

# MemoVAD: Resource-Efficient Video Anomaly Detection via Dynamic Semantic Memory in Edge Computing Scenarios

Author Name

Affiliation

email@example.com

## Abstract

Deploying Video Anomaly Detection (VAD) in real-world surveillance faces a fundamental tension between the demand for high-level semantics to ensure effectiveness and the limited computational resources of edge devices. Vision-Language Models (VLMs) provide rich open-vocabulary semantics, but their latency and computational cost preclude on-device deployment. To address the challenge, we propose MemoVAD, an edge–cloud collaborative framework that selectively incorporates VLM semantics into streaming VAD. MemoVAD runs most inference on the edge with a lightweight detector and a causal Temporal Context Encoder (TCE) to model temporal dependencies. Specifically, we introduce an Uncertainty-Aware Gating (UAG) policy grounded in Subjective Logic to model perceived uncertainty and query the cloud-based VLM only for high-uncertainty and semantically novel clips. Besides, a Dynamic Semantic Memory (DSM) is designed to cache VLM-verified prototypes for efficient retrieval, enabling the edge model to progressively absorb VLM-level semantics via a semantic adapter. Experiments on UCF-Crime and XD-Violence datasets via a real edge device show that MemoVAD substantially reduces communication overhead while surpassing state-of-the-art performance. The demo video is available at: <https://memovad2026.github.io/>.

## 1 Introduction

The widespread deployment of surveillance cameras in public spaces has resulted in an explosive growth of video data, creating an urgent demand for automated Video Anomaly Detection (VAD) [Ullah *et al.*, 2022; Shathik, 2025]. The primary objective of VAD is to identify abnormal events, such as accidents or criminal activities, within long sequences of normal activities. While recent advancements in deep learning have significantly improved detection accuracy, deploying the models in real-world scenarios remains a significant challenge. The difficulty arises primarily from the inherent conflict between the need for sophisticated semantic reason-

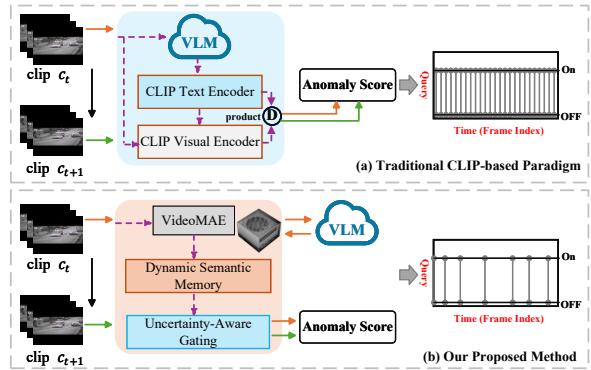


Figure 1: Comparison of different WSVAD paradigms. (a) Traditional methods suffer from high communication overhead due to per-clip VLM querying. (b) Our proposed framework reduces costs by querying the VLM only for uncertain and novel samples.

ing and the constrained computational resources available on edge devices [Ghasemi *et al.*, 2024; Guo *et al.*, 2025].

Traditional VAD approaches typically rely on reconstruction or prediction paradigms that operate under unsupervised or weakly supervised settings [Wu *et al.*, 2024b; Li *et al.*, 2025a]. The methods generally learn the distribution of normal patterns and treat deviations as anomalies. Although computationally efficient enough for some edge applications, they often struggle to distinguish between genuine anomalies and benign distributional shifts, such as dynamic background changes or camera jitter [Sultani *et al.*, 2018; Tian *et al.*, 2021]. Furthermore, the methods lack high-level semantic understanding, which restricts their ability to interpret complex scenes. Conversely, the emergence of Vision-Language Models (VLMs) has introduced a new paradigm where visual data is aligned with rich textual semantics through CLIP-based encoding [Deng *et al.*, 2025; Li *et al.*, 2025a]. VLMs demonstrate remarkable capability in zero-shot anomaly detection and semantic reasoning [Zanella *et al.*, 2024; Ye *et al.*, 2025; Shao *et al.*, 2025]. However, their massive parameter counts and heavy computational requirements render them unsuitable for direct deployment on resource-constrained edge hardware.

To leverage the semantic power of VLMs without overwhelming edge resources, edge-cloud collaboration offers

66 a promising direction [Zhang *et al.*, 2025; Sangwan *et al.*,  
67 2025]. A naive solution is to transmit all video data to a  
68 cloud-based VLM for processing [Kumar *et al.*, 2024; Wang  
69 et al., 2025; Sharshar *et al.*, 2025]. Nevertheless, it incurs  
70 huge bandwidth costs and high latency, which are unacceptable  
71 for real-time surveillance applications [Khan *et al.*, 2022;  
72 Mondal *et al.*, 2024; Li *et al.*, 2025b]. Consequently, a critical  
73 research gap exists in developing a framework that can maintain  
74 the low latency of edge processing while selectively integrating  
75 the high-level semantic capabilities of cloud-resident large models.  
76

To bridge the gap, we propose MemoVAD, a resource-efficient framework for video anomaly detection via dynamic semantic memory in edge computing scenarios. As shown in Fig. 1, anomalies typically exhibit recurring semantic patterns and manifest as temporally continuous segments. Therefore, continuous reliance on a heavy VLM is unnecessary. Instead, we introduce a novel collaborative mechanism where a lightweight edge detector handles the majority of the inference workload and queries the remote VLM only when necessary. Specifically, MemoVAD introduces three components to balance semantic capability and edge efficiency. First, it adopts a lightweight detector that uses a frozen Video-MAE [Tong *et al.*, 2022] backbone for motion feature extraction, together with a casual Temporal Context Encoder (TCE) to model local temporal dependencies. Second, we propose an Uncertainty-Aware Gating (UAG) mechanism grounded in Subjective Logic [Sensoy *et al.*, 2018], which estimates perceived uncertainty and triggers a VLM query when the edge model lacks sufficient evidence for a reliable decision. Finally, we develop a Dynamic Semantic Memory (DSM) that caches VLM-verified semantic prototypes on device. For semantically similar events, the model retrieves relevant knowledge from the memory instead of contacting the remote server, enabling efficient online knowledge distillation and progressively transferring VLM-level semantics to the edge model.

In summary, our main contributions are as follows:

- We propose MemoVAD, a resource-efficient collaborative framework that harmonizes the speed of edge computing with the semantic reasoning of large VLMs.
- We introduce an uncertainty-aware subjective-logic gate to query VLMs only under high perceived uncertainty, and a dynamic semantic memory caching VLM insights to cut communication overhead while preserving accuracy.
- Our MemoVAD demonstrates remarkable improvements compared to various benchmarks across two publicly datasets, validating its superior performance.

## 115 2 Related Works

### 116 2.1 Weakly Supervised Video Anomaly Detection 117 with VLMs

118 Due to the huge cost of frame-level annotations, current re-  
119 search predominantly focuses on Weakly Supervised Video  
120 Anomaly Detection (WSVAD), where only video-level la-  
121 bels are available [Mishra *et al.*, 2024; Abdalla *et al.*, 2025;

Liu *et al.*, 2024; Wu *et al.*, 2024a; Karim *et al.*, 2024;  
122 Zhang *et al.*, 2023; Pu *et al.*, 2024]. Multiple Instance Learn-  
123 ing (MIL) serves as the dominant framework in the domain,  
124 treating videos as bags of clips to distinguish anomalous bags  
125 from normal ones [Zhang *et al.*, 2022; Tang *et al.*, 2023;  
126 Fang *et al.*, 2024]. However, standard MIL-based detec-  
127 tors typically rely on discriminative feature embeddings that  
128 lack explicit semantic interpretability. To mitigate the is-  
129 sue, recent studies have integrated Vision-Language Mod-  
130 els (VLMs), leveraging the zero-shot capabilities of mod-  
131 els like CLIP [Radford *et al.*, 2021] and GPT-4V to iden-  
132 tify anomalies via textual prompting [Wu *et al.*, 2024b;  
133 Ye *et al.*, 2025; Yang *et al.*, 2024; Zanella *et al.*, 2024;  
134 Shao *et al.*, 2025]. While VLMs demonstrate superior per-  
135 formance in recognizing semantically complex events, their  
136 substantial computational overhead poses a significant bar-  
137 rier to real-time deployment. Our work builds upon the effi-  
138 cient MIL formulation but enhances it by integrating explicit  
139 semantic knowledge. Specifically, to mitigate the computa-  
140 tional cost of VLMs, we adopt an online knowledge distilla-  
141 tion strategy that continuously updates the edge model via a  
142 dynamic memory mechanism.

## 144 2.2 Efficient Edge-Cloud Collaboration

Deploying deep learning models on edge devices requires  
145 careful optimization to balance accuracy and efficiency [Guo  
146 et al., 2025; Ghasemi *et al.*, 2024]. Common techniques  
147 include model compression methods such as quantization,  
148 pruning, and lightweight architecture design. While the tech-  
149 niques reduce inference latency, they often degrade model  
150 capacity and performance. Edge-cloud collaborative intel-  
151 ligence seeks to mitigate the by offloading heavy computa-  
152 tion to the cloud [Zhang *et al.*, 2025; Sangwan *et al.*, 2025;  
153 Shathik, 2025]. Traditional offloading strategies decide  
154 which parts of a model to execute locally and which to trans-  
155 mit based on bandwidth and battery constraints. However,  
156 most existing collaborative frameworks focus on partition  
157 points within a fixed network architecture rather than dy-  
158 namic interaction based on sample difficulty. Our approach  
159 differs by employing an uncertainty-driven policy that dy-  
160 mically determines the necessity of cloud interaction. By  
161 utilizing Subjective Logic to quantify evidence sufficiency,  
162 MemoVAD ensures that communication resources are ex-  
163 pended only on hard and novel samples, thereby achieving  
164 a superior balance between resource efficiency and detection  
165 performance.

## 166 3 Methodology

### 167 3.1 Problem Definition

We formulate WSVAD as a Multiple Instance Learning  
168 (MIL) task. Let  $\mathcal{X} = \{(\mathcal{V}_i, Y_i)\}_{i=1}^{|\mathcal{X}|}$  denote the training set,  
169 where  $\mathcal{V}_i$  is an untrimmed video and  $Y_i \in \{0, 1\}$  is the video-  
170 level label.  $Y_i = 0$  indicates a normal video, while  $Y_i = 1$   
171 implies the video contains at least one anomalous event. Each  
172 video  $\mathcal{V}_i$  is divided into a sequence of  $T$  non-overlapping  
173 clips  $\{c_1, c_2, \dots, c_T\}$ .

The objective is to learn a clip-level anomaly scoring func-  
174 tion  $f_\theta(c_t) \rightarrow y_t \in [0, 1]$ , such that  $y_t$  is high for anomalous  
175

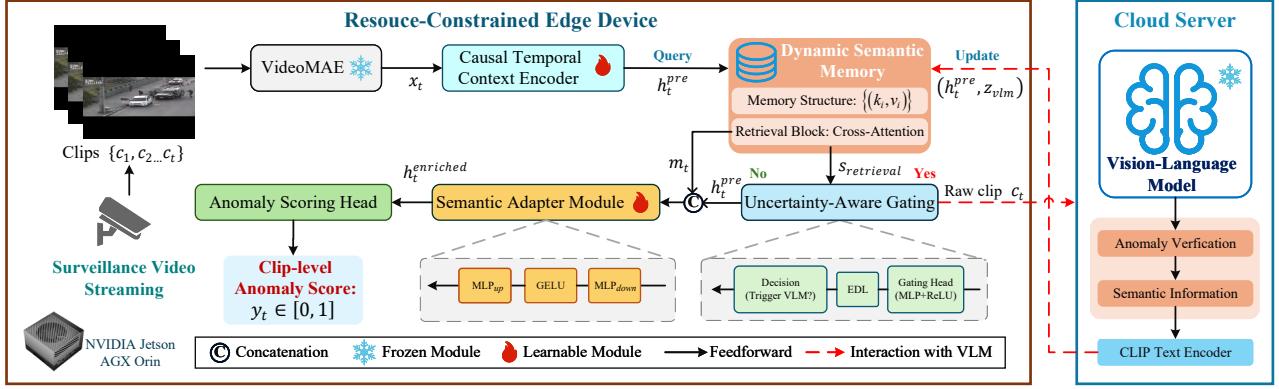


Figure 2: The architecture of the MemoVAD system.

178 clips and low for normal ones. Under the MIL assumption,  
 179 the relationship between clip scores and the video label is de-  
 180 fined as:

$$Y_i = \max_{t=1}^T y_t. \quad (1)$$

181 Specifically, for a normal bag  $Y_i = 0$ , all clips are normal  
 182  $y_t \approx 0$ ; for an anomalous bag  $Y_i = 1$ , at least one clip is  
 183 anomalous  $y_t \approx 1$ .

184 Unlike traditional centralized VAD, we consider an edge-  
 185 cloud collaborative scenario. The local edge device has lim-  
 186 ited computational resources and must minimize communica-  
 187 tion with the remote VLM. The goal is to maximize detection  
 188 performance while adhering to a strict communication bud-  
 189 get:

$$\min \sum_{t=1}^T \mathbb{I}(\text{Query}_t) \quad \text{s.t. Accuracy} \geq \delta, \quad (2)$$

190 where  $\mathbb{I}(\cdot)$  is the indicator function for triggering a remote  
 191 VLM query, and  $\delta$  is the desired performance threshold.

## 192 3.2 Overview

193 We propose MemoVAD, a resource-efficient framework de-  
 194 signed to bridge the gap between lightweight edge detection  
 195 and semantic reasoning. As illustrated in Fig. 2, the frame-  
 196 work operates under an edge-cloud collaborative paradigm to  
 197 identify anomalies from the input video clips defined in Sec.  
 198 3.1. MemoVAD comprises three key components:

- 199 1. **Edge-Resident Detector:** A lightweight network utilizing  
 200 a frozen VideoMAE backbone for initial feature ex-  
 201 traction and temporal modeling.
- 202 2. **Uncertainty-Aware Gating (UAG):** A policy module  
 203 that determines whether to query the VLM based on per-  
 204 ceived uncertainty.
- 205 3. **Dynamic Semantic Memory (DSM):** A continuously  
 206 updating memory bank that stores VLM-verified proto-  
 207 types, enabling the edge model to retrieve high-level  
 208 semantic knowledge without recurring communication  
 209 costs.

## 210 3.3 Edge-Resident Network

### 211 Feature Extraction

212 To guarantee real-time inference on resource-constrained  
 213 edge devices, we adopt VideoMAE-Small [Tong *et al.*, 2022]  
 214 as our visual backbone. We freeze the pre-trained weights to  
 215 preserve its generalizable motion features and prevent cata-  
 216 strophic forgetting during domain adaptation. Given an input  
 217 video clip  $c_t \in \mathbb{R}^{C \times F \times H \times W}$ , the backbone extracts a com-  
 218 pact feature vector  $x_t$ :

$$x_t = \mathcal{F}_{\text{backbone}}(c_t) \in \mathbb{R}^{D_{\text{stu}}}, \quad (3)$$

219 where  $D_{\text{stu}}$  denotes the feature embedding dimension. For  
 220 the input preprocessing, we resize frames to a standard spa-  
 221 tial resolution of  $H \times W$  and sample  $F$  frames with a fixed  
 222 temporal stride  $\tau$ .

### 223 Causal Temporal Context Encoder

224 Since the backbone processes clips independently, the ex-  
 225 tracted feature  $x_t$  lacks temporal context required for detect-  
 226 ing complex anomalies. To bridge the gap, we introduce a  
 227 lightweight Temporal Context Encoder (TCE) implemented  
 228 as a 2-layer causal Transformer Encoder, which aggregates  
 229 information from a fixed-length history window of  $L$  clips.  
 230 To retain sequential order information, learnable positional  
 231 embeddings are added to the input sequence before encoding:

$$h_t^{\text{pre}} = \text{TCE}_{\text{causal}}([x_{t-L+1}, \dots, x_t] + P), \quad (4)$$

232 where  $P$  denotes learnable positional embeddings. The out-  
 233 put  $h_t^{\text{pre}} \in \mathbb{R}^{D_{\text{stu}}}$  is the pre-enrichment contextual student  
 234 feature for clip  $t$ . In deployment, we implement the causal  
 235 Transformer with KV-cache for streaming inference, enabling  
 236 amortized constant-time updates per incoming clip.

## 237 3.4 Dynamic Semantic Memory (DSM)

238 DSM is designed to realize *online knowledge distillation* by  
 239 caching VLM reasoning results and enabling semantic re-  
 240 trieval on the edge.

### 241 Memory Structure

242 We structure the memory as a set of key-value pairs  $\mathcal{M} =$   
 243  $\{(k_i, v_i)\}_{i=1}^N$  with memory size  $N$ .

- **Key**  $k_i \in \mathbb{R}^{D_{\text{stu}}}$  stores the pre-enrichment student feature  $h_t^{\text{pre}}$  for a VLM-confirmed event. It ensures that keys share the same distribution as incoming queries.
- **Value**  $v_i \in \mathbb{R}^{D_{\text{vlm}}}$  stores the corresponding semantic embedding produced by the teacher VLM, serving as high-level knowledge to be distilled.

## Memory Retrieval

For each incoming clip feature  $h_t^{\text{pre}}$ , we retrieve relevant semantic information from  $\mathcal{M}$  using multi-head cross-attention. The query is derived from the student feature, while the memory keys and values serve as  $K$  and  $V$ :

$$Q = h_t^{\text{pre}} W_Q, \quad K = \mathcal{M}_{\text{key}} W_K, \quad V = \mathcal{M}_{\text{val}} W_V. \quad (5)$$

The retrieved semantic information  $m_t$  is computed via the attention mechanism:

$$m_t = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (6)$$

Simultaneously, to quantify retrieval confidence for gating, we compute the maximum cosine similarity between  $h_t^{\text{pre}}$  and stored keys:

$$s_{\text{retrieval}} = \max_{i \in \{1, \dots, N\}} \left( \frac{h_t^{\text{pre}} \cdot k_i}{\|h_t^{\text{pre}}\|_2 \|k_i\|_2} \right) \in [-1, 1]. \quad (7)$$

The metric  $s_{\text{retrieval}}$  measures how well the current input matches the known anomalous prototypes in the memory.

## Semantic Adapter Module (SAM)

Directly fusing semantic prototypes into the student space may cause distribution mismatch. We introduce a lightweight Semantic Adapter Module (SAM) to align and fuse  $m_t$  into the student feature space.

We first project the retrieved semantic prototype to the student dimension via a lightweight projection  $\phi : \mathbb{R}^{D_{\text{vlm}}} \rightarrow \mathbb{R}^{D_{\text{stu}}}$  (e.g., a single linear layer). Then we concatenate and pass through a bottleneck adapter:

$$\Delta h_t = \text{MLP}_{\text{up}} \left( \sigma \left( \text{MLP}_{\text{down}} \left( \text{Concat} [h_t^{\text{pre}}, \phi(m_t)] \right) \right) \right), \quad (8)$$

where  $\text{MLP}_{\text{down}}$  compresses by ratio  $r = 4$ ,  $\sigma$  is GELU, and  $\text{MLP}_{\text{up}}$  restores to  $D_{\text{stu}}$ . The enriched student feature is obtained by a gated residual connection:

$$h_t^{\text{enriched}} = \text{LayerNorm} (h_t^{\text{pre}} + \alpha \cdot \Delta h_t), \quad (9)$$

where  $\alpha$  is a learnable scalar initialized to 0 to prevent negative transfer in early training.

## 3.5 Uncertainty-Aware Gating (UAG)

Continuous VLM querying is extremely expensive. UAG requests VLM assistance only for *hard or novel* samples, using perceived uncertainty and memory confidence.

### Gating Policy

Softmax confidence cannot reliably separate aleatoric and perceived uncertainty. We adopt Subjective Logic [Sensoy *et al.*, 2018] to estimate perceived uncertainty via evidential learning.

Given the pre-enrichment feature  $h_t^{\text{pre}}$ , the gating head predicts non-negative evidence:

$$\mathbf{e}_t = \text{ReLU} (\text{MLP} (h_t^{\text{pre}})), \quad \mathbf{e}_t = [e_{t,0}, e_{t,1}], \quad (10)$$

for the Normal and Anomalous classes. Evidence defines a Dirichlet distribution  $\text{Dir}(\alpha_t)$ :

$$\alpha_t = \mathbf{e}_t + \mathbf{1}, \quad S_t = \sum_k \alpha_{t,k}, \quad (11)$$

where  $S_t$  represents evidence strength. The perceived uncertainty is defined as inverse total evidence:

$$u_t = \frac{K}{S_t} = \frac{2}{\alpha_{t,0} + \alpha_{t,1}}, \quad (K = 2). \quad (12)$$

To improve robustness under domain shift where models can be over-confident yet wrong, we trigger VLM queries when the case is uncertain and novel. The triggering condition is:

$$\text{Trigger}_t = \mathbb{I}(u_t > \tau_{\text{unc}} \wedge s_{\text{retrieval}} < \tau_{\text{sim}}), \quad (13)$$

where  $\tau_{\text{unc}}, \tau_{\text{sim}}$  are hyper-parameters trading off accuracy and communication cost.

## VLM Inference and Memory Update

When triggered, the raw clip  $c_t$  is transmitted to the remote VLM. We uniformly sample  $F$  frames from  $c_t$  and provide a lightweight prompt for anomaly verification. The VLM outputs: (1) Anomaly Verification, deciding whether the clip truly contains an anomaly; and (2) Semantic Extraction, producing a text-aligned embedding  $z_{\text{vlm}} \in \mathbb{R}^{D_{\text{vlm}}}$ .

If the VLM confirms an anomaly, we write the new knowledge into memory:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{(h_t^{\text{pre}}, z_{\text{vlm}})\}. \quad (14)$$

**Online Semantic-Diversity Replacement.** To keep a fixed memory budget  $N_{\text{max}}$  while avoiding quadratic recomputation, we maintain a redundancy score for each key:

$$\rho_i = \max_{j \neq i} \cos(k_i, k_j). \quad (15)$$

When inserting a new key  $k_{\text{new}}$ , we compute  $\cos(k_{\text{new}}, k_i)$  for all stored keys and update affected  $\rho_i$  incrementally. If the memory is full, we remove the most redundant key  $\arg \max_i \rho_i$ , yielding an online replacement with amortized  $O(ND_{\text{stu}})$  per insertion.

## 3.6 Objective Function

The model is trained with weak video-level labels. Let  $S(\cdot)$  denote the anomaly classifier head followed by a Sigmoid. The predicted anomaly score for the  $t$ -th clip is:

$$y_t = S(h_t^{\text{enriched}}) \in [0, 1]. \quad (16)$$

### MIL Ranking Loss

We formulate weakly supervised VAD as a Multiple-Instance Learning task. Normal videos are negative bags  $\mathcal{B}_n$  and anomalous videos are positive bags  $\mathcal{B}_a$ . Using the standard MIL assumption, the video-level score is represented by the maximum clip score. The ranking loss is:

$$\mathcal{L}_{\text{mil}} = \max \left( 0, m - \max_{t \in \mathcal{B}_a} y_t + \max_{t \in \mathcal{B}_n} y_t \right), \quad (17)$$

where  $m \in (0, 1)$  is the margin hyperparameter.

## 325 Semantic Distillation Loss

326 To distill semantic knowledge from the VLM, we minimize  
 327 the distance between student and teacher representations. We  
 328 introduce a projection head  $\psi : \mathbb{R}^{D_{\text{stu}}} \rightarrow \mathbb{R}^{D_{\text{vlm}}}$  (two-layer  
 329 MLP) and apply supervision only when teacher signals are  
 330 available. The distillation loss is:

$$\mathcal{L}_{\text{distill}} = \frac{1}{\sum_{t=1}^T \mathbb{M}_t + \epsilon} \sum_{t=1}^T \mathbb{M}_t \cdot \|\psi(h_t^{\text{enriched}}) - \text{stop\_grad}(z_{\text{vlm}})\|_2^2, \quad (18)$$

331 where  $\epsilon$  is a small constant for numerical stability. Here,  
 332  $\mathbb{M}_t = 1$  if the VLM is triggered for clip  $t$ , and 0 otherwise.  
 333 The stop-gradient operator treats the VLM as a fixed semantic  
 334 anchor.

## 335 Temporal Smoothness Loss

336 We impose temporal continuity and sparsity priors:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=1}^{T-1} (y_t - y_{t+1})^2 + \lambda_{sp} \sum_{t=1}^T y_t, \quad (19)$$

337 where the first term enforces smoothness, and the second en-  
 338 courages sparsity of anomaly predictions.  $\lambda_{sp}$  is a sparsity  
 339 hyper-parameter that balances the two constraints. It prevents  
 340 the trivial solution where the model predicts high scores for  
 341 all clips.

## 342 Gating Head Supervision

343 We supervise the evidential gating head following Evidential  
 344 Deep Learning [Sensoy *et al.*, 2018]. Let  $Y_{\text{bag}} \in \{0, 1\}$  be the  
 345 video-level label. To reduce label noise under MIL, we apply  
 346 the loss only to the top-scoring clip(s) in each bag:

$$\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{EDL}}(\alpha_t, Y_{\text{bag}}). \quad (20)$$

## 347 Total Loss

348 The final objective function is a weighted sum of the compo-  
 349 nents:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mil}} + \lambda_1 \mathcal{L}_{\text{distill}} + \lambda_2 \mathcal{L}_{\text{smooth}} + \lambda_3 \mathcal{L}_{\text{gate}}, \quad (21)$$

350 where  $\lambda_1, \lambda_2$ , and  $\lambda_3$  are pre-defined hyperparameters that  
 351 balance the trade-off between anomaly detection, semantic  
 352 alignment, and uncertainty estimation.

## 353 4 Experiments

### 354 4.1 Experimental Settings

355 **Datasets.** We evaluate MemoVAD on two large-scale  
 356 weakly supervised video anomaly detection benchmarks. **(1)**  
 357 **UCF-Crime** [Sultani *et al.*, 2018] contains 1,900 real-world  
 358 surveillance videos spanning 13 anomaly categories and nor-  
 359 mal activities. **(2) XD-Violence** [Wu *et al.*, 2020] consists  
 360 of 4,754 videos collected from movies and games, providing  
 361 audio-visual signals. In our experiments, we use only the vi-  
 362 sual modality.

363 **Metrics.** Following standard protocols, we employ the Area  
 364 Under the Receiver Operating Characteristic Curve (**AUC**)

Method	Ref.	Backbone	AUC (%)↑	AP (%)↑
Sultani et al.	CVPR'18	C3D	77.92	73.20
RTFM	ICCV'21	I3D	84.30	77.81
CRFD	TIP'21	I3D	84.89	75.90
MSL	AAAI'22	I3D	85.62	78.58
MGFN	AAAI'23	I3D	86.67	80.11
UR-DMU	AAAI'23	I3D	86.97	81.66
CLIP-TSA	ICIP'23	CLIP	87.58	82.17
TPWNG	CVPR'24	CLIP	87.79	83.68
VadCLIP	AAAI'24	CLIP	88.02	84.51
OVVAD	CVPR'25	CLIP	86.40	69.31
EventVAD	MM'25	CLIP	87.51	64.04
<b>MemoVAD (Ours)</b>	-	VideoMAE-S	<b>89.45</b>	<b>85.97</b>

Table 1: Comparison results on UCF-Crime (AUC) and XD-Violence (AP). Best results are highlighted in **bold**.

Method	FPS ↑	Latency (s) ↓	CR <sub>UCF</sub> (%) ↓	CR <sub>XD</sub> (%) ↓
I3D baseline	23.7	0.868	0.00	0.00
CLIP-based baseline	11.5	1.512	100.0	100.0
<b>MemoVAD (Ours)</b>	<b>36.2</b>	<b>0.475</b>	<b>8.63</b>	<b>15.72</b>

Table 2: Efficiency on Jetson AGX Orin (streaming, batch=1). FPS and Latency are measured end-to-end. CR: Percentage of clips querying the remote VLM.

for UCF-Crime and Average Precision (**AP**) for XD-Violence to evaluate detection performance. In addition to detection performance, we strictly evaluate resource efficiency in edge scenarios. Specifically, we report Throughput (**FPS**) and Latency (s) to verify real-time capabilities, and define Communication Rate (**CR**) as the percentage of clips that trigger a VLM query.

**Baselines.** For comparison, Sultani et al. [Sultani *et al.*, 2018], RTFM [Tian *et al.*, 2021], CRFD [Wu and Liu, 2021], MSL [Li *et al.*, 2022], MGFN [Chen *et al.*, 2023], UR-DMU [Zhou *et al.*, 2023], CLIP-TSA [Joo *et al.*, 2023], TPWNG [Yang *et al.*, 2024], VadCLIP [Wu *et al.*, 2024b], OVVAD [Li *et al.*, 2025a], and EventVAD [Shao *et al.*, 2025] are chosen as baselines.

The dataset distributions, implementation details and VLM prompts are presented in Appendix 1 and 2, respectively.

## 381 4.2 Comparison with State-of-the-Art Methods

**Main Results.** As presented in Table 1, MemoVAD achieves 382 89.45% AUC on UCF-Crime and 85.97% AP on XD- 383 Violence, consistently outperforming prior approaches with 384 C3D/I3D or CLIP-based feature backbones. Such 385 consistent gains across two widely-used benchmarks indicate 386 the strong effectiveness and robustness of our method under 387 different anomaly categories and evaluation metrics. In 388 particular, MemoVAD improves the best competitors TPWNG 389 and VadCLIP by an absolute margin of 1.66% and 1.43% AUC 390 on UCF-Crime, and 2.29% and 1.46% AP on XD-Violence, 391 respectively. Notably, the gains are achieved by utilizing 392 the frozen VideoMAE-S as a resource-efficient student backbone, 393 demonstrating that accurate anomaly detection does not 394 strictly require computationally intensive foundation models. 395

TCE	DSM	UAG	SAM	AUC (%)↑	FPS ↑	Latency (s)↓	CR (%)↓
×	×	×	×	72.33	40.6	0.266	0.00
✓	✗	✗	✗	79.56	39.8	0.353	0.00
✓	✓	✗	✗	90.10	12.8	1.481	100.0
✓	✓	✓	✗	87.85	37.5	0.462	8.63
✓	✓	✓	✓	<b>89.45</b>	<b>36.2</b>	<b>0.475</b>	<b>8.63</b>

Table 3: Ablation study of key components on UCF-Crime. ✓ and ✗ denote the inclusion and exclusion of each module, respectively.

Gating Strategy	Metric	AUC (%)↑	CR (%)↓
Softmax Entropy	Confidence	86.12	12.45
Evidential Uncertainty (Only)	Uncertainty	87.05	9.80
Retrieval Similarity (Only)	Similarity	88.20	14.20
<b>Hybrid (Ours)</b>	Unc. + Sim.	<b>89.45</b>	<b>8.63</b>

Table 4: Comparison of different gating policies on UCF-Crime.

**Efficiency Results.** MemoVAD is explicitly designed for real-time anomaly detection under an edge-collaborative setting, where both on-device computation and communication costs are critical. As detailed in Table 2, MemoVAD maintains real-time capability on the Jetson AGX Orin by achieving a throughput of 36.2 FPS. The performance represents a substantial improvement over the I3D baseline, which operates exclusively on the edge at 23.7 FPS. Furthermore, MemoVAD significantly outpaces the CLIP-based baseline that relies continuously on remote VLM support, running at only 11.5 FPS. The throughput improvement is also reflected in the latency metrics, where MemoVAD achieves a rapid response time of 0.475 seconds, markedly outperforming both the CLIP-based baseline and the I3D-based baseline. Such a performance gain validates the efficacy of utilizing the lightweight VideoMAE-S architecture, which avoids the computational bottlenecks often associated with the heavier semantic backbones employed by competing methods. Beyond local computational efficiency, MemoVAD significantly minimizes the dependency on remote VLM access. While the CLIP-based baseline necessitates a 100% query rate to function, our method initiates VLM queries for merely 8.63% and 15.72% of the clips on the UCF-Crime and XD-Violence datasets respectively. By processing the vast majority of clips locally, MemoVAD effectively circumvents the bandwidth limitations and latency penalties often associated with expensive remote inference.

### 4.3 Ablation Studies

**Effectiveness of Key Components.** Table 3 summarizes the incremental contributions of each module. Starting with the baseline in Row 1, the frozen VideoMAE backbone achieves 72.33% AUC at 40.6 FPS, but is constrained by the lack of temporal modeling. Incorporating the causal TCE as shown in Row 2 effectively aggregates historical motion cues and boosts AUC to 79.56% with a negligible latency overhead resulting in 39.8 FPS. To gauge the maximum potential of semantic reasoning, we evaluate a teacher forcing variant with DSM in Row 3 that queries the VLM for every clip.

Update Policy	AUC (%)↑	AP (%)↑
First-In-First-Out (FIFO)	87.34	83.10
Random Replacement	86.90	82.55
Least Recently Used (LRU)	88.10	84.20
<b>Semantic-Diversity (Ours)</b>	<b>89.45</b>	<b>85.97</b>

Table 5: Ablation on memory update policies. AUC and AP are selected as metrics.

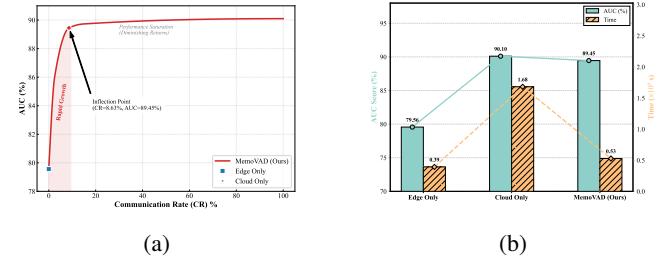


Figure 3: Trade-off evaluation on UCF-Crime. (a) AUC vs. Communication Rate. (b) AUC vs. Time

While the configuration yields a peak AUC of 90.10%, it incurs huge costs including a 100% communication rate and a throughput drop to 12.8 FPS with a high latency of 1.481 seconds, which renders it impractical for edge deployment. To mitigate the issue, the UAG mechanism in Row 4 selectively triggers queries only for clips with high-uncertainty and novelty, reducing communication to 8.63% and restoring a real-time throughput of 37.5 FPS with 0.462 seconds, while incurring only a modest AUC drop to 87.85%. Finally, the introduced SAM alleviates the feature distribution mismatch between the student and teacher. By aligning retrieved semantics, it recovers the AUC to 89.45% and closely matches the upper bound while maintaining a high efficiency of 36.2 FPS and a marginal latency of 0.475 seconds.

**Analysis of Gating Policies.** To validate the design of our Uncertainty-Aware Gating, we compare the proposed method against standard baselines as detailed in Table 4. The baseline utilizing solely Softmax confidence yields suboptimal performance of 86.12%, primarily due to its inability to discriminate between hard samples and out-of-distribution data. While employing evidential uncertainty alone improves precision, it fails to capture novel semantic events that are statistically confident yet semantically unfamiliar. Consequently, our hybrid approach achieves the optimal synergy by triggering VLM queries exclusively when the model exhibits perceived uncertainty or when the input remains semantically distinct from existing memory prototypes.

**Effectiveness of Memory Dynamics.** We further investigate the impact of memory management on long-term learning stability in Table 5. Naive strategies, such as FIFO or Random replacement, result in the catastrophic forgetting of rare anomaly prototypes, leading to a performance degradation exceeding 2%. In contrast, our Online Semantic-Diversity Replacement ensures the retention of a spanning set of diverse abnormal information, thereby maximizing the

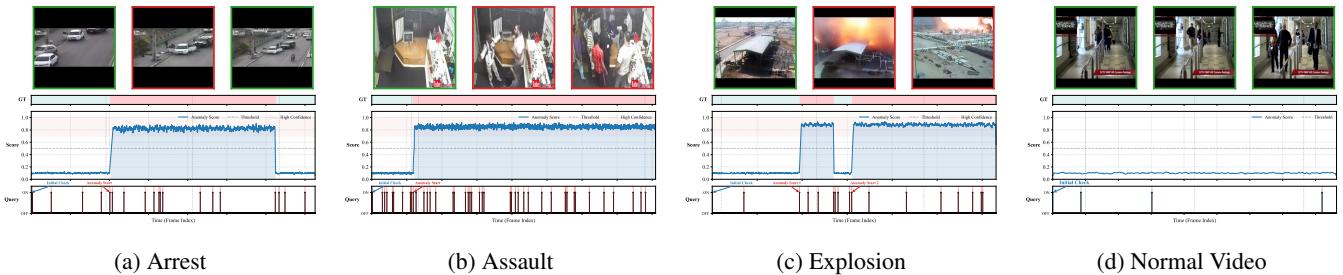


Figure 4: Qualitative visualization of MemoVAD on abnormal and normal scenarios. The rows from top to bottom illustrate: input frames with detection status (Green: Normal, Red: Anomaly), Ground Truth (GT), predicted Anomaly Score, and the VLM Query Trigger signal.

utility of the fixed storage budget of 2,048 slots.

**Efficiency-Accuracy Trade-off.** Fig. 3 illustrates the balance between computational efficiency and detection accuracy. To characterize the operational envelope of MemoVAD, we systematically modulate the gating hyperparameters, specifically the uncertainty threshold  $\tau_{\text{unc}}$  and the similarity threshold  $\tau_{\text{sim}}$ . Fig. 3a reveals a smooth Pareto frontier where the proposed method maintains an AUC exceeding 89% even when the communication overhead is constrained to less than 10%. Furthermore, we also provide a comparative analysis against the baseline deployment paradigms in Fig. 3b. Specifically, while exhibiting a marginal degradation in AUC compared to the Cloud-only approach, our method significantly reduces the inference latency. Notably, the total runtime of MemoVAD is approximately one-third of that required by the Cloud-only paradigm, demonstrating its practicality for real-time edge computing scenarios. Such performance significantly surpasses the baselines and underscores the adaptability of our framework to fluctuating network bandwidths in real-world deployments.

More experimental results (e.g. Effects with different VLMs, Parameter sensitivity analysis) can be found in Appendix 3.

#### 4.4 Qualitative Analysis

**Qualitative Visualization.** Fig. 4 details the inference dynamics of MemoVAD across diverse scenarios. The blue curves denote the predicted anomaly scores and the black vertical stems indicate VLM query triggers. As presented in Fig. 4a, the query frequency spikes immediately at the anomaly onset to resolve high perceived uncertainty. Upon the termination of the event, the system rapidly validates the return to a normal state, resulting in a rapid decay of the anomaly score. Similarly, in Fig. 4b, where the anomaly persists until the end of the video, the system maintains robust recognition throughout the abnormal event duration. Notably, in Fig. 4c, the MemoVAD successfully validates the interval between events to confirm the temporary restoration of normality, demonstrating its ability to capture complex temporal dependencies. In the normal video depicted in Fig. 4d, our system maintains consistently low anomaly scores with extremely sparse updates, effectively minimizing computational costs in the absence of semantic shifts. Overall, MemoVAD precisely localizes events matching ground-truth red borders,

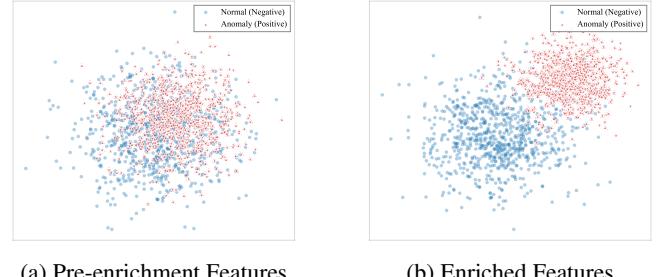


Figure 5: Feature visualization via t-SNE on UCF-Crime dataset.

while meeting rate limits via content-adaptive resources.

**Feature Discrimination Visualization.** To intuitively evaluate the quality of learned representations, we visualize the feature distributions on the UCF-Crime dataset using t-SNE, as depicted in Fig. 5. The visualization in Fig. 5a reveals that the baseline VideoMAE exhibits a high degree of feature entanglement, where anomalous samples are heavily overlapped with normal patterns due to the lack of discriminative semantic guidance. In contrast, as illustrated in Fig. 5b, benefiting from specific optimization and semantic enrichment of MemoVAD, the learned features form distinct and compact clusters with clear decision boundaries. It further demonstrates that MemoVAD effectively disentangles anomalous features away from the normal distribution.

## 5 Conclusion

In this work, we proposed MemoVAD, a framework that strategically bridges a lightweight edge detector with a powerful remote VLM. Instead of continuously relying on the cloud, our system learns to seek remote guidance only when uncertain, caching the insights locally to improve over time. Our experiments confirm that the uncertainty-driven collaboration yields state-of-the-art performance with minimal bandwidth usage, demonstrating a practical path for deploying foundation model capabilities in real-world surveillance without being overwhelmed by their computational weight.

Future work will explore extending the paradigm to collaborative edge computing scenarios, enabling distributed devices to share semantic knowledge.

## 540 References

- [Abdalla *et al.*, 2025] Moshira Abdalla, Sajid Javed, Muaz Al Radi, Anwaar Ulhaq, and Naoufel Werghi. Video anomaly detection in 10 years: A survey and outlook. *Neural Computing and Applications*, pages 1–44, 2025.
- [Chen *et al.*, 2023] Yingxian Chen, Zhengze 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, pages 387–395, 2023.
- [Deng *et al.*, 2025] Huilin Deng, Hongchen Luo, Wei Zhai, Yanming Guo, Yang Cao, and Yu Kang. Vmad: Visual-enhanced multimodal large language model for zero-shot anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 2025.
- [Fang *et al.*, 2024] Heng Fang, Sheng Huang, Wenhao Tang, Luwen Huangfu, and Bo Liu. Sam-mil: A spatial contextual aware multiple instance learning approach for whole slide image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6083–6092, 2024.
- [Ghasemi *et al.*, 2024] Mahshid Ghasemi, Zoran Kostic, Javad Ghaderi, and Gil Zussman. Edgecloudai: Edge-cloud distributed video analytics. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1778–1780, 2024.
- [Guo *et al.*, 2025] Siyan Guo, Cong Zhao, Shusen Yang, Yingying Liang, Yimeng Wang, and Qing Han. Edge-cloud collaborative real-time video object detection for industrial surveillance systems. *IEEE Intelligent Systems*, 2025.
- [Joo *et al.*, 2023] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023.
- [Karim *et al.*, 2024] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6848–6856, 2024.
- [Khan *et al.*, 2022] Muhammad Asif Khan, Ridha Hamila, Aiman Erbad, and Moncef Gabbouj. Distributed inference in resource-constrained iot for real-time video surveillance. *IEEE Systems Journal*, 17(1):1512–1523, 2022.
- [Kumar *et al.*, 2024] Tajinder Kumar, Purushottam Sharma, Jaswinder Tanwar, Hisham Alsghier, Shashi Bhushan, Hesham Alhumyani, Vivek Sharma, and Ahmed I Alutaibi. Cloud-based video streaming services: Trends, challenges, and opportunities. *CAAI Transactions on Intelligence Technology*, 9(2):265–285, 2024.
- [Li *et al.*, 2022] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403, 2022.
- [Li *et al.*, 2025a] Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29203–29212, 2025.
- [Li *et al.*, 2025b] Guo Li, Jiandian Zeng, Zihao Peng, Yuzhu Liang, Xi Zheng, and Tian Wang. E2ec: Edge-to-edge collaboration for efficient real-time video surveillance inference. *IEEE Transactions on Mobile Computing*, 2025.
- [Liu *et al.*, 2024] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7):1–38, 2024.
- [Mishra *et al.*, 2024] Pratik K Mishra, Alex Mihailidis, and Shehroz S Khan. Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(2):1073–1085, 2024.
- [Mondal *et al.*, 2024] Manash Kumar Mondal, Sourav Banerjee, Debasish Das, Uttam Ghosh, Mohammed S Al-Numay, and Utpal Biswas. Toward energy-efficient and cost-effective task offloading in mobile edge computing for intelligent surveillance systems. *IEEE Transactions on Consumer Electronics*, 70(1):4087–4094, 2024.
- [Pu *et al.*, 2024] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024.
- [Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.
- [Sangwan *et al.*, 2025] Amit Sangwan, Priyangshu Sen, Arjun Singh, Shaila Zaman, Yu Zhou, and Michael Medley. From cloud to edge: Enabling offline iot with small-scale language models. *IEEE Internet of Things Magazine*, 2025.
- [Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.
- [Shao *et al.*, 2025] 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*, pages 2586–2595, 2025.
- [Sharshar *et al.*, 2025] Ahmed Sharshar, Latif U Khan, Waseem Ullah, and Mohsen Guizani. Vision-language

- 649 models for edge networks: A comprehensive survey. *IEEE*  
 650 *Internet of Things Journal*, 2025.
- 651 [Shathik, 2025] J Anvar Shathik. Smart vision systems for  
 652 public safety: Real-time crowd monitoring and anomaly  
 653 detection in urban spaces using deep learning and edge  
 654 computing. *International Journal of Applied Mathematics*,  
 655 38(6s):720–743, 2025.
- 656 [Sultani *et al.*, 2018] Waqas Sultani, Chen Chen, and  
 657 Mubarak Shah. Real-world anomaly detection in surveil-  
 658 lance videos. In *Proceedings of the IEEE Conference on*  
 659 *Computer Vision and Pattern Recognition*, pages  
 660 6479–6488, 2018.
- 661 [Tang *et al.*, 2023] Wenhao Tang, Sheng Huang, Xiaoxian  
 662 Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple in-  
 663 stance learning framework with masked hard instance min-  
 664 ing for whole slide image classification. In *Proceedings of*  
 665 *the IEEE/CVF International Conference on Computer Vi-*  
 666 *sion*, pages 4078–4087, 2023.
- 667 [Tian *et al.*, 2021] Yu Tian, Guansong Pang, Yuanhong  
 668 Chen, Rajvinder Singh, Johan W Verjans, and Gustavo  
 669 Carneiro. Weakly-supervised video anomaly detection  
 670 with robust temporal feature magnitude learning. In *Pro-*  
 671 *ceedings of the IEEE/CVF International Conference on*  
 672 *Computer Vision*, pages 4975–4986, 2021.
- 673 [Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang,  
 674 and Limin Wang. Videomae: Masked autoencoders are  
 675 data-efficient learners for self-supervised video pre-  
 676 training. *Advances in Neural Information Processing Sys-*  
 677 *tems*, 35:10078–10093, 2022.
- 678 [Ullah *et al.*, 2022] Waseem Ullah, Amin Ullah, Tanveer  
 679 Hussain, Khan Muhammad, Ali Asghar Heidari, Javier  
 680 Del Ser, Sung Wook Baik, and Victor Hugo C De Al-  
 681 buquerque. Artificial intelligence of things-assisted two-  
 682 stream neural network for anomaly detection in surveil-  
 683 lance big video data. *Future Generation Computer Sys-*  
 684 *tems*, 129:286–297, 2022.
- 685 [Wang *et al.*, 2025] Hanling Wang, Qing Li, Li Chen,  
 686 Haidong Kang, Fei Ma, and Yong Jiang. Holotrace:  
 687 Llm-based bidirectional causal knowledge graph for edge-  
 688 cloud video anomaly detection. In *Proceedings of the*  
 689 *33rd ACM International Conference on Multimedia*, pages  
 690 6510–6519, 2025.
- 691 [Wu and Liu, 2021] Peng Wu and Jing Liu. Learning causal  
 692 temporal relation and feature discrimination for anomaly  
 693 detection. *IEEE Transactions on Image Processing*,  
 694 30:3513–3527, 2021.
- 695 [Wu *et al.*, 2020] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun,  
 696 Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only  
 697 look, but also listen: Learning multimodal violence detec-  
 698 tion under weak supervision. In *European Conference on*  
 699 *Computer Vision*, pages 322–339. Springer, 2020.
- 700 [Wu *et al.*, 2024a] Peng Wu, Xuerong Zhou, Guansong  
 701 Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yan-  
 702 ning Zhang. Weakly supervised video anomaly detection  
 703 and localization with spatio-temporal prompts. In *Pro-*  
 704 *ceedings of the 32nd ACM International Conference on*  
 705 *Multimedia*, pages 9301–9310, 2024.
- 706 [Wu *et al.*, 2024b] Peng Wu, Xuerong Zhou, Guansong  
 707 Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yan-  
 708 ning Zhang. Vadclip: Adapting vision-language models  
 709 for weakly supervised video anomaly detection. In *Pro-*  
 710 *ceedings of the AAAI Conference on Artificial Intelligence*,  
 711 volume 38, pages 6074–6082, 2024.
- 712 [Yang *et al.*, 2024] Zhiwei Yang, Jing Liu, and Peng Wu.  
 713 Text prompt with normality guidance for weakly super-  
 714 vised video anomaly detection. In *Proceedings of the*  
 715 *IEEE/CVF Conference on Computer Vision and Pattern*  
 716 *Recognition*, pages 18899–18908, 2024.
- 717 [Ye *et al.*, 2025] Muchao Ye, Weiyang Liu, and Pan He.  
 718 Vera: Explainable video anomaly detection via verbalized  
 719 learning of vision-language models. In *Proceedings of*  
 720 *the Computer Vision and Pattern Recognition Conference*,  
 721 pages 8679–8688, 2025.
- 722 [Zanella *et al.*, 2024] Luca Zanella, Willi Menapace, Massi-  
 723 miliano Mancini, Yiming Wang, and Elisa Ricci. Harness-  
 724 ing large language models for training-free video anomaly  
 725 detection. In *Proceedings of the IEEE/CVF Conference on*  
 726 *Computer Vision and Pattern Recognition*, pages 18527–  
 727 18536, 2024.
- 728 [Zhang *et al.*, 2022] Hongrun Zhang, Yanda Meng, Yitian  
 729 Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland,  
 730 and Yalin Zheng. Dtdf-mil: Double-tier feature distillation  
 731 multiple instance learning for histopathology whole slide  
 732 image classification. In *Proceedings of the IEEE/CVF*  
 733 *Conference on Computer Vision and Pattern recognition*,  
 734 pages 18802–18812, 2022.
- 735 [Zhang *et al.*, 2023] Chen Zhang, Guorong Li, Yuankai Qi,  
 736 Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-  
 737 Hsuan Yang. Exploiting completeness and uncertainty of  
 738 pseudo labels for weakly supervised video anomaly de-  
 739 tection. In *Proceedings of the IEEE/CVF Conference on*  
 740 *Computer Vision and Pattern Recognition*, pages 16271–  
 741 16280, 2023.
- 742 [Zhang *et al.*, 2025] Yang Zhang, Hanling Wang, Qing Bai,  
 743 Haifeng Liang, Peican Zhu, Gabriel-Miro Muntean, and  
 744 Qing Li. Vavlm: Toward efficient edge-cloud video ana-  
 745 lytics with vision-language models. *IEEE Transactions on*  
 746 *Broadcasting*, 2025.
- 747 [Zhou *et al.*, 2023] Hang Zhou, Junqing Yu, and Wei Yang.  
 748 Dual memory units with uncertainty regulation for weakly  
 749 supervised video anomaly detection. In *Proceedings of*  
 750 *the AAAI Conference on Artificial Intelligence*, volume 37,  
 751 pages 3769–3777, 2023.