Universiteit Leiden

# Master Computer Science

Towards Robust and Generalizable Video Anomaly Understanding with Anomaly-Aware Context Tokens for MLLMs


| | |
|---|---|
| Name: | Arsen Ignatosyan |
| Student ID: | s4034538 |
| Date: | 20/08/2025 |
| Specialisation: | Artificial Intelligence |
| 1st supervisor: | Hazel Doughty |
| 2nd supervisor: | Lu Cao |


Master's Thesis in Computer Science


Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Towards Robust and Generalizable Video Anomaly Understanding with Anomaly-Aware Context Tokens for MLLMs

Arsen Ignatosyan, Hazel Doughty, Lu Cao
*Leiden Institute of Advanced Computer Science*
*Leiden University*
*Leiden, The Netherlands*
*a.ignatosyan@umail.leidenuniv.nl*

***Abstract*—Video Anomaly Understanding (VAU) extends traditional anomaly detection by not only identifying irregular events in real-world video streams but also providing interpretable explanations of these events. This capability is increasingly important in wide range of real-world applications where enhanced trust and transparency are essential. Recent progress in multi-modal large language models (MLLMs) has shown potential for building more generalizable and explainable VAU systems, but two key challenges remain: (1) high sensitivity to prompt variations in instruction-tuned models, and (2) the computational inefficiency of current MLLM architectures for real-time scenarios. In this work, we propose a novel framework to address both challenges. First, we introduce a prompt learning strategy that integrates learnable, instance-conditional context tokens into textual prompts. This approach overcomes prompt sensitivity without requiring partial or full model finetuning, significantly improving robustness in open-world anomaly detection settings. Second, we adopt a lightweight video encoder, VideoMamba, which preserves the structure of Vision Transformers while leveraging the Mamba architecture for efficient sequence modeling. This enables significantly faster inference and adaptation. Our combined approach improves generalization to unseen anomalies and reduces latency, making MLLM-based VAU more viable for real-world use. Extensive results demonstrate that our method offers a solid balance between strong performance, efficiency, and interpretability.**

## 1. Introduction

Video Anomaly Understanding (VAU) aims to detect irregular events in unstructured, real-world video streams and provide textual explanations for them [1]. Compared to traditional Video Anomaly Detection (VAD), which is concentrated on identifying whether an event is normal or abnormal, VAU extends the scope toward understanding by incorporating contextual reasoning and explanation. This shift reflects the practical demands of applications such as public surveillance [2, 3, 4, 5, 6], autonomous driving [7, 8], and industrial monitoring [9, 10], where both accurate detection and meaningful interpretation are essential.

Because of the sensitive nature of these domains, explainability and generalization are critical. A system that only signals anomalies without offering interpretable reasoning may limit trust and restrict practical deployment. The emergence of multi-modal large language models (MLLMs) [11, 12, 13, 14, 15], alongside new benchmarks for video anomaly understanding (VAU) [16, 17, 18], has created new opportunities for building more accurate and explainable anomaly detection systems via instruction-tuning [17, 18].

Despite recent progress in video anomaly understanding, our study identifies a critical limitation of instruction-tuned MLLMs: their sensitivity to prompt variations. This issue becomes particularly noticeable when the number of question–answer pairs in the training data is limited, causing models to overfit to specific phrasings or question formats [19, 20]. In crucial applications such as video anomaly understanding, this is a serious concern. Minor changes in prompt wording could lead to the misclassification of an abnormal event as normal, potentially resulting in missed responses or safety failures.

To address this, we propose adapting prompt learning to the MLLM setting as an alternative to partial or full model finetuning. Prompt learning addresses the limitations of manual prompt engineering by learning the optimal "context" for prompts, and it has been successfully applied to both natural language processing and vision-language tasks [21, 22, 23, 24, 25]. In contrast to hand-crafted prompts, learnable prompts provide a more parameter-efficient approach, enabling task-specific adaptation without the need to finetune the model and alter the pre-trained weights. While prompt learning has proven effective in CLIP-style vision-language models [26] and large language models (LLMs) [27], it has not been explored in MLLMs, particularly for temporally complex tasks like video anomaly understanding. In this work, we propose integrating learnable, instance-conditional context tokens, drawing inspiration from Conditional Context Optimization (CoCoOp) [23], into the textual prompt. These context tokens are optimized to become anomaly-aware while preserving the general knowledge of the base MLLM. This lightweight prompt optimization strategy reduces sensitivity to prompt variations in MLLMs compared to their instruction-tuned counterparts, improving the robustness of anomaly understanding and lowering the

risk of misclassifying abnormal events.

Another key challenge is the real-time application of MLLMs. The computational complexity of many current architectures makes them unsuitable for latency-sensitive environments, limiting their practical applicability in real-world video surveillance systems. To address this, we propose integrating an efficient and state-of-the-art video encoder, VideoMamba [28], into the MLLM framework. Our approach adopts the commonly used "ViT-MLP-LLM" paradigm [29, 30, 31, 32], replacing the standard Vision Transformer (ViT) [33] with VideoMamba, which follows the ViT architecture closely and benefits from the computational efficiency of the Mamba design [34].

To address these limitations, and to further improve the practicality and generalization of VAD with MLLMs, we contribute the following:

- **Prompt Learning for MLLMs:** We introduce a novel prompt learning strategy for MLLMs in video anomaly detection. By integrating learnable, instance-conditional context tokens into textual prompts, our method improves robustness to prompt variations without modifying the underlying model parameters.
- **Efficient Video Encoder Integration:** We propose the integration of a lightweight and high-performance VideoMamba encoder into an MLLM framework. This design follows the "ViT-MLP-LLM" paradigm while leveraging the efficiency of the Mamba architecture, enabling real-time inference in latency-sensitive scenarios.
- **Anomaly-Aware Generalization:** Our approach enhances the model's ability to generalize to unseen anomalies by integrating anomaly-aware context tokens into prompts, while leaning on the broad visual-textual knowledge present in the base MLLM.

## 2. Related Work

**Video Anomaly Detection.** Video Anomaly Detection (VAD) aims to detect unusual events or frames in long, raw videos [3, 35, 36, 37, 38, 39], which presents a significant challenge due to the scarcity of labeled anomalous data and, as the name suggests, the rarity of these events. Traditional approaches to VAD were based on handcrafted features [2, 3, 35, 36, 40, 41], but recent developments have seen the dominance of deep learning approaches, with different types of unsupervised, fully-supervised and weakly-supervised methods.

Unsupervised Video Anomaly Detection (UVAD) [37] has been widely studied due to the difficulty of collecting and annotating large-scale anomalous videos, focusing on learning the normality of videos and detecting deviations as a one-class classification problem [42]. The two main UVAD approaches are reconstruction-based [43, 44, 45] and regression-based methods [46, 47, 48, 49].

Fully-supervised video anomaly detection [50] has received comparatively less attention in the literature due to the challenges that come with obtaining fine-grained data and precise temporal annotations for anomalies.

Weakly-Supervised Video Anomaly Detection (WS-VAD) relies on video-level annotations, since these are more viable to obtain than fine-grained ones. Because of this, weakly-supervised approach has gained increasing attention in recent years. Multiple-Instance Learning (MIL) [51, 52] is the mainstream paradigm used for various weakly-supervised learning methods [4, 53, 54, 55, 56, 57, 58]. In these approaches, anomaly classifiers are trained using bags of positive (anomalous) and negative (normal) samples.

These methods have two main drawbacks: they provide limited semantic understanding of anomalies and often depend on either outdated models or computationally intensive transformers. In contrast, our approach leverages VideoMamba, an efficient video understanding model, for feature extraction and enhances semantic interpretation by integrating it with an MLLM.

**Large Models in Video Understanding.** The rapid advancement of proprietary and open-source Large Language Models (LLMs) such as ChatGPT [59], LLaMA [60], and Mistral [61] has sparked significant interest, based on their conversational and text generation capabilities. These models are pretrained on vast amounts of text data, making them highly effective in a variety of natural language processing tasks. More recently, the scope of LLMs has expanded into multi-modal domains, with Multi-modal LLMs (MLLMs) [29, 31, 62, 63, 64, 65, 66, 67] incorporating visual understanding into the model's capabilities. The introduction of large-scale video-text pair datasets, such as WebVid [68] and Valley [69], has enabled the addition of video modality into MLLMs, improving their capacity for video understanding. This area has significant attention from the research community, leading to the development of various models, including VideoChat [67], Video-ChatGPT [12], LLaMA-Adapter [66], Qwen2.5-VL [11], Intern2.5-VL [15], InternVideo2.5 [70], LLaVA-Next-Video [71], Video-LLaMA [13], and Video-LLaVA [14]. This list is not exhaustive and continues to grow rapidly, with each model continuously improving its capabilities through newer versions and updates.

Adapting general-purpose large models to sensitive tasks such as anomaly detection and understanding is crucial. In this work, we introduce instance-conditional anomaly-aware tokens that guide these models toward improved anomaly understanding, without requiring any modification of the pre-trained models themselves.

**Large Models in Video Anomaly Detection.** Recent advancements in video anomaly detection have been significantly influenced by the integration of large pre-trained models and MLLMs. Several recent works [72, 73, 74, 75] have leveraged pre-trained vision-language CLIP [26] model improving the anomaly detection process by including textual information alongside visual features. Zanella et al. [76] introduces a training-free framework that uses the pre-trained BLIP-2 [77] captioning model to generate captions for each frame and prompts LLM to estimate an anomaly

score. Lv et al. [78] introduced video-based large language model in a weakly supervised setting, which is able to detect anomalies and explain their details, highlighting the potential of combining video data with language models for anomaly detection tasks. In addition to textual and visual modalities, Tang et al. [17] introduces a motion modality, calculated from temporal and spatial information in video frames using Gunnar Farneback's algorithm [79]. Tang et al. [17] also introduces a dataset of anomalous videos with language descriptions and question-answer pairs. Zhang et al. [18] introduces a novel large-scale hierarchical video anomaly dataset for multi-granular anomaly comprehension, enabling multi-modal instruction tuning for more detailed anomaly detection, and also presents an Anomaly-focused Temporal Sampler (ATS) to select relevant frames for feeding into MLLMs based on respective anomaly scores.

These approaches to video anomaly understanding [17, 18, 78] typically rely on partial or full finetuning of generalist models to improve their ability to understand anomalies. However, such finetuning often leads to catastrophic forgetting, where the model loses its general knowledge. To address this issue, we propose the use of learnable, instance-conditional anomaly-aware context tokens that guide the generalist model toward anomaly understanding without sacrificing its broader capabilities.

**Prompt Learning.** Prompt learning, originating from the natural language processing domain, builds on the idea of knowledge probing, where cloze-style "fill-in-the-blank" prompts are used to get answers from pre-trained language models [21]. While effective for adapting language models to downstream tasks such as sentiment analysis, their manually crafted prompts are often suboptimal. To address this, continuous prompt learning was introduced, where continuous vectors in the embedding space are optimized to better exploit the capabilities of language models, without being limited to the discrete word representations [80, 81, 82].

This concept has been extended to vision-language models (VLMs) such as CLIP [26], with the objective of adapting these models to new tasks, surpassing the performance of zero-shot model with hand-crafted prompts, and ultimately achieving domain generalization. Instead of updating the model parameters, prompt learning inserts a small set of learnable embeddings, known as prompt tokens, into the input space, offering efficiency in both parameter count and convergence rate [22, 23, 24, 25]. Zhou et al. [22] optimizes continuous sets of prompt vectors that replace the context words in a prompt as an input for the language encoder of a frozen CLIP model [26], while, in a subsequent work, Zhou et al. [23] improves upon this approach by introducing conditional prompts on visual features to address generalization issues, specifically overfitting on base classes when handling unseen classes. Similarly, Khattak et al. [25] optimizes the prompts in both the vision and language branches of CLIP [26] to enhance alignment. On the other hand, Khattak et al. [83] addresses previous challenges by adding regularization to optimize both task-specific and task-agnostic representations.

In contrast to existing prompt learning methods, which are increasingly tailored to dual-encoder CLIP-like models, we propose a method designed specifically for video-based question answering (VideoQA), with a focus on anomaly understanding. To achieve this, we extend the idea of prompt learning to MLLMs. To our knowledge this is the first prompt learning application for MLLMs.

## 3. Preliminaries

In this section, we review the key components used in this work: BatchNorm-based Weakly Supervised Video Anomaly Detection (BN-WVAD) [58], Anomaly Focused Temporal Sampler (ATS) [18], Context Optimization (CoOp) [22], and Conditional Context Optimization (CoCoOp) [23]. These are organized into subsections based on their interactions and collectively form the foundation of our approach.

### 3.1. Review of BN-WVAD & ATS

We review the weakly supervised anomaly detection method along with an anomaly-focused temporal sampler, as these components will directly interact and heavily depend on each other in the following sections.

**BatchNorm-based Weakly Supervised Video Anomaly Detection (BN-WVAD).** We adopt the BN-WVAD [58] model as our anomaly detection algorithm for relevant frame filtering. An overview of its architecture is shown in Fig. 1. This method leverages the statistical properties of anomalous events, which typically deviate from normality, by exposing this information through a BatchNorm layer [85]. The input batch to the vision encoder consists of equal bags of normal and abnormal videos, following the Multi-Instance Learning (MIL) paradigm [4, 51, 52]. The vision encoder, as proposed in the original paper, is a frozen I3D [84], followed by a feature enhancer comprising a Global and Local Multi-Head Self-Attention layer [57]. Zhou et al. [58] introduced a new abnormality criterion called the Divergence of Feature from Mean (DFM), which quantifies the divergence of a feature from the BatchNorm mean vector using a Mahalanobis-inspired distance [86]. The DFM is formulated as follows:

$$\mathrm{DFM}\left(X, \mu, \sigma^2\right) = \sqrt{(X-\mu)^{\mathrm{T}} \Sigma^{-1} (X-\mu)}, \quad (1)$$

where $X$ represents the hidden features, $\mu$ is the mean calculated in a BatchNorm layer, and $\Sigma = \mathrm{diag}(\sigma^2)$ is the covariance matrix, with $\sigma^2$ being the variance.

The paper also introduces a novel loss function based on the DFM criterion, called the Mean-based Pull-Push (MPP) loss. This loss function optimizes the model by pulling normal features closer to the mean and pushing abnormal features further away. The MPP is formulated as follows:

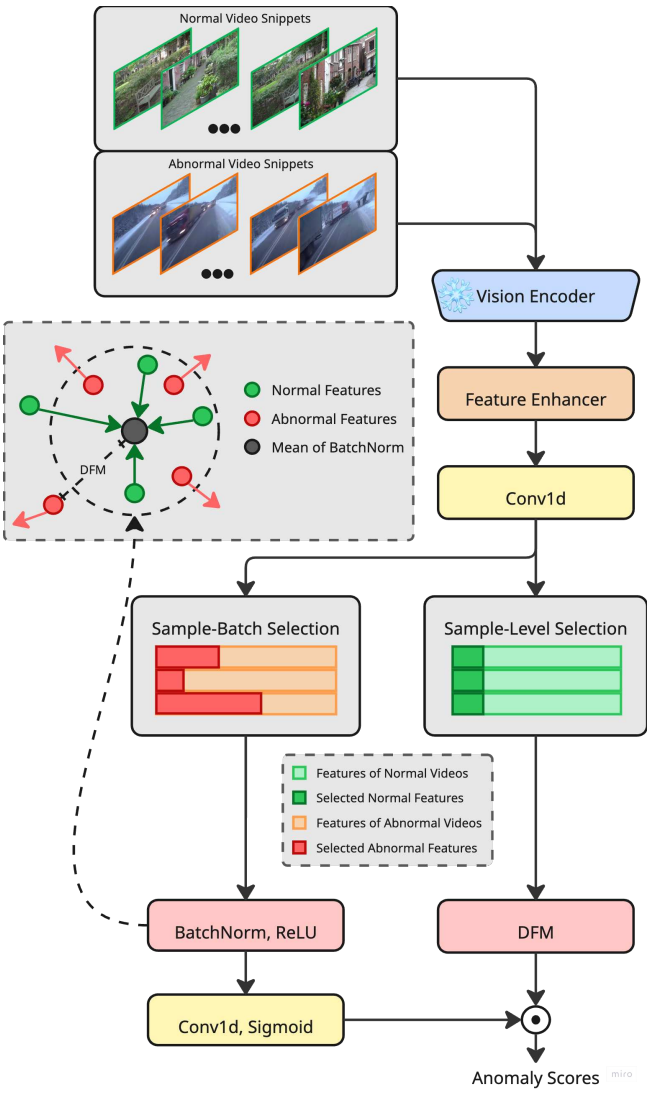



Figure 1: **Overall architecture of the Anomaly Scorer (BatchNorm-based Weakly Supervised Video Anomaly Detection [58]).** The input mini-batch consists of half normal and half abnormal videos, which are embedded by a frozen I3D [84] followed by a Transformer-based enhancer [57], producing enhanced features. "DFM" refers to the Divergence from Mean criterion (Eq. 1).

$$\mathcal{L}^{\text{mpp}}\left(X^n, X^a, \hat{\mu}, \hat{\sigma}^2\right) = \frac{1}{K}\sum_{k=1}^{K}[1 + \text{DFM}\left(X^n, \hat{\mu}, \hat{\sigma}^2\right) - \text{DFM}\left(X^a, \hat{\mu}, \hat{\sigma}^2\right)],$$

where $X^n$ and $X^a$ represent the normal and abnormal features, respectively, each containing $K$ instances. The $\hat{\mu}$ and $\hat{\sigma}^2$ are the running mean and variance, calculated from $\mu$ and $\sigma^2$, and updated using the exponential moving average (EMA) [87] to capture the long-term statistics of the feature distribution.

Zhou et al. [58] also propose two new abnormal snippet selection strategies, namely, batch-level selection (BLS) and sample-batch selection (SBS). BLS uses the statistical properties of BatchNorm to identify potential abnormal snippets across the entire mini-batch, rather than individual videos. This strategy adjusts the proportion of abnormal snippets selected from both the video and the mini-batch, making it more flexible to varying abnormality ratios across different videos. SBS strategy simply combines the selected snippets from both commonly used sample-level selection [54, 88, 89] and BLS strategies, addressing the limitations of each.

The anomaly score is computed by aggregating the DFM scores and the predictions of an anomaly classifier through element-wise multiplication. The anomaly classifier (represented by the [Conv1d, Sigmoid] block in Fig. 1) is trained solely on normal snippets from normal videos, minimizing the impact of label noise.

The overall training objective for this model is as follows:

$$\mathcal{L} = \mathcal{L}^{\text{nor}} + \lambda_1 \mathcal{L}_1^{\text{mpp}} + \lambda_2 \mathcal{L}_2^{\text{mpp}},$$

where $\mathcal{L}^{\text{nor}}$ is the $L_2$ norm loss for the anomaly classifier, $\mathcal{L}_1^{\text{mpp}}$ and $\mathcal{L}_2^{\text{mpp}}$ are the MPP losses computed for the hidden features from the first and second Conv1d layers, respectively. $\lambda_1$ and $\lambda_2$ are hyperparameters used to weight the respective losses.

**Anomaly-focused Temporal Sampler (ATS).** The density-aware sampler ATS [18] selects frames based on the anomaly scores provided by the anomaly scorer, giving priority to frames with higher scores and minimizing the use of less relevant snippets. Anomalous frames usually contain more diverse and relevant information than normal frames [54], making them more suitable for processing by heavier MLLMs. This approach ensures that the selected frames capture both critical anomaly frames and essential contextual information. It overcomes the limitations of methods such as dense window sampling that can introduce redundancy [76], and uniform sampling that may overlook key anomalies [16, 17], especially in short videos. The method proposes that anomaly scores are treated as a probability mass function, accumulated along the temporal axis to compute the cumulative distribution function. Frames are then sampled uniformly based on this cumulative distribution.

### 3.2. Reviews of CoOp & CoCoOp

We review both prompt learning techniques, as one is a continuation of the other, and both are crucial for understanding the main contribution of this paper.

**Context Optimization (CoOp).** CoOp [22] presents a straightforward approach for adapting CLIP-like vision-language models [26] to specific downstream tasks. It addresses the challenge of manual prompt engineering by

automatically learning the optimal "context" for prompts. This is achieved by replacing the context words of a prompt with continuous learnable vectors, while keeping all pre-trained parameters of the CLIP-like model frozen.

The core idea behind CoOp is to replace fixed, hand-crafted context (such as "a photo of a") with a set of continuous, learnable vectors. The paper [22] explores this approach for the image recognition task specifically. These vectors, denoted as $\{v_1, v_2, ..., v_M\}$, are designed to capture the most effective prompt context for a given downstream vision task, with each of these $M$ vectors having the same dimensionality as the word embeddings in the model. For a specific class $i$, the prompt $t_i$ is then constructed by concatenating these learnable context vectors with the word embedding of the class name $c_i$. This forms the prompt $t_i = \{v_1, v_2, ..., v_M, c_i\}$. The prompt is passed down the text encoder of CLIP into the shared embedding space of vision and text encoder. These continuous context vectors are learned directly from the downstream task data, allowing the model to learn task-specific prompt contexts, as opposed to relying on manual prompt engineering.

**Conditional Context Optimization (CoCoOp).** CoCoOp [23] builds on the foundation of CoOp [22] to address its limitations, more specifically, the issue of overfitting to the seen classes, which results in poor performance on unseen classes within the task. This is done with the introduction of the Meta-Net, a learnable lightweight neural network, which takes as an input the embedding of the image and outputs a conditional token, which is added to the context tokens introduced by CoOp [22]. The Meta-Net is composed of a two-layer bottleneck architecture (Linear-ReLU-Linear), where the hidden layer decreases the input dimension by 16 times. In this method, the trainable parameters include both the context vectors and the Meta-Net. Unlike CoOp's static prompts, this approach adapts to each new instance, making it less sensitive to unseen classes within the task, and more dependent on the visual information of each instance. The results in the paper demonstrate improved generalization from base classes to new classes within the same task, as well as improved cross-dataset transfer and domain generalization, when compared to both CoOp and the out-of-the-box CLIP model [26].

## 4. Methodology

In this section, we describe the methodology proposed in this study, outlining the overall approach, the design choices made, and the key contributions that distinguish it from existing work. This provides a clear foundation for understanding how the proposed framework addresses the challenges discussed earlier. An overview of our approach is shown in Fig. 2. Our method builds upon the framework proposed by Holmes-VAU [18], incorporating several modifications and improvements to address its limitations. Specifically, we target two key issues: first, we overcome the original method's sensitivity to prompt variations and improve its domain generalization; second, we improve the model's efficiency to move explainable anomaly detection closer to real-time streaming applications.

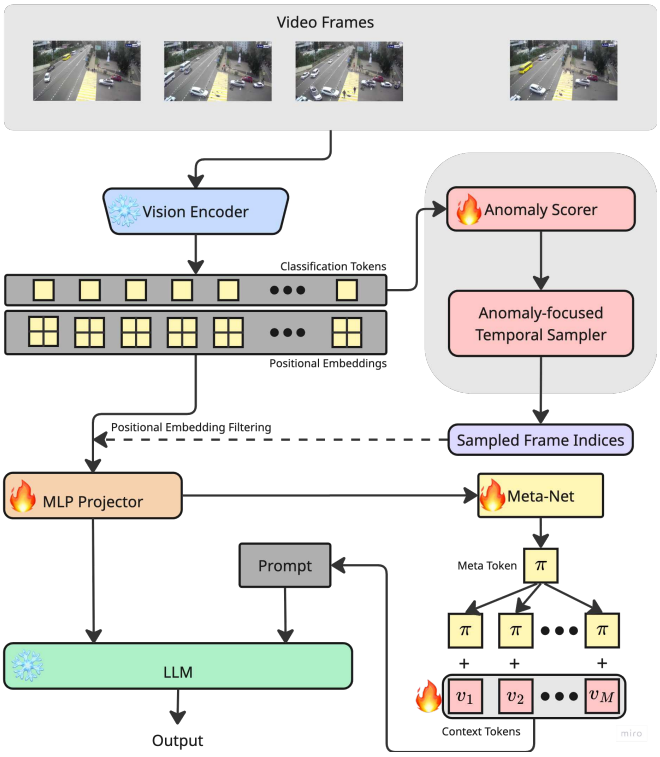



Figure 2: **Architecture of the proposed method.** The model processes an input video through the anomaly scorer for relevant frame sampling. The extracted visual features are fed into an MLP projector and combined with instance-conditional context tokens (combination of a meta token and context tokens) and the prompt are passed to the LLM for textual output. Training is divided into three phases, with different modules being trainable or frozen in each phase. The training phases are shown in Fig. 3.

### 4.1. Model Architecture Details

For the vision encoder, we use a frozen VideoMamba [28]. The MLP projector is initialized from scratch and the large language model (LLM) used is the pre-trained InternLM2 [90] with 1.8 billion parameters. For anomaly detection, we use the BN-WVAD [58] as the anomaly scorer and use the ATS method for anomaly-focused temporal sampling, as described in Section 3.1. In our implementation of BN-WVAD, the vision encoder is replaced with VideoMamba instead of I3D [84]. The Meta-Net and context tokens follow the same configuration as detailed in Section 3.2 for CoOp [22] and CoCoOp [23].

### 4.2. Integration of Instance-conditional Context Tokens

The primary contribution of this work lies in adapting the key idea of CoCoOp [23], instance-conditional context tokens, to an MLLM architecture. While the conceptual motivation of our implementation remains the same with the original method, using the instance-specific context embeddings, the technical implementation differs due to the nature of the multi-modal input (video-text pairs) and the downstream task. It is also important to note that the original CoCoOp was designed for a dual-encoder architecture such as CLIP, where the vision and text encoders are independent. In contrast, the architecture used in our setup features a strongly connected encoder design, where the text encoder depends heavily on the outputs of the vision encoder. Because of this fundamental difference we need a different integration strategy for instance-conditional context tokens.

In our adaptation, the instance-specific context tokens do not replace the entire prompt. Instead, the learnable instance-conditional context tokens ⟨ctx_token⟩ are appended at the end of the prompt. This design choice comes from the fact that our prompt can contain variable types of questions (e.g., descriptive, judgmental, or analytical), which cannot be substituted fully by a fixed set of learnable vectors. Therefore, these context embeddings serve more as task-adaptive vectors that bias the model toward specific focus areas, in our case, abnormalities in the input video. This design addresses the challenge of prompt sensitivity that can occur when finetuning the LLM component of an MLLM. It is important to note that the pre-trained LLM already possesses generalist knowledge and does not exhibit prompt sensitivity issues due to its vast pre-training on multiple datasets. Rather than finetuning the LLM, which could lead to overfitting on specific questions and their phrasing, task-specific information is incorporated by updating the learnable context tokens. This approach avoids modifying the model weights and lowering the likelihood of overfitting to the limited training set. To further enhance robustness to prompt sensitivity and improve domain generalization, these learnable context tokens are conditioned on the visual input, specifically, the output of the Meta-Net. This ensures that the context tokens takes into account both the visual information and the corresponding question-answer pairs, reducing the risk of overfitting to specific formulations of question-answer pairs during the training process.

The final prompt follows to the formatting convention used in InternVL2 [91], and is structured as follows:

$$\text{“Frame 1: } \langle \text{img}_1 \rangle, \ldots, \text{Frame n: } \langle \text{img}_n \rangle$$
$$\text{Please provide a detailed description of the video.}$$
$$\langle \text{ctx\_token}_1 \rangle \ldots \langle \text{ctx\_token}_M \rangle \text{”},$$

for visual input consisting of $n$ frames and the pre-defined $M$ context tokens. This modified prompt is used for forward pass through the tokenizer and subsequently into the LLM.

The Meta-Net receives as input the projected positional embeddings generated by the vision encoder. Let the input be denoted as $x \in \mathbb{R}^{B\times(T\cdot H\cdot W)\times C}$, where $B$ is the batch size, $T$ is number of frames, $H = W$ is the height and width of the positional embedding, and $C$ is the context dimension. Note that the context embedding dimension $C$ must be equal to the LLM's token embedding size, to make it compatible during concatenation. Firstly, we apply a mean pooling across the spatio-temporal dimension to obtain:

$$\bar{x} = \frac{1}{T \cdot H \cdot W} \sum_{k=1}^{T\cdot H\cdot W} x^{(k)} \in \mathbb{R}^{B\times 1\times C}$$

This pooled feature $\bar{x}$ is passed to the Meta-Net as input. Let $h_{\boldsymbol{\theta}}(\cdot)$ denote the Meta-Net, parametrized by $\boldsymbol{\theta}$, and let $\{v_1, v_2, \ldots, v_M\}$ be a set of learnable context vectors. Each instance-conditional context token is computed as:

$$v_m(\bar{x}) = v_m + \pi, \text{ where } \pi = h_{\boldsymbol{\theta}}(\bar{x}),\ m = \{1, 2, \ldots, M\}$$

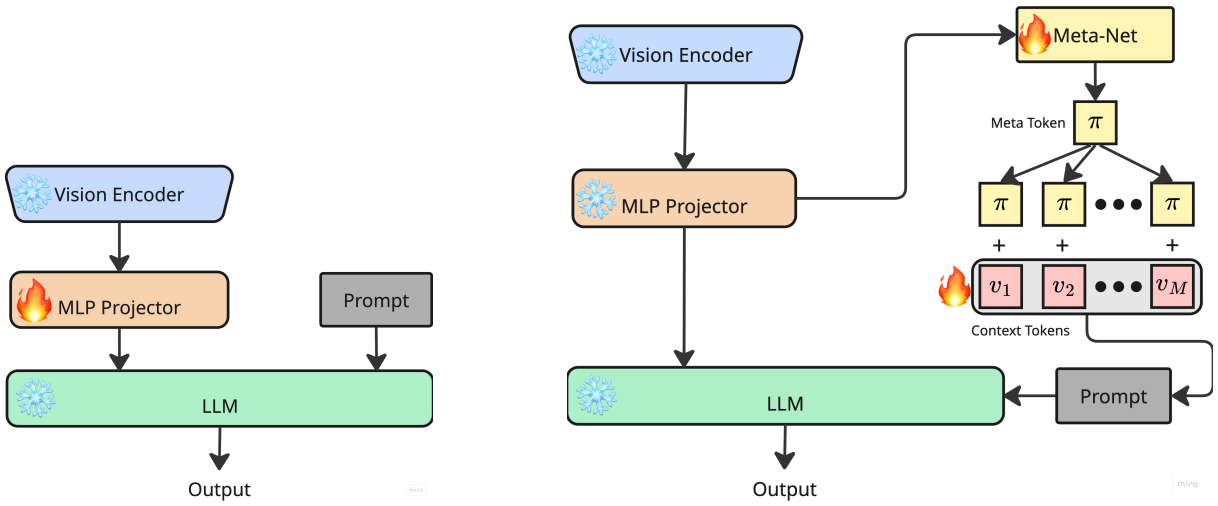
These learnable instance-conditional context vectors replace the placeholders ⟨ctx_token⟩ in the prompt string after the tokenization step.

This differs from the original CoCoOp pipeline, where the input was a single image embedding. Here, the input contains an explicit spatio-temporal dimension, which we reduce via pooling to maintain compatibility with the context tokens and, hence, the LLM architecture.

### 4.3. Integration of VideoMamba

VideoMamba [28] is a state-of-the-art video understanding model based on the Mamba architecture [34], specifically designed to efficiently handle long and fine-grained action sequences. In this work, we utilize the medium-sized variant of VideoMamba, pre-trained on the Kinetics-400 dataset [92], which processes video clips with fixed 32 frames at a resolution of $224 \times 224$ pixels, and has approximately 74 million parameters. The decision to use a 32-frame input configuration is driven by the nature of the anomaly detection task, where anomalies can occur at any moment, and missing critical frames could significantly impact performance. Additionally, a larger input frame size generally leads to better results across all tasks. By capturing a longer temporal context, the model's ability to detect subtle or short-lived deviations from normal behavior is improved [93, 94]. Therefore, a larger number of input frames increases the likelihood of identifying anomalies without skipping key moments.

In our setup, VideoMamba replaces InternViT [91] as the visual encoder, which was previously used in Holmes-VAU [18]. The lightweight 300-million-parameter variant of InternViT, named as InternViT-300M, is used in the lighter versions of InternVL2 [91] and its iterations. It is built on a Transformer architecture with self-attention [95]. It is pre-trained using various strategies to support general-purpose visual understanding. While it has strong performance across various tasks, its self-attention mechanism




(a) **Stage 1:** MLP warmup (b) **Stage 2:** Instance-conditional context tokens training

Figure 3: **MLLM incremental training stages.**

introduces quadratic computational and memory costs with respect to the number of input frames, which limits its efficiency for processing longer video sequences. In contrast, VideoMamba's Mamba-based architecture scales linearly with sequence length, making it more suitable for scenarios that require efficient handling of extended temporal information.

Since VideoMamba closely follows the Vision Transformer (ViT) architecture [33] and replaces ViT's self-attention mechanism with the bidirectional Mamba block [96] for video sequences, its use as a vision encoder in our approach aligns with the widely used "ViT-MLP-LLM" paradigm [29, 30, 31, 32]. The proposed MLLM would be referred to as **InternMambaVL**.

### 4.4. Training Details

In this section, we describe the two primary training processes involved in our method: the training of the anomaly scorer and the multi-stage training of MLLM with instance-specific context tokens.

**Training of Anomaly Scorer.** The anomaly scoring method BN-WVAD [58] uses the I3D [84] vision encoder to extract visual features for model training. In our approach, since the overall architecture is based on VideoMamba [28], we use it instead to extract features from the anomaly detection datasets and train the anomaly scorer accordingly. The visual features are extracted prior to the training, as the method itself requires an input of equally sized instances of normal and abnormal videos, with all of their snippets, which is computationally expensive. The extracted features are sampled at a rate of 1, as previously mentioned, to ensure that no critical frames are missed, considering the nature of the downstream task. During training, upon receiving the input features, we apply the distortion strategy known as "Noise All Patches", as introduced in GeneralAD [97], that essentially adds random Gaussian noise to the embeddings. This technique is used to improve the model's robustness to noise and serves as a sort of augmentation in the embedding space.

**Training of MLLM with Instance-Conditional Context Tokens.** The training of InternMambaVL begins with a Multi-Layer Perceptron (MLP) warmup stage, illustrated in Fig. 3a. During this phase, only the MLP component is trainable, while both the vision and language models are kept frozen. As it can be seen from Fig. 3a, the MLP serves as a bridge between the vision and language representations. Since the vision and language models are pre-trained on distinct datasets for different tasks, this warmup stage is essential to align their feature spaces. It enables the MLP to adapt the outputs of the vision model into a representation that is more understandable and technically compatible with the input of the language model.

Using the model trained in the previous stage, we proceed to integrate the Meta-Net and context tokens into the architecture for further training, as shown in Fig. 3b. The core MLLM remains frozen, with all of its components kept constant and only the newly added Meta-Net and context tokens are trainable. During this stage, the Meta-Net and context tokens learn to better handle the anomaly explanation task, providing the necessary additional cues to enhance the LLM's response to the user's instruction. This approach is aimed to replace the LoRA finetuning strategy [98], previously used in [18], which suffers from prompt sensitivity and catastrophic forgetting of the model's general knowledge [99]. Once this step is completed, the training process is concluded. It is important to note that during the experimentation phase, the InternVL2 [91] model with instance-conditional context tokens will also be utilized. For this, the InternVL2 model is trained only in the second

stage, as its vision and text representations are already well aligned from its vast pre-training. Additionally, to increase the model's robustness to frame sampling strategies, frames used in both training stages are sampled using segment-based random sampling technique.

## 5. Data

In this section we introduce the datasets used in this paper.

**UCF-Crime.** UCF-Crime [4] is a large-scale indoor and outdoor video surveillance dataset containing 1,900 real-world videos labeled across 13 types of anomalous events, including abuse, arrest, arson, assault, burglary, explosion, fighting, road accidents, robbery, shooting, shoplifting, stealing, and vandalism. The training set includes 800 normal and 810 abnormal videos annotated at the video level. The testing set consists of 140 normal and 150 abnormal videos, each with precise temporal annotations, for frame-level evaluation.

**XD-Violence.** XD-Violence [100] is a multi-modal, multi-source dataset featuring 4,754 untrimmed videos collected from diverse environments including movies, games, live scenes, and surveillance footage. It captures six types of anomalous events: abuse, car accident, explosion, fighting, riot and shooting. The training set contains 2,049 normal and 1,905 abnormal videos with video-level annotations, while the testing set includes 300 normal and 500 abnormal videos annotated at the frame level. XD-Violence has both video and audio modalities.

**HIVAU-70k.** HIVAU-70k [18] is a large-scale benchmark developed for hierarchical instruction-based video anomaly understanding across multiple granularities. The dataset introduces 70,000 multi-granular annotations at the clip, event, and video levels, including 5,443 videos, 11,076 events, and 55,806 clips. The final form of the dataset is a video-based question answering (VideoQA) task focused on anomalies.

The dataset is constructed using a semi-automated annotation pipeline that combines manual video segmentation with recursive free-text annotations generated by LLMs. A video perception model is first used to extract detailed captions for each clip, that is expanded into an event and video summary with the use of LLMs. Based on these summaries, structured prompts are created to extract various types of responses: captions, judgments, descriptions, and analytical answers. All LLM-generated outputs are manually verified to ensure high-quality annotations. To build instruction-tuning data for video anomaly understanding, the method matches free-text annotations with pre-designed anomaly-related user instructions.

Built upon the UCF-Crime [4] and XD-Violence [100] datasets, HIVAU-70k provides a detailed and extensive dataset for the field of advancing video anomaly understanding. Including annotations at multiple granularities should enable the model to develop both short-term and long-term reasoning capabilities.

**HAWK.** HAWK [17] is a benchmark designed to improve instruction-based video anomaly understanding by introducing rich language annotations and question-answer pairs for anomaly scenes across seven widely-used video datasets: UCF-Crime [4], ShanghaiTech [5], CUHK Avenue [3], UCSD Ped1 [101] and Ped2 [2], DoTA [8], and UBnormal [102].

The dataset is constructed through a semi-automated annotation process that begins with generating textual captions of anomalous events using perception tools. These captions are then refined and expanded into detailed descriptions using LLMs, followed by manual verification for quality assurance. Using these verified descriptions, HAWK generates open-ended question–answer pairs for each scenario is aimed to improve the model's ability to handle diverse user queries, which are then manually checked.

**CUVA.** CUVA [16] is a benchmark designed to improve understanding of video anomalies, and offers a rich collection of question-answer pairs and detailed annotations for unusual events. The dataset includes 1,000 video clips, totaling 32.46 hours, and features 6,000 question-answer pairs. These videos are drawn from popular platforms like Bilibili and YouTube, with a focus on VideoQA tasks related to anomalies.

The dataset is broken down into several key subtasks, that answer the following questions: "Why did this anomaly occur?", "What caused the anomaly?", and "How serious is this anomaly?". It covers 11 broad categories, including incidents like animals hurting people, pedestrian accidents, traffic violations, fires, fights, thefts, and vandalism, covering 42 fine-grained types of anomalies.

What sets CUVA apart from other video anomaly datasets, like HAWK [17] and HIVAU-70K [18], is that it is fully hand-annotated, ensuring the highest quality and accuracy of the labels possible.

## 6. Experiments

Our approach is primarily evaluated across the following key experimental settings: 1) performance of the proposed model on the same dataset (Section 6.1); 2) analysis of prompt sensitivity (Section 6.2); 3) assessment of domain generalization (Section 6.3); 4) anomaly detection results with a comparison to relevant methods (Section 6.4); and 5) ablation study for different components of the proposed approach (Section 6.5). In Section 6.4, we provide an implementation overview for the video anomaly detection task separately. Below, we outline the experimental setup for the video anomaly understanding task.

**Implementation Details.** Our proposed method is initialized as described in Section 4.1, using the VideoMamba visual encoder [28], a newly initialized MLP projection layer, and InternLM2 [90] with 1.8 billion parameters as an LLM. It is important to note that all models for instruction-tuning described in this section are trained on the HIVAU-70k dataset. For all experiments, a Stage 1 warmed-up InternMambaVL

model is used, unless stated otherwise. The Stage 1 MLP warmup, shown in Fig. 3a, is conducted for 1 epoch with a batch size of 4 and a gradient accumulation step of 8.

To comprehensively evaluate the impact of integrating instance-conditional context tokens into the MLLM, all experiments also involve training the InternVL2 model [91] with 2 billion parameters under the same settings. Both the InternMambaVL and InternVL2 models, when integrated with instance-specific context tokens (shown in Fig. 3b), are trained for 3 epochs, as loss convergence slows significantly before the third epoch. The naming convention for instance-conditional context tokens integrated into the MLLMs InternMambaVL and InternVL2 is **InternMambaVL+C** and **InternVL2+C**, respectively. In these settings, the batch size is 1, with a gradient accumulation step of 32. It is important to note that InternMambaVL samples 16 to 32 frames per video and InternVL2 samples from 6 to 12 frames.

Training for both Stage 1 and Stage 2 uses the AdamW optimizer [103] with cosine learning rate decay [104], a warm-up period, and a learning rate of $4e{-}5$. All experiments are conducted on a NVIDIA A100 40GB GPU.

**Baselines.** To assess the impact of learnable instance-conditional context tokens as an alternative to traditional or LoRA finetuning [98], we compare the developed models for InternVL2 MLLM against two baselines: the out-of-the-box InternVL2, which serves as a zero-shot model, and Holmes-VAU [18], which has its LLM finetuned using LoRA [98] on the same dataset. Similarly, for the InternMambaVL MLLM, we compare the developed models to two baselines: the stage 1 warmed-up InternMambaVL (used as a zero-shot model), and InternMambaVL-LoRA, which undergoes LoRA finetuning under the same settings as the Holmes-VAU method. It is important to note that stage 1 pre-trained InternMambaVL is used for comparison purposes. However, it cannot be considered a fully developed MLLM, as it has not undergone the multi-staged extensive training on large-scale general datasets that InternVL2 has. We make the comparison for both MLLMs to better assess the effect that learnable instance-conditional anomaly-aware tokens have on the task of video anomaly understanding.

**Evaluation Metrics.** To evaluate the quality of reasoning texts generated by the baseline models and the proposed approach for the task of video anomaly understanding, we adopt several standard text generation metrics: BLEU [105], which measures n-gram precision between generated and reference texts; CIDEr [106], originally proposed for image captioning, but here repurposed to assess similarity with human-annotated reasoning using TF-IDF-weighted n-gram similarity; METEOR [107], which evaluates unigram matches with consideration for synonymy, stemming, and word order; and ROUGE [108], which measures the recall of overlapping n-grams and sequences. It is important to note that there are several BLEU scores, based on the different levels of n-grams specified (1-4), and we use the cumulative of these. These metrics are computed by comparing the model's outputs with the annotated ground truth textual explanations.

In addition to traditional evaluation metrics, we use model-based scorers to assess the quality of generated responses. These evaluators judge responses based on how effectively they follow the given instructions, considering not only the prompt but also additional information such as reference answers and visual inputs. These evaluators can provide more detailed feedback than traditional metrics, which primarily focus on word overlap.

MLLM-as-a-Judge [109] is an evaluation framework that uses powerful MLLMs as evaluators for vision-language tasks. Given the original user prompt, the model-generated response, and a predefined scoring rubric, the MLLM evaluates how effectively the response aligns with the instruction and associated visual content. Since the original framework does not cover video-instruction scenarios, we adapt it for this purpose using Qwen2.5-VL [11], a 7-billion parameter MLLM. Our evaluation involves providing the MLLM with 16 representative frames sampled from the video, the corresponding instruction, and the model's response. The prompt template used for this setup is illustrated in Prompt Template 1.

Prometheus 2 [110] is a state-of-the-art open-source evaluator language model (LM) with 7 billion parameters, specifically trained to serve as an automatic evaluator for assessing the quality of responses generated by various language models. It demonstrates strong correlation with both human judgments and proprietary LM-based evaluators. Given the original user prompt, the model-generated response, a reference answer, and a predefined scoring rubric, Prometheus 2 evaluates how effectively the response aligns with the instruction and the reference. Unlike the MLLM-as-a-Judge framework, which repurposes existing multi-modal models, Prometheus 2 is specifically trained for the purpose of evaluation and explicitly incorporates a reference answer, which is treated as the gold standard. The prompt template used for Prometheus 2 is shown in Prompt Template 2.

### 6.1. Anomaly Reasoning Analysis

In this subsection, we examine the impact of integrating prompt learning on the downstream HIVAU-70k dataset [18]. Specifically, we compare the performance of the zero-shot model, the proposed model with instance-conditional context tokens, and the LoRA finetuned version. As previously mentioned, the experiments were conducted using both the InternVL2 and InternMambaVL MLLMs. The results for each model will be presented separately. All metrics are computed based on the test partition of the HIVAU-70k dataset.

**Results for InternVL2 MLLM.** As shown in Table 1, the LoRA finetuned version Holmes-VAU [18] excels in the same downstream task it was trained on. This result is expected, as finetuning specifically adjusts the model's parameters to produce output that aligns with the annotations present in the training data. The first thing that can be

**Template prompt for MLLM-as-a-Judge Scoring Evaluation**

**(System Input)**
You are a helpful assistant proficient in analyzing vision reasoning problems.
**(Instruction)**
Please serve as an unbiased judge in assessing the quality of the responses from AI assistants regarding the user's instruction and a video.
**Evaluation Steps**
Please examine the provided video attentively. Begin by conducting a detailed analysis of the responses provided. Capture your comprehensive observations and insights in the 'Analysis' section. Following your analysis, move on to the judgement phase, where you will make informed decisions or conclusions based on the analysis conducted. Give your final judgements in the 'Judgement' section. Ensure that your final output is in a JSON format with keys ``Analysis" for the initial response analysis, and ``Judgement" for your final judgement only. Ensure that the content under each key does not contain any nested JSON structures.
**Evaluation Method**
You will receive a single response from the AI assistant to the user's instruction. Use scores to show the quality of the response. Here is the detailed scoring rubric for evaluating the quality of responses from AI assistants:
**Poor (1):** The response significantly deviates from the user's instruction and fails to address the query effectively. It shows a lack of relevance, accuracy, and comprehensiveness. Creativity and granularity are absent or poorly executed.
**Fair (2):** The response addresses the user's instruction partially, with evident shortcomings in relevance, accuracy, or comprehensiveness. It lacks depth in creativity and granularity, indicating a superficial understanding of the user's inquiry.
**Average (3):** The response adequately addresses the user's instruction, showing a fair level of relevance, accuracy, and comprehensiveness. It reflects a basic level of creativity and granularity but may lack sophistication or depth in fully capturing the user's inquiry.
**Good (4):** The response is well-aligned with the user's instruction, demonstrating a high degree of relevance, accuracy, and comprehensiveness. It shows creativity and a nuanced understanding of the topic, with detailed granularity that enhances the response quality.
**Excellent (5):** The response perfectly adheres to the user's instruction, excelling in relevance, accuracy, comprehensiveness, creativity, and granularity. It provides an insightful, detailed, and thorough answer, indicating a deep and nuanced understanding of the user's inquiry.
**Notice**
Your assessment should identify whether the assistant effectively adheres to the user's instruction and addresses the user's inquiry.
In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses.
Do not allow the length of the responses to influence your evaluation.
Do not favor certain names or positions of the assistants. Be as objective as possible.
**Here is the input:**
```
[The Start of User Instruction]
{item['instruction']}
[The End of User Instruction]
[The Start of Assistant's Answer]
{item['response']}
[The End of Assistant's Answer]
```

Prompt Template 1: MLLM-as-a-Judge

**Template prompt for Prometheus Direct Assessment**

**(Instruction)**
You are a fair judge assistant tasked with providing clear, objective feedback based on specific criteria, ensuring each assessment reflects the absolute standards set for performance.
**Task Description:**
An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows:
``Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 5)"
4. Please do not generate any other opening, closing, and explanations.
```
The instruction to evaluate:
{item['instruction']}
Response to evaluate:
{item['response']}
Reference Answer (Score 5):
{item['reference_answer']}
```
**Score Rubrics:**
**"criteria"**: Is the model proficient in following a detailed process of analyzing and judging AI assistant responses based on user instructions and a video?
**"score1_description"**: The model fails to examine the video attentively and provide a coherent analysis. It gives judgments that are uninformed or irrelevant, lacking a structured approach to assessing the responses.
**"score2_description"**: The model intermittently follows the instructions and attempts an analysis, but often misses key insights or provides superficial judgments. The response lacks consistency in evaluating the quality of the responses.
**"score3_description"**: The model usually follows the steps in examining the video and analyzing the responses, providing a basic level of understanding. However, the analysis and judgment may lack depth or nuance in some areas.
**"score4_description"**: The model provides a thorough analysis and judgment, demonstrating a strong understanding of the process. It effectively identifies key aspects of the responses and offers a well-structured evaluation, though there may be occasional gaps in granularity or insight.
**"score5_description"**: The model excels in conducting a detailed, insightful analysis of the video and responses, showcasing an excellent grasp of the process. It provides a nuanced, well-rounded judgment that addresses all aspects of the task with precision and clarity.

Prompt Template 2: Prometheus

noticed is that InternVL2+$C$ achieves results that are closer to those of Holmes-VAU than to the off-the-shelf InternVL2 model. A more interesting comparison can be made between the out-of-the-box generalist model, InternVL2 [91], and InternVL2+$C$. Both models share the same underlying parameters, but InternVL2+$C$ differs by having additional instance-conditional context tokens, which are specifically trained for anomaly detection.

When comparing these models, it becomes evident that there is an inconsistency between the model-based metrics and traditional text generation metrics. Traditional metrics BLEU, ROUGE, METEOR, and CIDEr indicate that InternVL2+$C$ clearly outperforms InternVL2, with the former showing a performance improvement ranging from $2\times$ to $10\times$ across all granularities. These traditional metrics suggest that the words or word sequences generated

| Method | Granularity | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|---|
| InternVL2+$C$ | Clip | 0.0694 | 0.2958 | 0.3137 | 0.2619 | 2.2609 | 2.8164 |
| | Event | 0.1020 | 0.3145 | 0.3274 | 0.8857 | 1.8895 | 2.3464 |
| | Video | 0.1252 | 0.3100 | 0.3130 | 1.1248 | 1.6859 | 2.1901 |
| InternVL2 [91] | Clip | 0.0282 | 0.2122 | 0.2632 | 0.0777 | 2.3128 | 3.1533 |
| | Event | 0.0044 | 0.1055 | 0.1739 | 0.0119 | 1.7439 | 2.5559 |
| | Video | 0.0052 | 0.1146 | 0.1741 | 0.0186 | 1.7085 | 2.5442 |
| Holmes-VAU† [18] | Clip | **0.1294** | **0.3479** | **0.3635** | **0.5173** | **2.3248** | **3.1964** |
| | Event | **0.1414** | **0.3843** | **0.4129** | **1.0841** | **2.5852** | **2.8507** |
| | Video | **0.1655** | **0.3935** | **0.3886** | **1.2489** | **2.3879** | **2.8580** |

TABLE 1: **Comparison of anomaly reasoning performance on the HIVAU-70k dataset [18] for InternVL2+$C$ and its corresponding baselines.** "+$C$" denotes the integration of instance-conditional context tokens into the model and "†" denotes reproduced results. Results are reported across traditional and model-based evaluation metrics and at multiple granularities to assess anomaly reasoning in the downstream task, with higher scores indicating better anomaly understanding.

| Method | Granularity | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|---|
| InternMambaVL+$C$ | Clip | 0.0249 | 0.2189 | 0.2366 | 0.0831 | 2.0844 | 1.8801 |
| | Event | 0.0156 | 0.1753 | 0.2228 | 0.1006 | 1.8842 | 2.1739 |
| | Video | 0.0240 | 0.1837 | 0.2197 | 0.1199 | 1.8291 | 2.1797 |
| InternMambaVL | Clip | 0.0129 | 0.1532 | 0.1641 | 0.0338 | 1.5133 | 1.5851 |
| | Event | 0.0030 | 0.1137 | 0.1237 | 0.0716 | 1.3079 | 1.7833 |
| | Video | 0.0110 | 0.1074 | 0.1198 | 0.0583 | 1.2563 | 1.7107 |
| InternMambaVL-LoRA | Clip | **0.0488** | **0.2558** | **0.2729** | **0.1656** | **2.1742** | **2.0995** |
| | Event | **0.0577** | **0.2583** | **0.3017** | **0.4259** | **2.0150** | **2.2603** |
| | Video | **0.0731** | **0.2639** | **0.2906** | **0.5499** | **1.9621** | **2.3405** |

TABLE 2: **Comparison of anomaly reasoning performance on the HIVAU-70k dataset [18] for InternMambaVL+$C$ and its corresponding baselines.** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across traditional and model-based evaluation metrics and at multiple granularities to assess anomaly reasoning in the downstream task, with higher scores indicating better anomaly understanding.

by InternVL2+$C$ are more aligned on average with the annotated descriptions of anomalies. On the other hand, model-based metrics Prometheus 2 and MLLM-as-a-Judge show a contrasting results, with Prometheus 2 reporting comparable performance between the two models across all granularities, while MLLM-as-a-Judge favors the out-of-the-box InternVL2, suggesting that it performs better than InternVL2+$C$.

This disagreement in results also appears when comparing InternVL2 to Holmes-VAU. While Prometheus 2 clearly favors Holmes-VAU, suggesting a significantly better performance, MLLM-as-a-Judge indicates that the performance of both models is comparable across the clip granularity. This discrepancy may point to a potential bias in the MLLM-as-a-Judge metric toward out-of-the-box models, given that it operates in a similar manner to the zero-shot model. Since MLLM-as-a-Judge has not been trained on anomaly detection tasks and takes as input the original model's response, frames sampled from the video, and the user's instruction, it is highly likely that both this model and the out-of-the-box InternVL2 reference the same visual cues unrelated to the anomaly itself. This could explain why MLLM-as-a-Judge assigns a higher score to the out-of-the-box model.

**Results for InternMambaVL MLLM.** As shown in Table 2, the LoRA finetuned InternMambaVL-LoRA model outperforms in the same downstream task it was trained on, as expected. Interestingly, we notice a larger gap between InternMambaVL-LoRA and InternMambaVL+$C$ in terms of text generation metrics, while the model-based metrics are more aligned. In terms of text generation metrics, InternMambaVL+$C$ is closer to InternMambaVL than InternMambaVL-LoRA, but in model-based scores, the reverse is true.

This further highlights the disagreement observed in the anomaly reasoning results for the InternVL2-based models. The poor performance of both InternMambaVL and InternMambaVL+$C$ was expected, as discussed in previous subsections. Specifically, InternMambaVL does not contain well-aligned general knowledge, and consequently, InternMambaVL+$C$ also lacks this alignment. Nevertheless, InternMambaVL+$C$ performs better across all metrics

compared to InternMambaVL, indicating that the instance-conditional context tokens have effectively learned the necessary information to improve the model's ability for the task of anomaly reasoning.

## 6.2. Evaluation of Prompt Sensitivity

Combating the prompt sensitivity issue of finetuned models is the main focus of this research. To evaluate the proposed approach and its corresponding baselines for prompt sensitivity, we utilize the Qwen2.5 LLM [111] to paraphrase the original prompts from the HIVAU-70k dataset's test partition, which were designed by the annotators. Examples of the paraphrased questions can be seen below:

**Examples of Paraphrased Questions from HIVAU-70K**

1) **Original question:** Are there anomalies observed in the video clip?
   **Paraphrased question:** Is anything unusual or unexpected detected in the recorded footage?
2) **Original question:** Could you provide a summary of the anomaly events in this video?
   **Paraphrased question:** Could you give me an overview of the unusual occurrences depicted in this video?
3) **Original question:** How do the characteristics of this event support its classification as an anomaly?
   **Paraphrased question:** What factors contribute to this occurrence being classified as unusual or abnormal?

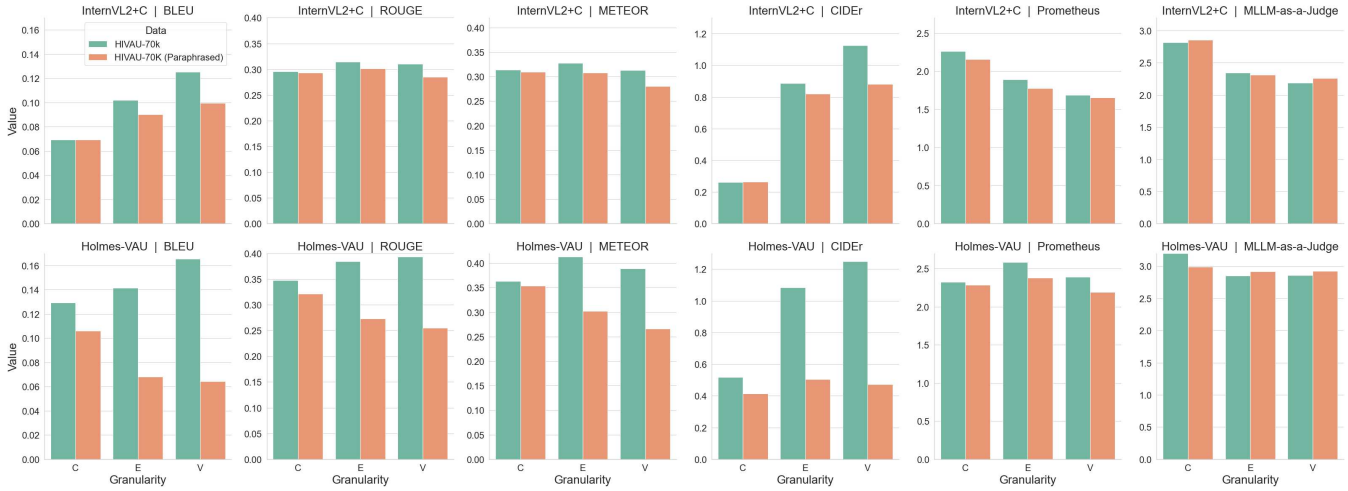
This test aims to determine whether rewording the same question, without changing its context, affects the performance of models trained on the downstream dataset. Similar to the previous subsection, we compare InternVL2+$C$ (integrated instance-conditional context tokens) with the out-of-the-box InternVL2, LoRA-finetuned Holmes-VAU, and InternMambaVL+$C$ with stage 1 warmed-up InternMambaVL, and LoRA-finetuned InternMambaVL-LoRA.

| Method | Granularity | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|---|
| InternVL2+$C$ | Clip | 0.0694 | 0.2936 | 0.3099 | 0.2643 | 2.1574 | 2.8561 |
| | Event | **0.0903** | **0.3014** | **0.3081** | **0.8206** | 1.7762 | 2.3141 |
| | Video | **0.0996** | **0.2852** | **0.2800** | **0.8817** | 1.6549 | 2.2573 |
| InternVL2 [91] | Clip | 0.0220 | 0.1759 | 0.2389 | 0.0396 | **2.3260** | 2.9396 |
| | Event | 0.0058 | 0.1203 | 0.1894 | 0.0115 | 1.8583 | 2.7126 |
| | Video | 0.0057 | 0.1229 | 0.1796 | 0.0080 | 1.7581 | 2.7123 |
| Holmes-VAU [18] | Clip | **0.1060** | **0.3216** | **0.3540** | **0.4135** | 2.2848 | **2.9899** |
| | Event | 0.0682 | 0.2732 | 0.3019 | 0.5036 | **2.3770** | **2.9208** |
| | Video | 0.0643 | 0.2548 | 0.2657 | 0.4723 | **2.1935** | **2.9246** |

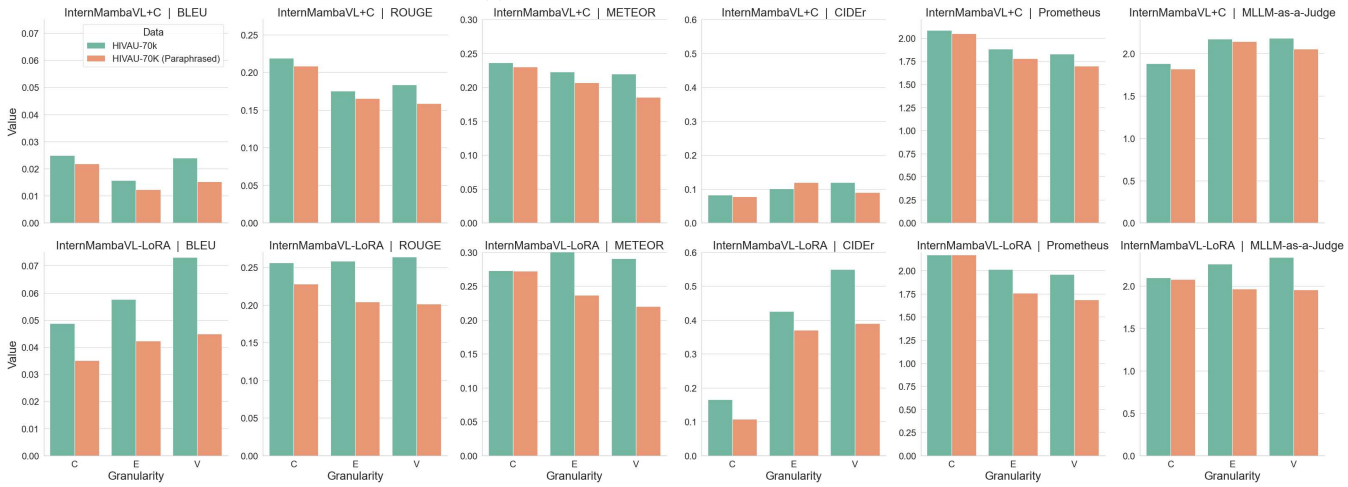
TABLE 3: **Evaluation of anomaly reasoning and prompt sensitivity for InternVL2+$C$ and its corresponding baselines on the HIVAU-70k dataset [18] using paraphrased prompts.** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across traditional and model-based evaluation metrics and at multiple granularities to assess both the robustness of anomaly reasoning and the models' sensitivity to variations in prompt wording, with higher scores indicating better anomaly understanding.

| Method | Granularity | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|---|
| InternMambaVL+$C$ | Clip | 0.0219 | 0.2083 | 0.2300 | 0.0775 | 2.0471 | 1.8177 |
| | Event | 0.0123 | 0.1651 | 0.2069 | 0.1195 | **1.7779** | **2.1417** |
| | Video | 0.0153 | 0.1587 | 0.1858 | 0.0901 | **1.6960** | **2.0516** |
| InternMambaVL | Clip | 0.0079 | 0.1132 | 0.1404 | 0.0196 | 1.4621 | 1.7044 |
| | Event | 0.0015 | 0.0881 | 0.0953 | 0.0505 | 1.2674 | 1.6364 |
| | Video | 0.0018 | 0.0826 | 0.0884 | 0.0364 | 1.1738 | 1.6395 |
| InternMambaVL-LoRA | Clip | **0.0351** | **0.2277** | **0.2723** | **0.1073** | **2.1701** | **2.0792** |
| | Event | **0.0422** | **0.2041** | **0.2365** | **0.3704** | 1.7589 | 1.9696 |
| | Video | **0.0449** | **0.2014** | **0.2202** | **0.3896** | 1.6859 | 1.9578 |

TABLE 4: **Evaluation of anomaly reasoning and prompt sensitivity for InternMambaVL+$C$ and its corresponding baselines on the HIVAU-70k dataset [18] using paraphrased prompts.** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across traditional and model-based evaluation metrics and at multiple granularities to assess both the robustness of anomaly reasoning and the models' sensitivity to variations in prompt wording, with higher scores indicating better anomaly understanding.




(a) InternVL2+$C$ vs Holmes-VAU [18]




(b) InternMambaVL+$C$ vs InternMambaVL-LoRA

Figure 4: **Performance comparison of InternVL2+$C$ and InternMambaVL+$C$ with their respective LoRA-finetuned versions on the HIVAU-70k dataset [18] and its paraphrased variant.** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across all metrics to evaluate both anomaly reasoning performance and sensitivity to prompt variations, with higher scores indicating better anomaly understanding.

**Results for InternVL2 MLLM.** As shown in Table 3, the results observed in the previous subsection, where the Holmes-VAU outperforms all other models with every metric, has shifted. This change can be attributed to the fine-tuned model's high sensitivity to prompts, which is clearly reflected in the results.

For this dataset, the noticeable discrepancy between model-based and traditional text generation metrics persists. Specifically, when the text generation metrics improve, the model-based metrics tend to decrease. Interestingly, both the event and video granularities perform best with the InternVL2+$C$ model, which integrates instance-conditional context tokens. In contrast, the clip granularity performs best with the Holmes-VAU model. This is likely because clip granularity focuses solely on the simpler caption question type, as well as the dominance of clips in the dataset compared to other granularities. The divergence between model-based and traditional metrics is most apparent at this granularity. According to model-based metrics, the InternVL2+$C$ performs similarly to both InternVL2 and Holmes-VAU. However, traditional metrics show that InternVL2+$C$ outperforms InternVL2, but still lags behind Holmes-VAU.

A clear illustration of Holmes-VAU's sensitivity, compared to InternVL2+$C$, can be seen in Fig. 4a. There, the traditional metrics reveal a significant score decrease between the paraphrased and original data, while the InternVL2+$C$ model remains stable and consistent across all metrics.

**Results for InternMambaVL MLLM.** As shown in Table 4, the performance of InternMambaVL-LoRA drops significantly when tested with paraphrased prompts com-

| Method | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|
| InternVL2+$C$ | 0.0027 | 0.1582 | 0.1589 | **0.0510** | 1.5184 | 2.5476 |
| InternVL2 [91] | **0.0045** | **0.1693** | **0.2182** | 0.0099 | **2.8698** | **3.5935** |
| Holmes-VAU [18] | 0.0005 | 0.1438 | 0.0946 | 0.0128 | 1.5050 | 2.4972 |

TABLE 5: **Evaluation of anomaly reasoning and domain generalization for InternVL2+$C$ and its corresponding baselines on the HAWK dataset [17].** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across traditional and model-based evaluation metrics to assess the anomaly reasoning and the capability of the domain generalization, with higher scores indicating better anomaly understanding.

| Method | BLEU [105] | ROUGE [108] | METEOR [107] | CIDEr [106] | Prometheus 2 [110] | MLLM-as-a-Judge [109] |
|---|---|---|---|---|---|---|
| InternVL2+$C$ | 0.0012 | **0.1408** | **0.1637** | **0.0299** | 1.4733 | 2.0777 |
| InternVL2 [91] | 0.0007 | 0.0632 | 0.1250 | 0.0000 | **2.2517** | **2.8517** |
| Holmes-VAU [18] | **0.0013** | 0.1365 | 0.1257 | 0.0183 | 1.8297 | 2.3354 |

TABLE 6: **Evaluation of anomaly reasoning and domain generalization for InternVL2+$C$ and its corresponding baselines on the CUVA dataset [16].** "+$C$" denotes the integration of instance-conditional context tokens into the model. Results are reported across traditional and model-based evaluation metrics to assess the anomaly reasoning and the capability of the domain generalization, with higher scores indicating better anomaly understanding.

pared to the original test prompts. In contrast, as seen in Fig. 4b, the InternMambaVL+$C$ model demonstrates more consistent results. Moreover, InternMambaVL+$C$ outperforms InternMambaVL-LoRA in model-based metrics for both event and video granularities. Although the performance of InternMambaVL+$C$ was suboptimal for anomaly reasoning for the downstream dataset, as discussed in the previous subsection, it shows greater consistency across the paraphrased dataset. This suggests that InternMambaVL-LoRA, like the other LoRA finetuned model Holmes-VAU, is sensitive to prompt variations.

In this benchmark, the stable performance of InternMambaVL+$C$ and the fluctuating performance of InternMambaVL-LoRA lead to InternMambaVL+$C$ performing closer to InternMambaVL-LoRA than to the base InternMambaVL. Again, the poor performance of InternMambaVL is due to its lack of a well-aligned general knowledge base, which disrupts its overall performance.

### 6.3. Domain Generalization

Generalization to out-of-distribution data is critical for machine learning models to function reliably in real-world settings, where distributional shifts are common [112]. Unlike humans, who adapt quickly to new or unseen conditions, machine learning models often struggle with such scenarios.

In previous subsections, we observed that including instance-conditional context vectors enables models to retain their general knowledge while improving their anomaly understanding. This setup also shows greater robustness to changes in prompts compared to finetuned baselines.

To assess whether this combination of general knowledge preservation and anomaly-aware context tokens transfer effectively to other anomaly understanding datasets, we evaluate on the test splits of the HAWK [17] and CUVA [16] datasets. For CUVA, we use the original test partition, whereas for HAWK, we retain the anomaly descriptions and append relevant prompts. Importantly, we exclude the UCF-Crime subset from HAWK to prevent potential data leakage, as our models were partially trained on videos from UCF-Crime. In contrast to the previous subsections, we exclude the InternMambaVL model from this evaluation. This decision is based on the observation that the stage 1 warmed-up InternMambaVL, along with its other variants, lacks sufficiently well-aligned and generalizable knowledge due to the lack of pre-training on diverse data. Including it in this comparison could introduce distortions in the results and complicate the evaluation of generalization to out-of-distribution anomalies.

As shown in Table 5, the out-of-the-box InternVL2 model outperforms both InternVL2+$C$ and Holmes-VAU across the majority of both traditional and model-based metrics. Furthermore, InternVL2+$C$ outperforms Holmes-VAU across all metrics. These results confirm that adding instance-conditional context tokens can improve domain generalization compared to finetuned models. However, the inclusion of instance-conditional tokens can also mislead the model in some cases. This may be due to the significant differences in anomaly types between the datasets: HAWK (excluding the UCF-Crime subset, as used in our case) focuses on general, socially unusual, and contextual anomalies such as loitering, unusual movements, and the presence of non-pedestrian objects. In contrast, HIVAU-70k [18], on which the instance-conditional context tokens are trained, primarily focuses on violence-related anomalies. The generalist InternVL2 model handles this broader range of anomalies better than the version with instance-conditional context tokens. While there is a limited number of video anomaly understanding datasets, the experiments conducted on the HAWK dataset suggest that, overall, instance-conditional context tokens outperform finetuned models.

For the CUVA dataset (Table 6), InternVL2+$C$ surpasses both finetuned Holmes-VAU and out-of-the-box In-

ternVL2 baselines on most traditional metrics. However, for the model-based metrics, it performs the worst, with the out-of-the-box InternVL2 achieving the best results. This further supports the observed bias of model-based metrics toward the generalist InternVL2 model, regardless of the dataset it is being evaluated on. The stronger performance of InternVL2+$C$ over both finetuned Holmes-VAU and the base InternVL2 on CUVA may be attributed to CUVA's anomaly categories being more closely aligned with those in HIVAU-70k than with HAWK's categories.

Taken together, results across both datasets indicate that instance-conditional context tokens improve domain generalization over finetuned models and, in certain cases, can even exceed the performance of a strong out-of-the-box generalist model.

### 6.4. Anomaly Detection Results

This subsection presents the anomaly detection results obtained from our experiments and analyzes them. We first provide the implementation details for the video anomaly detection task.

**Implementation Details.** The anomaly scorer is trained on the UCF-Crime and XD-Violence datasets following the original settings proposed in [58]. Videos are divided into 100 snippets, and features are extracted and linearly interpolated to this fixed length. Two Conv1D layers with a kernel size of 1 output feature dimensions of 32 and 16, respectively. Hyperparameters are set to $\lambda_1 = 5$ and $\lambda_2 = 20$, with selection ratios aligned to dataset distributions: $\rho_s = 0.1$ and $\rho_b = 0.2$. Optimization is performed using Adam [113] with a learning rate of $1\text{e}{-4}$ and a weight decay of $5\text{e}{-5}$. The model is trained for 3000 iterations using a mini-batch of 64 normal and abnormal videos. Feature extraction uses 5-crop augmentation for XD-Violence and 10-crop for UCF-Crime. All experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU.

**Evaluation Protocols.** We follow established evaluation protocols to ensure fair comparisons with previous methods. Specifically, for UCF-Crime, we use the area under the curve (AUC) of the frame-level receiver operating characteristic (ROC) curve as the primary metric. For XD-Violence, the frame-level average precision (AP) is the key metric for assessment.

**Results.** We compare our method against state-of-the-art weakly supervised anomaly detection approaches. The results are summarized in Table 7. Unlike prior works that train models separately on UCF-Crime and XD-Violence, our method is trained on a combined dataset of both, which introduces significant domain variability. Despite this challenge, our model achieves an AP of 81.46% for XD-Violence and an AUC of 84.01% on UCF-Crime, which are competitive results considering the added complexity from dataset mixture.

| Method | Backbone | XD-Violence AP/% | UCF-Crime AUC/% |
|---|---|---|---|
| RTFM [114] | I3D [84] | 77.81 | 84.30 |
| MSL [115] | I3D [84] | 78.28 | 85.30 |
| S3R [88] | I3D [84] | 80.26 | 85.99 |
| CU-Net [89] | I3D [84] | 78.74 | 86.22 |
| MGFN [116] | I3D [84] | 79.19 | 86.98 |
| UR-DMU [57] | I3D [84] | 81.66 | 86.97 |
| CLIP-TSA [72] | ViT [33] | 82.19 | 87.58 |
| VadCLIP [74] | ViT [33] | 84.51 | 88.02 |
| Yang et al. [75] | ViT [33] | 83.68 | 87.79 |
| BN-WVAD [58] | I3D [84] | 84.93 | 87.24 |
| BN-WVAD† | I3D [84] | 80.62 | 83.92 |
| Ours (separate) | VideoMamba [28] | 82.96 | 85.91 |
| Ours (combined) | VideoMamba [28] | 81.46 | 84.01 |

TABLE 7: **Comparison of video anomaly detection performance under combined and separate training settings with state-of-the-art weakly supervised methods.** "†" denotes reproduced results.

Our model uses a VideoMamba [28] backbone for feature extraction, which performs better in the same setup as the I3D [84] counterpart, as it can be seen from our reproduced results for BN-WVAD and our separately trained models results, displayed in Table 7. However, combining datasets like UCF-Crime and XD-Violence introduces challenges due to their different video characteristics and definitions of anomaly. This results in a domain shift and, consequently, negative transfer, where learning from both domains interferes with the model's ability to perform well on either one individually [117]. The drop in performance observed when training on the combined dataset compared to separate training (Table 7) further confirms the negative effect of domain shift.

### 6.5. Ablation Study

**Context Length.** The ablation study on context length is conducted across all datasets discussed in the previous subsections. We assess context lengths of 64, 512, and 1024, with results presented in Fig. 5. The choice of 1024 is based on its proximity to the average length of input tokens being fed into the LLM, taking into account both the prompt and the additional tokens appended during processing, such as image start and end tokens. To explore how context length influences performance, we select three values, ranging from 0 to 1024, to cover a broad range. The metrics shown in the figure are averaged across various granularities. As demonstrated, a context length of 512 consistently outperforms the other settings across most metrics and datasets. Moreover, it shows the least performance variation when prompt questions are paraphrased (second setting on the x-axis), highlighting that it is both the most effective and the most stable configuration across different datasets.

**Instance-Conditional Context Tokens vs Context Tokens.** An ablation study on the instance-conditional network

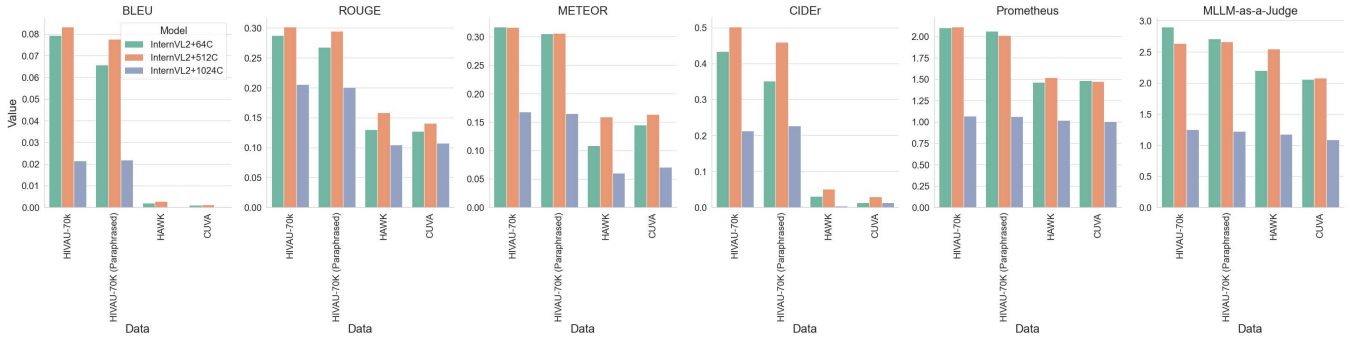



Figure 5: **Ablation on context lengths.** "$+64C$", "$+512C$", "$+1024C$" denote the integration of 64, 512, and 1024 instance-conditional context tokens into the model, respectively. Results are reported across traditional and model-based metrics and datasets, averaged over all granularities.

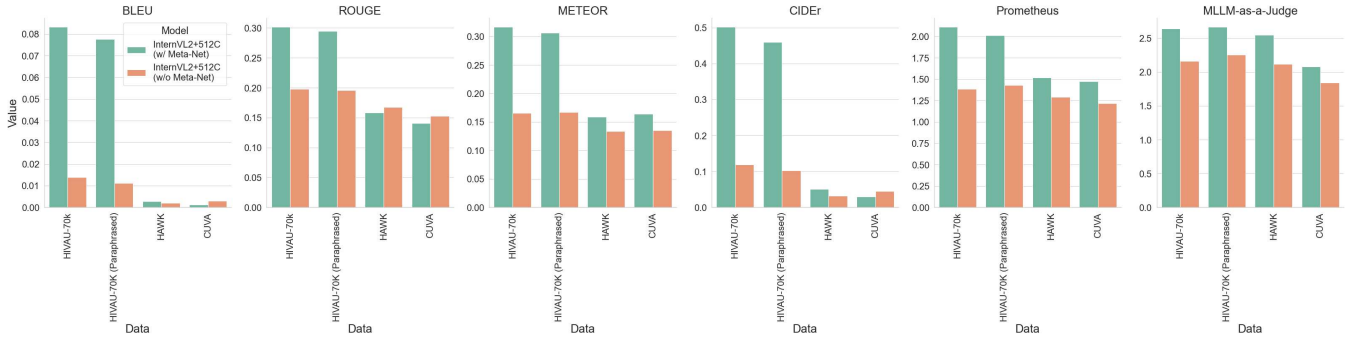



Figure 6: **Ablation on instance-conditional network (Meta-Net).** "$+C$", denotes the integration of context tokens into the model. Results are reported across traditional and model-based metrics and datasets, averaged over all granularities.

(Meta-Net) was conducted across all datasets discussed in the previous subsections. As shown in Fig. 6, incorporating instance-conditional context tokens consistently outperforms the pure context tokens without Meta-Net. The metrics shown in the figure are averaged across all granularities. This improvement is noticeable across all evaluation settings, including downstream task performance, prompt sensitivity, and domain generalization. These results suggest that conditioning context tokens on visual information improves overall model performance in anomaly detection.

| Vision Encoder | Input Size (T, H, W) | # params | FLOPS |
|---|---|---|---|
| VideoMamba [28] | (32, 224, 224) | **74M** | **900G** |
| InternViT [91] | (32, 224, 224) | 300M | 4977G |
| InternViT [91] | (12, 448, 448) | 300M | 7443G |

TABLE 8: **Comparison of vision encoders in terms of computational cost (GFLOPs) and model size (number of parameters).** Lower values indicate faster models.

**InternVL2 vs InternMambaVL.** An ablation study on the integration of the VideoMamba vision encoder is conducted by evaluating the performance of InternVL2+$C$ and InternMambaVL+$C$ across all settings discussed in the previous subsections. As shown in Fig. 7, InternVL2+$C$ generally outperforms InternMambaVL+$C$, which is expected given that InternVL2 benefits from multi-stage and multi-strategy pre-training, unlike InternMambaVL. However, a notable observation from the comparison is that InternMambaVL+$C$ exhibits more consistent performance than InternVL2+$C$, which shows a significant drop in performance for domain generalization task across all metrics, specifically when transitioning from standard downstream tasks to the HAWK and CUVA datasets. This may be associated with InternMambaVL's use of 32 input frames, providing richer temporal information than InternVL2.

Another advantage of InternMambaVL lies in its efficiency. Despite VideoMamba having a fixed large temporal size, it is significantly lighter than InternViT. This is illustrated in Table 8, where the training configurations for InternMambaVL+$C$ and InternVL2+$C$ correspond to the first and third rows, respectively. Notably, the vision encoder in InternMambaVL is over eight times more computationally efficient than that of InternVL2.

## 7. Conclusion

Our research tackles key challenges in adapting large pre-trained multi-modal large language models to downstream tasks such as video anomaly understanding (VAU). Although foundation models have demonstrated impressive generalization capabilities across both vision and language domains [11, 12, 13, 14, 15], their deployment in critical ap-

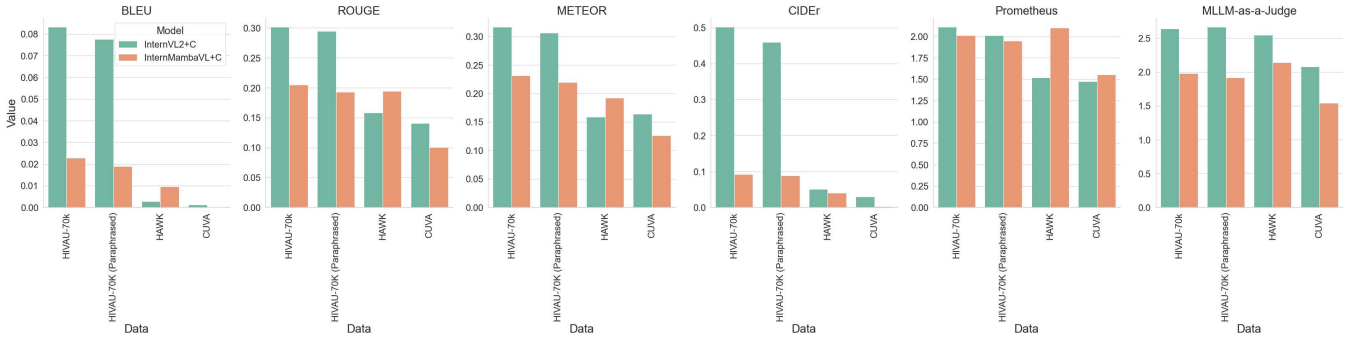



Figure 7: **Ablation on VideoMamba vision encoder.** "$+C$", denotes the integration of context tokens into the model. Results are reported across traditional and model-based metrics and datasets, averaged over all granularities.

plications like VAU is slowed down by high computational demands and sensitivity to prompt formulation. Instruction-tuning is often required for such adaptations and it can introduce over-reliance on specific prompt formats, cause reduced generalizability, and even catastrophic forgetting.

To address these limitations, we propose a parameter-efficient and computationally practical adaptation framework for MLLM-based VAU. Building on the success of dual-encoder vision-language models for prompt learning [22, 23, 25], we introduce a learnable, instance-conditional context tokens design. These tokens are appended to user instructions and help minimize prompt sensitivity and overfitting, and maintain the foundational model's general knowledge. Our approach avoids full or partial model finetuning of the core model and improves robustness across domains, datasets, and previously unseen anomaly types.

Additionally, by integrating a lightweight vision encoder VideoMamba [28] we reduce the computational overhead substantially, achieving a favorable balance between performance and efficiency.

**Limitations and Future Work.** A key limitation of our approach lies in training efficiency. For instance, in InternVL2 [91], the variable number of video frames per input constrains the batch size to one, which complicates training due to inconsistent temporal dimensions of visual inputs. InternMambaVL overcomes this issue, as the visual backbone of VideoMamba [28] enforces a fixed temporal size, making a mini-batch training possible.

Another limitation is that InternMambaVL lacks the broad general knowledge embedded in InternVL2, primarily because it does not undergo multi-stage pre-training on large-scale datasets. Future work could address this by conducting comprehensive video understanding pre-training on InternMambaVL, following the strategies used for InternVL2 [91], InternVL2.5 [15], or leveraging large video-text datasets such as WebVid [68] or Valley [69]. Moreover, an interesting direction would be to construct a fully state space model-based MLLM by replacing InternLM2 [90] in InternMambaVL with the Mamba language model [34], and, again having a vast pre-training. Such a design could significantly improve efficiency over transformer-based MLLMs, making real-time usage of MLLMs more viable.

Finally, while instance-conditional context tokens enhance domain generalization compared to models finetuned solely on related downstream tasks, their performance can still lag behind their respective out-of-the-box models that do not use context tokens. This indicates potential overfitting to the downstream dataset and limited generalization in anomaly understanding. To address this issue, one promising direction is to regulate learned prompts by maximizing the agreement between prompted and frozen MLLMs (Mutual Agreement Maximization), as introduced in [83]. This can be achieved by introducing a regularization term in the training loss that minimizes the distance between the outputs of the frozen MLLM and the prompted MLLM, preventing the model with context tokens from drift too far from the generalist model. Another complementary direction is to replace mean pooling before the Meta-Net with explicit spatio-temporal modeling, such as incorporating a 3D convolution layer [118] to serve as a more expressive spatio-temporal aggregator. This modification can improve the Meta-Net's spatio-temporal comprehension and result in a more informative outputs from this model. Addressing these limitations is crucial for improving anomaly understanding and achieving robust generalization across diverse anomaly types.

## Acknowledgments

This work was performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

## References

[1] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song, "Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models," *ACM Computing Surveys*, vol. 56, no. 7, 4 2024. [Online]. Available: /doi/pdf/10.1145/3645101?download=true

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010.

[3] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2013, pp. 2720–2727.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world Anomaly Detection in Surveillance Videos," Tech. Rep. [Online]. Available: http://crcv.ucf.edu/projects/real-world/

[5] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection – A New Baseline," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 12 2017. [Online]. Available: https://arxiv.org/pdf/1712.09867

[6] Y. Xu, H. Hu, C. Huang, Y. Nan, Y. Liu, K. Wang, Z. Liu, and S. Lian, "TAD: A Large-Scale Benchmark for Traffic Accidents Detection From Video Surveillance," *IEEE Access*, vol. 13, pp. 2018–2033, 2025.

[7] D. Bogdoll, M. Nitsche, and J. M. Zollner, "Anomaly Detection in Autonomous Driving: A Survey," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2022-June. IEEE Computer Society, 2022, pp. 4487–4498.

[8] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall, "DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 444–459, 1 2023.

[9] J. Liu, Y. Yan, J. Li, W. Zhao, P. Chu, X. Sheng, Y. Liu, and X. Yang, "IPAD: Industrial Process Anomaly Detection Dataset," 4 2024. [Online]. Available: http://arxiv.org/abs/2404.15033

[10] Z. Huang and Y. Wu, "A Survey on Explainable Anomaly Detection for Industrial Internet of Things," *5th IEEE Conference on Dependable and Secure Computing, DSC 2022 and SECSOC 2022 Workshop, PASS4IoT 2022 Workshop SICSA International Paper/Poster Competition in Cybersecurity*, 2022.

[11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL Technical Report," 2 2025.

[12] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1. Association for Computational Linguistics (ACL), 2024, pp. 12 585–12 602.

[13] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," 10 2023. [Online]. Available: http://arxiv.org/abs/2306.02858

[14] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection," 10 2024. [Online]. Available: http://arxiv.org/abs/2311.10122

[15] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, L. Gu, X. Wang, Q. Li, Y. Ren, Z. Chen, J. Luo, J. Wang, T. Jiang, B. Wang, C. He, B. Shi, X. Zhang, H. Lv, Y. Wang, W. Shao, P. Chu, Z. Tu, T. He, Z. Wu, H. Deng, J. Ge, K. Chen, K. Zhang, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang, "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling," 1 2025. [Online]. Available: http://arxiv.org/abs/2412.05271

[16] H. Du, S. Zhang, B. Xie, G. Nan, J. Zhang, J. Xu, H. Liu, S. Leng, J. Liu, H. Fan, D. Huang, J. Feng, L. Chen, C. Zhang, X. Li, H. Zhang, J. Chen, Q. Cui, and X. Tao, "Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 18 793–18 803, 4 2024. [Online]. Available: https://arxiv.org/pdf/2405.00181

[17] J. Tang, H. Lu, R. Wu, X. Xu, K. Ma, C. Fang, B. Guo, J. Lu, Q. Chen, and Y.-C. Chen, "Hawk: Learning to Understand Open-World Video Anomalies," 5 2024. [Online]. Available: http://arxiv.org/abs/2405.16886

[18] H. Zhang, X. Xu, X. Wang, J. Zuo, X. Huang, C. Gao, S. Zhang, L. Yu, and N. Sang, "Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity," 3 2025. [Online]. Available: http://arxiv.org/abs/2412.06171

[19] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen, "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs," *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, pp. 1950–1976, 10 2024. [Online]. Available: https://arxiv.org/pdf/2410.12405

[20] A. Razavi, M. Soltangheis, N. Arabzadeh, S. Salamat, M. Zihayat, and E. Bagheri, "Benchmarking Prompt Sensitivity in Large Language Models," *Lecture Notes in Computer Science*, vol. 15574 LNCS, pp. 303–313, 2 2025. [Online]. Available: https://arxiv.org/pdf/2502.06065

[21] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, 2019, pp. 2463–2473.

[22] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," 10 2022. [Online]. Available: http://arxiv.org/abs/2109.01134http://dx.doi.org/10.1007/s11263-022-01653-1

[23] ——, "Conditional Prompt Learning for Vision-Language Models," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 16 795–16 804, 3 2022. [Online]. Available: https://arxiv.org/pdf/2203.05557

[24] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned Gradient for Prompt Tuning," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 15 613–15 623.

[25] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal Prompt Learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June. IEEE Computer Society, 2023, pp. 19 113–19 122.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2 2021. [Online]. Available: http://arxiv.org/abs/2103.00020

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1. Association for Computational Linguistics (ACL), 2019, pp. 4171–4186.

[28] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "VideoMamba: State Space Model for Efficient Video Understanding," 3 2024. [Online]. Available: http://arxiv.org/abs/2403.06977

[29] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," 5 2024. [Online]. Available: http://arxiv.org/abs/2310.03744

[30] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 24 185–24 198, 12 2023. [Online]. Available: https://arxiv.org/pdf/2312.14238

[31] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models," in *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR, 2024.

[32] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, and C. Ruan, "DeepSeek-VL: Towards Real-World Vision-Language Understanding," 3 2024. [Online]. Available: https://arxiv.org/pdf/2403.05525

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. [Online]. Available: https://arxiv.org/pdf/2010.11929

[34] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," 12 2023. [Online]. Available: https://arxiv.org/pdf/2312.00752

[35] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 555–560, 3 2008.

[36] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model." Institute of Electrical and Electronics Engineers (IEEE), 3 2010, pp. 935–942.

[37] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 18–32, 1 2014.

[38] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December. IEEE Computer Society, 12 2016, pp. 733–742.

[39] J. Wang and A. Cherian, "GODS: Generalized one-class discriminative subspaces for anomaly detection," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 10 2019, pp. 8200–8210.

[40] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates." Institute of Electrical and Electronics Engineers (IEEE), 3 2010, pp. 2921–2928.

[41] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2011, pp. 3313–3320.

[42] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep One-Class Classification," 2018.

[43] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 12 2018, pp. 3379–3388.

[44] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 13 568–13 577.

[45] M. Zaigham Zaheer, A. Mahmood, M. Haris Khan, M. Segu, F. Yu, and S. I. Lee, "Generative Cooperative Learning for Unsupervised Video Anomaly Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June. IEEE Computer Society, 2022, pp. 14 724–14 734.

[46] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June. IEEE Computer Society, 6 2019, pp. 481–490.

[47] H. Park, J. Noh, and B. Ham, "Learning Memory-guided Normality for Anomaly Detection," 3 2020.

[48] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2020, pp. 10 536–10 544.

[49] O. Hirschorn and S. Avidan, "Normalizing Flows for Human Pose Anomaly Detection," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 13 499–13 508.

[50] F. Landi, C. G. M. Snoek, and R. Cucchiara, "Anomaly Locality in Video Surveillance," 1 2019.

[51] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," *Advances in Neural Information Processing Systems*, vol. 15, 2002.

[52] W. Li and N. Vasconcelos, "Multiple instance learning for soft bags via top instances," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 4277–4285, 10 2015.

[53] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, pp. 29 573–29 588, 11 2018.

[54] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 4955–4966.

[55] P. Wu and J. Liu, "Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3513–3527, 2021.

[56] S. Li, F. Liu, and L. Jiao, "Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, vol. 36. Association for the Advancement of Artificial Intelligence, 6 2022, pp. 1395–1403.

[57] H. Zhou, J. Yu, and W. Yang, "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection," 2 2023. [Online]. Available: http://arxiv.org/abs/2302.05160

[58] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, "BatchNorm-based Weakly Supervised Video Anomaly Detection," 11 2023. [Online]. Available: http://arxiv.org/abs/2311.15367

[59] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto,

B. Jonn, H. Jun, T. Kaftan, Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "GPT-4 Technical Report," 3 2024.

[60] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," 2 2023.

[61] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," 10 2023.

[62] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action," 3 2023.

[63] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "MM-VID: Advancing Video Understanding with GPT-4V(ision)," 10 2023.

[64] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems*, vol. 36. Neural information processing systems foundation, 2023.

[65] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," 10 2023.

[66] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, and Y. Qiao, "LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model," 4 2023.

[67] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "VideoChat: Chat-Centric Video Understanding," 5 2023.

[68] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1708–1718.

[69] R. Luo, Z. Zhao, M. Yang, Z. Yang, M. Qiu, T. Wang, Z. Wei, Y. Wang, and C. Chen, "Valley: Video Assistant with Large Language model Enhanced abilitY," 6 2023.

[70] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao, M. Dou, K. Chen, W. Wang, Y. Qiao, Y. Wang, and L. Wang, "InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling," 1 2025.

[71] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models," 7 2024.

[72] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: Clip-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection," in *Proceedings - International Conference on Image Processing, ICIP*. IEEE Computer Society, 2023, pp. 3230–3234.

[73] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning Prompt-Enhanced Context Features for Weakly-Supervised Video Anomaly Detection," *IEEE Transactions on Image Processing*, vol. 33, pp. 4923–4936, 2024.

[74] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38. Association for the Advancement of Artificial Intelligence, 3 2024, pp. 6074–6082.

[75] Z. Yang, J. Liu, and P. Wu, "Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 18 899–18 908, 4 2024. [Online]. Available: https://arxiv.org/pdf/2404.08531

[76] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing Large Language Models for Training-free Video Anomaly Detection," 4 2024. [Online]. Available: http://arxiv.org/abs/2404.01014

[77] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *Proceedings of Machine Learning Research*, vol. 202. ML Research Press, 2023, pp. 20 351–20 383.

[78] H. Lv and Q. Sun, "Video Anomaly Detection and Explanation via Large Language Models," 1 2024. [Online]. Available: http://arxiv.org/abs/2401.05702

[79] G. Farnebäck, "Fast and accurate motion estimation using orientation tensors and parametric motion models," *Proceedings - International Conference on Pattern Recognition*, vol. 15, pp. 135–139, 2000.

[80] Z. Zhong, D. Friedman, and D. Chen, "Factual Probing Is [MASK]: Learning vs. Learning to Recall," *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pp. 5017–5033, 4 2021. [Online]. Available: https://arxiv.org/pdf/2104.05240

[81] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, vol. 1, pp. 4582–4597, 1 2021. [Online]. Available: https://arxiv.org/pdf/2101.00190

[82] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 3045–3059, 4 2021. [Online]. Available: https://arxiv.org/pdf/2104.08691

[83] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating Prompts: Foundational Model Adaptation without Forgetting," Tech. Rep. [Online]. Available: https://github.com/muzairkhattak/PromptSRC.

[84] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2 2018. [Online]. Available: http://arxiv.org/abs/1705.07750

[85] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, 2 2015. [Online]. Available: https://arxiv.org/pdf/1502.03167

[86] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 1 2000. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0169743999000477

[87] E. S. Gardner, "Exponential smoothing: The state of the art," *Journal of Forecasting*, vol. 4, no. 1, pp. 1–28, 1 1985. [Online]. Available: /doi/pdf/10.1002/for.3980040103https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980040103https://onlinelibrary.wiley.com/doi/10.1002/for.3980040103

[88] J. C. Wu, H. Y. Hsieh, D. J. Chen, C. S. Fuh, and T. L. Liu, "Self-supervised Sparse Representation for Video Anomaly Detection," *Lecture Notes in Computer Science*, vol. 13673 LNCS, pp. 729–745, 2022. [Online]. Available: https://dl.acm.org/doi/10.1007/978-3-031-19778-9_42

[89] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, and M. H. Yang, "Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 16 271–16 280, 12 2022. [Online]. Available: https://arxiv.org/pdf/2212.04090

[90] Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, X. Dong, H. Duan, Q. Fan, Z. Fei, Y. Gao, J. Ge, C. Gu, Y. Gu, T. Gui, A. Guo, Q. Guo, C. He, Y. Hu, T. Huang, T. Jiang, P. Jiao, Z. Jin, Z. Lei, J. Li, J. Li, L. Li, S. Li, W. Li, Y. Li, H. Liu, J. Liu, J. Hong, K. Liu, K. Liu, X. Liu, C. Lv, H. Lv, K. Lv, L. Ma, R. Ma, Z. Ma, W. Ning, L. Ouyang, J. Qiu, Y. Qu, F. Shang, Y. Shao, D. Song, Z. Song, Z. Sui, P. Sun, Y. Sun, H. Tang, B. Wang, G. Wang, J. Wang, J. Wang, R. Wang, Y. Wang, Z. Wang, X. Wei, Q. Weng, F. Wu, Y. Xiong, C. Xu, R. Xu, H. Yan, Y. Yan, X. Yang, H. Ye, H. Ying, J. Yu, J. Yu, Y. Zang, C. Zhang, L. Zhang, P. Zhang, P. Zhang, R. Zhang, S. Zhang, S. Zhang, W. Zhang, W. Zhang, X. Zhang, X. Zhang, H. Zhao, Q. Zhao, X. Zhao, F. Zhou, Z. Zhou, J. Zhuo, Y. Zou, X. Qiu, Y. Qiao, and D. Lin, "InternLM2 Technical Report," 3 2024. [Online]. Available: https://arxiv.org/pdf/2403.17297

[91] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, J. Ma, J. Wang, X. Dong, H. Yan, H. Guo, C. He, B. Shi, Z. Jin, C. Xu, B. Wang, X. Wei, W. Li, W. Zhang, B. Zhang, P. Cai, L. Wen, X. Yan, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang, "How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites," *Science China Information Sciences*, vol. 67, no. 12, 4 2024. [Online]. Available: https://arxiv.org/pdf/2404.16821

[92] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," 5 2017. [Online]. Available: https://arxiv.org/pdf/1705.06950

[93] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, pp. 1933–1941, 10 2017. [Online]. Available: /doi/pdf/10.1145/3123266.3123451?download=true

[94] M. Abdalla, S. Javed, M. A. Radi, A. Ulhaq, and N. Werghi, "Video Anomaly Detection in 10 Years: A Survey and Outlook," 5 2024. [Online]. Available: https://arxiv.org/pdf/2405.19387

[95] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention Is All You Need," p. 1, 6 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762

[96] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model," *Proceedings of Machine Learning Research*, vol. 235, pp. 62 429–62 442, 1 2024. [Online]. Available: https://arxiv.org/pdf/2401.09417

[97] L. P. J. Sträter, M. Salehi, E. Gavves, C. G. M. Snoek, and Y. M. Asano, "GeneralAD: Anomaly Detection Across Domains by Attending to Distorted Features," 7 2024. [Online]. Available: http://arxiv.org/abs/2407.12427

[98] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR 2022 - 10th International Conference on Learning Representations*, 6 2021. [Online]. Available: https://arxiv.org/pdf/2106.09685

[99] X. Li, W. Ren, W. Qin, L. Wang, T. Zhao, and R. Hong, "Analyzing and Reducing Catastrophic Forgetting in Parameter Efficient Tuning," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2 2024. [Online]. Available: https://arxiv.org/pdf/2402.18865

[100] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision," Tech. Rep. [Online]. Available: https://roc-ng.github.io/XD-Violence/.

[101] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909–926, 5 2008.

[102] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection," 4 2023. [Online]. Available: http://arxiv.org/abs/2111.08644

[103] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *7th International Conference on Learning Representations, ICLR 2019*, 11 2017. [Online]. Available: https://arxiv.org/pdf/1711.05101

[104] ——, "SGDR: Stochastic Gradient Descent with Warm Restarts," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 8 2016. [Online]. Available: https://arxiv.org/pdf/1608.03983

[105] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, p. 311, 2001. [Online]. Available: https://dl.acm.org/doi/pdf/10.3115/1073083.1073135

[106] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 4566–4575, 11 2014. [Online]. Available: https://arxiv.org/pdf/1411.5726

[107] A. Lavie and A. Agarwal, "Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," pp. 228–231, 2007. [Online]. Available: https://dl.acm.org/doi/pdf/10.5555/1626355.1626389

[108] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," pp. 74–81, 2004. [Online]. Available: https://aclanthology.org/W04-1013/

[109] D. Chen, R. Chen, S. Zhang, Y. Wang, Y. Liu, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun, "MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark," *Proceedings of Machine Learning Research*, vol. 235, pp. 6562–6595, 2 2024. [Online]. Available: https://arxiv.org/pdf/2402.04788

[110] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models," *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 4334–4353, 5 2024. [Online]. Available: https://arxiv.org/pdf/2405.01535

[111] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 Technical Report," 12 2024. [Online]. Available: https://arxiv.org/pdf/2412.15115

[112] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 8 2022. [Online]. Available: http://arxiv.org/abs/2103.02503http://dx.doi.org/10.1109/TPAMI.2022.3195549

[113] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 12 2014. [Online]. Available: https://arxiv.org/pdf/1412.6980

[114] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4955–4966, 1 2021. [Online]. Available: https://arxiv.org/pdf/2101.10030

[115] S. Li, F. Liu, and L. Jiao, "Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 1395–1403, 6 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20028

[116] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y. C. Wu, "MGFN: Magnitude-Contrastive Glance-and-Focus Network for Weakly-Supervised Video Anomaly Detection," *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, vol. 37, pp. 387–395, 11 2022. [Online]. Available: https://arxiv.org/pdf/2211.15098

[117] M. Cho, T. Kim, M. Shim, D. Wee, and S. Lee, "Towards Multi-Domain Learning for Generalizable Video Anomaly Detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50 256–50 284, 12 2024.

[118] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 10 2015. [Online]. Available: http://arxiv.org/abs/1412.0767