This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Joint Learning of General and Diverse Patterns with Mixture of Memory Experts for Weakly-Supervised Video Anomaly Detection

Bo Sun[1,2]*, Junxi Chen[1]*, Zhe Wu[2,3]†, Feng Gao[4], Fan Yang[4], Li Su[1]†, Yaowei Wang[2,5],
[1]University of Chinese Academy of Sciences
[2]Institute of Perceptual Intelligence, Pengcheng Laboratory
[3]Pazhou Lab, [4]Peking University, [5]Harbin Institute of Technology, Shenzhen,
{sunbo24, chenjunxi22}@mails.ucas.ac.cn, wuzh02@pcl.ac.cn,
{gaof, fyang.eecs}@pku.edu.cn, suli@ucas.ac.cn, yaoweiwang@gmail.com

Abstract

*Weakly-supervised Video Anomaly Detection (wVAD) aims to detect abnormal events using only binary labels, making it challenging to capture both the diversity of anomalies and their shared semantic cues. Existing methods either focus on a generic anomaly pattern, achieving strong generalization but weak discrimination, or rely on class-level diversity modeling, which ignores shared semantics and suffers from limited generalization. To overcome these limitations, we propose the **Mixture of Memory Experts (MoME)**, a unified framework that jointly learns general and diverse patterns. Each expert in MoME possesses an internal memory for fine-grained specialization and shares an external memory for general knowledge aggregation. To enhance semantic diversity and improve generalization beyond coarse class-level supervision, we introduce an **Anomaly Prototype Router (APR)** that leverages large language models to construct generalized anomaly prototypes for semantically guided expert routing. Moreover, balanced routing, expert diversity, and enhanced pattern discriminability are ensured by the regularization loss for APR, the distinctiveness loss for experts, and reconstruction together with memory tasks, respectively. Extensive experiments on UCF-Crime and XD-Violence demonstrate that our approach achieves state-of-the-art performance, validating the effectiveness of jointly modeling generality and diversity for robust anomaly detection.*

1. Introduction

Video Anomaly Detection (VAD) aims to automatically identify abnormal events, thereby alleviating the heavy workload of human monitors and enabling timely responses. This capability is crucial for a wide range of applications, including security surveillance [5, 15], traffic monitoring [24] and industrial process control [7]. Since collecting frame-level annotations for training are arduous, recent studies are mainly turn to weakly-supervised VAD (wVAD) [4, 9, 14, 19, 28] which only leverages video-level labels.

To identify abnormal events, traditional wVAD methods [9,18,21,28,45], as in Fig. 1(a), typically learn a *generic*

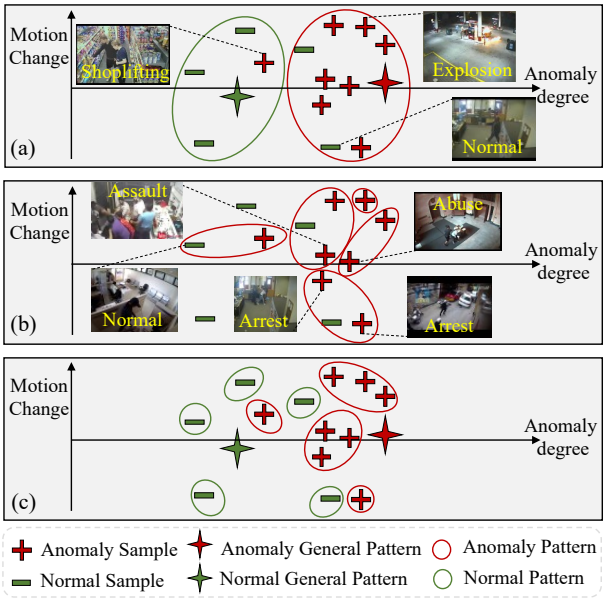


Figure 1. (**a**) Conventional wVAD methods learn a single generic anomaly pattern, ignoring anomaly diversity. (**b**) Class-specific methods rely on coarse labels for forced classification, improving separability but missing global semantics and causing confusion. (**c**) Our method models both *general* and *diverse* patterns and encourages semantic diversity in the feature space.

*Equal contribution.
†Corresponding authors.

anomaly pattern by treating all abnormal samples as homogeneous under binary supervision. Such a general pattern promotes cross-scene generalization but lacks discriminative ability, as real-world anomalies are inherently diverse in their visual characteristics and semantic cues. For example, anomalies like *shoplifting* involve subtle action semantics, whereas events such as *explosions* exhibit drastic scene dynamics and salient global changes. A single unified pattern is therefore too coarse to represent such diverse phenomena effectively.

Recent works [2, 37, 38] begin to model anomaly diversity under explicit class supervision, as illustrated in Fig. 1(b). These methods partition the feature space according to class labels [37, 38] or assign different anomaly categories to dedicated subnetworks [2]. Although such designs enhance discriminative capability, they inevitably compromise generalization. First, predefined class labels lack broad conceptual information and often cause overfitting within specific class clusters. For example, the model may overfit to the class *shoplifting*, leading normal actions such as *picking up an item*, which share similar visual contexts, to be misclassified as anomalies. Second, class labels is typically too coarse and lacks semantic richness, ignoring intra-class variations and inter-class similarities.

We argue that a robust anomaly detection model should simultaneously learn both general and diverse patterns, where the former enhances generalization ability and the latter provides more discriminative representations. Moreover, diversity should be modeled through semantic guidance that captures fine-grained contextual cues rather than relying on coarse class labels. To this end, we propose the Mixture of Memory Experts (MoME) network, which consists of a Mixture of Normal Memory Experts (MoNE) for normal patterns and a Mixture of Anomaly Memory Experts (MoAE) for anomalies. These components jointly model general and diverse patterns, achieving stronger discriminability while preserving generalization, as illustrated in Fig. 1(c). Specifically, each expert in MoME possesses an internal memory to capture feature diversity and shares an external memory to encode general representations. To obtain more expressive representations beyond coarse class labels, we introduce the Anomaly Prototype Router (APR), which leverages Large Language Models (LLMs) to obtain generalized anomaly prototypes and employs a learnable expert-prototype matrix constrained by a regularization loss to encourage semantic-aware routing. Furthermore, a sparse MoE design is incorporated to encourage each expert to specialize in distinct representations, with a distinctiveness loss enforcing inter-expert diversity.

In summary, our main contributions are as follows:

- We propose a Mixture of Memory Experts network, where each expert's internal memory captures diverse patterns, while a shared external memory encodes general knowledge. This design enables the joint learning of general and diverse patterns, preserving generalization ability while enhancing discriminative capacity.
- We introduce an Anomaly Prototype Router, which assigns LLM-generated semantic prototypes to each frame and allocates them through a learnable expert-prototype matrix constrained by a regularization loss to model generalizable semantic diversity.
- Extensive experiments on two benchmarks demonstrate that our method achieves state-of-the-art performance and strong generalization.

2. Related Work

2.1. Weakly-Supervised Video Anomaly Detection

Weakly-supervised Video Anomaly Detection (wVAD) aims to detect anomalous events using only video-level supervision. Most wVAD methods [9, 18, 21, 28, 41] follow the Multiple Instance Learning (MIL) paradigm, where a video is modeled as a bag of snippets and most anomaly instances in positive bags are encouraged to outscore normal ones in negative bags through a ranking-based loss. To improve discriminability, feature-based approaches [31, 44] exploit intrinsic feature statistics, such as feature magnitude [31] or deviations from the BatchNorm mean [44]. Meanwhile, multimodal methods [4, 22, 29, 37] incorporate auxiliary modalities, including optical flow [6, 30], audio [4, 34], and textual information [22, 29], to enhance robustness and semantic understanding. More recently, Large Vision-Language Models (LVLMs) have been explored [30, 37, 40] to exploit rich cross-modal semantic knowledge, improving interpretability and anomaly reasoning capability.

2.2. Mixture of Experts

The MoE framework, first introduced by Jacobs et al. [11], decomposes complex problems into simpler subtasks handled by specialized experts, while a router assigns inputs to relevant expert. Later, Shazeer et al. [26] demonstrated its scalability for large-scale deep learning with thousands of feed-forward sub-networks, and Lepikhin et al. [13] extended it to vision via V-MoE with sparse activation. Recently, large-scale models such as GPT [1], Qwen [3], and DeepSeek [8, 16] have adopted MoE mechanisms to enhance scalability and efficiency. MoE has only recently been explored in video anomaly detection. D'Amicantonio et al. [2] employed a dense MoE with class-specific experts trained on class-labeled data, where the expert count was tied to anomaly classes, limiting generalization and diversity due to coarse class granularity. In contrast, we adopt a sparse MoE that integrates anomaly semantics and decouples the number of experts from anomaly classes, leading to improved generalization and flexibility.

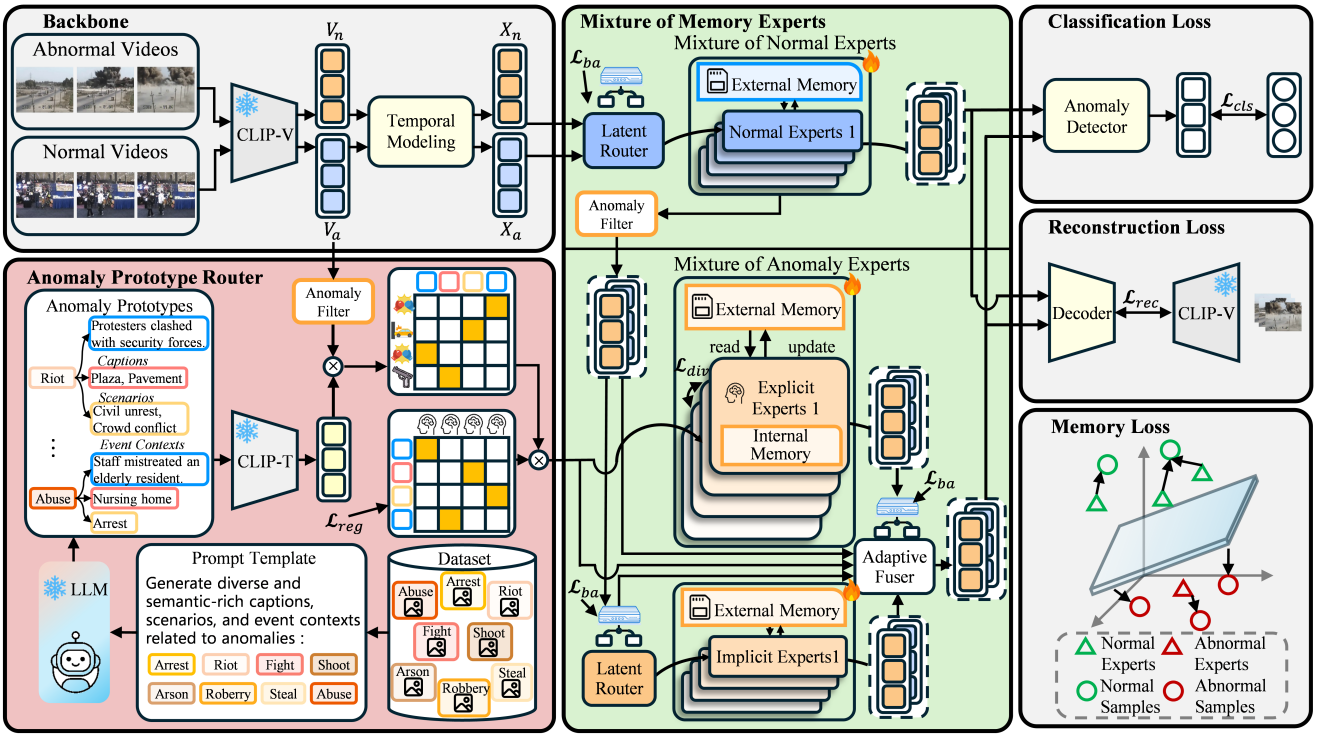


Figure 2. Overall proposed framework. (**1**) The MoME framework jointly learns general patterns via the external memory and diverse patterns through multiple experts. (**2**) The Anomaly Prototype Router (APR) leverages LLM-derived semantic prototypes to capture anomaly semantic diversity beyond coarse class labels. (**3**) The regularization loss L_{reg} and expert balance loss L_{ba} ensure balanced routing; the distinctiveness loss L_{div} encourages expert diversity; and the reconstruction and memory tasks enhance pattern discriminability.

3. Method

3.1. Overview

Given normal and anomaly training videos with a fixed length of T frames, denoted as $\mathcal{V}_n = \{\mathcal{V}_n^i\}_{i=1}^T$ and $\mathcal{V}_a = \{\mathcal{V}_a^i\}_{i=1}^T$, our training objective is to identify the anomalous subset $\mathcal{V}_{\hat{a}} = \{\mathcal{V}_{\hat{a}}^i\}_{i=1}^{T_a}$ containing T_a frames, where $\hat{a}$ denotes truly anomalous. Each frame is labeled with a binary label $Y \in \{0, 1\}$, where $Y = 1$ for anomalous and $Y = 0$ for normal. At inference, the model produces frame-level anomaly scores ranging from 0 to 1. As illustrated in Fig. 2, frame-level features are first extracted using a frozen visual encoder [23], which outputs visual features of dimension d, denoted as $V_n, V_a \in \mathbb{R}^{T\times d}$. These features are then temporally modeled following [37] to obtain temporal representations $X_n, X_a \in \mathbb{R}^{T\times d}$. The proposed Mixture of Normal Memory Experts (MoNE) and Mixture of Anomaly Memory Experts (MoAE) jointly learn general and diverse patterns, generating memory-augmented features that are subsequently fed into a detector to produce anomaly scores. Shared memory blocks capture general patterns, while the routing mechanism, which comprises the router and its corresponding experts, models diverse ones. Moreover, regularization and expert balance losses ensure balanced routing, distinctiveness loss encourages experts diversity, and reconstruction and memory tasks enhance pattern discriminability.

3.2. Mixture of Memory Experts Network

Mixture of Normal Memory Experts. MoNE is proposed to jointly learn general and diverse normal patterns. We capture diverse normal patterns via a *Latent Router* (LR), a linear layer that adaptively distributes input features to multiple latent normal experts, due to the absence of explicit supervision of normal data. Given normal temporal features $X_n \in \mathbb{R}^{T\times d}$, the latent routing scores are computed as:

$$G_n^l = W_l X_n^T + b, \quad R_n^l = \text{Softmax}\big(\text{TopK}_\text{v}(G_n^l, k)\big), \quad (1)$$

where $W_l \in \mathbb{R}^{E_n\times d}$ and $b \in \mathbb{R}^1$ are learnable parameters, and E_n denotes the number of normal experts. $G_n^l \in \mathbb{R}^{T\times E_n}$ represents the routing scores for all latent normal experts, while $R_n^l \in \mathbb{R}^{T\times k}$ denotes the normalized routing weights of the top k selected experts, where $\text{TopK}_\text{v}(\cdot)$ returns the top k values.

Using the routing scores G_n^l, each normal snippet is assigned to its most relevant expert, resulting in expert-specific subsets $X_n = \{X_n^i \in \mathbb{R}^{T_i\times d}\}_{i=1}^{E_n}$, where T_i denotes the number of snippets assigned to the i-th expert. Details of the snippets-to-experts assignment algorithm are

provided in the supplementary material due to space limitations. Each expert processes its subset using a shared external normal memory $M^N_{\text{ext}} \in \mathbb{R}^{m\times d}$, consisting of m memory slots, to jointly learn general and diverse normal patterns. The pattern-augmented outputs are defined as:

$$S^{\text{N}}_n, H^{\text{N}}_n = \sum_{i=1}^{E_n} R^l_{n,i}\Phi^l_{n,i}(X^i_n, M^N_{\text{ext}}), \tag{2}$$

where $R^l_{n,i}$ is the routing weight of the i-th latent normal expert, and $\Phi^l_{n,i}(\cdot)$ denotes its output given X^i_n and shared memory M^N_{ext}. $S^{\text{N}}_n \in \mathbb{R}^{T\times m}$ and $H^{\text{N}}_n \in \mathbb{R}^{T\times d}$ denote the weighted memory similarities across all normal experts and aggregated normal representations, respectively.

Anomaly Filter. Given the anomaly features $X_a \in \mathbb{R}^{T\times d}$, they are fed into MoNE to produce S^N_a and H^N_a. Anomalous snippets are identified as those with the lowest similarity to normal patterns:

$$I_{\hat{a}} = \text{TopK}_{\text{I}}(1 - S^N_a, T_a), \tag{3}$$

where $I_{\hat{a}} \in \mathbb{R}^{T_a}$ represents the indices of the most T_a anomalous snippets, with $\text{TopK}_{\text{I}}(\cdot)$ returns the top k indices. The corresponding refined anomaly-related features and similarity matrices are denoted as:

$$\begin{aligned} V_{\hat{a}} &= V_a[I_{\hat{a}}] \in \mathbb{R}^{T_a\times d}, \quad X_{\hat{a}} = X_a[I_{\hat{a}}] \in \mathbb{R}^{T_a\times d}, \\ H^N_{\hat{a}} &= H^N_a[I_{\hat{a}}] \in \mathbb{R}^{T_a\times d}, \quad S^N_{\hat{a}} = S^N_a[I_{\hat{a}}] \in \mathbb{R}^{T_a\times m}. \end{aligned} \tag{4}$$

Mixture of Anomaly Memory Experts. Given the anomaly temporal feature $X_{\hat{a}}$ filtered by MoNE, the final anomaly memory-augmented representation is computed as:

$$S^{\text{A}}_{\hat{a}}, H^{\text{A}}_{\hat{a}} = \sum_{j\in\{l,e\}}\sum_{i=1}^{E_j} F^j_{\hat{a},i}\Phi^j_{a,i}(X^i_{\hat{a}}, M^{\text{A}}_{\text{ext}}), \tag{5}$$

where $F^e_{\hat{a},i}$ and $F^l_{\hat{a},i}$ represent the Adaptive-Fuser weights associated with the i-th explicit and latent anomaly experts, respectively, which will be detailed in Section 3.4. Their corresponding outputs, $\Phi^e_{a,i}(\cdot)$ and $\Phi^l_{a,i}(\cdot)$, are derived from the input $X_{\hat{a}}$ and the shared anomaly memory $M^{\text{A}}_{\text{ext}}$. Here, $S^{\text{A}}_{\hat{a}} \in \mathbb{R}^{T_a\times m}$ represents the weighted memory similarity across all anomaly experts, and $H^{\text{A}}_{\hat{a}} \in \mathbb{R}^{T_a\times d}$ denotes the aggregated anomaly-aware features. Similarly, we compute S^{A}_n and H^{A}_n for normal temporal features X_n.

3.3. Memory Expert

As shown in Fig. 3, the memory expert models general patterns via the external memory M_{ext} and specific patterns via the internal memory M_{in}, both in $\mathbb{R}^{m\times d}$, where m and d denote the number of memory slots and the feature dimension, respectively. All experts share the same architecture.

Memory Reading. As an example, consider the i-th normal expert with temporal feature sequence $X^i_n \in \mathbb{R}^{T_i\times d}$.

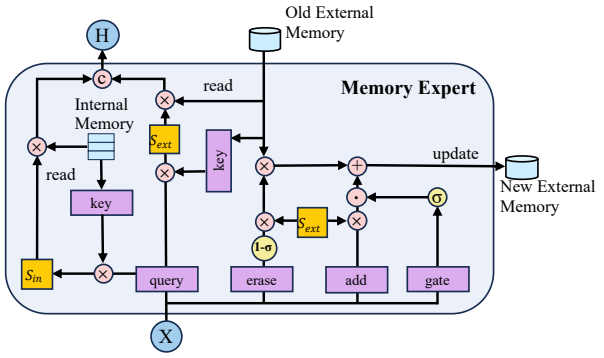


Figure 3. Illustration of the architecture of a memory expert. Each expert contains an internal memory module for learning specialized patterns and an external memory input for learning general patterns.

Both X^i_n and the memory matrices are projected into a shared embedding space:

$$Q = X^i_n W_q, \quad K_{in} = M_{in}W_k, \quad K_{ext} = M_{ext}W_k, \tag{6}$$

where $W_q, W_k \in \mathbb{R}^{d\times d}$ are learnable matrices. Here, $Q \in \mathbb{R}^{T_i\times d}$ is the query, and $K_{in}, K_{ext} \in \mathbb{R}^{m\times d}$ are the keys for M_{in} and M_{ext}, respectively. Memory similarities are computed via scaled dot-product attention:

$$S_{in} = \sigma\left(\frac{QK^\top_{in}}{\sqrt{d}}\right), \quad S_{ext} = \sigma\left(\frac{QK^\top_{ext}}{\sqrt{d}}\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function applied along the memory dimension. The final memory-augmented representation is obtained by reading from both memories and concatenating their outputs:

$$H = \text{concat}(S_{in}M_{in}, S_{ext}M_{ext}), \tag{8}$$

where $H \in \mathbb{R}^{T_i\times 2d}$ encodes both general and expert-specific knowledge.

Memory Updating. The internal memory is trained via backpropagation to capture expert-specific features, while the external memory is dynamically updated with an erase-and-write mechanism to store more knowledge [27, 33]. Given X^i_n, we compute three features:

$$V_e = \sigma(f_e(X^i_n)), V_g = \sigma(f_g(X^i_n)), V_a = \tanh(f_a(X^i_n)), \tag{9}$$

where f_e, f_g, f_a are linear layers and $V_e, V_g, V_a \in \mathbb{R}^{T_i\times d}$ are the erase, gate, and add features. $\sigma(\cdot)$ constrains values to $[0, 1]$ for selective erasure and writing, while $\tanh(\cdot)$ maps values to $[-1, 1]$, enabling positive and negative updates. The external memory similarity scores S_{ext} weight the erase and add features via outer products:

$$V_E = S_{ext} \otimes V_e, \quad V_A = S_{ext} \otimes V_a, \tag{10}$$

where $V_E, V_A \in \mathbb{R}^{T_i\times m\times d}$ denote the weighted erase and add features, respectively. The add features is modulated by the gate features and normalized for stability:

$$\hat{V}_A = \frac{V_A \odot V_g}{||V_A \odot V_g||_2}, \tag{11}$$

Averaging across all snippet features yields fused erase and add features $\bar{V}_E, \bar{V}_A \in \mathbb{R}^{m\times d}$. Finally, the external memory is updated via the erase-add operation with learning rate α:

$$M_{ext} = M_{ext} \odot (1 - \alpha\bar{V}_E) + \alpha\bar{V}_A. \tag{12}$$

3.4. Anomaly Prototype Router

Anomaly videos include class labels that provide explicit guidance for learning diverse anomaly patterns. To exploit this supervision, we introduce the *Anomaly Prototype Router* (APR). Although class-specific information helps model anomaly diversity, directly using class labels, as in [2], is often suboptimal since anomalies may span multiple classes, making fixed class-dependent experts difficult to train.

To address this issue, we prompt several large language models with the original anomaly class names and manually select the most generalizable outputs, referred to as *Anomaly Prototypes* (APs). APR aims to enhance semantic representations beyond discrete labels, encouraging the model to capture intra-class diversity and align semantically related patterns across categories.

Given the APs, we extract semantic features $T \in \mathbb{R}^{n\times d}$ and video features $V_{\hat{a}} \in \mathbb{R}^{T_a\times d}$ using CLIP's text and visual encoders, where n is the number of prototypes. Both modalities are normalized, and the cross-modal similarity is computed as:

$$S = \mathrm{Softmax}\left(\frac{1}{\tau}\cdot\frac{V_{\hat{a}}}{\|V_{\hat{a}}\|_2}\left(\frac{T}{\|T\|_2}\right)^\top\right), \tag{13}$$

where τ is a temperature parameter controlling the sharpness of similarity scores, and $S \in \mathbb{R}^{T_a\times n}$ denotes the normalized visual-textual similarity distribution.

To associate each AP with its corresponding experts, we introduce a learnable expert-prototype allocation matrix $A \in \mathbb{R}^{E_e\times n}$, where E_e is the number of explicit anomaly experts. The expert routing scores are then computed as:

$$G^e_{\hat{a}} = SA^\top, \quad R^e_{\hat{a}} = \mathrm{Softmax}(\mathrm{TopK_v}(G^e_{\hat{a}}, k)), \tag{14}$$

where $G^e_{\hat{a}} \in \mathbb{R}^{T_a\times E_e}$ denotes the routing scores for all explicit experts, and $R^e_{\hat{a}} \in \mathbb{R}^{T_a\times k}$ represents the normalized top k routing weights.

To ensure balanced expert routing, we introduce intra- and inter-expert regularizations on the routing matrix A. The intra-expert regularization promotes sparsity and diversity among expert activations:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{N}\sum_{i=1}^{E_e} H(A_i) + \left\|AA^\top - I_E\right\|_F^2, \tag{15}$$

where the entropy term encourages sparsity, and the orthogonality term enforces diversity across experts. The inter-expert regularization encourages balanced utilization of all experts:

$$\mathcal{L}_{\text{inter}} = \sum_{k=1}^{N}\left(\mu_k - \frac{1}{N}\right)^2 + 2\sum_{k=1}^{N}\sum_{i<j}|A_{i,k}A_{j,k}|, \tag{16}$$

where μ_k denotes the mean of the k-th column of A, $||\cdot||_F$ is the Frobenius norm. The total regularization loss $\mathcal{L}_{reg}$ is computed as the sum of the intra and inter terms.

While the APR explicitly routes samples based on semantic information, some latent anomaly patterns may remain unexplored. To capture these latent relation we also introduce an *latent Router* (LR), which adaptively learns latent anomaly diversity:

$$G^l_{\hat{a}} = W_l X^T_{\hat{a}} + b,\ R^l_{\hat{a}} = \mathrm{Softmax}\big(\mathrm{TopK_v}(G^l_{\hat{a}}, k)\big), \tag{17}$$

where $W_l \in \mathbb{R}^{E_l\times d}$ and $b \in \mathbb{R}$ are learnable parameters, with E_l representing the number of latent anomaly experts. $G^l_{\hat{a}} \in \mathbb{R}^{T_a\times E_l}$ denotes the routing scores for all latent experts, and $R^l_{\hat{a}} \in \mathbb{R}^{T_a\times k}$ represents the normalized routing weights of the selected top k experts.

To effectively integrate explicit and latent cues, we design an *Adaptive Fuser* (AF), a conditional network that adaptively fuses the routing results from APR and LR. Given the temporal anomaly features $X_{\hat{a}}$, AF re-evaluates the routing weights of the selected explicit and latent experts.

$X_{\hat{a}}$ is first projected via a linear layer to obtain auxiliary routing features $Z_x \in \mathbb{R}^{T_a\times 2k}$. The explicit scores $R^e_{\hat{a}}$ are then z-score normalized to match the mean and standard deviation of $R^l_{\hat{a}}$, and concatenated with it to form the joint routing representation $Z_v \in \mathbb{R}^{T_a\times 2k}$. In parallel, we retrieve expert identity embeddings $Z_i \in \mathbb{R}^{T_a\times 2k\times d}$ from a learnable identity matrix $I \in \mathbb{R}^{(E_e+E_l)\times d}$. The fused representation is defined as:

$$Z = \mathrm{concat}(Z_v, Z_i, Z_x), \tag{18}$$

which is passed through a lightweight network to produce the final routing weights:

$$F^e_{\hat{a}}, F^l_{\hat{a}} = F_{\hat{a}} = W_f\,\mathrm{ReLU}(ZW_z + b_z) + b_f, \tag{19}$$

where $W_z \in \mathbb{R}^{(d+2)\times d}$, $W_f \in \mathbb{R}^{d\times 1}$, and $b_z \in \mathbb{R}^d$, $b_f \in \mathbb{R}^1$ are learnable parameters. Here, $F^e_{\hat{a}}, F^l_{\hat{a}} \in \mathbb{R}^{T_a\times k}$ denote the re-evaluates routing weights for the Top-k explicit and latent experts, respectively.

3.5. Model Optimization

We directly utilize the discriminative memory-augmented features from MoME for anomaly detection. The final frame-level anomaly scores are computed as:

$$y_n = \mathrm{MLP}\big(H^N_n\big), \quad y_{\hat{a}} = \mathrm{MLP}\big(H^A_{\hat{a}}\big), \tag{20}$$

where MLP denotes a lightweight anomaly detector shared across normal and abnormal features, optimized with a bi-

nary classification loss:

$$\mathcal{L}_{\text{cls}} = \text{BCE}(y_n, 0) + \text{BCE}(y_{\hat{a}}, 1). \tag{21}$$

To guide MoNE and MoAE in learning respective patterns, we introduce a reconstruction task. Two linear decoders $d_n(\cdot)$ and $d_a(\cdot)$ reconstruct features: $R_n^N = d_n(H_n^N), R_n^A = d_a(H_n^A), R_{\hat{a}}^A = d_a(H_{\hat{a}}^A), R_{\hat{a}}^N = d_n(H_{\hat{a}}^N)$, and the reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \sum_{i \in P} \text{MSE}(R_i, V_i) + \delta\big(1 - \sum_{i \in N} \text{MSE}(R_i, V_i)\big), \tag{22}$$

where MSE denotes the mean squared error, $\delta(\cdot)$ is the ReLU function, V_i are CLIP features, and $P = \{n^N, \hat{a}^A\}$ and $N = \{n^A, \hat{a}^N\}$ denote the positive and negative sets, respectively. Besides, following [43], we employ memory-based supervision:

$$\begin{aligned} \mathcal{L}_{\text{m}} = \text{BCE}(S_n^N, 1) + \text{BCE}(S_{\hat{a}}^N, 0) \\ + \text{BCE}(S_n^A, 0) + \text{BCE}(S_{\hat{a}}^A, 1), \end{aligned} \tag{23}$$

To ensure expert diversity, we employ an expert balancing loss [26]. Given routing score matrix $G \in \mathbb{R}^{l \times e}$, where l denotes the feature sequence length and e denotes the number of experts, we first aggregate routing scores over all tokens to obtain $\hat{G} \in \mathbb{R}^e$. The balancing loss is then defined as:

$$\mathcal{L}_{\text{ba}} = \text{CV}\big(\hat{G}\big), \tag{24}$$

where $\text{CV}(\cdot)$ denotes the squared coefficient of variation. This loss is applied to the normal latent routing score (G_n^l), the anomaly latent routing score ($G_{\hat{a}}^l$), and the fused weight ($F_{\hat{a}}$):

$$\mathcal{L}_{\text{b}} = \mathcal{L}_{\text{ba}}\big(G_n^l\big) + \mathcal{L}_{\text{ba}}\big(G_{\hat{a}}^l\big) + \mathcal{L}_{\text{ba}}\big(F_{\hat{a}}\big). \tag{25}$$

As load balancing does not ensure diversity—similar inputs may still be assigned to overlapping experts—we introduce a *memory distinctiveness loss* that maximizes the pairwise ℓ_1 distances among expert memories. Let $E_a = E_e + E_l$ be the total number of anomaly experts, and $M_{\text{in}}^{a,i}$ and $M_{\text{in}}^{n,i}$ denote the i-th anomaly and normal internal memory matrices, respectively. The distinctiveness loss is defined as:

$$\mathcal{L}_{\text{div}} = \sum_{k \in \{a,n\}} \frac{E_k(E_k - 1)}{2\sum_{i<j} \|M_{\text{in}}^{k,i} - M_{\text{in}}^{k,j}\|_1}. \tag{26}$$

4. Experimental

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate MoME on two large-scale video anomaly detection benchmarks: UCF-Crime [28] and XD-Violence [34]. **UCF-Crime** contains 1,900 untrimmed surveillance videos across 13 anomaly categories (e.g., *Assault*, *Burglary*, *Explosion*), with video-level weak annotations. Its long normal segments, cluttered backgrounds, and large intra-class variation make it highly challenging. **XD-Violence** consists of 4,754 untrimmed videos with multimodal annotations (visual, audio, and video-level labels). It covers diverse sources including movies, sports events, and surveillance footage, and defines violence broadly, providing a comprehensive benchmark for multimodal and robust anomaly detection methods.

Evaluation Metrics. Consistent with prior work [4, 12, 28, 34, 36], we report the frame-level area under the ROC curve (AUC) on UCF-Crime. For XD-Violence, we adopt the average precision (AP) metric following the standard evaluation protocol.

4.2. Implementation Details

We employ the visual and text encoders of the pre-trained CLIP model (ViT-B/16) [23] to extract video and text features. During training, the batch size is set to 128. The initial learning rate is 1×10^{-4} for UCF-Crime and 4×10^{-5} for XD-Violence. The MoNE network comprises 24 latent experts, while the MoAE includes both explicit and latent experts, each with 12 experts. For all three types of experts, the number of activated experts is fixed at $k = 1$ during both training and inference. The video length T is set to 256, and the number of anomalous frames T_a is fixed at 16. Additional implementation details are provided in the supplementary material.

Table 1. Comparison with other methods. The best results are in **bold**, and the second-best are underlined.

	Methods	Feature	UCF (AUC %)	XD (AP %)
Semi	SVM baseline	CLIP	50.10	50.80
	OCSVM [25]	CLIP	63.20	28.63
	Hasan et al. [10]	CLIP	51.20	31.25
Weakly	Sultani et al [28]	C3D	75.41	75.68
	RTFM [31]	I3D	84.30	77.81
	UR-DMU [43]	I3D	86.97	81.77
	GS-MOE [2]	I3D	**91.58**	**82.89**
Weakly	UMIL [18]	CLIP	86.75	-
	CLIP-TSA [12]	CLIP	87.58	82.19
	OVVAD [35]	CLIP	86.40	66.53
	IFS-VAD [42]	CLIP	86.57	84.14
	TPWNG [38]	CLIP	87.79	83.68
	VadCLIP [37]	CLIP	88.02	84.51
	FederatedVAD [32]	CLIP	85.07	74.23
	DEN-VAD [39]	CLIP	87.86	86.13
	ITC [17]	CLIP	**89.04**	85.45
	MoME	CLIP	88.32	**86.15**

4.3. Comparisons with SOTA methods

The proposed method is compared with the following SOTA approaches, which can be categorized into semi-supervised methods [10, 25] and weakly supervised methods [2, 12, 18, 28, 31, 32, 35, 37–39, 42, 43]. The comparison results are summarized in Table 1. Our method achieves the best performance on the XD-Violence dataset, reaching a peak AP of 86.15%. On the UCF-Crime dataset, our MoME remains highly competitive among CLIP-based approaches, achieving an AUC of 88.32%. The ITC and GS-MoE methods learn class-specific representations. Although they perform strongly on UCF-Crime, their perfor-

Event Count	1	2	3	4	5	6	7	8	9	10	11	12	14	16	17	18	23
Video Number	237	105	58	26	15	23	17	7	3	1	2	1	1	1	1	1	1
VadCLIP	91.62	73.22	73.18	71.47	54.24	52.03	66.82	55.18	42.90	82.03	70.01	32.76	75.21	53.93	51.80	60.30	74.36
MoME	95.10	81.76	76.72	82.76	70.32	60.99	75.40	58.92	46.06	74.90	68.42	38.35	73.74	66.81	56.26	41.96	65.70

Table 2. Comparison between MoME and VadCLIP across different anomalous event counts per video (%).

mance drops on XD-Violence, which may be attributed to the lower diversity and higher class separability of UCF-Crime. In contrast, our method maintains a more balanced performance across both datasets compared with most previous approaches. These results demonstrate the effectiveness of the proposed joint learning framework in capturing both general and semantically diverse patterns.

MoNE	MoAE	M_{ext}	AF	APR	LR	UCF (AUC %)	XD (AP %)
✓		✓				69.72	26.95
	✓	✓	✓	✓	✓	60.67	33.15
✓	✓		✓	✓	✓	87.08	80.79
✓	✓	✓		✓	✓	87.69	81.94
✓	✓	✓	✓		✓	86.57	82.86
✓	✓	✓	✓	✓		86.78	80.49
✓	✓	✓	✓	✓	✓	**88.32**	**86.15**

Table 3. Ablation studies of proposed modules.

$\mathcal{L}_{cls}$	$\mathcal{L}_m$	$\mathcal{L}_b$	$\mathcal{L}_{div}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{rec}$	UCF (AUC %)	XD (AP %)
✓						63.73	26.11
✓	✓					87.58	85.40
✓	✓	✓				87.92	85.55
✓	✓	✓	✓			86.80	85.58
✓	✓	✓		✓	✓	87.72	84.65
✓	✓	✓	✓	✓		88.20	85.88
✓	✓	✓	✓	✓	✓	**88.32**	**86.15**

Table 4. Ablation studies of proposed loss functions.

4.4. Ablation Studies

Effect of proposed modules. Table 3 summarizes the influence of each module. As our method directly performs anomaly detection on augmented features, using only MoNE or MoAE leads to unstable learning and poor performance. Removing the external memory M_{ext} leads to a notable drop (-1.24% AUC on UCF, -5.36% AP on XD), highlighting its role in capturing general patterns. Further excluding the routing modules (APR, LR) and their experts substantially degrades performance, showing that both explicit semantic and latent diverse anomaly patterns are indispensable. The Adaptive Fuser (AF) also proves essential by balancing these routing pathways. With all modules enabled, our framework achieves the best results of 88.32% AUC on UCF-Crime and 86.15% AP on XD-Violence, validating the overall design.

Effect of proposed tasks. Table 4 summarizes the contribution of each loss term on UCF-Crime and XD-Violence. Using only the classification loss $\mathcal{L}_{cls}$ yields poor results (63.73% AUC / 26.11% AP), indicating limited supervision. Adding the memory loss $\mathcal{L}_m$ dramatically boosts performance to 87.58% / 85.40%, confirming its key role in memory-guided learning. The balance loss $\mathcal{L}_b$ and regularization loss $\mathcal{L}_{reg}$ provide additional gains by encouraging balanced expert utilization. The distinctiveness loss $\mathcal{L}_{div}$ promotes more diverse representations, and the reconstruction loss $\mathcal{L}_{rec}$ further refines feature alignment. When all losses are jointly optimized, the model achieves the best results (88.32% AUC / 86.15% AP), demonstrating the effectiveness of multi-objective optimization.

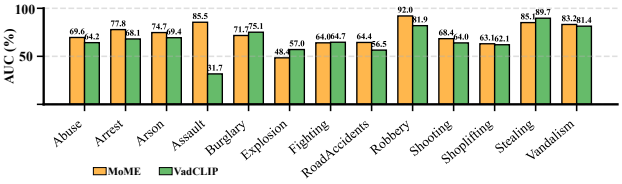


Figure 4. Category-wise performance comparison with VadCLIP.

4.5. Discussions

Robustness Analysis on Multi-Event Videos. Real-world videos often contain multiple anomalies, making it essential to evaluate a model's ability to detect multiple events within a single video. We assess this capability on the XD-Violence dataset, which includes videos with varying numbers of anomalous events, as shown in Table 2. As expected, detection performance generally decreases as the number of events increases. For single-event videos, both MoME and VadCLIP can achieve AP above 90%, while multi-event cases are more challenging. Although MoME performs slightly lower than VadCLIP in a few cases with more than nine events, such videos are extremely rare and have limited statistical significance. Overall, MoME consistently outperforms the baseline across most event counts, demonstrating stronger robustness in multi-event detection.

Per-Class Results. Figure 4 presents per-class results of MoME and the baseline VadCLIP on UCF-Crime. MoME achieves superior performance in most categories, with notable improvements in those involving complex dynamics. For instance, the AUC increases by 53.8% in *Assault* and by 10.1% in *Robbery*. These gains demonstrate the advantage of jointly modeling general and semantically diverse patterns. Further analyses are provided in the supplementary material.

Cross-dataset results. For a video anomaly detector to be practical in real-world scenarios, strong generalization across different domains is crucial. To assess this capability, we perform cross-dataset evaluations between UCF-Crime and XD-Violence, as reported in Table 5. Transferring from UCF to XD is notably more challenging due to the higher diversity and complex anomaly scenes in XD-Violence, whereas the reverse direction (XD to UCF) is relatively easier. Despite these challenges, MoME consis-

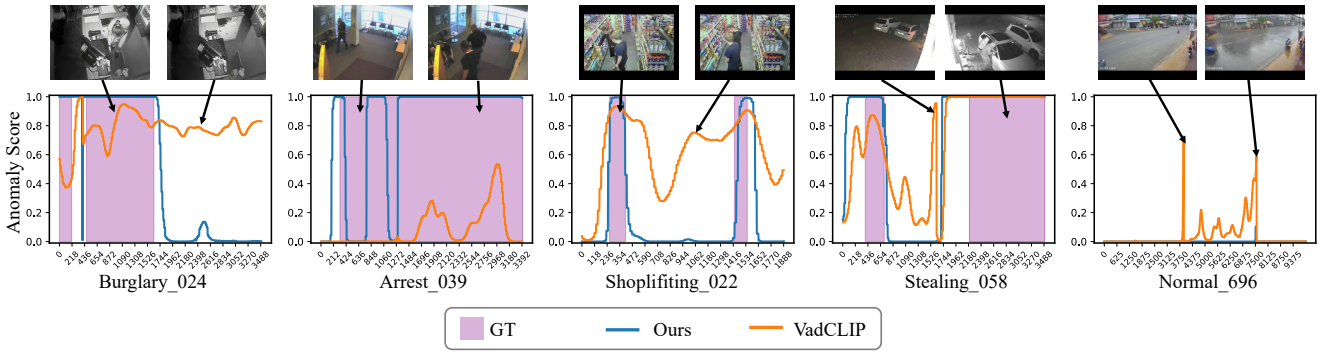


Figure 5. Qualitative results of our method and VadCLIP on the UCF-Crime test videos. The purple shaded regions indicate segments where abnormal events occur. The Y-axis denotes the anomaly scores, and the X-axis corresponds to the video frame indices.

tently achieves better performance and exhibits comparable performance drops in both transfer directions, indicating stronger robustness and cross-domain generalization.

Source	UCF-Crimes	XD-Violence	XD-Violence	UCF-Crimes
Target	UCF-Crimes		XD-Violence	
VadCLIP	88.02	79.95 (↓9.17%)	84.51	53.41 (↓36.80%)
MoME	88.32	**80.77** (↓8.55%)	86.15	**54.08** (↓37.23%)

Table 5. Cross-dataset comparison results (%).

4.6. Qualitative Results

Anomaly Scores. We plot the anomaly scores of MoME and VadCLIP in Fig. 5. In particular, our method results in a high abnormal score even for a *shoplifting* scene where action-related cues is important. Moreover, MoME maintains stable low scores on normal videos, whereas VadCLIP occasionally produces spurious spikes. Additional examples are provided in the supplementary material.

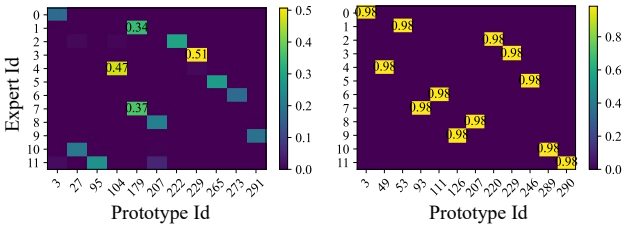


Figure 6. Visualization of expert-prototype allocations. Left: without regularization; Right: with regularization.

Expert-prototype Allocation Matrix. To illustrate the effect of the regularization constraint on the allocation matrix A, we visualize the experts with its most related prototypes in Fig. 6. The left shows training without regularization, while the right includes it. Without regularization, multiple experts focus on similar prototypes, leading to redundant allocations and weak expert-prototype associations. With regularization, experts are encouraged to specialize in distinct prototypes, resulting in more strong and diverse allocations. The full results are available in the supplementary material.

Embedding features. To further evaluate our model's ability to learn discriminative and semantically diverse representations, we visualize the feature distributions on the UCF-Crime test set using t-SNE [20], including the original CLIP features, our model's augmented features grouped by class labels, and those grouped by expert IDs, as illustrated in Fig. 7. The visualization shows that the original CLIP features exhibit substantial overlap and confusion between normal and anomalous samples. In contrast, features extracted by our model achieve a much clearer separation between normal and anomalous patterns. However, grouping features by class labels proves suboptimal, as semantically similar samples may be artificially divided. By comparison, grouping by expert IDs yields more coherent and semantically consistent representations.

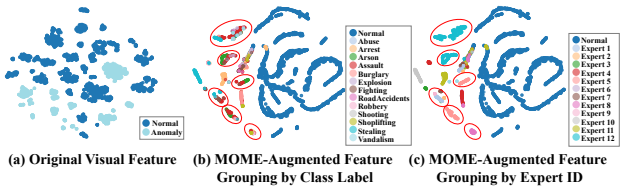


Figure 7. Visualization of augmented features on UCF-Crimes.

5. Conclusions

In this paper, we introduced MoME, a sparse Mixture-of-Memory-Experts framework that unifies the learning of general and diverse patterns for weakly-supervised video anomaly detection. By combining shared external memory for global regularities with expert-specific internal memories for specialization, MoME effectively balances generalization and discrimination. The proposed Anomaly Prototype Router (APR), guided by LLM-derived semantic prototypes, enhances explicit diversity learning, while the Latent Router enables implicit diversity learning and the Adaptive Fuser integrates both explicit and implicit cues. Comprehensive experiments and analyses demonstrate that MoME consistently outperforms prior approaches in both accuracy and generalization.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (U25A20441, 62472238, 62441232) and in part by the Project of Peng Cheng Laboratory (PCL2025A14).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10275–10285, 2025. 2, 5, 6

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[4] Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18319–18329, 2024. 1, 2, 6

[5] Junxi Chen, Liang Li, Yunbin Tu, Li Su, Zhe Xue, and Qingming Huang. Generalizing single-frame supervision to event-level understanding for video anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[6] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12137–12146, 2023. 2

[7] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE Access*, 11:55297–55315, 2023. 1

[8] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 2

[9] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021. 1, 2

[10] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 6

[11] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[12] 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. 6

[13] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 2

[14] 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*, pages 1395–1403, 2022. 1

[15] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 1

[16] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. 2

[17] Tianshan Liu, Kin-Man Lam, and Bing-Kun Bao. Injecting text clues for improving anomalous event detection from weakly labeled videos. *IEEE Transactions on Image Processing*, 33:5907–5920, 2024. 6

[18] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023. 1, 2, 6

[19] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021. 1

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[21] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2665–2674, 2023. 1, 2

[22] 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. 2

[23] 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 supervi-

sion. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 6

[24] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, and Partha Pratim Roy. Anomaly detection in road traffic using visual surveillance: A survey. *Acm Computing Surveys (CSUR)*, 53(6):1–26, 2020. 1

[25] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999. 6

[26] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2, 6

[27] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015. 4

[28] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1, 2, 6

[29] Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, and Xiaojin Gong. Tdsd: Text-driven scene-decoupled weakly supervised video anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5055–5064, 2024. 2

[30] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024. 2

[31] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 2, 6

[32] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21017–21025, 2025. 6

[33] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 4

[34] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 2, 6

[35] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024. 6

[36] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9301–9310, 2024. 6

[37] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 2, 3, 6

[38] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18899–18908, 2024. 2, 6

[39] Chen Zhang, Guorong Li, Yuankai Qi, Hanhua Ye, Laiyun Qing, Ming-Hsuan Yang, and Qingming Huang. Dynamic erasing network with adaptive temporal modeling for weakly supervised video anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 6

[40] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the computer vision and pattern recognition conference*, pages 13843–13853, 2025. 2

[41] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 2

[42] Yuanhong Zhong, Ruyue Zhu, Ge Yan, Ping Gan, Xuerui Shen, and Dong Zhu. Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal mlp. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6

[43] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777, 2023. 6

[44] Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Heng Tao Shen. Batchnorm-based weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[45] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019. 1