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.

Fine-VAD: Towards Fine-Grained Video Anomaly Detection via Progressive Cross-Granularity Learning

Menghao Zhang[1], Yiyan Zhu[1], Pengfei Ren[1], Haifeng Sun[1], Qi Qi[1,2,*], Zirui Zhuang[1], Huazheng Wang[1], Lei Zhang[3], Jianxin Liao[1], Jingyu Wang[1,4,*]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
[2]E-Byte.com, Beijing, China [3]China Unicom, Beijing, China
[4]Pengcheng Laboratory, Shenzhen, China

{zhangmenghao,zhuyiyan,rpf,hfsun,qiqi8266,zhuangzirui}@bupt.edu.cn
zhangl83@chinaunicom.cn; {wanghz,liaojx,wangjingyu}@bupt.edu.cn

Abstract

In this paper, we explore video anomaly detection (VAD) from a fine-grained perspective, which aims not only to detect anomalous events but also to identify their specific categories. Due to the limited number of examples per category, existing methods either fail to handle intra-class variation across diverse contexts or struggle with inter-class confusion caused by shared visual primitives. To address these challenges, we propose a ***progressive cross-granularity learning paradigm*** *that leverages coarse- and fine-grained labels in a complementary manner to progressively refine representations from generic anomaly patterns to category-specific semantics. Building on this paradigm, we develop* ***Fine-VAD****, a progressive alignment framework that aligns video features with supervision signals at multiple granularities. Extensive experiments on two benchmark datasets demonstrate that Fine-VAD achieves up to 47.7% relative improvement in fine-grained anomaly classification. Notably, our paradigm generalizes well across diverse model architectures, offering an adaptable and effective solution for real-world fine-grained VAD.*

1. Introduction

Video Anomaly Detection (VAD) aims to automatically identify events in videos that deviate from expected patterns and has been widely adopted in intelligent surveillance. With the rapid advancement of vision-language technologies [18] and the increasing demands of the real world, research attention is shifting towards ***fine-grained VAD*** [8, 27, 33]. Unlike traditional approaches that merely detect the presence of anomalies, fine-grained VAD aims to further distinguish their specific types, such as arson, assault, or car accidents. This finer categorization is critical in practical scenarios like public safety, where different anomalies require distinct response measures.

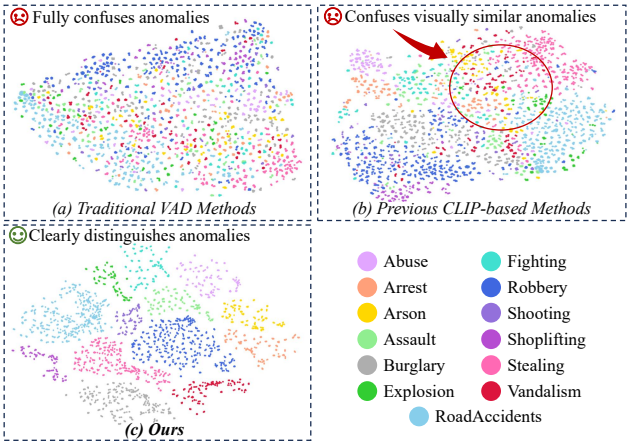

Figure 1. t-SNE visualizations of fine-grained anomaly classification features from the traditional VAD method [21], the CLIP-based method [29], and the proposed Fine-VAD, on UCF-Crime.

Compared to generic fine-grained classification tasks [10, 14], fine-grained VAD presents unique challenges due to its context-dependent nature. Anomalies are typically defined as behaviors that deviate from the contextual norm, meaning that the same visual pattern may correspond to different categories in different scenes. Consequently, two key challenges arise: ***(i) Inter-class confusion from shared visual primitives.*** Semantically distinct anomalies are often

*Corresponding authors.

composed of similar visual or motion primitives. For example, both arson and explosion frequently involve fire and smoke, despite belonging to different categories. This compositional overlap leads to entangled feature representations and blurs the semantic boundaries between categories. ***(ii) Intra-class variation from diverse spatio-temporal contexts.*** The same anomaly category may exhibit drastically different appearances across scenes. For instance, arson may range from a small ignition in an isolated corner to a large-scale blaze engulfing a vehicle in an open area. Such variations in scale, background, and motion make it difficult to learn a consistent category representation.

A straightforward solution for fine-grained VAD is to train models on large-scale datasets with detailed annotations. However, the context-dependent nature of anomalous events makes it nearly infeasible to collect sufficient samples for each category. Earlier works [21, 26, 27] leverage video-level category annotations for supervised learning, but the limited examples per class hinder the ability to capture intra-class variation (see Figure 1(a)). Recent methods [8, 28, 29, 34, 37] seek to alleviate data scarcity by exploiting pre-trained vision-language models (VLMs) such as CLIP [18]. Nevertheless, CLIP lacks fine-grained understanding of object attributes, action dynamics, and interaction semantics [11, 30], leading to inter-class confusion among visually similar anomalies (see Figure 1(b)).

To tackle the challenges of fine-grained VAD, we propose a novel ***progressive cross-granularity learning*** paradigm, grounded in a key observation: although annotated samples for each anomaly category are scarce, coarse-grained labels (i.e., binary anomaly labels) are relatively abundant. Instead of directly learning discriminative features for every category, this paradigm progressively guides the model from capturing generic anomaly representations to acquiring category-specific semantics. On one hand, coarse-grained annotations provide broader coverage and enable the model to learn stable features shared across diverse anomaly types, thereby forming a unified anomaly representation space. On the other hand, fine-grained labels serve to refine this space by imposing semantic boundaries between categories. Through this learning process, the model reduces intra-class variation by anchoring features in shared anomaly patterns, while alleviating inter-class confusion via targeted category-level discrimination.

Building on the proposed progressive cross-granularity learning paradigm, we develop a unified framework for fine-grained VAD, termed **Fine-VAD**. It employs CLIP to align video features with supervision signals of varying granularity across three levels. At the *coarse level*, the model learns generic anomaly representations by aligning all anomalous videos with binary anomaly labels, establishing a unified anomaly space. At the *intermediate level*, visually or semantically related categories are clustered into macro-categories, allowing the model to delineate coarse category boundaries through pseudo-labels. At the *fine level*, ground-truth category labels are introduced to refine representations and strengthen category-specific semantics. Notably, the proposed paradigm is model-agnostic and can be instantiated with either vision-language models or conventional architectures, as validated in Section 4.3.

To evaluate the effectiveness of Fine-VAD, we conduct extensive experiments on two benchmark datasets, UCF-Crime [21] and XD-Violence [26]. Compared with state-of-the-art methods, Fine-VAD achieves superior performance in fine-grained anomaly detection, yielding a remarkable 47.7% relative improvement. These results not only highlight the superiority of our framework, but also demonstrate the robustness of our paradigm across diverse architectures.

Our contributions are as follows:

- A **progressive cross-granularity learning paradigm** is proposed for fine-grained VAD, which gradually shifts learning from generic to category-specific anomaly representations, bridging the granularity gap in supervision.
- A simple yet effective framework, **Fine-VAD**, is designed to instantiate the proposed paradigm, aligning video features with multi-granularity supervision to jointly reduce intra-class variation and inter-class confusion.
- Extensive experiments show that Fine-VAD **outperforms SOTA methods in fine-grained VAD by up to 47.7%**, achieves superior coarse-grained detection performance, and generalizes well across architectures.

2. Related Work

2.1. Fine-Grained Video Anomaly Detection

Most existing VAD studies focus on determining whether a video is anomalous, without further categorization. This includes both semi-supervised methods trained solely on normal data [15, 19, 25, 31, 38, 39] and weakly supervised methods relying on video-level labels [12, 13, 16, 23, 36].

In real-world scenarios, however, identifying the specific type of anomalous event is equally critical, which has motivated the development of fine-grained VAD. Sultani et al. [21] and Wu et al. [26] introduce weakly supervised datasets with category annotations, UCF-Crime and XD-Violence, respectively. These datasets enable models to classify anomalies into specific categories but remain limited by the small number of samples per class. Early work [27] introduces audio information to complement the limited visual data. With the advent of vision-language models (VLMs), especially CLIP [18], subsequent studies attempt to leverage their pretrained feature representations for fine-grained VAD. For instance, Zanella et al. [34] and Wu et al. [29] directly use label texts to enhance visual representation, while later methods [17, 28] integrate CLIP encodings to synthesize pseudo-anomalies. Nevertheless, since CLIP is pre-

trained via image-text alignment, it primarily captures visual appearance rather than fine-grained semantics [11], often leading to inter-class confusion among visually similar anomalies. To overcome this limitation, we propose a progressive cross-granularity learning paradigm that progressively guides the model from learning generic anomaly patterns to acquiring category-specific semantics.

2.2. Multi-Task and Staged Learning in VAD

Several semi-supervised VAD studies have explored multi-task or staged learning strategies that involve multiple objectives, partially resembling our paradigm. These works primarily aim to learn more robust representations of normal patterns. For instance, methods such as [4, 5, 22, 39] introduce auxiliary tasks like prediction or reconstruction to model normal behavior across varying spatial and temporal scales. Another line of research [20, 38] adopts staged training schemes, where tasks are executed in order to extract features of normality. In contrast, our progressive learning paradigm is not designed to model normality but to construct a discriminative feature space that facilitates the separation of different anomaly categories. Unlike multi-task learning, where auxiliary objectives serve for feature enrichment, our cross-granularity supervision provides progressive semantic guidance toward fine-grained VAD.

3. Method

3.1. Problem Definition

The fine-grained VAD task assumes that only video-level labels are available during training. Given a video v, we denote its binary anomaly label as $y_b \in \{0, 1\}$ and its category label as $y_m \in \{0, 1, \ldots, M\}$, where M represents the number of *anomalous* categories and the index 0 is reserved for the *normal* class. A video is labeled as normal if none of its frames contain anomalies, i.e., $y_b = 0$ and $y_m = 0$; otherwise, if at least one frame exhibits an anomalous event, it is labeled as anomalous ($y_b = 1$) with a specific anomaly category $y_m = m$, where $m \in \{1, \ldots, M\}$. Following prior fine-grained VAD settings [8, 27], we assume that each video contains at most one dominant anomaly type.

3.2. Progressive Cross-Granularity Framework

To address the dual challenges of intra-class variation and inter-class confusion, we propose a ***progressive cross-granularity learning*** paradigm that gradually shifts learning from generic to category-specific anomaly representations through hierarchical supervision signals of increasing semantic granularity. Given a video representation f, the learning proceeds as: (1) At the *coarse level*, the model is trained to distinguish anomalies from normal instances using abundant binary labels y_b. This step establishes a unified feature space for all anomaly types, enabling f to capture generic anomaly patterns and mitigate intra-class variation by anchoring diverse instances to shared representations. (2) At the *intermediate level*, we introduce pseudo macro-category labels y_p by grouping semantically or visually similar categories. This encourages f to learn coarse-grained category boundaries and prevents the model from collapsing under sparse fine-grained supervision. (3) At the *fine level*, we refine f using sparse ground-truth category labels y_m to enhance category-specific semantics and resolve inter-class confusion among visually similar anomalies.

Instantiation. To instantiate this paradigm in practice, we build on the pre-trained CLIP due to its strong cross-modal alignment ability, which provides a reliable foundation for constructing multi-granularity video-text correspondences. We therefore instantiate the proposed paradigm as a unified framework named **Fine-VAD**, as illustrated in Fig. 2. For visual feature extraction, following previous work [29], we uniformly sample dense video frames from each input video at intervals of 16 frames to obtain a frame sequence V. These frames are then encoded by the frozen CLIP image encoder E_{image} to produce frame-level visual features. To bridge the domain gap between images and videos in CLIP, we adopt the temporal modeling strategy of [29], which employs a Local-and-Global Temporal Adapter (LGT-Adapter) to capture temporal dependencies within and across frames:

$$f_V = \text{LGT}\big(E_{\text{image}}(V)\big). \tag{1}$$

It is worth noting that the LGT-Adapter serves only as a reference implementation. As demonstrated in Section 4, the proposed paradigm consistently improves performance across different architecture and adapter designs (Tables 1 and 2), highlighting its model-agnostic nature.

3.3. Progressive Cross-Granularity Learning

Building upon the framework, Fine-VAD implements the proposed paradigm through a three-level hierarchical alignment process that progressively incorporates supervision from binary, pseudo, and category labels.

Coarse Level. Based on the temporally enhanced representation f_V, the model is first guided to distinguish anomalous from normal videos using binary labels y_b. A feed-forward layer projects f_V into a frame-level feature sequence $f_{\text{video}} \in \mathbb{R}^{n \times d}$, where n denotes the number of frames and d is the feature dimension. Two text embeddings for the coarse classes (normal and anomalous) are obtained from the frozen CLIP text encoder, forming $E_{\text{coarse}} = [e_{\text{norm}}; e_{\text{anom}}] \in \mathbb{R}^{2 \times d}$. The coarse alignment map $M_{\text{coarse}} \in \mathbb{R}^{n \times 2}$ is computed via cosine similarity between the visual and textual features:

$$M_{\text{coarse}} = \text{sim}(f_{\text{video}}, E_{\text{coarse}}), \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity computed on normalized features. Each element $M_{\text{coarse}}(i, j)$ measures the

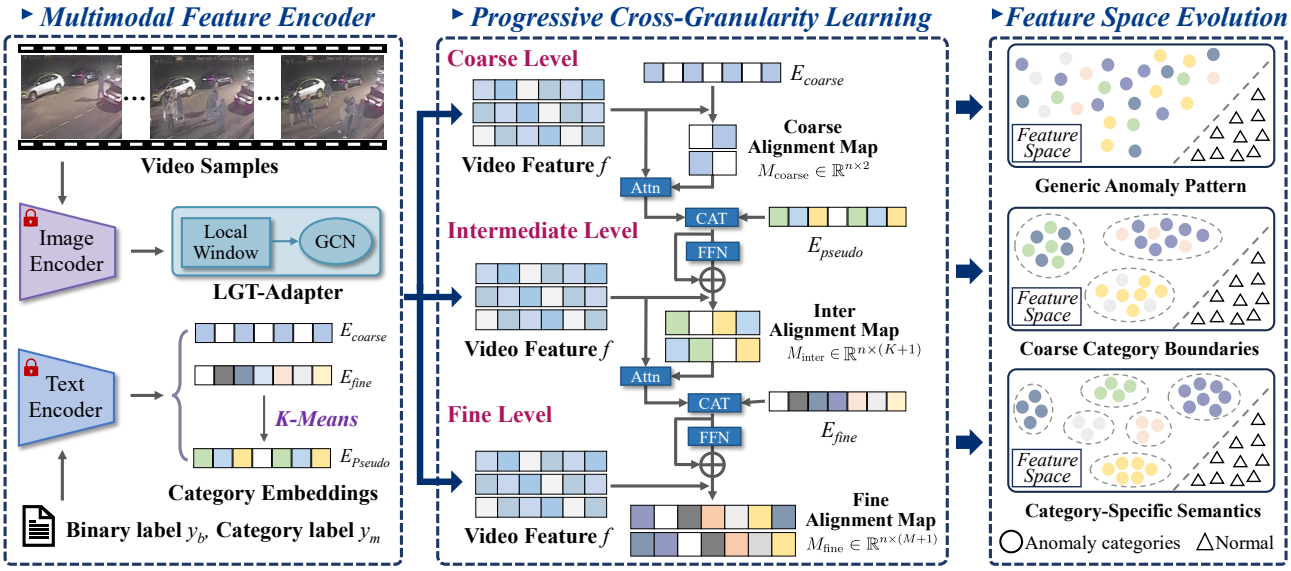

Figure 2. Overall framework of **Fine-VAD**. The proposed progressive cross-granularity paradigm aligns video features with multi-level supervision, guiding the feature space to evolve from generic anomaly patterns to category-specific semantics.

similarity between the i-th frame and the j-th coarse text embedding. This alignment establishes a unified anomaly representation space that captures generic anomaly patterns shared across diverse categories, providing a stable foundation for the subsequent hierarchical learning.

Intermediate Level. At the intermediate level, the model is guided to establish pseudo boundaries for anomaly categories by clustering semantically similar types into K macro-categories. We first encode the ground-truth category labels y_m into embeddings $E_{\text{fine}} = \{e_0, e_1, \ldots, e_M\}$, where each $e_i \in \mathbb{R}^d$ denotes the text embedding of an anomaly category. To group similar anomaly types, we apply standard K-means to partition the anomalous embeddings $\{e_1, \ldots, e_M\}$ into K pseudo macro-categories:

$$\arg\min_{\{C_k\}_{k=1}^K} \sum_{k=1}^{K} \sum_{e_i \in C_k} \|e_i - \mu_k\|^2, \mu_k = \frac{1}{|C_k|} \sum_{e_i \in C_k} e_i. \quad (3)$$

where μ_k is the centroid of cluster C_k. We set the pseudo embedding for the k-th macro-category as $E^k_{\text{pseudo}} := \mu_k$. Together with the normal prototype e_0, the augmented set $E^{\text{aug}}_{\text{pseudo}} = [\, e_0;\, E_{\text{pseudo}}\,]$ contains $K{+}1$ prototypes.

To inject coarse-level semantics, we employ a cross-attention mechanism that uses the coarse alignment map as the query and the video features as key-value pairs, producing guided embeddings $E^{\text{coarse}}_{\text{guide}} \in \mathbb{R}^{n\times d}$:

$$E^{\text{coarse}}_{\text{guide}} = \text{Attention}(M_{\text{coarse}}, f_{\text{video}}). \quad (4)$$

We then refine the pseudo-category embeddings through a lightweight feed-forward fusion:

$$\hat{E}_{\text{pseudo}} = \text{FFN}\big([\, E^{\text{aug}}_{\text{pseudo}}, \text{pool}(E^{\text{coarse}}_{\text{guide}})\,]\big) + E_{\text{pseudo}}, \quad (5)$$

where $[\cdot,\cdot]$ denotes vector concatenation and $\text{pool}(\cdot)$ averages frame-wise guidance into a global context vector.

Finally, the intermediate-level alignment map $M_{\text{inter}} \in \mathbb{R}^{n\times(K+1)}$ is computed as:

$$M_{\text{inter}} = \text{sim}(f_{\text{video}}, \hat{E}_{\text{pseudo}}). \quad (6)$$

Fine Level. At the fine level, the model focuses on category-specific semantics by aligning the ground-truth category labels y_m with the video representations. This stage is further informed by the intermediate-level alignment map M_{inter} learned previously. We compute the attention-guided embedding $E^{\text{inter}}_{\text{guide}}$ and the refined category embeddings $\hat{E}_{\text{fine}}$ following Eq. (4) and (5), respectively. The fine alignment map $M_{\text{fine}} \in \mathbb{R}^{n\times(M+1)}$ is computed as:

$$M_{\text{fine}} = \text{sim}(f_{\text{video}}, \hat{E}_{\text{fine}}). \quad (7)$$

This fine-level alignment enables the model to capture subtle category distinctions and resolve inter-class confusion among visually similar anomalies. In summary, each alignment stage inherits guidance from the previous one via attention-based embedding refinement (Eqs. 4-5), forming a continuous semantic evolution $M_{\text{coarse}} \to M_{\text{inter}} \to M_{\text{fine}}$.

3.4. Training and Inference

Training Objective. The training objective follows the progressive nature of the proposed paradigm, ensuring stable optimization from coarse to fine granularities. During training, both the CLIP image and text encoders are kept frozen. At each stage, the model generates an alignment map $M \in \mathbb{R}^{n\times m}$ that measures frame-wise similarities between video features and category embeddings. A top-T

pooling operation is then applied to each row of M, and the selected values are averaged to produce a category-level similarity vector $S = \{s_0, s_1, \ldots, s_m\}$, which serves as the input to all subsequent loss functions.

Coarse Level. At the *coarse level*, where data is abundant, we jointly optimize a binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ and a contrastive loss $\mathcal{L}_{\text{cts}}$. The binary loss is formulated as

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log s_{i,1} + (1-y_i)\log(1-s_{i,1})\,], \quad (8)$$

where $s_{i,1}$ denotes the anomaly score after top-T pooling. The contrastive term is implemented following the InfoNCE [7], encouraging compact anomaly representations that are well separated from the normal manifold. This stage establishes a stable anomaly-aware feature space, providing a foundation for finer-grained learning.

Intermediate and Fine Levels. As supervision becomes sparser, the optimization focus gradually shifts from general separation to targeted discrimination. At the *intermediate level*, no explicit supervised loss is applied to avoid overfitting to noisy pseudo-labels. Instead, this stage serves as a structural regularizer that refines the feature space through the soft attention guidance, and provides macro-category priors $\mathcal{C}_{p(i)}$ used in the weighting scheme of Eq. (9).

At the *fine level*, we adopt a weighted cross-entropy loss to enhance discrimination among similar categories:

$$\begin{aligned} \mathcal{L}_{\text{refine}} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{M}\alpha_{i,j}\, y_{i,j}\log s_{i,j}, \\ \alpha_{i,j} &= \begin{cases} \omega, & \text{if } j \in \mathcal{C}_{p(i)}, \\ 1, & \text{otherwise}, \end{cases} \end{aligned} \quad (9)$$

where $y_{i,j}$ is the ground-truth indicator, $s_{i,j}$ the similarity score, and $\mathcal{C}_{p(i)}$ the pseudo macro-category of video i. The weighting factor $\omega > 1$ amplifies intra-group discrimination within visually related categories, encouraging finer separation across closely aligned semantics. Contrastive loss is not applied at this stage, as its push-pull mechanism may disrupt intra-class compactness under limited supervision, as further validated in Section 4.

Overall Loss. The final training objective is given by:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \lambda_1 \mathcal{L}_{\text{cts}} + \lambda_2 \mathcal{L}_{\text{refine}}, \quad (10)$$

where λ_1 and λ_2 are balancing hyperparameters.

Inference. During inference, we apply a threshold σ to the fine-level alignment map M_{fine}. Each element in M_{fine} denotes the similarity between a frame and a fine-grained anomaly category embedding. A frame is assigned to a specific anomaly category if its similarity score exceeds a predefined threshold σ; otherwise, it is classified as normal.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our method on two benchmarks that provide category-level annotations for anomalous events: **UCF-Crime** [21] and **XD-Violence** [26]. UCF-Crime contains 1,900 surveillance videos collected from 13 scenes, covering 13 types of anomalies and normal events. XD-Violence consists of 4,754 untrimmed videos from diverse sources such as movies, web videos, and CCTV recordings, focusing on six violence-related categories.

Evaluation Metrics. For *fine-grained* VAD, we adopt the standard video action detection protocol and compute the mean Average Precision (mAP) under multiple Intersection-over-Union (IoU) thresholds from 0.1 to 0.5. The average mAP (AVG) is also reported, and mAP is computed only for abnormal videos in the test set.

4.2. Implementation Details

For the backbone, we adopt the frozen image and text encoders from the pre-trained CLIP (ViT-B/16) [18]. The LGT-Adapter follows the original implementation of [29] without any modification to ensure a fair comparison. For hyperparameter settings, the number of pseudo macro-categories K in Eq. (3) is set to 4, and the MIL aggregation parameter T is fixed at 16. In Eq. (10), we set $\lambda_1 = 1$ and $\lambda_2 = 1$ for UCF-Crime, and $\lambda_1 = 1 \times 10^{-4}$ and $\lambda_2 = 1$ for XD-Violence. All models are trained using the AdamW optimizer with a batch size of 64 and a learning rate of 1×10^{-5}. Training is conducted for 10 epochs on a single NVIDIA RTX 4090 GPU.

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

We compare the proposed Fine-VAD framework with existing *fine-grained* VAD methods [21, 27–29, 34]. Given the limited number of prior works addressing fine-grained VAD, we further include representative approaches from three related paradigms: (1) SOTA *image-based* anomaly detection [6, 9], (2) *weakly supervised* VAD [2, 16, 24, 40], and (3) *MLLM-based* approaches [1, 8, 32, 35]. Here, MLLMs denote pretrained multimodal large language models with strong generalization and reasoning capabilities.

Fine-Grained VAD Results. Table 1 presents a quantitative comparison between Fine-VAD and prior methods on UCF-Crime and XD-Violence. Fine-VAD achieves an average mAP of 14.99% on UCF-Crime and 31.87% on XD-Violence, consistently outperforming all competing methods. These results further suggest that existing approaches, including recent MLLM-based methods, still struggle to distinguish semantically similar anomaly categories, underscoring the challenge of fine-grained VAD. Notably, on UCF-Crime, which contains more anomaly categories and greater intra-class variation, Fine-VAD surpasses the previ-

Table 1. **Comparison of fine-grained video anomaly detection performance**. Methods are grouped by paradigm from top to bottom: image-based, weakly supervised, fine-grained, MLLM-based, and our proposed variants.

Method	UCF-Crime (mAP@IoU%)						XD-Violence (mAP@IoU%)					
	0.1	0.2	0.3	0.4	0.5	AVG	0.1	0.2	0.3	0.4	0.5	AVG
WinCLIP [9] CVPR'23	2.14	0.94	0.48	0.30	0.12	0.80	4.69	2.26	1.25	0.84	0.23	1.85
UniVAD [6] CVPR'25	3.22	1.98	1.27	0.88	0.31	1.53	5.19	3.02	1.84	1.07	0.69	2.36
UR-DMU [40] AAAI'23	4.94	2.62	1.09	0.92	0.64	2.04	1.17	0.82	0.58	0.35	0.18	0.62
MGFN [2] AAAI'23	2.88	2.69	1.67	1.55	1.49	2.06	6.24	4.86	3.67	2.31	1.70	3.76
PEL4VAD [16] TIP'24	2.91	2.75	1.57	1.48	1.45	2.03	5.29	4.08	2.64	1.87	1.53	3.08
FedCLIP [24] AAAI'25	2.72	2.16	1.43	1.17	0.98	1.70	4.68	4.21	2.14	1.52	1.49	2.81
RealAD [21] CVPR'18	5.73	4.41	2.69	1.93	1.44	3.24	22.72	15.57	9.98	6.20	3.78	11.65
AVVD [27] TMM'23	10.27	7.01	6.25	3.42	3.29	6.05	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP [29] AAAI'24	11.72	7.83	6.40	4.53	2.93	6.68	37.03	30.84	23.38	17.90	14.31	24.70
AnomalyCLIP [34] CVIU'24	6.29	3.64	2.04	0.88	0.35	2.64	17.26	14.29	10.54	6.57	3.28	10.39
OVVAD [28] CVPR'24	10.23	7.88	5.34	4.43	2.95	6.17	19.57	13.90	9.91	6.35	4.16	10.78
LAVAD [35] CVPR'24	10.21	7.88	5.52	4.05	3.02	6.14	33.60	28.19	20.13	13.26	7.89	20.61
ExVAD [8] ICML'25	16.51	12.35	9.41	7.82	4.65	10.15	40.14	32.75	28.78	20.15	18.35	28.23
VERA [32] CVPR'25	8.22	6.09	4.84	3.56	2.01	4.94	25.13	18.14	15.31	13.29	10.54	16.48
Qwen2.5-VL-7B [1] 2025	7.63	5.61	4.20	3.19	1.86	4.50	19.26	15.43	13.26	10.04	7.49	13.09
Fine-VAD (Ours)	**21.43**	**18.62**	**16.25**	**11.41**	**7.26**	**14.99**	**44.06**	**36.69**	**31.22**	**25.80**	**21.58**	**31.87**
Fine_{CNN}	11.58	7.99	5.31	3.86	2.26	6.21	24.69	17.88	11.93	8.10	5.28	13.58
$\text{Fine}_{\text{MLLM}}$	19.24	14.92	12.87	10.39	7.42	12.97	39.35	32.24	27.90	21.58	19.03	28.02

ous best method [8] by 47.7% in average mAP. Although the method in [8] builds on the large-scale LLaMA 3.1 [3], our approach achieves consistently better performance across all IoU thresholds without relying on any large multimodal model, further validating the effectiveness of the proposed progressive cross-granularity design.

Paradigm Adaptability across Architectures. As shown in Table 1, we further examine the adaptability of the proposed progressive cross-granularity learning paradigm across two distinct architectures: a conventional CNN-based feature extractor and a MLLM. Specifically, we construct $\textbf{Fine}_{\textbf{CNN}}$ by incorporating the three-stage learning scheme into the I3D backbone [21], which improves the average mAP over the I3D baseline by 2.97% on UCF-Crime and 1.93% on XD-Violence. We also build $\textbf{Fine}_{\textbf{MLLM}}$ by instantiating the same paradigm with a progressive prompting strategy on Qwen2.5-VL-7B, yielding average mAP gains of 8.47% and 14.93% on the two datasets, respectively. Despite their substantially different modeling mechanisms, both variants benefit consistently from the coarse-to-fine supervision process. These results indicate that the effectiveness of our approach originates from the learning paradigm itself, rather than from any particular model architecture.

4.4. Ablation Studies

We conduct comprehensive ablation studies on the UCF-Crime dataset to evaluate the contribution of each hierarchical level, component, and loss term. UCF-Crime is selected for its larger number of anomaly categories.

Table 2. **Ablation study on hierarchical levels and temporal adapters** on the UCF-Crime dataset.

ID	Temporal Adapter	Hierarchical Level			AVG mAP%
		Coarse	Inter	Fine	
1				✓	8.64
2	LGT	✓		✓	10.93
3	-Adapter [29]		✓	✓	11.60
4		✓	✓	✓	**14.99**
5	-	✓	✓	✓	5.62
6	Transformer	✓	✓	✓	12.01
7	GCN	✓	✓	✓	12.13
8	Axial TF [34]	✓	✓	✓	13.88
9	TA [28]	✓	✓	✓	14.56

Ablation on Each Hierarchical Level. Table 2 (ID 1-4) summarizes the ablation results of each level in our progressive cross-granularity paradigm. The baseline (ID 1), trained solely with sparse category labels, achieves only 8.64% average mAP, indicating that limited supervision is insufficient to learn discriminative representations. Introducing coarse-level alignment (ID 2) yields a +2.29% mAP improvement, demonstrating that modeling general anomaly patterns provides a strong foundation for subsequent learning. Further incorporating intermediate-level

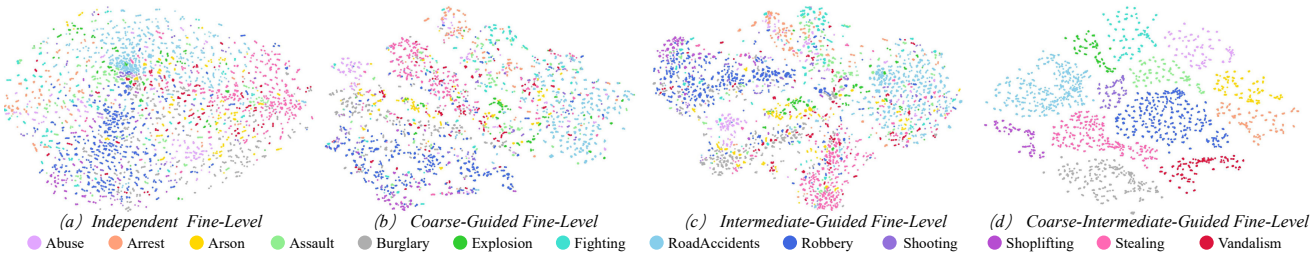

Figure 3. t-SNE visualization of feature distributions on UCF-Crime under different supervision levels: (a) fine-only, (b) +coarse, (c) +intermediate, and (d) full three-level supervision.

guidance (ID 3) brings an additional +2.96% gain, underscoring the importance of coarse category structuring before fine-level refinement. The full paradigm (ID 4) reaches 14.99% mAP, verifying our key insight that the stages are not merely additive but mutually reinforcing: coarse-level alignment reduces intra-class variation, intermediate guidance introduces semantic structure, and fine-level learning resolves inter-class confusion. As shown in Fig. 3, the feature distributions become progressively more compact and separable across supervision stages, validating the effectiveness of hierarchical guidance.

Robustness to the Temporal Adapter. Table 2 (ID 4-9) compares Fine-VAD equipped with different temporal modeling modules. Removing the temporal adapter entirely (ID 5) causes a drastic performance drop to 5.62% mAP, confirming that temporal reasoning is essential for fine-grained anomaly understanding. Among all variants, the LGT-Adapter yields the best performance, benefiting from its ability to jointly model local and global temporal dependencies. Nevertheless, when replacing it with other adapters,including a vanilla Transformer encoder, GCN, Axial Transformer [35], or TA module [28], our framework still achieves over 12% mAP, surpassing previously reported state-of-the-art methods. This demonstrates that Fine-VAD's effectiveness stems from the generality of the proposed progressive cross-granularity learning paradigm, rather than any specific temporal architecture.

Ablation on Loss Functions. Table 3 presents the ablation results for different loss configurations across hierarchical levels, validating the effectiveness of our loss design. We first examine the coarse-level contrastive loss $\mathcal{L}_{\text{cts}}$. Comparing ID 1 (without $\mathcal{L}_{\text{cts}}$, 10.93% mAP) and ID 2 (with $\mathcal{L}_{\text{cts}}$, 12.90% mAP) reveals a clear gain of +1.97%, indicating that forming a compact anomaly-aware feature space early in training provides a stronger foundation for subsequent stages. At the ***intermediate level***, the full model (ID 6, 14.99% mAP) omits supervised losses at this stage, avoiding overfitting to noisy pseudo-labels. Adding $\mathcal{L}_{\text{mce}}$ (ID 7) or $\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$ (ID 8) leads to substantial performance drops of -2.41% and -3.03%, respectively, confirming that this stage functions best as a structural regularizer rather than a supervised branch. At the ***fine level***, adding $\mathcal{L}_{\text{cts}}$ on top of $\mathcal{L}_{\text{refine}}$ (ID 9) slightly reduces performance (-0.13%), suggesting that $\mathcal{L}_{\text{refine}}$ already provides sufficient fine-grained discrimination, while additional contrastive supervision may introduce instability under label sparsity.

Table 3. **Ablation of loss functions at each hierarchical level**. $\mathcal{L}_{\text{bce}}$, $\mathcal{L}_{\text{mce}}$, $\mathcal{L}_{\text{cts}}$, and $\mathcal{L}_{\text{refine}}$ represent binary cross-entropy, multi-class cross-entropy, contrastive loss, and the proposed refinement loss, respectively.

ID	Coarse Level	Inter Level	Fine Level	AVG mAP%
1	$\mathcal{L}_{\text{bce}}$	–	$\mathcal{L}_{\text{mce}}$	10.93
2	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	–	$\mathcal{L}_{\text{mce}}$	12.90
3	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	–	$\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$	12.93
4	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{mce}}$	$\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$	13.84
5	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$	13.92
6	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	–	$\mathcal{L}_{\text{refine}}$	**14.99**
7	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{mce}}$	$\mathcal{L}_{\text{refine}}$	12.58
8	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{mce}}+\mathcal{L}_{\text{cts}}$	$\mathcal{L}_{\text{refine}}$	11.96
9	$\mathcal{L}_{\text{bce}}+\mathcal{L}_{\text{cts}}$	–	$\mathcal{L}_{\text{cts}}+\mathcal{L}_{\text{refine}}$	14.86
10	$\mathcal{L}_{\text{bce}}$	–	$\mathcal{L}_{\text{refine}}$	12.07

Table 4. Effect of K on AVG (%) performance.

Value of K	2	3	4	5
UCF-Crime	9.14	12.96	**14.99**	12.37
XD-Violence	27.74	30.92	**31.87**	25.88

Overall, the results demonstrate that our loss design effectively addresses the challenge of sparse fine-grained labels: a strong coarse-level objective builds a robust representation base, a loss-free intermediate stage regularizes the feature space, and a targeted fine-level refinement loss ensures precise anomaly categorization.

Analysis of K. We study the effect of K, the number of pseudo macro-categories that controls the granularity of intermediate supervision. This hyperparameter reflects a trade-off between semantic discrimination and sample sufficiency: a small K may merge semantically distinct anomaly

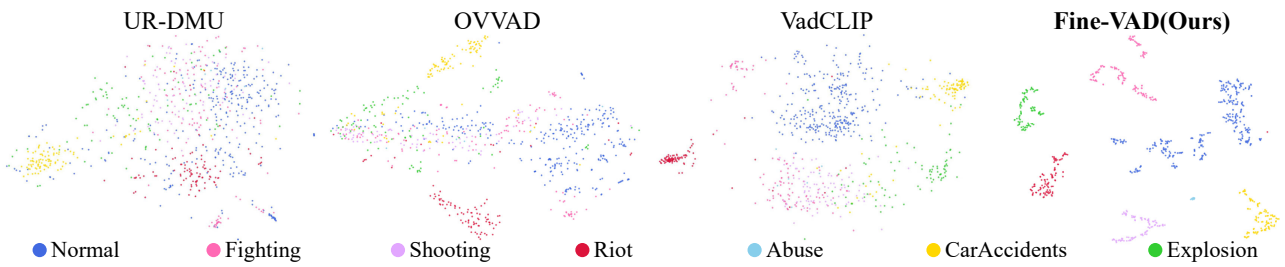

Figure 4. Qualitative comparison of t-SNE feature visualizations on the XD-Violence dataset across different methods.

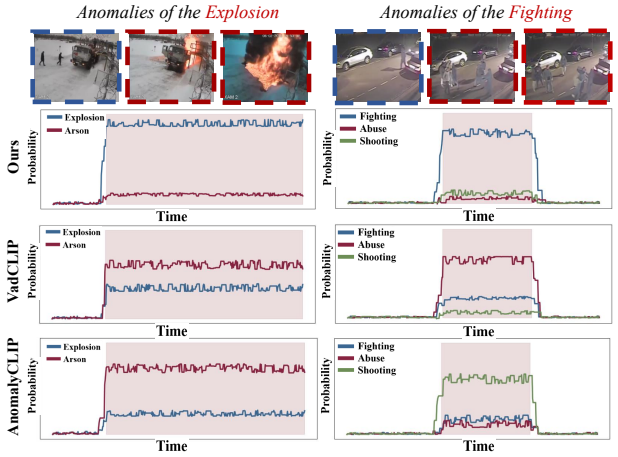

Figure 5. Qualitative comparison in fine-grained VAD. From top to bottom: predictions by Ours, VadCLIP [29], and AnomalyCLIP [34]. Pink regions denote ground-truth anomalies.

categories, while a large K can lead to data sparsity and unstable clustering due to the limited number of samples in each group. As shown in Table 4, the results empirically confirm this trade-off and show that K=4 achieves the best average mAP on both UCF-Crime and XD-Violence, which is therefore adopted as the default setting. Interestingly, on XD-Violence, which contains fewer anomaly categories, K=3 performs comparably to the optimal setting. This observation indicates that the optimal intermediate granularity is dataset-dependent and closely related to the complexity of the category space.

4.5. Qualitative Results

In addition to the qualitative comparisons on UCF-Crime shown in Fig. 1, Fig. 4 visualizes feature distributions on XD-Violence and compares them with state-of-the-art methods [28, 29, 40]. Fine-VAD yields a feature space with clearer inter-class separation and more compact intra-class clusters, whereas the compared methods exhibit more severe category overlap. Fig. 5 further shows that Fine-VAD can effectively distinguish visually similar anomaly categories, while existing methods frequently confuse them. For example, VadCLIP [29] misclassifies *Explosion* and *Fighting* as *Arson* and *Abuse*, respectively, whereas AnomalyCLIP [34] predicts them as *Arson* and *Shooting*. These results further verify that Fine-VAD learns more discriminative fine-grained semantics than existing approaches.

5. Conclusion

In this paper, we tackle the problem of fine-grained VAD, where the scarcity of per-category samples limits the learning of robust category-specific features. To address this challenge, we introduce a progressive cross-granularity learning paradigm that leverages abundant coarse-grained labels to construct a generic anomaly representation space, which in turn facilitates fine-grained classification. We implement this paradigm through a hierarchical alignment framework, termed Fine-VAD. Extensive experiments demonstrate that Fine-VAD not only outperforms existing methods across multiple benchmarks but also generalizes well across different architectures, confirming the model-agnostic nature of the proposed paradigm.

Limitations and Future Work. Our current study focuses on videos containing a single anomaly type. In future work, we will extend Fine-VAD to more realistic scenarios involving multiple co-occurring anomalies and investigate how interactions among anomalies affect fine-grained recognition.

Acknowledgments

This work was supported in part by the National Key R&D Program of China 2024YFE0200800,the National Natural Science Foundation of China under Grants (62406039, 62321001, 62471055, U23B2001, 62101064, 62171057, 62201072), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China(JYB2025XDXM107), the High-Quality Development Project of the MIIT(2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the Fundamental Research Funds for the Central Universities (2024PTB-004), the 2025 Education and Teaching Reform Project Funding at Beijing University of Posts and Telecommunications (2025YZ005).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 5, 6

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 5, 6

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 6

[4] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*, pages 12742–12752, 2021. 3

[5] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4505–4523, 2022. 3

[6] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *CVPR*, pages 15194–15203, 2025. 5, 6

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. 5

[8] Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *ICML*, pages 25750–25761, 2025. 1, 2, 3, 5, 6

[9] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pages 19606–19616, 2023. 5, 6

[10] Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In *AAAI*, pages 2570–2578, 2024. 1

[11] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based CLIP for better fine-grained understanding. In *NeurIPS*, 2024. 2, 3

[12] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, pages 1395–1403, 2022. 2

[13] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, pages 8022–8031, 2023. 2

[14] Zhen-Xiang Ma, Zhen-Duo Chen, Tai Zheng, Xin Luo, Zixia Jia, and Xin-Shun Xu. Few-shot fine-grained image classification with progressively feature refinement and continuous relationship modeling. In *AAAI*, pages 6036–6044, 2025. 1

[15] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *CVPR*, pages 8310–8320, 2025. 2

[16] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.*, 33:4923–4936, 2024. 2, 5, 6

[17] Zhangbin Qian, Jiawei Tan, Zhilong Ou, and Hongxing Wang. Clip-driven multi-scale instance learning for weakly supervised video anomaly detection. In *ICME*, pages 1–6, 2024. 2

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 5

[19] Nicolae-Catalin Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, and Mubarak Shah. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *CVPR*, pages 15984–15995, 2024. 2

[20] Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. In *ICCV*, pages 10296–10306, 2023. 3

[21] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 2, 5, 6

[22] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *CVPR*, pages 22846–22856, 2023. 3

[23] 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 *ICCV*, pages 4955–4966, 2021. 2

[24] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *AAAI*, pages 21017–21025, 2025. 5, 6

[25] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *ECCV*, pages 494–511, 2022. 2

[26] 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 *ECCV*, pages 322–339, 2020. 2, 5

[27] Peng Wu, Xiaotao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685, 2023. 1, 2, 3, 5, 6

[28] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *CVPR*, pages 18297–18307, 2024. 2, 6, 7, 8

[29] 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 *AAAI*, pages 6074–6082, 2024. 1, 2, 3, 5, 6, 8

[30] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. In *ICML*, pages 68777–68793, 2025. 2

[31] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023. 2

[32] Muchao Ye, Weiyang Liu, and Pan He. VERA: explainable video anomaly detection via verbalized learning of vision-language models. In *CVPR*, pages 8679–8688, 2025. 5, 6

[33] Hongyang Yu, Xinfeng Zhang, Yaowei Wang, Qingming Huang, and Baocai Yin. Fine-grained accident detection: Database and algorithm. *IEEE Trans. Image Process.*, 33: 1059–1069, 2024. 1

[34] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into CLIP latent space for video anomaly recognition. *Comput. Vis. Image Underst.*, 249:104163, 2024. 2, 5, 6, 8

[35] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *CVPR*, pages 18527–18536, 2024. 5, 6, 7

[36] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *CVPR*, pages 16271–16280, 2023. 2

[37] 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 *CVPR*, pages 13843–13853, 2025. 2

[38] Menghao Zhang, Jingyu Wang, Qi Qi, Pengfei Ren, Haifeng Sun, Zirui Zhuang, Huazheng Wang, Lei Zhang, and Jianxin Liao. Video anomaly detection via progressive learning of multiple proxy tasks. In *ACM Multimedia*, pages 4719–4728, 2024. 2, 3

[39] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *CVPR*, pages 17385–17394, 2024. 2, 3

[40] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, pages 3769–3777. AAAI, 2023. 5, 6, 8