# Just Dance with $\pi$!
# A Poly-modal Inductor for Weakly-supervised Video Anomaly Detection

Snehashis Majhi[1,2*], Giacomo D'Amicantonio[5*], Antitza Dantcheva[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4], Egor Bondarev[5], François Brémond[1,2]

[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe [5] Eindhoven University of Technology

* Joint first authors. Code: https://github.com/snehashismajhi/PI-VAD

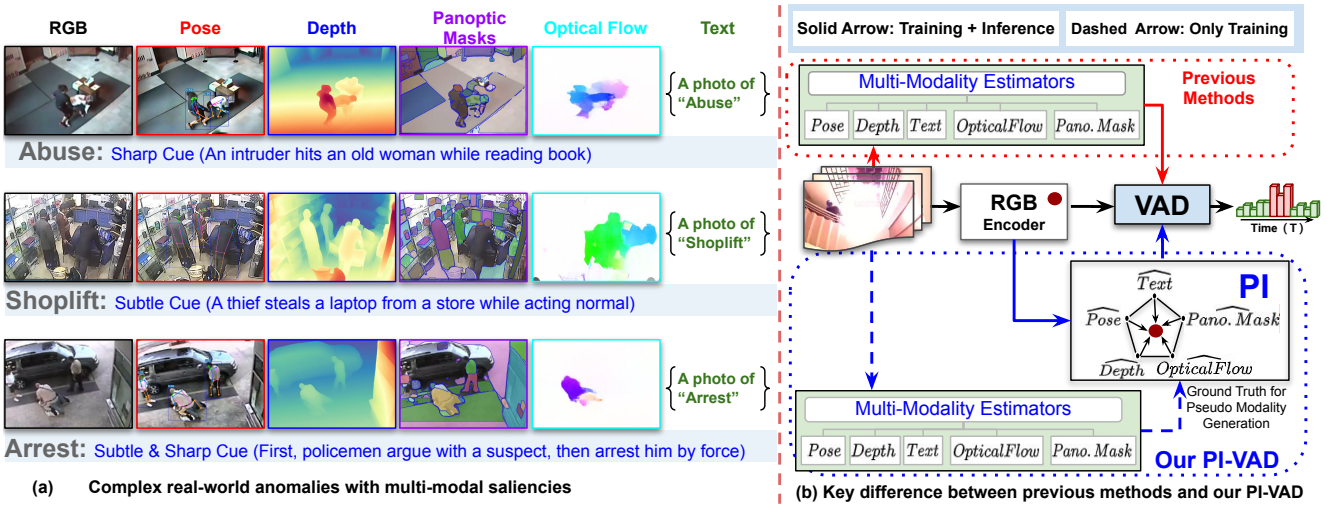



Figure 1. **a):** Illustration of abnormal frames and respective multi-modal saliencies in complex real-world scenes. *Optical flow* captures distinct abnormal motion in "Abuse" and "Arrest", while *depth* and *pose* detect subtle movements that *optical flow* may miss. *Panoptic masks* and *text* provide overall scene context. **b):** Comparison of multi-modal methods with our PI-VAD. PI-VAD requires the five modalities only during training, significantly reducing computation and enabling real-world applicability.

## Abstract

*Weakly-supervised methods for video anomaly detection (VAD) are conventionally based merely on RGB spatio-temporal features, which continues to limit their reliability in real-world scenarios. This is due to the fact that RGB-features are not sufficiently distinctive in setting apart categories such as shoplifting from visually similar events. Therefore, towards robust complex real-world VAD, it is essential to augment RGB spatio-temporal features by additional modalities. Motivated by this, we introduce the Poly-modal Induced framework for VAD: "**PI-VAD**" (or $\pi$-VAD), a novel approach that augments RGB representations by five additional modalities. Specifically, the modalities include sensitivity to fine-grained motion (Pose), three dimensional scene and entity representation (Depth), surrounding objects (Panoptic masks), global motion (optical flow), as well as language cues (VLM). Each modality represents an axis of a polygon, streamlined to add salient cues to RGB. $\pi$-VAD includes two plug-in modules, namely Pseudo-modality Generation module and Cross Modal Induction module, which generate modality-specific prototypical representation and, thereby, induce multi-modal information into RGB cues. These modules operate by performing anomaly-aware auxiliary tasks and necessitate five modality backbones – only during training. Notably, $\pi$-VAD achieves state-of-the-art accuracy on three prominent VAD datasets encompassing real-world scenarios, without requiring the computational overhead of five modality backbones at inference.*

## 1. Introduction

Weakly supervised video anomaly detection (WSVAD) aims to predict frame-level anomaly scores using only video-level labels, avoiding the need for detailed frame-

by-frame annotation. WSVAD methods [5, 23, 26, 31, 39] are effective for detecting large-scale scene anomalies, like explosions or road accidents, by training on both normal and anomalous videos to improve generalization in diverse real-world settings. However, they often struggle with more complex, human-centered anomalies such as "shoplifting", "stealing", and "abuse", where human interactions and subtle actions are involved. This limitation stems from the fact that most current methods rely on single-modality (video-only) features, which may not fully capture the complexity of these scenarios. Towards improving WSVAD in real-world settings, we place emphasis on including additional modalities, such as *pose*, *depth*, *panoptic masks*, *optical flow*, and *language semantics*, to facilitate a more nuanced scene representation. The additional information provided by these modalities describes detailed human movements, entity distances, motion dynamics, and context, rendering WSVAD more effective for detecting complex anomalies.

Despite affirmative implications of multi-modal semantics, their adaptation to WSVAD remains under-explored in current research. This is majorly due to three reasons: **(i) limited data with limited supervision:** while recent multi-modal foundation models like CLIP [24], IMAGEBIND [9] necessitates more than 400 million images for multi-modal association, the anomaly-detection task inherently deals with sparse and limited data (e.g 810 anomaly videos in UCF-Crime dataset [26]). Further, the absence of frame-level labels in WSVAD can lead to ambiguous multi-modal association; **(ii) disparity among modalities:** since each modality captures unique characteristics at various semantic levels (i.e. from contextual to fine-grained), there exists an underlying disparity among modalities that brings to the fore additional challenges in associating the modalities meaningfully; **(iii) increased inference overhead:** common multi-modal foundation models presume the availability of all modalities during inference as well, thereby linearly adding multiple modalities to the framework increases the inference overhead significantly, hindering real-time applicability. These challenges lead us to the main question: **what is the best strategy to combine multiple disparate modalities to RGB with limited data and supervision, without compromising the latency?**

Motivated by the above, we introduce a novel Polymodal Induced Transformer for weakly-supervised video anomaly detection, called **PI-VAD** (or $\pi$**-VAD**). Deviating from all WSVAD benchmarks, $\pi$-VAD synthesizes latent embeddings from five complementary modalities — pose, depth, panoptic segmentation, optical flow, and language semantics — to augment and enrich RGB-based analysis. $\pi$-VAD comprises two novel plugin modules that integrate seamlessly into a WSVAD framework: **(i)** the Pseudo Modality Generation (PMG) module, and **(ii)** the Cross Modal Induction (CMI) module. The PMG module generates synthetic, modality-specific prototype embeddings directly from RGB features, capturing each modality's distinctive characteristics. This approach mitigates inference latency by circumventing the need for individual modality backbones, thus preserving $\pi$-VAD operational efficiency.

The CMI module aligns uncoupled modalities within a unified, RGB-anchored embedding space through a double-alignment process. Initially, it constructs semantic associations between each modality and RGB via a contrastive alignment objective, ensuring cohesive integration of multi-modal embeddings. CMI leverages a pre-trained VAD model to guide the aligned multi-modal representations towards a unified task-aware and aligned representation, ensuring that the learned alignments are contextually relevant to anomaly detection. This distillation process injects $\pi$-VAD with a nuanced, semantically grounded multi-modal representation, enabling robust anomaly detection even under limited data and supervision. Moreover, the architecture of $\pi$-VAD facilitates the scalable incorporation of additional modalities without exacerbating latency constraints. To our knowledge, $\pi$-VAD is the first framework to harness the full spectrum of multi-modal representations within WSVAD, setting a new paradigm for complex anomaly detection in video analysis.

To summarize, our contributions are three-fold.

- We introduce $\pi$-VAD, a novel multi-modal method that harnesses five or more modalities to seamlessly infuse critical multi-modal cues into RGB cues, thereby enhancing the weakly-supervised video anomaly detection.
- We present two-plugin modules that are designed to synthesize multi-modal prototypes and learn effective associations to RGB. These plugin modules perform anomaly-aware auxiliary task to generate and bind meaningful multi-modal representations
- We provide an exhaustive experimental analysis to validate the robustness of $\pi$-VAD on UCF-Crime [26], XD-Violence [31], and MSAD [41]. The results suggest that $\pi$-VAD outperforms previous prominent approaches.

## 2. Related Work

**Weakly supervised video anomaly detection methods** [5, 16, 19, 20, 23, 26, 29, 31, 36, 38, 39, 42] rely on training models with video-level weak annotations, which include both normal and anomalous data. The foundational work by Sultani et al. [26] introduced a deep multiple instance learning (MIL) ranking framework for video anomaly detection. Since then, numerous adaptations of this approach have been developed. For instance, Tian et al. [28] introduced a feature magnitude learning function to better identify anomalous instances. Chen et al. [4] proposed a feature amplification mechanism with a amplitude contrast loss to enhance the discriminative capabilities of features. Lv et al. [18] introduced an Unbiased Multiple Instance Learn-

ing (UMIL) framework to create unbiased anomaly classifiers. However, these one-stage methods often concentrate on highly discriminative segments while overlooking the ambiguous and subtle ones. To address this, recent work has shifted towards pseudo-label-based, two-stage self-training methods [5, 14] to improve the accuracy of anomaly scores. Li et al. [14] introduced a multi-sequence learning technique to iteratively refine anomaly scores by progressively shortening selected sequences. However, these methods rely on single-modal video information and do not incorporate corresponding multi-modal data. Recently, cross-modal approaches have started to incorporate information from multiple modalities to improve the accuracy of discriminative features and pseudo-labels, although they primarily use text-based anomaly categories and miss the richer semantic information of anomalous events.

**Multi-modal video representation learning** leverages multiple modalities—such as RGB, depth, text, audio, and poses—to create richer representations. This approach is commonly based on two techniques: contrastive loss and knowledge distillation. Contrastive loss, as used in models like CLIP [24], creates a shared embedding space across different modalities by aligning similar features closely. Knowledge distillation, on the other hand, transfers knowledge between modalities, allowing a salient modality like text to help a modality like RGB learn more effectively. Some methods, like ViFiCLIP [25] and CoCLR [15], combine contrastive learning with knowledge distillation to fine-tune alignments across modalities, while also making cross-modal learning more efficient. However, these techniques require large-scale datasets to effectively learn multi-modal representations, while video anomaly datasets are inherently sparse and small scale. To leverage cross-modal information with limited data, we generate pseudo-modalities during training and apply contrastive loss and knowledge distillation to guide the shared feature space semantically.

## 3. Preliminaries: Uni-modal WSVAD Method

In this section, we briefly describe the working principle of existing uni-modal WSVAD methods. Uni-modal WSVAD focuses solely on the RGB modality for both training and inference. The video $V$ is first divided into non-overlapping snippets of 16 frames, resulting in $T$ snippets. A pre-trained 3D convolutional network (e.g., I3D [1]) is then used to extract features from each snippet, forming a feature map $\mathcal{F}_{RGB} \in \mathbb{R}^{T\times D}$, where $D$ is the feature dimension. Given $\mathcal{F}_{RGB}$, the goal of the uni-modal WSVAD method is to train an RGB task encoder that can predict frame-level anomaly scores while only having access to video-level labels during training.

Deviating from standard uni-modal WSVAD, in this work we explore the multi-modal (*i.e.* two or more modalities) representation learning ability in WSVAD task. We aim to answer questions such as: **how many modalities are required to represent real-world complex anomalies? With limited dataset and supervision, is it possible for a model to effectively learn from more than five modalities and use only RGB for inference?** While multi-modal methods of action understanding [6–8] can be applicable to the WSVAD task, their effectiveness depends on the amount of labeled data available. Therefore, we propose a novel multi-modal framework that can effectively associate more than five modalities to RGB for the WSVAD task.

## 4. Proposed $\pi$-VAD

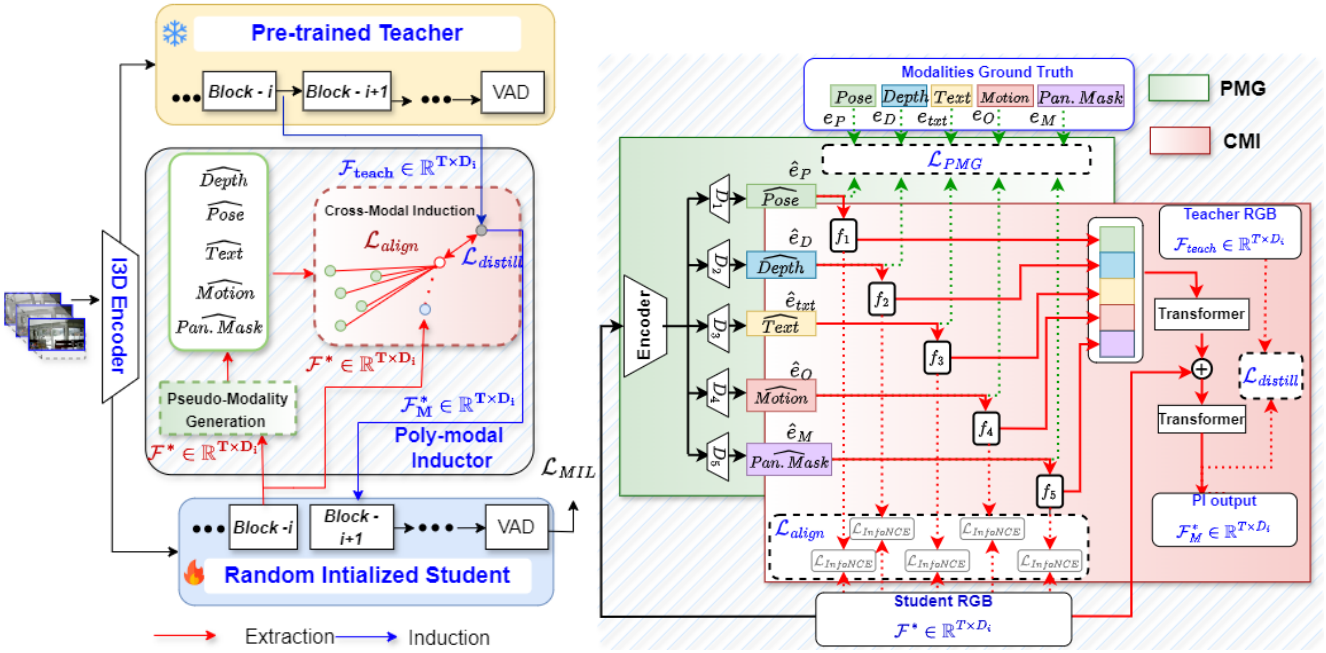
In this section, we introduce our Poly-modal Induced Transformer for weakly-supervised video anomaly detection, referred to as $\pi$-VAD (illustrated in Figure 2). $\pi$-VAD adopts a teacher-student architecture incorporating a novel poly-modal inductor. While the teacher and student share an identical functional architecture, the teacher is pre-trained on the WSVAD task and remains frozen, and the student is randomly initialized.

### 4.1. Poly-modal Inductor (PI)

The objective of the poly-modal inductor (illustrated in Figure 2a) is to enhance the student's RGB representation by promoting the learning of discriminative features for anomalous events within a cohesive multi-modal feature space. This is enforced by two key modules of poly-modal inductor: **(i) Pseudo Modality Generation (PMG) module** learns anomaly relevant synthetic approximation of the actual modalities component, **(ii) Cross Modal Induction (CMI) module** facilitates the semantic alignment between the multi-modal encodings from PMG and the RGB embeddings of the student while ensuring that the alignment is pertinent to WSVAD. As visible in Figure 2b, the poly-modal inductor is adaptable and can be integrated at various blocks of the teacher-student architecture; however, we deliberately position it in the initial and final blocks to capture both low and high-level multi-modal features effectively. Further, regardless of the student's specific block, the poly-modal inductor processes the output representation from *Block-i* of the student $\mathcal{F}^* \in \mathbb{R}^{T\times D_i}$ and injects the refined multi-modal feature $\mathcal{F}^*_M \in \mathbb{R}^{T\times D_i}$ into *Block i+1* of the student, thereby enhancing the student's ability to learn discriminative representation for anomaly detection.

### 4.2. Pseudo Modality Generation Module

The Pseudo Modality Generation (PMG) module aims to synthetically derive embeddings for pose ($\hat{e_P} \in \mathbb{R}^{T\times d_P}$), depth ($\hat{e_D} \in \mathbb{R}^{T\times d_D}$), panoptic masks ($\hat{e_M} \in \mathbb{R}^{T\times d_M}$), optical flow ($\hat{e_O} \in \mathbb{R}^{T\times d_O}$), and text ($\hat{e_{txt}} \in \mathbb{R}^{T\times d_{txt}}$) directly from the student's intermediate RGB feature representation $\mathcal{F}^* \in \mathbb{R}^{T\times D_i}$. This approach fulfills two key objectives: **(i)** eliminating the reliance on multi-modal back-




(a) Overview of $\pi$-VAD (b) Poly-modal Inductor (PI)

Figure 2. **(a) Overview of Poly-modal Induced VAD ($\pi$-VAD):** In the *training phase*, $\pi$-VAD uses a teacher-student approach, where a poly-modal inductor enhances the student's RGB representation by generating and associating five distinct modalities. Note that the teacher's weights remain fixed during training. At inference, the student and poly-modal inductor operate independently to detect video anomalies. **(b) Poly-modal Inductor (PI):** PI refines the student's intermediate feature, $\mathcal{F}^*$, by generating pseudo-modalities through a modality generation module (PMG). These generated modalities are then combined with $\mathcal{F}^*$ to produce an enhanced feature set, $\mathcal{F}^*_M$.

bones (e.g., SAM, yolov7, etc.) within the poly-modal inductor during inference; **(ii)** since multi-modal embeddings can introduce redundancy, noise, or even conflicting information, this approach selectively retains only the multi-modal cues essential to the WSVAD task.

To enforce these objectives, we design PMG that can be seen as an encoder-decoder structure. As shown in Figure 2, PMG has one encoder and five parallel decoders $D_1$, $D_2$, ..., $D_5$. We deliberately kept one encoder to learn shared RGB features for all the modalities. The six decoders operate in a mutually exclusive manner to generate the six modalities. The encoder has a 1D-convolutional layer to project RGB embeddings to a low-dimensional latent space.

For each modality decoder, a single linear layer translates the RGB latent representation to a modality-specific RGB representation, maintaining the latent space's dimensionality. By doing this, we generate diverse views of the same RGB embeddings, enhancing the information contained in the embeddings that is relevant to a specific modality while suppressing possible noise. Subsequently, a 1D convolution layer is used as a decoder to generate the modality embeddings $\hat{e_j}$, where $j \in \{P, D, M, O, \text{txt}\}$.

Training the PMG requires *ground-truth* embeddings $e_j$, where $j \in \{P, D, M, O, \text{txt}\}$ from the corresponding modality decoders. We utilize the intermediate embeddings of YOLOV7-pose [30], DepthAnythingV2 [33], SAM [12], RAFT [27] and VifiCLIP [25] to represent pose, depth, panoptic mask, optical flow and text modality *ground-truths*. The combined training objective for the PMG is

$$\mathcal{L}_{PMG} = \sum_{j=1}^{5} \frac{1}{d_j} \sum_{k=1}^{d_j} (e_{\hat{j},k} - e_{j,k})^2, \text{where } j \in \{\text{P, D, M, O, txt}\}. \tag{1}$$

Once PMG is trained with $\mathcal{L}_{PMG}$, it has the ability to precisely generate pseudo modalities which are subsequently used in PI to augment the RGB representation of the student.

### 4.3. Cross Modal Induction Module

In this stage, Cross Modal Induction (CMI) combines the generated pseudo-modalities $\hat{e_j}$ with the RGB embeddings $\mathcal{F}^*$, aiming to create a shared representation space that promotes all relevant features for the task. It aligns the pseudo-modalities from the PMG with the RGB embeddings, which contain critical visual information from the current video snippet $T_i$. Our aim is to ensure that the most relevant modalities for $T_i$ converge with the RGB embeddings in a cohesive representation space, thereby strengthening multi-modal associations. By aligning these diverse modalities, the PMG produces modality embeddings informed by the RGB data, enhancing relevant information and filtering out

irrelevant details. This joint representation is essential for the WSVAD task, as it improves the model's ability to utilize multi-modal insights effectively.

To achieve this, we learn a shared latent space between each modality and the RGB embeddings by applying a snippet-level, bi-directional InfoNCE contrastive loss [22]. This loss is applied between each pseudo-modality embedding $\hat{e_j}$ (where $j \in \{P, D, M, O, \text{txt}\}$) and the RGB embedding $\mathcal{F}^*$. The bi-directional approach provides a more balanced measure of similarity between positive and negative pairs. Since the contrastive loss is applied at the snippet level, we treat representations from the same snippet index $T_i$ (i.e., $\mathcal{F}^*(T_i)$ and $\hat{e_j}(T_i)$) as positive pairs, and representations from different snippets as negative pairs. This encourages similarity in positive pairs and discourages similarity in negative pairs. The similarity between embeddings is computed as: $sim(\mathcal{F}^*(T_i), \hat{e_j}(T_i)) = \frac{\mathcal{F}^*(T_i) \cdot \hat{e_j}(T_i)}{\|\mathcal{F}^*(T_i)\| \|\hat{e_j}(T_i)\|}$ and the contrastive alignment loss is defined as

$$\mathcal{L}_{InfoNCE} = -\frac{1}{T} \sum_{i=1}^{T} \log \frac{\exp\left(\frac{sim(\mathcal{F}^*(T_i), \hat{e_j}(T_i)}{\tau}\right)}{\sum_{k=1, i \neq k}^{T} \exp\left(\frac{sim(\mathcal{F}^*(T_i), \hat{e_j}(T_k)}{\tau}\right)} \tag{2}$$

$$\mathcal{L}_{align} = \sum_{i=1}^{5} \mathcal{L}_{InfoNCE}, \quad i \in \{\text{P, D, M, O, txt}\} \tag{3}$$

Next, we aim to identify and prioritize the most relevant modalities for each snippet by reducing cross-modal conflicts and noise, resulting in task-oriented multi-modal embeddings. First, we *concatenate* the aligned embeddings from each modality along the embedding dimension. Then, we use a stack of *transformer* blocks to highlight the most pertinent modalities by explicitly encoding the cross-correlations among them. Additionally, the RGB embeddings $F^*$ from the student model are added between the transformer blocks to enhance the RGB representation with contextually relevant information from multiple modalities.

Second, we guide the final multi-modal output from the last transformer block, $\mathcal{F}_M^* \in \mathbb{R}^{T \times D_i}$, towards a task-specific representation for WSVAD. This ensures that relevant modalities are produced in the PMG with minimal noise, and that the alignment between salient modalities and RGB is optimized for WSVAD with minimal cross-modal conflict. This is achieved through a distillation process, which minimizes the difference between $\mathcal{F}_M^*$ and the teacher's pre-trained features at the same stage, $\mathcal{F}_{teach} \in \mathbb{R}^{T \times D_i}$. The distillation loss guiding this minimization is defined as:

$$\mathcal{L}_{distill} = \frac{1}{D_i} \sum_{k=1}^{D_i} (\mathcal{F}_{Mk}^* - \mathcal{F}_{teachk})^2. \tag{4}$$

### 4.4. $\pi$-VAD Optimization

$\pi$-VAD is optimized in two steps. In the *first step*, the student model, PMG module, and CMI module are warmed up with the $\mathcal{L}_{PMG}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{distill}$ respectively. This ensures that all the components are correctly initialized before optimizing for the actual task and thereby it avoids possible pitfalls in which one of the modalities overpowers the others independently to what information adds to the RGB embeddings. The loss function for the first step is:

$$\mathcal{L}_{first} = \mathcal{L}_{PMG} + \mathcal{L}_{align} + \mathcal{L}_{distill}. \tag{5}$$

In the *second step*, the model is trained on the WSVAD task with the standard MIL loss function used in UR-DMU [40]. In order to avoid the decoupling of the aligned modalities, the final training objective is:

$$\mathcal{L}_{second} = \mathcal{L}_{MIL} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{distill} + \mathcal{L}_{PMG} \tag{6}$$

where $\lambda_1$ and $\lambda_2$ are hyper parameters that allows us to balance the impact of the distillation and alignment components on the training process. The $\mathcal{L}_{PMG}$ is not balanced by a factor to ensure that the pseudo-modalities generated remain bounded to the ground-truth modalities throughout the training process.

## 5. Experiments

We conduct extensive experiments on two of the most common and challenging VAD datasets publicly available, UCF-Crime [26] and XD-Violence [31]. For each dataset we extract the pose, depth, semantic, textual and motion features. The audio features contained in the XD-Violence dataset are considered as an additional modality and aligned via PI with the others. Additionally, we test our approach on the MSAD dataset [41], a recently published dataset that contains real-world anomalous videos collected in a more diverse set of scenarios compared to UCF-Crime.

We follow the established evaluation protocols [18, 26, 31, 32] to measure the performance of $\pi$-VAD on the datasets. For additional information on the experimental settings, we refer to Section A of the appendix.

### 5.1. SoTA comparison and analysis

To evaluate $\pi$-VAD, we place inductor modules at early and late stages of a UR-DMU [40] model. The results of our experiments are shown in Table 1. Compared with current multi-modal SoTA approaches, $\pi$-VAD demonstrates superior capabilities in both UCF-Crime and XD-Violence datasets. On **UCF-Crime**, $\pi$-VAD marks a $+2.31\%$ improvement over multi-modal VadCLIP, and outperforms the best RGB-based model by $+2.75\%$. It is important to notice that the $AUC_A$ metric, which measures the capabilities of

| Model | Encoder | UCF-Crime AUC | UCF-Crime $AUC_A$ | XD-Violence AP | XD-Violence $AP_A$ |
|---|---|---|---|---|---|
| *SoTA with multi-modality at inference* | | | | | |
| HL-Net [31] | I3D | 82.44 | - | - | - |
| HSN [21] | I3D | 85.45 | - | - | - |
| MACIL-SD [35] | I3D+audio | - | - | 83.40 | - |
| UR-DMU | I3D+audio | - | - | 81.77 | - |
| TPWNG [34] | CLIP | 87.79 | - | 83.68 | - |
| PEMIL [2] | I3D+Text | 86.83 | - | 88.21 | - |
| VadCLIP [32] | CLIP | 88.02 | 70.23 | 84.15 | - |
| *SoTA with RGB-only at inference* | | | | | |
| MIL [26] | C3D | 75.41 | 54.25 | 75.68 | 78.61 |
| | I3D | 77.42 | - | - | - |
| RTFM [28] | I3D | 84.30 | 62.96 | 77.81 | 78.57 |
| CLAV [5] | I3D | 86.10 | - | - | - |
| UR-DMU [40] | I3D | 86.97 | 70.81 | 81.66 | 83.94 |
| SSRL [13] | I3D | 87.43 | - | - | - |
| MSL [14] | V-Swin | 85.30 | - | 78.28 | - |
| WSAL [17] | I3D | 85.38 | 67.38 | - | - |
| ECU [37] | V-Swin | 86.22 | - | - | - |
| MGFN [4] | V-Swin | 86.67 | - | - | - |
| UMIL [18] | CLIP | 86.75 | 68.68 | - | - |
| TSA [10] | CLIP | 87.58 | - | 82.17 | - |
| **$\pi$-VAD (Ours)** | **I3D** | **90.33** (+2.75%) | **77.77** (+6.96%) | **85.37** (+3.20%) | **85.79** (+1.85%) |

Table 1. State-of-the-art comparisons on UCF-Crime and XD-Violence in the WSVAD task. The best results are written in **bold**.

| Model | MSAD (NeurIPS'24) AUC | $AUC_A$ | AP | $AP_A$ |
|---|---|---|---|---|
| RTFM[28] | 86.65 | - | - | - |
| MGFN[4] | 84.96 | - | - | - |
| TEVAD[3] | 86.82 | - | - | - |
| UR-DMU[40] | 85.02 | - | - | - |
| UR-DMU * | 85.78 | 67.95 | 67.35 | 75.30 |
| **$\pi$-VAD (Ours)** | **88.68** | **71.25** | **71.26** | **77.86** |

Table 2. State-of-the-art comparisons on MSAD. * indicates our own implementation. The best results are written in **bold**.

the model to detect abnormal events, shows an even larger improvement over the previous methods. In fact, $\pi$-VAD outscores VadCLIP and UR-DMU, the best scoring multi-modal and RGB-based methods, by $+7.54\%$ and $+6.96\%$ respectively. In Figure 3 we compare the class-wise performance of $\pi$-VAD with the baseline model (*i.e.* UR-DMU) employed as a teacher. $\pi$-VAD improves upon the class-wise $AUC$ scores achieved by UR-DMU in all classes except "Abuse", "Assault" and "Robbery". Notably, the "Explosion" class proves to be the most challenging class for the UR-DMU, with a score of $47.25\%$. In this class, $\pi$-VAD almost doubles the performance of UR-DMU, which shows the ability of $\pi$-VAD to learn and leverage an extensive poly-modal scene representation towards the detection of short anomaly events. Significant improvements can be

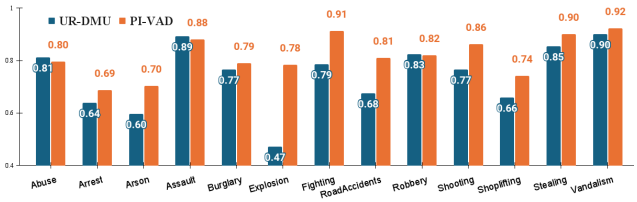



Figure 3. Class-wise $AUC$ comparison of $\pi$-VAD with UR-DMU [40] on the UCF-Crime dataset.

observed for other challenging classes, such as "Shoplifting" and "Shooting", where the anomalous events happen in subtle ways that are more difficult to be detected by an RGB-only model.

Similar results are observed on the **XD-Violence** dataset, where $\pi$-VAD improves by $+1.22\%$ over the $AP$ score of VadCLIP. The $AP_A$ shows comparable improvements, outperforming UR-DMU by $+1.85\%$. On the more recent **MSAD** dataset, $\pi$-VAD achieves a $+1.65\%$ performance improvement compared to three available SoTA. It is a significant boost on MSAD as it contains 14 diverse scenarios and distinct environmental conditions, presenting a poignant challenge and benchmark for real-world performance. We refer to Section B of the appendix for a more in-depth analysis of the class-wise performance and contributions from all modalities on these two datasets.

## 5.2. Components ablation study

We analyze the contribution of the two components, PMG and CMI of PI on the overall $AUC$. The core idea behind CMI is that it's essential to align disparate modalities to form comprehensive scene representation, which can then be adapted for anomaly detection. As shown in Table 3 both alignment and adaptation are necessary to fully harness the multi-modal cues. From **Row-1** it can be observed that when modalities remain decoupled and unguided for VAD, it under perform compared to the baseline model, UR-DMU [40]. Likewise, the model struggles to leverage aligned modalities effectively for the VAD task without a distillation mechanism. This is likely due to residual noise overwhelming salient cues, underscoring the critical role of distillation in filtering and refining the multi-modal cues.

| PMG | CMI $\mathcal{L}_{align}$ | CMI $\mathcal{L}_{distill}$ | AUC |
|---|---|---|---|
| ✓ | - | - | 84.66 |
| ✓ | ✓ | - | 85.84 |
| ✓ | - | ✓ | 86.29 |
| - | ✓ | ✓ | **90.58** |
| ✓ | ✓ | ✓ | 90.33 |

Table 3. Contribution of the reconstruction, alignment and distillation auxiliary tasks on the main VAD task for UCF-Crime.

Reconstructing the modalities at test time leads to a trade-off between computational cost at inference time and

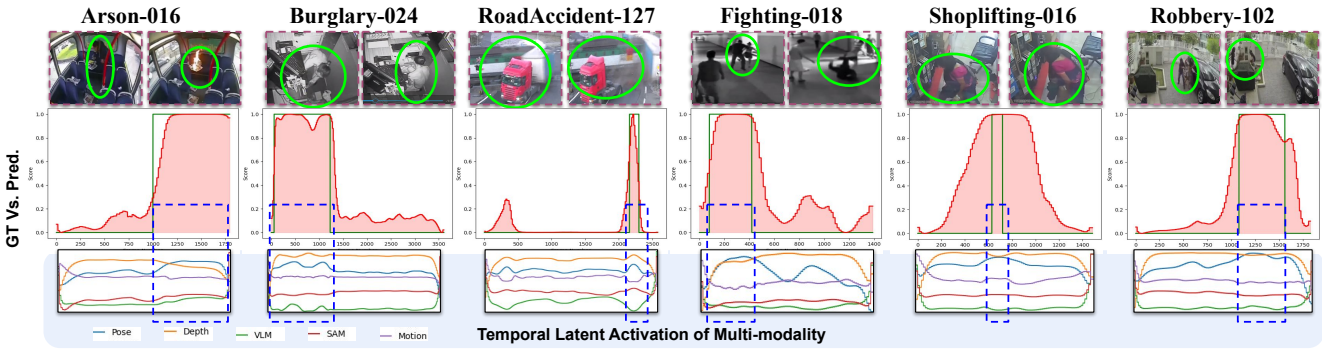



Figure 4. Visualization of sample frames and ground truth (green shed) vs. prediction scores (red shed) for various cases in Row-1 and Row-2. For each plot in Row-2, the X and Y axis denotes the number of frames and corresponding anomaly scores. Row-3 shows the latent activation learned by multi-modality. We plot the mean value of the normalized modalities activations from the first transformer block of the late PI module to show the alignment between modalities and their correlation to the predicted abnormal scores.

performance. Without reconstructing the modalities, PI uses the modality features as input for CMI and obtains a marginal $\simeq 0.25\%$ performance gain. The trade-off between computational costs and performances is illustrated in Table 5. Despite requiring more computational and memory resources than the baseline RGB-only method, $\pi$-VAD achieves real-time performance, processing at 30 frames per second, making it a viable option for practical deployment.

| | **Early** | **Late** | **Both** |
|---|---|---|---|
| **AUC** | 87.14 | 87.48 | **90.33** |

Table 4. Comparison of the effect of the early and late PI on performance for UCF-Crime with all modalities available at training.

| | **UR-DMU** | **Modality Backbones** | **$\pi$-VAD** |
|---|---|---|---|
| **GFLOPs** | 1.54 | 2,561.40 | 19.88 |
| **Params. (M)** | 6.16 | 2,406.50 | 82.81 |
| **FPS** | 110.09 | - | 30.51 |
| *AUC* | 86.97 | 90.58 | **90.33** |

Table 5. Computational cost comparison with the UR-DMU, the modalities backbones, and the proposed $\pi$-VAD.

Further, we observe a complementary effect between early and late PI, as shown in Table 4. Some anomalies rely on multi-modal association in low-level RGB features, while others need high-level associations. To fully capture all types of anomalies, both early and late PI are essential. Further discussion continues in Section C of the appendix.

### 5.3. Qualitative Analysis

To verify our method's effectiveness, we present qualitative results in Figure 4 that illustrate various types of anomalies, including those based on scenes, human actions, and differing durations. Across these scenarios, our method consistently detects anomalies with high confidence, as shown in **Row-2** of Figure 4. Furthermore, understanding the contribution of each modality to the anomaly scores is essential. **Row-3** shows two key aspects of multi-modal activation curves: **(i)** the curve's amplitude and **(ii)** its pattern in the abnormal regions. The amplitude generally reflects each modality's importance, with depth being critical in CCTV applications. Depth helps distinguish between foreground and background objects, supporting better interaction analysis and occlusion handling.

For scene-based anomalies like "Burglary-024" and "Arson-016", the pose and text modality activation pattern aligns closely with abnormal regions, due to its strong ability to capture global context. In cases like "RoadAccident-127", all modalities contribute significantly to identifying the anomaly. For human-based anomalies, such as "Fighting-018" and "Shoplifting-016", the pose modality strongly correlates with abnormal regions, and depth complements this by adding spatial context to the 2D key points. Overall, all five modalities are useful for detecting real-world CCTV anomalies, with text and panoptic masks being particularly important for scene-based anomalies, while pose and depth are key for human-based anomalies.

## 6. Modality Evaluation

To analyze the impact of each modality of the VAD task and the interaction between modalities, we focus on the performance of the UCF-Crime. We refer to Section B of the appendix for the modality evaluation of the other two datasets.

### 6.1. Single Modality Evaluation

To properly evaluate the contributions of the different modalities, we trained the model with each modality individually. Table 6 shows that each modality is able to enhance the RGB features and improve the baseline performance on the UCF-Crime dataset, with motion having the largest positive impact for the $AUC$ metric. This is coherent with the intuitive understanding that motion is often the most important factor in distinguishing between a normal and an abnormal action. However, the depth modality overperforms the others by a large margin on $AUC_A$. The class-wise evaluation for the individual modality contributions in

| Modality | | | | | UCF-Crime | |
|---|---|---|---|---|---|---|
| Pose | Depth | Text | Pan. | Motion | AUC | $AUC_A$ |
| - | - | - | - | - | 86.97 | 70.81 |
| ✓ | - | - | - | - | 87.65 | 73.24 |
| - | ✓ | - | - | - | 87.75 | **75.14** |
| - | - | ✓ | - | - | 87.89 | 69.45 |
| - | - | - | ✓ | - | 87.71 | 72.13 |
| - | - | - | - | ✓ | **87.92** | 72.04 |
| ✓ | ✓ | - | - | - | 88.14 | 74.06 |
| ✓ | ✓ | ✓ | - | - | 88.85 | 75.67 |
| ✓ | ✓ | ✓ | ✓ | - | 90.31 | 76.47 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **90.33** | **77.77** |

Table 6. Modality impact comparisons on UCF-Crime. The best results are written in **bold**.

Figure 5 shows that depth is the best modality for the majority of classes. We conjecture that the scene information contained in the depth features allows PI to better model the spatial interactions between entities in the scene. This is supported by the performance of the depth modality on the "Explosion" class: people and objects tend to move away quickly from the source of an explosion, leading to sharp changes in the depth features.

The text modality exhibits robust performance, second-best ranking in the $AUC$ metric and excelling in capturing normal scenarios. We hypothesize that this advantage derives from the ViFiCLIP training, optimized through video-text pairs in the Kinetics-600 dataset [11]. However, the text modality underperforms in $AUC_A$, likely due to its inclination to represent coarse-grained normalcy patterns more effectively than the details of anomaly events.

### 6.2. Poly-modal Evaluation

Table 6 shows the benefits of combining multiple modalities for WSVAD, where the performance of $\pi$-VAD across the dataset increases by sequentially adding modalities. Specifically, incorporating textual and panoptic modalities yields the largest performance gains, as they capture essential real-world information from large-scale training datasets, enriching $\pi$-VAD's implicit scene representation. We hypothesize that this representation is comprehensive enough to limit additional contributions from the motion modality for certain types of anomalies, a hypothesis supported by minimal class-wise performance gains in Figure 6 and qualitative examples such as "Burglary-024" and "Fighting-018" in Figure 4. The motion modality is nonetheless crucial on other classes, such as "Shoplifting", where anomalous events are usually subtle and best represented by motion-based cues after the pose cue.

Furthermore, Figure 6 suggests that the interaction between different modalities within $\pi$-VAD can lead to complementary or contrastive performance for specific anomaly types. In fact, the modality activations for the "Robbery-102" example in Figure 4 show that pose is the most relevant modality for the anomalous event in the video, while the contributions of the other modalities are marginal. This effect showcases the ability of $\pi$-VAD to leverage the relevant modalities for each anomaly type without over-relying on specific modalities.

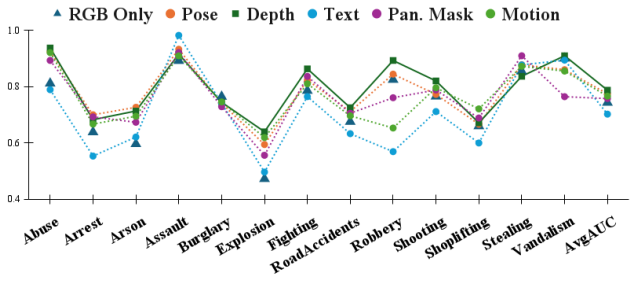



Figure 5. Class-wise $AUC$ comparison between the RGB model and RGB with one additional modality model on UCF-Crime.

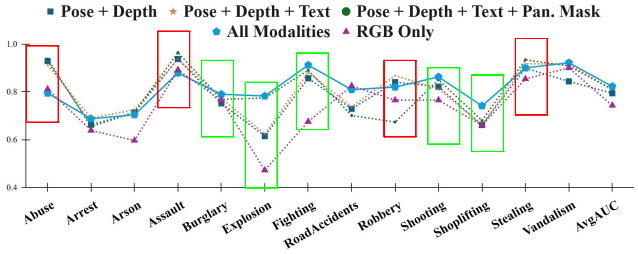



Figure 6. Comparison between the $AUC$ scores of different mixtures of modalities in the $\pi$-VAD framework for the UCF-Crime dataset. In red, we highlight the classes on which the modalities have contrastive features, in green the classes where the modalities are complementary.

## 7. Conclusion

This paper presents $\pi$-VAD, the first poly-modal framework for WSVAD, significantly advancing video anomaly detection by expanding beyond traditional RGB-based methods and integrating multiple modalities to address complex anomaly categories in real-world settings. $\pi$-VAD incorporates five auxiliary modalities, namely pose, depth, panoptic masks, optical flow, and text cues, which jointly enrich anomaly detection with diverse, fine-grained contextual cues. Both novel integrated modules, Pseudo-modality Generator and Cross Modal Induction, enable effective multi-modal learning during training, without imposing additional computational burden during inference. $\pi$-VAD sets a new benchmark for robust and efficient weakly-supervised anomaly detection in real-world applications by showcasing state-of-the-art results on three major datasets.

**Acknowledgements:** This work was supported by Toyota Motor Europe (TME) and the French government, through the 3IA Cote d'Azur Investments managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002

## References


[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[2] 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.

[3] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023.

[4] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 387–395, 2023.

[5] 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.

[6] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13053–13064, 2021.

[7] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living, 2020.

[8] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2022.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[10] 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.

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[13] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022.

[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.

[15] Zichao Li, Cihang Xie, and Ekin Dogus Cubuk. Scaling (down) clip: A comprehensive analysis of data, architecture, and training strategies. *arXiv preprint arXiv:2404.08197*, 2024.

[16] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin Chen. Social mil: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[17] 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.

[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.

[19] Snehashis Majhi, Srijan Das, François Brémond, Ratnakar Dash, and Pankaj Kumar Sa. Weakly-supervised joint anomaly detection and classification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021.

[20] Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021.

[21] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *Computer Vision and Image Understanding*, 241:103955, 2024.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang. Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 173–183, 2021.

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

[25] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023.

[26] 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.

[27] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[28] 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.

[29] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[30] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[31] 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.

[32] 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.

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

[34] 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 (CVPR)*, pages 18899–18908, 2024.

[35] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6278–6287, 2022.

[36] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020.

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023.

[38] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019.

[39] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023.

[41] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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