓ, and $\text{Top}_n(\cdot)$ returns at most n candidates whose similarity to the query is no lower than a preset threshold τ_s; if fewer than n such samples exist in the current pool, the shortfall is sought in the next memory level. Let $\ell_{\text{nor}}^* = \min\{\ell : |R_{\text{nor}}^{(\ell)}| \ge \kappa\}$. If $\ell_{\text{nor}}^* \le 3$, set $\widehat{R}_{\text{nor}} = R_{\text{nor}}^{(\ell_{\text{nor}}^*)}$; otherwise $\widehat{R}_{\text{nor}} = R_{\text{nor}}^{(3)}$. We apply an identical procedure to $\mathcal{M}_{\text{ano}}$ to obtain $\widehat{R}_{\text{ano}}$. The final reference set is

$$\widehat{R}_{\text{final}} = \widehat{R}_{\text{nor}} \cup \widehat{R}_{\text{ano}}, \tag{4}$$

containing up to 2κ segments for downstream anomaly scoring. This progressive retrieval first exploits temporally adjacent cues, then expands into same-scene context, and finally to semantically similar events across all videos. Such a design

retrieves the most relevant evidence with minimal overhead, ensuring robust support in ambiguous or rare scenarios. This can be particularly useful for difficult anomalies, where the current segment often contains only weakened, local, or discontinuous anomaly cues due to occlusion, blur, or low visibility. In such cases, hierarchical retrieval is effective not merely because it adds more context, but because the retrieved references may guide the model to focus on anomaly-relevant local evidence by relating these weak cues to more complete event evolution and anomaly semantics. As a result, anomaly-consistent interpretations become more competitive than distracting background or appearance variations, making the weak anomaly evidence more likely to contribute to the final judgment.

E. Vision–language Anomaly Scoring

As shown in Figure 3(c), we construct the image set $\mathcal{I}_t = \{v_t\} \cup \widehat{R}_{\text{final}}$, where $\widehat{R}_{\text{final}} = \{r_1, \ldots, r_{2\kappa}\}$ is the reference set for the current segment v_t, and feed $\mathcal{I}_t$ into the frozen VLM Φ_{VLM} along with a structured prompt. Each reference r_i is labeled as *normal* or *abnormal* to provide context, while the current sample v_t remains unlabeled. The model outputs

$$\alpha_t,\ C_t,\ \mathrm{e}_t = \Phi_{\text{VLM}}\big(\mathcal{I}_t, \text{prompt}\big), \tag{5}$$

where $\alpha_t \in [0,1]$ is the anomaly probability of v_t, $C_t \in [0,1]$ the associated model confidence, and e_t the natural-language explanation.

a) Evidence-aware Temporal Attention (EA): To suppress frame-wise noise, we refine raw anomaly scores within a temporal neighbourhood $\mathcal{N}_t = \{t - W_R, \ldots, t + W_R\}$. For each neighbour $\ell \in \mathcal{N}_t$, we recall its triple $(\alpha_\ell, C_\ell, U_\ell)$ and convert the confidence–uncertainty pair $(1 - C_\ell, U_\ell)$ into an *evidence score* $-\beta\, \mathbf{q}^\top \Lambda \mathbf{k}_\ell$. A softmax over these scores yields attention weights that aggregate the neighbourhood:

$$\alpha_t^{\text{final}} = \sum_{\ell \in \mathcal{N}_t} \omega_\ell\, \alpha_\ell, \qquad \omega_\ell = \mathrm{softmax}_{\ell \in \mathcal{N}_t}\big(-\beta\, \mathbf{q}^\top \Lambda \mathbf{k}_\ell\big), \tag{6}$$

where $\mathbf{q} = (1,1)^\top$, $\mathbf{k}_\ell = (1 - C_\ell,\, U_\ell)^\top$, and $\Lambda = \mathrm{diag}(\lambda, 1 - \lambda)$. Here $\beta > 0$ controls attention sharpness and $\lambda \in [0,1]$ balances confidence against uncertainty. We finally output $(\alpha_t^{\text{final}}, e_t)$, which suppresses sporadic noise, delivers more robust detection, and preserves interpretability.

b) Online Memory Update: To ensure that each memory bank contains only reliable references, we update the normal ($\mathcal{M}_{\text{nor}}$) and anomalous ($\mathcal{M}_{\text{ano}}$) banks immediately after each evaluation. For the current sample v_t, we pack its visual feature $\mathbf{f}_t^v$ together with the corresponding uncertainty score U_t, confidence score C_t, and anomaly score α_t into a memory entry $b_t = (\mathbf{f}_t^v, U_t, C_t, \alpha_t)$ and insert b_t into exactly one bank:

$$\begin{aligned} \mathcal{M}_{\text{nor}} &\leftarrow \mathcal{M}_{\text{nor}} \cup \{b_t\} \quad \text{if } C_t > \tau_c \ \wedge\ \alpha_t \le \varepsilon_{\text{low}}, \\ \mathcal{M}_{\text{ano}} &\leftarrow \mathcal{M}_{\text{ano}} \cup \{b_t\} \quad \text{if } C_t > \tau_c \ \wedge\ \alpha_t \ge \varepsilon_{\text{high}}. \end{aligned} \tag{7}$$

Samples not satisfying either condition are not inserted.

To bound the storage footprint and suppress noise accumulation, we impose a per-bank capacity limit K_{G} on $\mathcal{M}_{\text{nor}}$ and $\mathcal{M}_{\text{ano}}$. Taking the normal bank as an example, whenever $|\mathcal{M}_{\text{nor}}| > K_{\text{G}}$ we evict the entry with the smallest priority score

$$p_k = C_k - \gamma U_k, \qquad m_{\text{drop}}^{\text{nor}} = \underset{m_k \in \mathcal{M}_{\text{nor}}}{\arg\min}\ p_k, \tag{8}$$

where $\gamma > 0$ balances confidence against uncertainty. The bank is then updated by removing the dropped entry, $\mathcal{M}_{\text{nor}} \leftarrow \mathcal{M}_{\text{nor}} \setminus \{m_{\text{drop}}^{\text{nor}}\}$. The anomalous bank is handled identically by replacing $\mathcal{M}_{\text{nor}}$ with $\mathcal{M}_{\text{ano}}$. Here, $\mathcal{M} \setminus \{x\}$ denotes removing element x from set $\mathcal{M}$. In this way, both memory banks remain compact, retaining only high-confidence, low-uncertainty entries that provide the most informative references for future retrieval.

Importantly, the memory bank is not intended to exhaustively cover all possible complex or rare anomalies. Instead, by retaining only high-confidence and low-uncertainty entries, it stores stable semantic references that preserve clear anomaly-relevant patterns. These references can provide transferable guidance for more complex observations, helping the model focus on key anomalous interactions or relations even when the current sample is noisy, crowded, partially occluded, or visually incomplete.

IV. Experiments

A. Experimental Settings

a) Datasets: We evaluate our method in a training-free, zero-shot setting, using **only** the RGB frames from each **test set** across three public benchmarks that cover diverse scenes, durations, and anomaly types. **XD-Violence (XD)** [2] is a large-scale dataset of 4,754 untrimmed web videos (217 h) from films and YouTube. Videos are annotated with weak video-level labels, where violence is further divided into six categories. **UCF-Crime (UCF)** [1] is a large-scale dataset of real-world surveillance videos, consisting of 1,900 clips (128 h) from diverse environments and spanning 13 crime categories. **ShanghaiTech (SHTech)** [43] is a dataset of campus surveillance videos captured by 13 fixed cameras under diverse lighting conditions and viewpoints. It contains 437 videos (18 h) with both normal activities and anomalies such as fighting, running, and throwing. The original dataset was introduced for one-class VAD, where only normal videos are used for training. For fair comparison with prior weakly supervised methods, we adopt the re-organized split in [44], which consists of 238 training videos and 199 testing videos.

b) Evaluation Metrics: Following prior work [11], we report results at the frame level. For XD-Violence, we follow the common practice of reporting both Average Precision (AP) and Area Under the ROC Curve (AUC), while for UCF-Crime and ShanghaiTech we report AUC. Compared to AUC, AP is often more suitable for XD-Violence since the dataset is highly imbalanced, and AP places greater emphasis on the positive (violent) class.

c) Implementation Details: We adopt InternVL3-8B [45] as the backbone VLM and use a frozen InternVL-14B-224px encoder [46] to embed both frames and category prompts. The category prompts are predefined according to the official anomaly categories of each dataset rather than learned dynamically. For each anomaly category, we use GPT-4o [47] to

TABLE I
CATEGORY PROMPT SETS USED FOR UNCERTAINTY ESTIMATION ON THE THREE DATASETS. EACH SET CONTAINS ONE PROMPT FOR EACH ANOMALY CATEGORY AND ONE ADDITIONAL NORMAL PROMPT.

Dataset	Prompt Set
XD-Violence ($m = 7$)	(1) A fight is happening; (2) Someone is shooting; (3) A riot is taking place; (4) A person is being abused; (5) A car accident is occurring; (6) An explosion is happening; (7) It looks like an ordinary moment.
UCF-Crime ($m = 14$)	(1) A person is being abused; (2) Someone is being arrested; (3) A building is on fire; (4) Someone is being assaulted; (5) A burglary is taking place; (6) An explosion is happening; (7) A fight is happening; (8) A road accident is occurring; (9) A robbery is in progress; (10) Someone is shooting; (11) A shoplifting incident is occurring; (12) Someone is stealing something; (13) An act of vandalism is happening; (14) It looks like an ordinary moment.
ShanghaiTech ($m = 10$)	(1) Someone riding a bike, scooter, or sliding device in a pedestrian area; (2) Someone running or jumping quickly in a pedestrian area; (3) Multiple people chasing or fighting in a pedestrian area; (4) Someone falling down; (5) Someone carrying or pushing something like a stroller or equipment in a pedestrian area; (6) Someone throwing something into the air in a pedestrian area; (7) A car or truck entering a pedestrian area; (8) Someone walking in the wrong direction or loitering in a pedestrian area; (9) A group of people gathering in a pedestrian area; (10) People walking or standing normally in a pedestrian area.

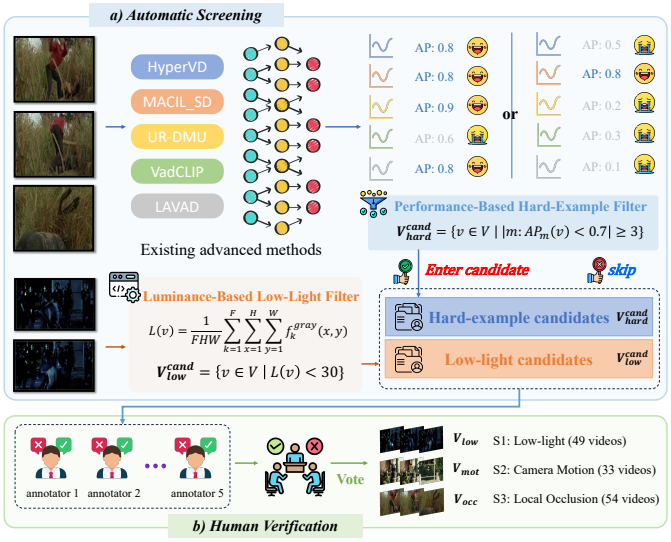


Fig. 4. Two-stage pipeline for identifying challenging cases in the XD-Violence [2] test split. In stage (a), automatic screening proposes low-light and other hard-example candidates, and in stage (b), five annotators verify and assign each video to the low-light, camera-motion, or local-occlusion subset.

rewrite the category name into a concise natural-language sentence while preserving its original semantics, and additionally include one normal prompt to represent the non-anomalous state. Therefore, the total number of prompts is $m = 7/14/10$ for XD-Violence/UCF-Crime/ShanghaiTech, respectively. The complete prompt sets used in our experiments are listed in Table I. Following LAVAD [11], we sample one of every 16 frames and split the stream into segments of η frames (η=4/3/5 for XD-Violence/UCF-Crime/ShanghaiTech). We downscale frames of XD-Violence to 448×448 to balance efficiency and fidelity, as it contains higher-resolution videos; UCF-Crime is already low-resolution and ShanghaiTech is smaller in scale, so both retain their native resolutions. Hierarchical retrieval is activated when segment entropy exceeds τ_u (0.40/0.55/0.60 for XD/UCF/SHTech). Segments with VLM confidence above τ_c=0.7 are stored, and a short-term pool retains the most recent K_S=10 segments. For each class, we retain up to κ references (1 for XD/UCF, 2 for SHTech). Memory maintenance employs insertion cut-offs ε_{low}=0.2 and $\varepsilon_{\text{high}}$=0.8. Evidence-aware temporal attention aggregates scores over W_R neighboring segments (8 for XD/UCF, 48 for SHTech). All experiments are run with seed 42 on an NVIDIA H800 NVL.

B. Robustness under Challenging Scenarios

1) Challenging subsets: To evaluate the robustness and generalization of our method, we curate three challenging subsets from the XD-Violence test set: Low-Light (49 videos), Camera-Motion (33 videos), and Local-Occlusion (54 videos). As shown in Figure 4, candidate videos are first identified through automatic filtering, and subsequently verified by five trained annotators to finalize each subset.

a) Automatic Screening: As illustrated in Figure 4(a), the automatic screening stage consists of two parts. (1) Performance-Based Hard-Example Filter. We evaluate five advanced methods [11], [22], [48]–[50] on every video and record each method's frame-level average precision, denoted $\text{AP}_m(v)$ for method index $m \in \{1, \ldots, 5\}$. A video is flagged as hard if at least three of the five methods score below 0.7: $V_{\text{hard}}^{\text{cand}} = \{ v \in V \mid |\{m : \text{AP}_m(v) < 0.7\}| \geq 3 \}$. (2) Luminance-Based Low-Light Filter. To efficiently estimate luminance, we sample frames with an adaptive stride depending on video length. Each sampled frame is converted to grayscale, and the average luminance of a video v is computed as $L(v) = \frac{1}{FHW} \sum_{k=1}^{F} \sum_{x=1}^{H} \sum_{y=1}^{W} f_k^{\text{gray}}(x, y)$, where F, H, and W denote the number of sampled frames, frame height, and frame width, respectively. Videos with $L(v) < 30$ are collected into the low-light candidate set $V_{\text{low}}^{\text{cand}} = \{ v \in V \mid L(v) < 30 \}$.

b) Human Verification: As shown in Figure 4(b), each candidate $v \in V_{\text{low}}^{\text{cand}} \cup V_{\text{hard}}^{\text{cand}}$ is independently assessed by five trained annotators. For each video, annotators vote on whether it belongs to one of the three challenging categories: low light, local occlusion, or camera motion. A video is assigned to a subset only if at least three annotators agree; otherwise, it is discarded. This procedure yields 49 low-light videos, 54 local-occlusion videos, and 33 camera-motion videos, which form the final challenging subsets V_{low}, V_{occ}, V_{mot}.

2) Performance in challenging scenarios: In Table II, we compare the proposed UGHR with several state-of-the-art weakly supervised methods [22], [48]–[53] and a training-free approach [11]. The evaluation covers three curated subsets—Low-Light (S1), Camera-Motion (S2), and Local-Occlusion (S3)—their union (Challenging), and the Standard split, which corresponds to the complementary portion of the XD-Violence test set after excluding the Challenging subset. **Scenario-wise analysis.** In the **low-light** scenario, UGHR performs strongly but trails HyperVD and MACIL_SD slightly, primarily because (1) both methods leverage the audio modality, which provides crucial cues when visual signals are severely degraded; and (2) they are fully trained on the dataset

TABLE II
AP (%) OF XD-VIOLENCE ON STANDARD AND CHALLENGING SPLITS FOR THREE SCENARIOS. OVERALL PERFORMANCE AGGREGATES ACROSS ALL SCENARIOS AND REPORTS THE DROP. WS: WEAKLY SUPERVISED; TF: TRAINING-FREE; ↑: HIGHER IS BETTER; ↓: LOWER IS BETTER.

Type	Method	Modality	S1: Low-light		S2: Camera Motion		S3: Local Occlusion		Overall Performance		
			well-lit ↑	low-light ↑	static ↑	motion ↑	none ↑	occlusion ↑	standard ↑	challenging ↑	drop ↓
WS	HyperVD [48]	A+V	85.26	**79.55**	86.10	<u>40.91</u>	86.39	34.52	87.31	<u>51.96</u> (2^{nd})	35.35
	MACIL_SD [49]	A+V	82.67	<u>75.31</u>	84.01	37.74	83.68	<u>40.61</u>	85.03	<u>50.93</u> (3^{rd})	34.10
	UR-DMU [50]	A+V	80.75	66.88	81.50	38.51	81.83	31.01	82.90	43.77	39.13
	PEL4VAD [51]	V	85.54	68.63	86.57	26.47	86.43	27.42	87.83	38.89	48.94
	BN-WVAD [52]	V	84.92	66.97	85.70	33.34	85.85	27.11	87.10	41.24	45.86
	VadCLIP [22]	V	84.86	52.16	85.45	30.81	85.39	26.00	86.73	38.95	47.78
	DSRL [53]	V	80.30	63.37	80.95	<u>38.80</u>	81.38	29.38	82.53	43.48	39.05
TF	LAVAD [11]	V	61.54	47.05	62.76	31.07	62.86	29.67	65.16	36.08	<u>29.08</u>
	Baseline	V	65.24	49.24	66.18	36.53	66.26	<u>35.05</u>	68.00	40.49	<u>27.51</u>
	UGHR (Ours)	V	77.46	<u>68.66</u>	78.15	**50.67**	78.28	**46.98**	79.20	**55.47** (1^{st})	**23.73**

and can thus learn dataset-specific statistics. By contrast, UGHR is entirely training-free and relies solely on the visual modality. Although its retrieval mechanism can introduce contextual cues, adjacent segments are also dark under low-light conditions and therefore offer limited utility, leaving UGHR at a mild disadvantage relative to multimodal, trained methods. In the **camera motion** and **local occlusion** scenarios, UGHR substantially outperforms all competing approaches. It not only recovers short-term visual evidence from nearby context, but also leverages high-confidence, semantically similar exemplars to guide inference when direct visual cues are unreliable or missing. Existing methods cannot effectively exploit such multi-level contextual information and therefore lag behind.

Overall analysis on the challenging split. On the aggregated Challenging split, weakly supervised methods generally outperform LAVAD, mainly because they can learn discriminative representations from abundant labeled data, whereas training-free approaches lack this advantage. Our baseline surpasses LAVAD by 4.41%, primarily benefiting from the supplementary information introduced by contextual retrieval. Nevertheless, two limitations remain: (i) ignoring scene uncertainty may cause indiscriminate retrieval and introduce distracting information; and (ii) unconstrained global retrieval can lead to semantic drift and inject noise. In contrast, UGHR achieves 55.47% AP on the Challenging split, exceeding the strongest weakly supervised competitor, HyperVD, by 3.51%. This superiority arises from its uncertainty-aware selective retrieval and hierarchical retrieval strategy, which not only avoids distracting retrievals, but also provides temporally and semantically alignable references for weak anomaly cues under occlusion, blur, or low visibility. As a result, the model is guided to focus more on anomaly-relevant local evidence, making anomaly-consistent interpretations more competitive than background distractions and thereby markedly improving robustness in challenging cases.

C. Cross-Dataset Generalization

In Table III, we systematically assess the cross-domain generalization of UGHR against several state-of-the-art weakly supervised methods [50]–[54] and a training-free approach [11]. Weakly supervised methods are trained on either XD-Violence or UCF-Crime, whereas training-free approaches require no target-domain data. Given the relatively small domain discrepancy between XD-Violence and UCF-Crime but the substantial shift to ShanghaiTech, we adopt the latter as a key evaluation benchmark to more faithfully reflect real-world cross-domain generalization.

TABLE III
CROSS-DATASET PERFORMANCE COMPARISON. WS: WEAKLY SUPERVISED; TF: TRAINING-FREE. TRAIN ⇓: ROWS FOR TRAINING DATASETS; TEST ⇒: COLUMNS FOR TESTING DATASETS.

Type	Method	Test ⇒ / Train ⇓	XD	UCF	SHTech
			AP (%)	AUC (%)	AUC (%)
WS	BN-WVAD [52]	XD	83.94	72.58	51.70
	PEL4VAD [51]	XD	84.64	75.32	41.50
	DSRL [53]	XD	77.19	67.32	54.92
	FedPrompt [54]	XD	74.23	77.35	54.62
	UR-DMU [50]	UCF	51.67	84.84	56.28
	PEL4VAD [51]	UCF	50.49	84.51	49.78
	FedPrompt [54]	UCF	55.91	84.03	<u>60.42</u>
TF	LAVAD [11]	–	62.01	80.28	59.80
	Baseline	–	64.58	75.57	<u>65.64</u>
	UGHR (Ours)	–	77.19	82.84	**86.37**

The results show that training-based methods excel on the source or closely related domains but degrade markedly under larger distribution shifts. For example, BN-WVAD attains 83.94% AP on XD-Violence, but its AUC drops to 72.58% on UCF-Crime and 51.70% on ShanghaiTech, indicating strong domain dependence. By contrast, training-free methods exhibit more stable performance across datasets, but LAVAD relies on a single-caption inference pipeline and lacks an active evidence-acquisition mechanism under information-sparse, challenging scenarios; meanwhile, our baseline's naive global retrieval tends to pull in weakly related references, amplifying cross-domain noise and undermining robustness.

In contrast, UGHR achieves leading performance on all three datasets (77.19% AP on XD; 82.84%/86.37% AUC on UCF/SHTech), notably surpassing all competitors on ShanghaiTech and demonstrating pronounced cross-domain generalization. We attribute this advantage to two factors: (i) the underlying VLM's rich pretraining confers strong open-set representations; and (ii) retrieving high-confidence easy exemplars as semantically aligned contextual anchors supplies

supportive evidence for harder cases, reinforcing weak cues and narrowing appearance gaps across datasets.

D. Plug-and-Play Enhancement of VLMs

Table IV presents the plug-and-play gains achieved by UGHR on five representative VLM backbones [45], [55]–[58], without requiring any weight updates. Specifically, on the Standard split, UGHR yields absolute AP improvements ranging from 20.41% to 35.79%. Under challenging conditions, it further provides gains of 10.01% to 29.14% over the respective backbones, effectively narrowing the gap to task-specific detectors. The consistent improvements across models of varying sizes, architectures, and pre-training objectives indicate that the benefits stem from UGHR itself rather than any particular backbone. This is because UGHR is designed for frozen, general-purpose VLMs, thereby avoiding the risk of overfitting from additional training. At the same time, it assesses scene uncertainty and hierarchically retrieves the most informative contextual cues, effectively strengthening weak visual signals and enabling general models to interpret anomalies more accurately.

TABLE IV
AP (%) OF FIVE VISION–LANGUAGE BACKBONES ON XD-VIOLENCE, WITH AND WITHOUT THE UGHR PLUG-IN.

Model	Standard			Challenging		
	w/o	w/	gain ↑	w/o	w/	gain ↑
DeepSeek-VL [55]	30.58	57.63	27.05	25.07	35.08	10.01
XComposer-2.5 [56]	45.33	69.65	24.32	27.17	37.26	10.09
Qwen2.5-VL [57]	46.26	66.67	20.41	27.13	39.01	11.88
ChemVLM [58]	45.11	73.37	28.26	29.45	44.91	15.46
InternVL3.0 [45]	43.41	79.20	**35.79**	26.33	55.47	**29.14**

E. Comparison with State-of-the-Art Methods

As shown in Table V, we compare UGHR against a broad spectrum of state-of-the-art approaches, including training-free methods [11], [12], [59], unsupervised methods [31]–[33], [60], [61], one-class methods [40], [62], and weakly supervised methods [22], [39], [51], [60], [63]–[65].

As a training-free approach, UGHR performs strongly against one-class and unsupervised baselines, achieving higher AUC across all datasets: on XD-Violence, it surpasses RareAnom by 23.48%; on UCF-Crime, it exceeds DyAnNet by 3.08%; and on ShanghaiTech, it outperforms GCL by 7.44% and further improves over the current best, AnomalyRuler, by 1.17%.

It is worth emphasizing that training-free video anomaly detection (VAD) is particularly challenging. Directly applying vision–language models (VLMs) to VAD—e.g., LLaVA-1.5—often degrades performance because VLMs tend to focus on foreground objects rather than actions or background cues that are critical for anomaly assessment, limiting generalization. LAVAD mitigates this via temporal scene descriptions, and EventVAD improves results with event segmentation and scoring, yet both still fall short of UGHR. Our method dynamically fuses temporal, scene, and semantic cues across multiple scales, assessing the abnormality of the current scene rather than relying on fixed-window context. Overall, UGHR sets a new training-free state of the art, particularly outperforming EventVAD on XD-Violence by 13.15%.

TABLE V
PERFORMANCE COMPARISON WITH FULLY-SUPERVISED (FS), WEAKLY-SUPERVISED (WS), ONE-CLASS (OC), UNSUPERVISED (US), AND TRAINING-FREE (TF) VAD METHODS ON THREE BENCHMARKS. * INDICATES OUR REPRODUCED RESULTS.

Type	Method	XD		UCF	SHTech
		AP	AUC	AUC	AUC
FS	Holmes-VAU [38] (CVPR'25)	87.68	–	88.96	–
WS	MSL [63] (AAAI'22)	78.58	–	85.62	97.32
	Cho *et al.* [64] (CVPR'23)	81.30	–	86.10	97.60
	OPVAD [65] (CVPR'24)	66.53	–	86.40	96.98
	VadCLIP [22] (AAAI'24)	84.51	–	88.02	97.49
	PEL4VAD [51] (TIP'24)	85.59	–	86.76	98.14
	VERA [39] (CVPR'25)	70.54	88.26	86.55	–
OC	BODS [62] (ICCV'19)	–	57.32	68.26	68.26
	GODS [62] (ICCV'19)	–	61.56	70.46	70.46
	AnomalyRuler [40] (ECCV'24)	–	–	–	85.20
	DiffVAD [60] (IJCAI'24)	–	–	77.10	–
US	GCL [31] (CVPR'22)	–	–	74.20	78.93
	RareAnom [32] (PR'23)	–	68.33	–	–
	DyAnNet [33] (WACV'23)	–	–	79.76	–
	DiffVAD [60] (IJCAI'24)	–	–	77.00	–
	FRD-UVAD [61] (TMM'24)	–	–	82.12	80.14
TF	LLaVA-1.5 [59] (CVPR'24)	50.26	79.62	72.84	–
	LAVAD [11] (CVPR'24)	62.01	85.36	80.28	59.80*
	EventVAD [12] (ACM MM'25)	64.04	87.51	82.03	–
	Baseline	64.58	89.85	75.57	65.64
	UGHR (Ours)	**77.19**	**91.81**	**82.84**	**86.37**

F. Ablation Studies

a) Effectiveness of each proposed module: As shown in Table VI, we conduct ablation studies on XD-Violence, UCF-Crime, and ShanghaiTech to assess the contributions of UGHR's three proposed components. Starting from a flat-retrieval baseline (Idx1: 64.58% AP on XD-Violence / 75.57% AUC on UCF-Crime / 65.64% AUC on ShanghaiTech), replacing it with the hierarchical context retriever (Idx2) yields improvements of 6.69%, 1.36%, and 5.97%, respectively, highlighting the value of near-to-far contextualization. Introducing the uncertainty-aware retrieval gate (Idx3), which skips retrieval for low-uncertainty segments, reduces noise and brings additional gains of 3.49%, 3.00%, and 10.66%. Finally, incorporating evidence-aware temporal attention (Idx4) suppresses score instability and lifts performance to 77.19% AP, 82.84% AUC, and 86.37% AUC, corresponding to overall improvements of 12.61%, 7.27%, and 20.73% over the baseline. These stepwise gains verify that hierarchical retrieval, adaptive gating, and temporal attention together transform a frozen VLM into a robust and generalizable anomaly detector.

b) Impact of hierarchical retrieval levels: As reported in Table VII, we conduct an ablation on XD-Violence to assess variants of the hierarchical context retriever. Using only short-term retrieval (Idx1)—which relies on the most temporally adjacent cues—the model attains 73.78% AP on the full test set, 75.64% on the standard subset, and 49.23% on

TABLE VI
ABLATION STUDY ON THREE BENCHMARKS. FR: FLAT RETRIEVAL; HR: HIERARCHICAL CONTEXT RETRIEVER; UG: UNCERTAINTY-AWARE RETRIEVAL GATE; EA: EVIDENCE-AWARE TEMPORAL ATTENTION.

Idx	Retrieval		Uncertainty		XD	UCF	SHTech
	FR	HR	UG	EA	AP (%)	AUC (%)	AUC (%)
1	✓	✗	✗	✗	64.58	75.57	65.64
2	✗	✓	✗	✗	71.27	76.93	71.61
3	✗	✓	✓	✗	74.76	79.93	82.27
4	✗	✓	✓	✓	**77.19**	**82.84**	**86.37**

the challenging subset. Augmenting with scene-level retrieval (Idx2) exposes broader context from earlier segments of the same video, improving AP by 1.22%, 1.65%, and 2.94% over Idx1. Combining short-term with global-level retrieval (Idx3) further introduces high-confidence semantic exemplars for recognizing novel or rare events, yielding gains of 1.68%, 2.15%, and 3.45% over Idx1. Finally, employing all three levels concurrently (Idx4) delivers the best results, with aggregate improvements of 3.41%, 3.56%, and 6.24% over Idx 1 on the full, standard, and challenging subsets, respectively. These results indicate that jointly leveraging precise temporal cues, scene-level context, and global semantic exemplars provides the most substantial gains, particularly under visually scarce scenarios.

TABLE VII
ABLATION OF SHORT-TERM, SCENE, AND GLOBAL RETRIEVAL LEVELS ON XD-VIOLENCE. FRAME-LEVEL AP (%) IS REPORTED FOR THE FULL TEST SET, THE STANDARD SPLIT, AND THE CHALLENGING SPLIT.

Idx	Retrieval Levels			AP (%)		
	Short	Scene	Global	Testing	Standard	Challenging
1	✓	✗	✗	73.78	75.64	49.23
2	✓	✓	✗	75.00	77.29	52.17
3	✓	✗	✓	75.46	77.79	52.68
4	✓	✓	✓	**77.19**	**79.20**	**55.47**

c) Comparison with alternative uncertainty measures: We further compare the Rényi-entropy-based uncertainty measure used in UG with several common retrieval-gating strategies. As shown in Table VIII, Rényi entropy achieves the best AP on both the full test set and the challenging subset, improving over the no-uncertainty variant by 2.73% and 3.85%, respectively. Although least-confidence and margin-sampling strategies remain competitive on the full test set, they are less effective on the challenging subset. This suggests that explicitly modeling the dispersion of the full prompt-matching distribution may provide a more informative signal for capturing multi-prompt competition and triggering retrieval when additional contextual support is likely beneficial.

d) Ablation on the Rényi order ρ: Table IX studies the effect of the Rényi order ρ in the uncertainty-aware retrieval gate. As ρ increases from 0.1 to around 1, the AP rises from 74.40% to 77.21%, and then remains stable within 77.14%-77.21% for $\rho \in [0.9, 1.3]$. This trend suggests that values near 1 appear to better preserve both dominant-prompt evidence and the competitive information carried by non-dominant prompts,

TABLE VIII
COMPARISON OF DIFFERENT UNCERTAINTY ESTIMATION STRATEGIES FOR RETRIEVAL GATING ON XD-VIOLENCE IN TERMS OF AP (%).

Uncertainty Strategy	Testing	Challenging
None	74.46	51.62
Random	75.62	53.60
Least Confidence [66]	76.76	54.35
Margin Uncertainty [67]	<u>77.17</u>	53.92
Rényi Entropy [15]	**77.19**	**55.47**

which is consistent with the goal of modeling competitive semantic uncertainty for retrieval triggering.

TABLE IX
ABLATION STUDY OF THE RÉNYI ORDER ρ IN UNCERTAINTY ESTIMATION ON XD-VIOLENCE.

ρ	0.1	0.3	0.5	0.7	0.9	0.99	1.01	1.1	1.3	1.5	1.7	1.9
AP (%)	74.40	75.53	76.64	77.11	77.19	**77.21**	<u>77.19</u>	77.14	77.19	76.90	76.96	76.98

e) Hyper-parameter Sensitivity: As shown in Figure 5, we assess UGHR's sensitivity on XD-Violence to three hyperparameters: the frame-window size (η), the number of reference exemplars (κ), and the uncertainty threshold (τ_u). Varying η from 1 to 5 changes the overall AP by only 0.53%—rising from 76.77% at η=2 to 77.30% at η=3. On the challenging subset, however, AP increases steadily with larger windows, from 51.46% at η=1 to a peak of 55.47% at η=4, followed by a slight decline at η=5, indicating that windows shorter than four frames discard crucial temporal cues whereas excessively long windows introduce redundancy. Adjusting κ has a limited effect on overall performance (76.23%-77.24%), but on the challenging subset AP rises from 55.47% at κ=1 to 56.61% at κ=2 before decreasing thereafter, suggesting that two references best balance semantic diversity and noise. Finally, sweeping τ_u over 0.35–0.55 yields a mild decrease in overall AP (77.24% to 76.92%), whereas the challenging subset attains its optimum at τ_u=0.40 (55.47%). Lower thresholds tend to over-trigger retrieval and amplify noise, while higher thresholds risk missing ambiguous segments.

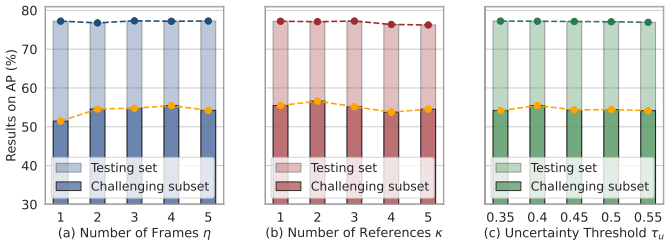


Fig. 5. Parameter ablation on XD-Violence. Semi-transparent bars denote AP (%) on the full test set; solid bars denote the challenging subset.

To further justify the dataset-specific parameter settings, we additionally conduct key hyper-parameter ablations on UCF-Crime and ShanghaiTech, as shown in Figure 6. Overall, the trends observed on these two datasets are largely consistent with those on XD-Violence, suggesting that UGHR is not overly sensitive to perturbations of key hyper-parameters

around the optimum and thus exhibits relatively good robustness. Specifically, UCF-Crime achieves the best performance at $(\eta, \kappa, \tau_u) = (3, 1, 0.55)$, suggesting that anomaly recognition in this dataset may often be supported by relatively short local temporal cues and a small number of high-confidence references. In contrast, the optimal configuration on ShanghaiTech is $(5, 2, 0.60)$, indicating that the subtler and more context-dependent anomalies in this dataset benefit from a slightly larger temporal window and more retrieved references. In addition, the sensitivity curves of τ_u on both datasets remain relatively smooth around the optimum, further suggesting that UGHR is relatively tolerant to the choice of the uncertainty threshold.

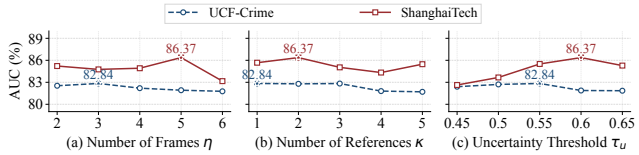


Fig. 6. Parameter ablation on UCF-Crime and ShanghaiTech. Results are reported as frame-level AUC (%) on the testing sets.

f) Impact of test video processing order: Since UGHR updates the memory banks online during inference, a natural question is whether changing the order of test videos would alter the final performance. To examine this, unless otherwise specified, we treat the default order of each test split as the standard evaluation protocol, and further conduct five additional experiments on XD-Violence by randomly shuffling the processing order of all test videos.

TABLE X
FRAME-LEVEL AP (%) ON XD-VIOLENCE UNDER THE DEFAULT VIDEO ORDER AND FIVE RANDOM SHUFFLED ORDERS. THE MAXIMUM FLUCTUATION IS ONLY 0.21%, INDICATING THAT THE ONLINE MEMORY UPDATE IS STABLE WITH RESPECT TO VIDEO ORDER.

Video Order	AP (%)
Default	77.19
Shuffle #1	77.23
Shuffle #2	77.06
Shuffle #3	77.27
Shuffle #4	77.20
Shuffle #5	77.21

As shown in Table X, the AP under the default order is 77.19%, and the maximum fluctuation across the five randomly shuffled orders is only 0.21%, with no obvious performance degradation associated with order changes. This stability mainly comes from two aspects. First, shuffling the test videos does not change the temporal context within each individual video, so the construction of short-term and scene-level memories remains largely consistent. Second, the order-affected global memory only serves as complementary evidence, and its candidate pool gradually becomes stable as more samples are processed. Therefore, although UGHR adopts online memory update, its inference process appears practically stable and not noticeably sensitive to the processing order of test videos.

G. Qualitative Analysis

a) Qualitative Analysis of Anomaly Detection Results: In Figure 7, we compare UGHR against four strong baselines over standard and challenging scenarios. Under standard conditions (row (a)), where anomalous cues are relatively salient, most methods detect anomalies reasonably well; however, their predictions still contain spurious peaks that do not align with the ground truth. HyperVD [48] and MACIL_SD [49] (weakly supervised) exhibit irregular oscillations and residual spikes; among training-free methods, LAVAD [11] degrades toward a near-noise trace, and the flat-retrieval baseline likewise produces false peaks. When visual evidence is scarce (row (b))—due to complex backgrounds, low illumination, or motion blur—these methods either miss anomalies or are corrupted by noise, often yielding isolated spikes or missing segments. In contrast, UGHR produces clean, rectangular pulses in both standard and challenging settings: the score rises sharply at true anomaly onset and falls promptly at offset, accurately delineating temporal boundaries even under limited visibility.

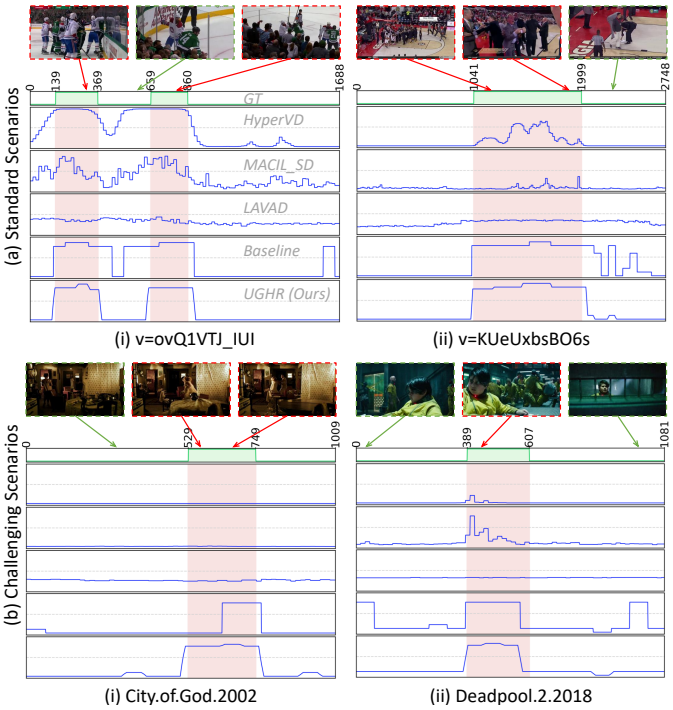


Fig. 7. Visualization of anomaly detection results on XD-Violence; The gray horizontal line at 0.5 indicates the anomaly detection threshold.

b) Qualitative Analysis of Contextual Support for Difficult Anomalies: Figure 8 further illustrates UGHR's contextual gains in two challenging scenarios and explains why contextual support is effective for difficult anomaly recognition. In the low-light indoor fight clip (a), scene-level retrieval first returns normal frames confirming that two people are typically present, while short-term retrieval surfaces abnormal frames from seconds earlier showing the initial clash. These cues help the model relate the weak and partially occluded evidence in the current segment to a more complete anomaly progression, while guiding attention toward anomaly-relevant

local interactions rather than background clutter, thereby contributing to correct recognition of the ongoing struggle. In the distant city explosion clip (b), short-term retrieval provides clear views of an aircraft in flight, while global-level retrieval brings in semantically similar explosion examples from other videos. Together, these cues help the model associate the smoke-obscured scene with anomaly-relevant event evolution and focus on the faint but critical evidence linked to the crash, supporting a more reliable inference of an aircraft crash and explosion. Across both cases, UGHR does not replace the current observation, but guides the model toward anomaly-relevant local evidence and suppresses distracting background variations when visual evidence is insufficient.

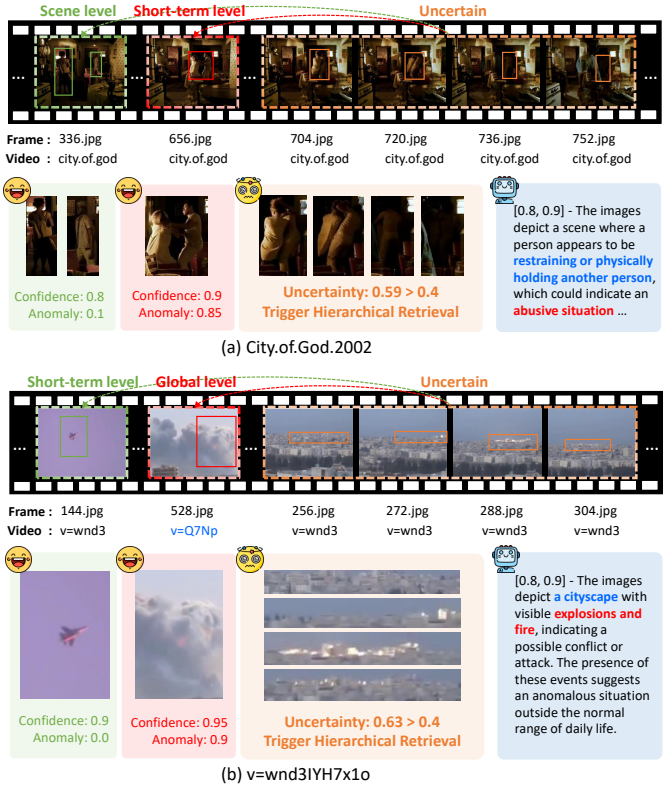


Fig. 8. Qualitative examples in challenging scenarios.

c) Qualitative Analysis of Generalization from Simple to Complex Anomalies: Figures 9 and 10 further illustrate why UGHR can generalize from high-confidence and relatively simple memory entries to more complex anomalies. Although the global memory bank mainly stores clear and reliable anomaly instances, these entries still preserve transferable anomaly-relevant patterns for noisy or visually complicated queries. In Figure 9, the query is challenged by crowding, small target scale, and the strong interference of a normal sports scene, making the local evidence easy to interpret as a regular game event. The retrieved global reference, while simpler and clearer, explicitly presents the core anomalous interaction of physical altercation, enabling the model to align the complex query with a stable anomaly cue rather than the dominant sports background.

This transferability is further supported by Figure 10, where the cross-image similarity concentrates on the players engaged

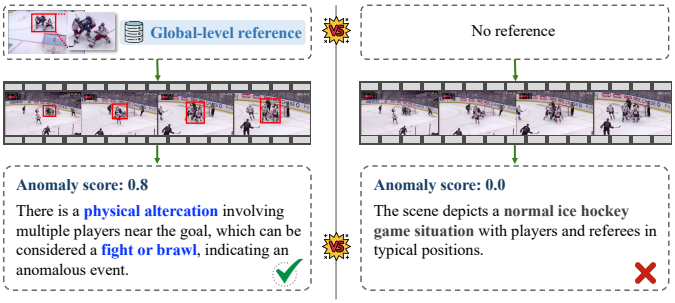


Fig. 9. Effect of global-level references on recognizing complex anomalies. Cross-video high-confidence references serve as semantic anchors for the query, allowing the model to focus on the underlying physical altercation rather than reducing the scene to a normal sports event.

in confrontation rather than on background texture or field appearance. This suggests that retrieval is driven by anomaly-discriminative interaction patterns instead of superficial visual similarity. Therefore, even when the current anomaly is more complex than the stored reference, UGHR may still leverage high-confidence memories to emphasize the key anomalous relations and contribute to more reliable generalization from simple to complex anomalies.

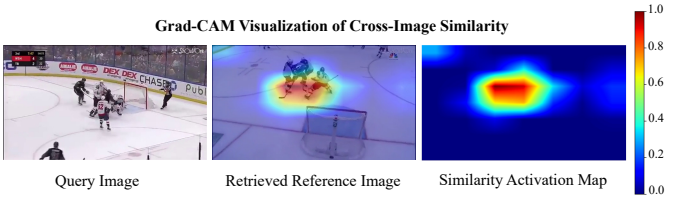
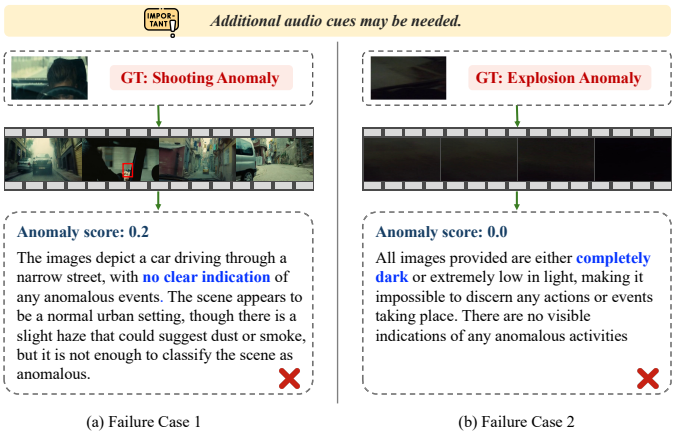


Fig. 10. Grad-CAM visualization of cross-image similarity between the query and retrieved global reference. Activations focus on physical confrontation rather than background regions, showing that retrieval relies on anomalous interactions instead of superficial appearance similarity.

H. Limitations and Future Work

a) Failure cases: Two representative failure cases are shown in Figure 11. In both examples, the current segment contains extremely weak visual evidence, making it difficult to reliably confirm the anomaly from RGB observations alone. In the first case, the visible cues are limited to partial occlusion, distant motion, and faint smoke, without clear evidence that can directly support a "shooting" anomaly. In the second case, the scene is almost entirely dark, leaving hardly any discriminative visual evidence for anomaly judgment.

b) Limitations: These cases clarify the main applicability boundary of UGHR. Although hierarchical retrieval may strengthen weak visual cues, it still relies on the existence of usable visual evidence in the current observation. When the decisive anomaly signal is intrinsically non-visual or is too severely degraded by darkness, occlusion, or restricted visibility, retrieval alone may only partially alleviate the difficulty. In such cases, reliable anomaly judgment often requires complementary evidence beyond RGB observations.



(a) Failure Case 1 (b) Failure Case 2

Fig. 11. Representative failure cases where visual evidence is extremely weak and complementary modalities such as audio would be needed for reliable judgment.

c) Future work: A promising future direction is to extend UGHR toward multimodal memory construction and retrieval, especially by incorporating complementary audio cues such as gunshots, explosions, or other event-related sounds. This could allow the retrieved context to provide not only visual references but also acoustic evidence when visual cues are insufficient, and may further improve robustness on extreme boundary cases while largely preserving the training-free and plug-and-play nature of the framework.

V. Conclusion

In this work, we propose UGHR, a training-free, plug-and-play framework designed to alleviate the pronounced performance degradation exhibited by existing methods in visually impoverished scenarios such as low-light or occluded scenes. To determine whether visual cues suffice for accurate anomaly detection, we design an uncertainty-aware retrieval gate that selectively triggers context retrieval only under high uncertainty. To supplement the most valuable contextual information, we design a hierarchical context retriever that gathers high-confidence examples at short-term, scene, and global levels, reinforcing weak visual cues.

Extensive experimental results validate the effectiveness of the components in our proposed framework and demonstrate that our method not only achieves new training-free state-of-the-art performance on three public benchmarks but also surpasses recent weakly supervised methods on the challenging subsets. A key limitation of UGHR is that it still depends on the presence of usable visual evidence in the query segment. A natural future direction is therefore to explore multimodal memory and retrieval, especially by incorporating audio cues, to better handle extremely challenging cases where visual evidence is severely degraded.

References

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.

[2] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 322–339.

[3] S. Li, J. Leng, C. Kuang, M. Tan, and X. Gao, "Video-level language-driven video-based visible-infrared person re-identification," *IEEE Transactions on Information Forensics and Security*, 2025.

[4] S. Li, J. Leng, J. Gan, M. Mo, and X. Gao, "Shape-centered representation learning for visible–infrared person re-identification," *Pattern Recognition*, vol. 167, p. 111756, 2025.

[5] X. Yu, N. Dong, L. Zhu, H. Peng, and D. Tao, "Clip-driven semantic discovery network for visible-infrared person re-identification," *IEEE Transactions on Multimedia*, 2025.

[6] M. Mo, X. Tong, M. Tan, J. Leng, J. Zheng, Y. Liu, H. Chen, J. Gan, W. Li, and X. Gao, "A2seek: Towards reasoning-centric benchmark for aerial anomaly understanding," *arXiv preprint arXiv:2505.21962*, 2025.

[7] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021.

[8] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall, "Dota: Unsupervised detection of traffic anomaly in driving videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 444–459, 2022.

[9] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4975–4986.

[10] J. Tang, H. Lu, R. Wu, X. Xu, K. Ma, C. Fang, B. Guo, J. Lu, Q. Chen, and Y. Chen, "Hawk: Learning to understand open-world video anomalies," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 139 751–139 785, 2024.

[11] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing large language models for training-free video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18 527–18 536.

[12] Y. Shao, H. He, S. Li, S. Chen, X. Long, F. Zeng, Y. Fan, M. Zhang, Z. Yan, A. Ma *et al.*, "Eventvad: Training-free event-aware video anomaly detection," *arXiv preprint arXiv:2504.13092*, 2025.

[13] J. G. Raaijmakers and R. M. Shiffrin, "Search of associative memory," *Psychol. Rev.*, vol. 88, no. 2, pp. 93–134, 1981.

[14] M. W. Howard and M. J. Kahana, "A distributed representation of temporal context," *J. Math. Psychol.*, vol. 46, no. 3, pp. 269–299, 2002.

[15] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. Probab.* Univ. California Press, 1961, pp. 547–562.

[16] S. Bai, Z. He, Y. Lei, W. Wu, C. Zhu, M. Sun, and J. Yan, "Traffic anomaly detection via perspective map based on spatial–temporal information matrix," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 117–124.

[17] G. Wang, X. Yuan, A. Zheng, H.-M. Hsu, and J.-N. Hwang, "Anomaly candidate identification and starting time estimation of vehicles from traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 382–390.

[18] G. Li, G. Cai, X. Zeng, and R. Zhao, "Scale-aware spatio-temporal relation learning for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 333–350.

[19] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8022–8031.

[20] H. Shi, L. Wang, S. Zhou, G. Hua, and W. Tang, "Abnormal ratios guided multi-phase self-training for weakly-supervised video anomaly detection," *IEEE Trans. Multimedia*, vol. 26, pp. 5575–5587, 2023.

[21] J. Chen, L. Li, L. Su, Z.-J. Zha, and Q. Huang, "Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18 319–18 329.

[22] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision–language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, 2024, pp. 6074–6082.

[23] C. Huang, W. Huang, Q. Jiang, W. Wang, J. Wen, and B. Zhang, "Multimodal evidential learning for open-world weakly-supervised video anomaly detection," *IEEE Trans. Multimedia*, 2025.

[24] J. Meng, H. Tian, G. Lin, J.-F. Hu, and W.-S. Zheng, "Audio-visual collaborative learning for weakly supervised video anomaly detection," *IEEE Trans. Multimedia*, 2025.

[25] S. Sun and X. Gong, "Hierarchical semantic contrast for scene-aware video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22 846–22 856.

[26] C. Yan, S. Zhang, Y. Liu, G. Pang, and W. Wang, "Feature prediction diffusion model for video anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5527–5537.

[27] C. Cao, Y. Lu, and Y. Zhang, "Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection," *IEEE Trans. Image Process.*, 2024.

[28] M. Zhang, J. Wang, Q. Qi, P. Ren, H. Sun, Z. Zhuang, H. Wang, L. Zhang, and J. Liao, "Video anomaly detection via progressive learning of multiple proxy tasks," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2024, pp. 4719–4728.

[29] M. Zhang, J. Wang, Q. Qi, H. Sun, Z. Zhuang, P. Ren, R. Ma, and J. Liao, "Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17 385–17 394.

[30] L. Zhu, L. Wang, A. Raj, T. Gedeon, and C. Chen, "Advancing video anomaly detection: A concise review and a new dataset," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 89 943–89 977, 2024.

[31] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14 744–14 754.

[32] K. V. Thakare, D. P. Dogra, H. Choi, H. Kim, and I.-J. Kim, "Rareanom: A benchmark video dataset for rare type anomalies," *Pattern Recognit.*, vol. 140, p. 109567, 2023.

[33] K. V. Thakare, Y. Raghuwanshi, D. P. Dogra, H. Choi, and I.-J. Kim, "Dyannet: A scene dynamicity guided self-trained video anomaly detection network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 5541–5550.

[34] A. O. Tur, N. Dall'Asen, C. Beyan, and E. Ricci, "Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations," in *Proc. Int. Conf. Image Anal. Process. (ICIAP)*, 2023, pp. 49–62.

[35] H. Shi, L. Wang, S. Zhou, G. Hua, and W. Tang, "Learning anomalies with normality prior for unsupervised video anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 163–180.

[36] H. Lv and Q. Sun, "Video anomaly detection and explanation via large language models," *arXiv preprint arXiv:2401.05702*, 2024.

[37] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang, "Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm," *arXiv preprint arXiv:2406.12235*, 2024.

[38] H. Zhang, X. Xu, X. Wang, J. Zuo, X. Huang, C. Gao, S. Zhang, L. Yu, and N. Sang, "Holmes-vau: Towards long-term video anomaly understanding at any granularity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 13 843–13 853.

[39] M. Ye, W. Liu, and P. He, "Vera: Explainable video anomaly detection via verbalized learning of vision–language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 8679–8688.

[40] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, "Follow the rules: Reasoning for video anomaly detection with large language models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 304–322.

[41] S. Gao, P. Yang, and L. Huang, "Suvad: Semantic understanding based video anomaly detection using mllm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[42] S. Ahn, Y. Jo, K. Lee, S. Kwon, I. Hong, and S. Park, "Anyanomaly: Zero-shot customizable video anomaly detection with lvlm," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2026, pp. 3026–3035.

[43] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 341–349.

[44] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1237–1246.

[45] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, Y. Duan, H. Tian, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[46] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual–linguistic tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24 185–24 198.

[47] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[48] X. Peng, H. Wen, Y. Luo, X. Zhou, K. Yu, P. Yang, and Z. Wu, "Learning weakly supervised audio–visual violence detection in hyperbolic space," *arXiv preprint arXiv:2305.18797*, 2023.

[49] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio–visual violence detection," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2022, pp. 6278–6287.

[50] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, 2023, pp. 3769–3777.

[51] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," *IEEE Trans. Image Process.*, 2024.

[52] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. T. Shen, "Batchnorm-based weakly supervised video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[53] J. Leng, Z. Wu, M. Tan, Y. Liu, J. Gan, H. Chen, and X. Gao, "Beyond euclidean: Dual-space representation learning for weakly supervised video violence detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024. [Online]. Available: https://openreview.net/forum?id=TbPv0qFnHO

[54] B. Wang, C. Huang, J. Wen, W. Wang, Y. Liu, and Y. Xu, "Federated weakly supervised video anomaly detection with multimodal prompt," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 20, 2025, pp. 21 017–21 025.

[55] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang *et al.*, "Deepseek-vl: Towards real-world vision–language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[56] P. Zhang, X. Dong, Y. Zang, Y. Cao, R. Qian, L. Chen, Q. Guo, H. Duan, B. Wang, L. Ouyang *et al.*, "Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output," *arXiv preprint arXiv:2407.03320*, 2024.

[57] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[58] J. Li, D. Zhang, X. Wang, Z. Hao, J. Lei, Q. Tan, C. Zhou, W. Liu, Y. Yang, X. Xiong *et al.*, "Chemvlm: Exploring the power of multimodal large language models in chemistry area," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 1, 2025, pp. 415–423.

[59] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26 296–26 306.

[60] M. Zhang, J. Wang, Q. Qi, P. Ren, H. Sun, Z. Zhuang, L. Zhang, and J. Liao, "Safeguarding sustainable cities: Unsupervised video anomaly detection through diffusion-based latent pattern learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024, pp. 7572–7580.

[61] C. Tao, C. Wang, S. Lin, S. Cai, D. Li, and J. Qian, "Feature reconstruction with disruption for unsupervised video anomaly detection," *IEEE Trans. Multimedia*, vol. 26, pp. 10 160–10 173, 2024.

[62] J. Wang and A. Cherian, "Gods: Generalized one-class discriminative subspaces for anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8201–8211.

[63] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 1395–1403.

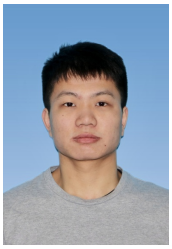
[64] M. Cho, M. Kim, S. Hwang, C. Park, K. Lee, and S. Lee, "Look around for anomalies: Weakly-supervised anomaly detection via context–motion relational learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12 137–12 146.

[65] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang, "Open-vocabulary video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18 297–18 307.

[66] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," *ACM SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

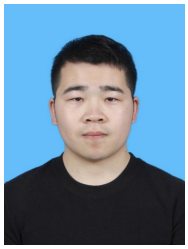
[67] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.*, 2001, pp. 309–318.

Jiankang Zheng received the B.E. degree in software engineering from North China Institute of Aerospace Engineering, China, in 2024. He is currently pursuing the M.E. degree in computer science and technology at Chongqing University of Posts and Telecommunications, China. His research interests include video anomaly understanding and vision-language reasoning models.



Mengjingcheng Mo received the B.E. and M.E. degrees from Chongqing University of Posts and Telecommunications, Chongqing, China, where he is currently pursuing the Ph.D. degree. His research interests include drone-view object detection, video anomaly understanding, and vision-language reasoning models.

Jiaxu Leng received the B.S. degree in electronic information science and technology from Tianshui Normal University, Gansu, China, in 2012, the M.S. degree in electrical engineering from the University of Electronic Science and Technology, Sichuan, China, in 2015, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2020. After obtaining his master's degree, he worked at Hisense for two years as an Algorithm Engineer. He is currently with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications. His current research interests include computer vision, deep learning, object detection, and stereo vision.



Mingpi Tan received the B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications. He is currently pursuing the Ph.D. degree in the School of Computer Science, Chongqing University of Posts and Telecommunications. His research interests include abnormal event analysis in videos.

Zhanjie Wu received the B.E. and M.E. degrees from Chongqing University of Posts and Telecommunications, China. He is currently pursuing the Ph.D. degree in electronic science and technology at Xidian University, China. His research interests include video anomaly analysis.

Ji Gan received the B.E. degree in software engineering from Huazhong University of Science and Technology, Wuhan, China, in 2015 and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include pattern recognition, human-computer interaction, and handwriting recognition and generation.

Haosheng Chen received the B.E. and M.E. degrees in computer science and technology and software engineering from Zhengzhou University, Zhengzhou, China, in 2014 and 2017, respectively, and the Ph.D. degree in computer science and technology from Xiamen University, Xiamen, China, in 2021. He is currently with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include pattern recognition, video analysis, and bio-inspired vision.

Xinbo Gao (Fellow, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a postdoctoral research fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 1999, he has been with the School of Electronic Engineering, Xidian University, where he is currently a Professor of Pattern Recognition and Intelligent Systems. Since 2020, he has also been a Professor of Computer Science and Technology with Chongqing University of Posts and Telecommunications. His current research interests include computer vision, machine learning, and pattern recognition. He has published seven books and around 300 technical articles in refereed journals and proceedings. Prof. Gao serves on the editorial boards of several journals, including Signal Processing and Neurocomputing. He has served as a General Chair or Co-Chair, Program Committee Chair or Co-Chair, or program committee member for around 30 major international conferences. He is a Fellow of the IEEE, IET, AAIA, CIE, CCF, and CAAI.