Information Fusion 126 (2026) 103388

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

A survey on multi-modal and weakly supervised approaches for robust anomaly detection in video data

Rui Z. Barbosa [a], Hugo S. Oliveira [b,*], João Manuel R.S. Tavares [c]

[a] University of Porto - Faculty of Engineering, Portugal
[b] University of Porto - Faculty of Sciences, Portugal
[c] Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Portugal

ARTICLE INFO

Keywords:
Video anomaly detection
Weakly-supervised
Abnormality criteria
Open-world
Feature representations
Benchmarks
Vision–language understanding
Multi-modal fusion
Cross-modal reasoning
Video anomaly understanding
Privacy-preserving

ABSTRACT

This survey provides a comprehensive overview of Video Anomaly Detection (VAD), focusing on identifying robust and interpretable methods for detecting anomalous events. It emphasizes the limitations of traditional supervised and unsupervised approaches and highlights the advantages of weakly supervised learning, in which models operate with minimal or no explicit anomaly annotations. A key contribution of this survey is its analysis of multi-modal data — integrating visual, audio, and textual modalities — for enhanced anomaly detection. It underscores how audio cues enrich visual features and how textual information during training fosters semantically richer representations. This multi-modal approach demonstrates improved generalization, better detection of subtle anomalies, and interpretable explanations of detected events, marking a paradigm shift in the field, Video Anomaly Understanding (VAU).

Synthesizing advancements from benchmarks to methodologies, this work advocates for a centralized platform to enable systematic comparisons across diverse datasets, standardized evaluation metrics, and reproducible ablation studies of novel components. Such a framework would streamline the integration of innovations, address version control and foster transparency, bridging the gap between isolated methodological advances and system-level robustness. By prioritizing contextual understanding, causal reasoning, and real-world interpretability, this initiative aims to elevate weakly supervised VAD beyond detection, ensuring models contextualize and explain anomalies in practical deployments.

1. Introduction

Detecting, identifying and understanding anomalies in video data stands as a critical challenge across diverse applications, from security surveillance and healthcare monitoring to autonomous systems and industrial safety [1].

This survey establishes a unified framework for weakly supervised video anomaly detection (WVAD) and its evolution into emerging paradigms, addressing the limitations of both unsupervised methods (reliant on normalcy assumptions) and fully supervised approaches (requiring costly frame annotations). We systematically trace how WVAD — powered by video-level labels — has matured through multi-modal fusion and vision–language models (VLMs) to tackle real-world challenges: data scarcity, open-world generalization, and interpretability demands [2,3].

Crucially, we position open-set [4,5], open-vocabulary [6,7], and training-free VAD [8–10] as natural extensions of WVAD aiming for generalization in dynamic, multi-scenario environments with class imbalance, and evolving anomaly patterns.

To address these challenges, this survey synthesizes innovations in feature representation, anomaly criteria, and learning paradigms, while emphasizing the role of new benchmarks and metrics in driving progress [3,11]. We highlight the transition from single-modality visual analysis to integrated frameworks that combine audio, text, and spatial–temporal reasoning—approaches exemplified by recent datasets [12–15]. These resources demand models to not only detect, classify and locate anomalies but also describe and reason about their context, causality, and implications, presenting a shift that underscores the growing intersection of VAD with video–language understanding (VLU).

1.1. Previous surveys

Early surveys [16,17] focused on traditional machine learning techniques, emphasizing handcrafted features such as optical flow, background subtraction, and motion trajectories in single-scene appearances. These approaches, while foundational, often needed to be improved in their ability to handle complex, high-dimensional video data.

* Corresponding author.
E-mail addresses: barbosa@tuta.com (R.Z. Barbosa), hugo.soares@fe.up.pt (H.S. Oliveira), tavares@fe.up.pt (J.M.R.S. Tavares).

https://doi.org/10.1016/j.inffus.2025.103388
Received 20 March 2025; Received in revised form 31 May 2025; Accepted 2 June 2025
Available online 11 August 2025

Table 1
Resume of the surveys for video anomaly detection.

Survey	VAD domain	Supervision	Multimodal	Description
A comprehensive review on deep learning-based methods for video anomaly detection (2020) [18]	Supervised, Semi Supervised and Unsupervised for Video Surveillance	✓	✗	Focused on Traditional Feature extractors (SIFT, SURF), Supervised, Semi-Supervised and Unsupervised.
A survey of single-scene video anomaly detection (2020) [16]	Supervised	✗	✗	Focused on Probabilistic, Temporal CNN Patterns, Hidden Markov methods
A Comprehensive Survey of Machine Learning Methods for Surveillance Videos Anomaly Detection (2023) [19]	Supervised, Semi-Supervised and Unsupervised for Urban Video	✓	✗	Focus on traditional Machine Learning models, such as Support Vector Machines, Recurrent Neural Networks and CNN.
Weakly Supervised Anomaly Detection: A Survey (2023) [20]	Tabular & Video	✓	✗	Focus on Weakly Supervised, With Multiple Instance, Active Learning and Graph Learning approaches for VAD.
Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models (2023) [2]	Generalized methods approaches	✓	✓(9 works)	Identifies several works across unsupervised, weakly-supervised, and supervised paradigms, establishing an application-oriented taxonomy.
Networking Systems for Video Anomaly Detection: A Tutorial and Survey (2024) [1]	Video IoT and Smart cities	✓	✓(13 works)	Systematically organizes unsupervised/weakly-supervised methodologies while analysing future trajectories to advance networked video anomaly detection.

The advent of deep learning marked a significant shift in the field, with more recent surveys (Table 1) exploring the use Generative Adversarial Network (GAN)s, Long Short Term Memory (LSTM)s, Convolutional Neural Network (CNN)s, more recently, transformers for VAD. These studies highlighted the transition from supervised learning, which relies heavily on labelled data, to unsupervised (relying solely on normal data) and weakly supervised methods designed to mitigate the scarcity of anomaly annotations. Some surveys have also addressed domain-specific challenges, such as detecting anomalies in crowded scenes, medical imaging videos, and autonomous driving scenarios.

While prior surveys catalogue techniques and routes (Table 1), they lack a unified framework to connect traditional WVAD with modern LLM-driven paradigms addressing open-world challenges. They often focus narrowly on either feature extraction techniques or specific learning paradigms, without fully integrating the broader multi-modal aspects of anomaly detection. Furthermore, the role of interpretability, particularly for sensitive applications where explainable decisions are crucial, still needs to be explored in prior literature.

This survey builds on the foundational insights of GVAD and NSVAD by shifting the focus toward the evolution of VAD methodologies from unimodal visual analysis to multimodal frameworks that integrate audio, text, and contextual reasoning. While GVAD systematically categorized deep learning approaches for generalized anomaly detection and NSVAD pioneers the first comprehensive tutorial bridging AI, IoT, and computing communities, this work addresses the evolution of weakly supervised methods towards interpretable solutions, emphasizing the integration of multi-modal data, open-world settings and train-free inference.

The Table compares recent surveys in VAD, highlighting their use of unimodal and multimodal data. Most surveys discussed here adopt multimodal approaches, incorporating visual and additional data (e.g., audio, text) to improve anomaly detection.

Unimodal methods typically use only visual data. On the other hand, other surveys explore how weak supervision can be applied to unimodal and multimodal data in anomaly detection tasks.

Overall, the surveys reflect a clear trend towards integrating multiple modalities (visual, audio, textual) in anomaly detection to enhance performance, improve generalization, and address limitations inherent in using a single modality.

1.2. Contributions

This survey advances VAD research through three key contributions:

(1) Methodological Evolution: Tracing the progression from early ranking-based techniques to modern VLM-driven frameworks, with a focus on weak supervision, under an *abnormality criterion*-centric taxonomy. This organizes methods by their core anomaly-defining principles (e.g., feature magnitude deviation, spatiotemporal inconsistency, semantic misalignment), eliminating redundancy across paradigms. This synthesis integrates multi-modal advances (audio, text, vision) and benchmark-driven demands for contextual reasoning.
(2) Unified Concepts: Organizing VAD components, including feature extractors, modulators, learning and optimization strategies, and benchmarking practices, to facilitate reproducibility and comparative analysis.
(3) Gap Analysis & Future Pathways: Highlighting the interplay between methods and unresolved challenges, such as cross-domain generalization, privacy-sensitive and edge deployments. The survey concludes with a roadmap for the development of adaptive, reproducible, and explainable VAD systems grounded in real-world constraints.

By integrating advances in feature representation, benchmark design, this survey highlights how recent datasets and evaluation metrics — such as GPT-Guided Reasonability scores and QA-based anomaly localization — demand models to detect, describe, and reason about anomalies, not merely classify them.

This work serves as a bridge between foundational WVAD research and the next frontier of multimodal, context-aware anomaly understanding. It advocates for a standardized evaluation framework that considers development and deployment in complex real-world scenarios.

Section 2 defines WVAD fundamentals and details literature search methodology; Sections 3 and 4 analyse feature extractors and benchmarks; Section 5 presents our taxonomy-driven review; Sections 6 and 7 discuss edge and privacy challenges; Section 8 concludes.

2. Foundations/definition of video anomaly detection

Video Anomaly Detection (VAD) aims to identify events in video sequences that deviate from normal or typical patterns, that may cause human and animal threats, security problems and economic losses. Anomalies are typically defined relative to contextual norms, making their detection inherently dependent on environmental and temporal factors. Anomalies can manifest as:

- **Behavioural:** Deviations in the behaviour of individuals or groups, such as loitering, fights, or sudden movements.
- **Object-Based:** The presence of unexpected objects, such as abandoned luggage or unauthorized vehicles.
- **Scene-Based:** Sudden changes in the scene, such as fire or accidents.

These context-dependent anomalies make their detection crucial in various applications such as security surveillance, healthcare monitoring, autonomous driving, and industrial process control [1,2]. Some key characteristics of VAD are:

- **Context-Dependent:** An event may be anomalous in one context but normal in another. For example, a person running might be typical in a park but unusual in a hospital corridor.
- **Rarity:** Video anomalies are typically infrequent, leading to challenges in collecting sufficient labelled data for supervised learning.
- **Representation Diversity:** Anomalies can span multiple modalities, such as visual cues (e.g., sudden motion or unusual objects), audio signals (e.g., loud crashes or screams), or textual information (e.g., subtitles or metadata).
- **Dynamic Nature:** Anomalies often involve temporal variations, making it necessary to analyse sequences of frames rather than individual ones.
- **Uncertainty:** Anomalies may not have well-defined boundaries, making them harder to detect and categorize.
- **Real-World Deployment:** often adds spatial–temporal complexity (e.g., dynamic backgrounds, occlusions) and domain shifts (e.g., lighting, camera angles).

This context sensitivity anomalies necessitate robust modelling of both local patterns (e.g., motion trajectories) and global semantics (e.g., scene-specific expectations), posing several challenges for their correct detection, namely:

- **Class Imbalance:** Normal events dominate datasets, while anomalies are scarce.
- **Lack of Annotations:** Labelling anomalies in videos is time-consuming and subjective.
- **Multi-Modality Integration:** Combining visual, audio, and textual data effectively.
- **Interpretability:** Providing clear explanations for why an event is considered anomalous.

These challenges highlight the research inclinations towards weakly supervised VAD (WVAD), which operates on coarse-grained annotations [12,21,22], balancing practicality and performance, making it preferable for real-world deployment where detailed annotations are scarce. It states a paradigm shift from early approaches reliance solely on normality modelling to detect deviations, often failing in complex, real-world scenarios due to limited semantic understanding.

However, these methods struggle with abnormality generalization and open-world anomalies. Recent advancements include open-set VAD [4,5] distinguishes seen vs. unseen anomalies, open-vocabulary VAD [6,7] incorporates the classification into its specific semantic labels, and train-free paradigm [8–10], adapting pre-trained models without retraining. These approaches address WVAD's limitations while emphasizing interpretability and real-world robustness.

2.1. WVAD

WVAD methods typically operate under three core assumptions: (1) normality is defined by learned priors from majority-class data, (2) anomalies are rare and distinct in feature space, and (3) weak supervision (e.g., video-level labels) can guide models to infer fine-grained anomaly patterns.

The following two seminal works have significantly shaped the field of Weakly-supervised Video Anomaly Detection (WVAD).

Multiple Instance Regressor (MIR) [21] (Fig. 1), frames the WVAD as a Multiple Instance Learning (MIL) task under a weak regression problem [23], alongside the release of the weakly supervised UCF-Crime (UCFC) dataset. The authors trained a 3-layer dense Multi Layer Perceptron (MLP) model to assign anomaly scores directly to segment-level features extracted from a Convolutional 3D Network (C3D) [24], pre-trained on Sports1M [25] dataset.

The training is guided by a modified hinge-based loss named ranking loss, which encourages the maximum anomaly score over instances in a positive bag (video containing anomalies) to be higher than in a negative bag (anomaly-free video). This ensures a considerable distance between selected abnormal and normal samples from each pair of bags. However, this strategy has inherent limitations:

- *Focus on a Single Anomaly:* The ranking loss only considers the segment with the highest anomaly score, disregarding other potential anomalies within a video. This can be problematic when multiple anomalous events occur or an anomaly spans multiple segments.
- *Neglecting Temporal Context:* The ranking loss does not explicitly consider the temporal relationships between segments, potentially missing important contextual information for anomaly detection.
- *Sensitivity to Initial Scores:* Relying solely on the initial anomaly scores from the network can be problematic, as these scores might be unreliable in the early stages of training.

The overall cost function in MIR also includes sparsity constraints and temporal smoothness terms to encourage the model to focus on a few anomalous segments and produce temporally coherent anomaly scores. The MIL paradigm has been widely adopted in subsequent works, drawing inspiration from research in related fields such as weakly supervised object detection [26,27] and weakly supervised temporal action localization [28,29].

The work of **Background Bias of AR models in WVAD (Background Bias)** [30] explores the nature of anomalous events and their impact in different Action Recognition (AR) models/datasets. UCFC is re-annotated with spatiotemporal anomaly labels. NLNet [31] trained on original/trimmed versions shows high classification errors on anomaly-free test sets, indicating background bias regardless of thresholds.

Identical evaluation of T-C3D [32], Two-Stream Inflated 3D CNN (I3D) [33], Temporal Segment Networks (TSN) [34,35], and 3DResNet [36] confirms background misclassification at clip/video levels. The Class Activation Mapping (CAM) [37] verifies non-anomalous spatial regions drive high scores.

Bias originates from anomalies occupying minimal/non-foreground frame areas while background dominates. Weakly supervised models exploit background as predictive shortcuts, ignoring subtle anomaly cues, causing the background bias of from AR's vanilla features for WVAD.

In response to these challenges, subsequent WVAD methods have explored various strategies to overcome the limitations of weak supervision and the background bias problem, prompted researchers to critically examine the fundamental assumptions of WVAD. This introspection led to a series of thought-provoking questions that have guided the development of more sophisticated methods:

(1) How can raw features be effectively modulated to enhance the separability between irregular-length anomalous and normal segments within a video while minimizing false alarms.
(2) How can confident features and scores be selected for optimization, ensuring alignment with the relevant spatial–temporal cues of both abnormal and normal events, while accounting for the characteristics of real-world anomalies under weak supervision?

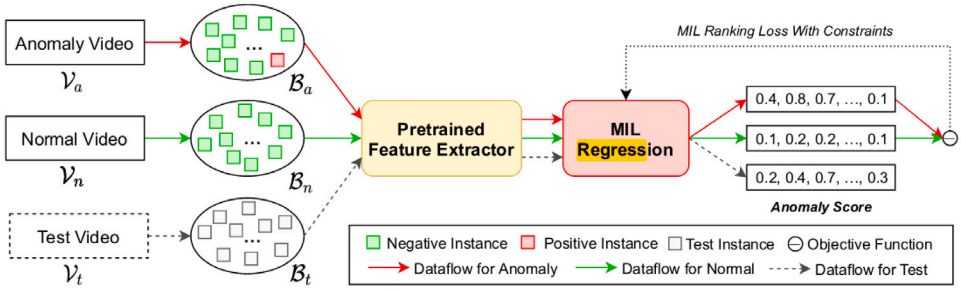


Fig. 1. MIR framework [21].
Source: Taken from [1].

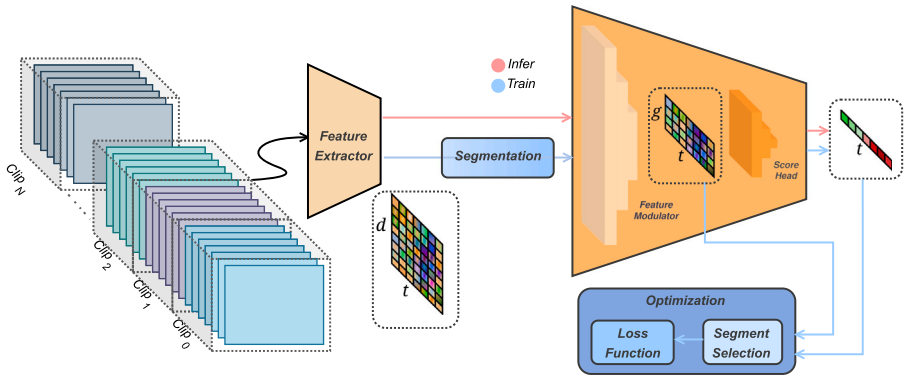


Fig. 2. Overall pipeline of (UWS4VAD) [38].

(3) How can the loss function be designed to effectively guide the learning process, addressing the limitations of the ranking loss and incorporating additional information or constraints to improve anomaly detection performance?

Driven by these questions, researchers have explored diverse methods, leading to a rich landscape of WVAD. These methods often leverage a combination of techniques to effectively learn from weakly labelled data and distinguish between normal and abnormal events (Fig. 2). These techniques include:

- *Feature Modulation* modules, often employing convolutional or self-attention layers, transform raw features extracted from backbone networks into more discriminative representations highlighting anomalous patterns.
- *Attention Mechanisms* to focus on the most relevant segments or features, guiding the model's attention towards salient cues for anomaly detection.
- *MIL*, commonly used in WVAD, treats videos as bags of instances (segments) and learns to distinguish between normal and abnormal bags. This approach effectively leverages weak video-level labels to guide the training process.
- *Novel Optimization techniques* to better separate normal and abnormal events in a latent space, often using pair-based loss functions to encourage separation.

These techniques are often combined in hybrid frameworks, creating systems that balance efficiency and robustness. Recent advancements emphasize semantic integration (e.g., text-based prompts [39–41]), cross-modality alignment (e.g., audio–visual fusion [42]) or additional annotation signals (e.g. glance [43]) to improve interpretability and generalization.

This progression highlights the shift from rigid, task-specific models to flexible, multimodal frameworks, setting the stage for the survey's focus on abnormality criteria as a unifying taxonomy.

2.2. Abnormality criteria as a guiding principle

In this study, we group the most prominent works in VAD using a taxonomy that clusters common approaches, identifies main differences, and highlights potentialities and limitations, enabling systematic comparison and identification of research gaps.

The concept of the *Abnormality Criteria (AC)* defines the characteristics to exploit in the context of a video with abnormal events present at random. It depicts the model's priorities and focuses on the input features to score confidently. Therefore, it can portray the anomaly definition through the model's lens. It guides the design of data curation/sampling, model's architecture, feature engineering and training objectives, and ultimately shaping the anomaly detection, classification and reasoning process.

These methods can be categorized based on their underlying AC, providing a structured overview of the field:

- *Magnitude:* Methods that leverage the assumption that abnormal events exhibit higher feature magnitudes than normal events.
- *Background & Normality:* Methods that focus on suppressing background information or explicitly modelling normality to detect deviations.
- *2-Stage & Label Noise:* Methods that employ two-stage training schemes to generate and refine pseudo-labels, mitigating the impact of label noise.
- *Anomaly Erase, Suppress, and Salience:* Methods that emphasize the extraction and analysis of salient features, often incorporating erasure or suppression techniques to focus on less conspicuous anomalies.
- *Temporal Dynamics:* Methods that exploit temporal variations and dynamic changes in video features as indicators of anomalous events.
- *Multi-modal:* Methods that integrate audio, visual, and potentially textual information to enhance anomaly detection.

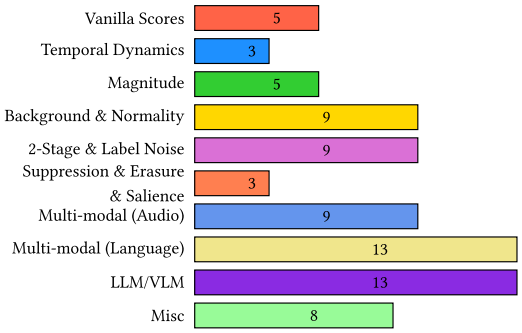


Fig. 3. Distribution of the works into the subgroups according to the defined taxonomy.

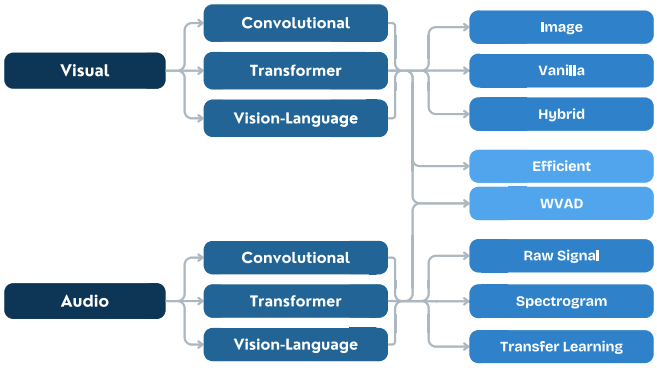


Fig. 4. Explored feature extraction architectures paradigms in visual and audio modalities, and their adaption in WVAD.

- *LLM/VLM:* Methods that leverage Large Language Models (LLMs) or Vision–Language Models (VLMs) to enhance semantic understanding. Utilizes natural language prompts, often with minimal supervision, to detect diverse and unseen anomalies via Image or Video Captioning or Visual Question Answering.
- *Misc:* Methods that do not neatly fit into the previous categories or represent promising new directions for future research.

This structured categorization defined in Fig. 3 reveals how WVAD methods evolve from isolated modality modelling to holistic frameworks that balance detection accuracy with semantic understanding. It synthesizes current methodologies and enables a deeper understanding of the trade-offs between different paradigms. This taxonomy not only organizes existing methods but also identifies gaps in privacy-preserving design and edge deployability.

Moreover, our survey extends beyond conventional reviews by emphasizing the integration of multi-modal data and advocating for more interpretable and generalizable anomaly detection models, aligning with new benchmarks. Through this lens, we aim to bridge existing gaps and provide a roadmap for future research in VAD.

The following sections will delve into each criterion, dissecting methodological nuances and their implications for real-world deployment. Sections 3 and 4 analyse feature extractors and benchmarks, while Section 5 presents our taxonomy-driven review. Edge and privacy challenges are discussed in Sections 6 and 7, with concluding insights in Section 8.

3. Feature representations

Deep learning models for image, audio, or video understanding (VU) tasks typically consist of two core components: a backbone network that extracts hierarchical features from raw inputs (images/video frames), and a task-specific head optimized for the target task.

This modular design enables flexible adaptation to diverse modalities and tasks, including VAD, which transforms raw video data into meaningful, task-agnostic embeddings that capture spatial, temporal, and semantic patterns. These Feature Extractor (FE)s often draw inspiration from and adapt architectures developed in the fields of image and Video Understanding (VU), where significant advancements have been made in tasks such as image classification, object detection, and Action Recognition/Location.

This section establishes the foundational landscape of audio and video feature extraction in multimodal learning, tracing their evolution from classical architectures to modern multimodal approaches, and their adaption in WVAD systems (Fig. 4). By dissecting the architectural paradigms in both modalities, it clarifies how advancements in feature representation, from convolutional networks to vision–language models, inform the design and performance in VAD methods. This evolution is identified as a critical enabler for WVAD and VAU.

3.1. Visual representation backbones

The evolution of backbone architectures for VU has followed a similar trajectory to traditional image classification, with adaptations to accommodate the extra temporal dimension. This trend is visible in both CNN or transformer-based networks, where backbone architectures are adapted from those for image-related tasks. The first works try to extend the modelling through the temporal axis. In contrast, others factorize spatial and temporal domains to achieve a better speed–accuracy tradeoff since joint spatio-temporal modelling is challenging to optimize.

CNNs have been the standard foundation for computer vision tasks since the introduction of AlexNet [44]. Since then, deeper and more effective convolutional neural networks have been proposed, such as VGG [45], Inception [46–49] and ResNet [50,51]. Apart from structural improvements, individual convolution layers have also received enhancements, such as depthwise convolution present in MobileNet [52–54] and EfficientNet [55,56], and deformable convolution DefConv [57,58].

Recent advances in Natural Language Processing (NLP) [59] have brought backbone innovations in the computer vision field with the appearance of Vision Transformer (ViT)s models. Self-attention mechanism has been used to replace 2D convolution layers in ResNet [60–62] or to complement CNN backbones [31,63–66].

ViT [67] takes this idea further by directly applying the Transformer architecture to image inputs. ViT splits an image into fixed-sized patches and provides the sequence of linear embeddings of these patches as input to a Transformer. However, the global attention mechanism in transformers leads to quadratic complexity concerning the input size and lacks inductive bias. Several new improvements to the original implementation have been proposed to address these limitations, even hybrid approaches that incorporate explicit convolution or desirable properties of convolution. DeiT [68] is one example that applies a teacher–student strategy to allow training on smaller datasets.

SwinTransformer [69,70] introduces a hierarchical representation and computes self-attention locally within non-overlapping windows, achieving linear computational complexity. Other works extend the idea of local self-attention by designing different shapes of attention windows or introducing soft local constraints to attention maps [71,72].

Adopting depth-wise separable convolution from [52] is also commonly employed as a feature backbone in [73–75]. The work of [76] extends the Inception family into the Transformers structure, while [77] incorporates insights from convolutional and Transformer design patterns. Other works focus on tackling the memory-inefficient operations and redundancy present Multi-Head Self Attention (MHSA) with deployment requirements in mind [78–83]

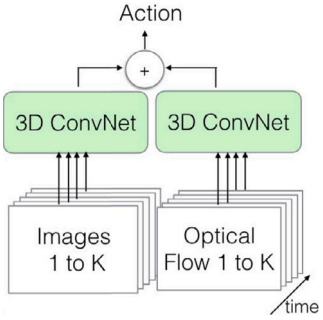


Fig. 5. I3D network from [33].

3.1.1. Convolution temporal based models

This section delves into the application of CNN's for capturing temporal video information. It will explore how traditional 2D CNN architectures have been adapted to handle the temporal dimension, leading to the development of 3D CNNs and other innovative approaches for video feature extraction.

The most straightforward approach to model temporal relationships uses 3D convolutional layers, such as C3D [24], which directly performs convolution in the space formed by frame height, width, and video duration or frames by a direct extension of 2D convolution for joint spatial and temporal modelling at the operator level. WVAD introductory work, MIR [21] used a C3D pre-trained on Kinetics-400 dataset [33] to extract video features from *fc6* layer activations. Several works employ the same methodology such as [84–98]

Another common architecture is the I3D proposed by [33], which builds upon the success of 2D CNNs for image classification to capture video information [47]. The Kinetics-400 dataset for AR is also introduced. The method starts by adding an extra dimension over the temporal axis of 2D backbone convolution filters and then converting the images into a video sequence with repeated copies. This enables effective bootstrap 3D filters from 2D pre-trained CNNs models trained on ImageNet (Fig. 5). Moreover, the network has two streams [99], trained separately, with the final being averaged, enabling the model to harness scene and object structure from the RGB branch and motion information from the optical-flow branch.

Despite the superior performance of I3D compared to C3D, 3D CNNs have high computational complexity and memory requirements, hindering their practical application in resource-constrained environments [100]. Solutions to overcome these memory requirements include combining 2D and 3D convolutions [101–103], decomposing 3D convolution into 2D convolution by a temporal operation [104,105], using zero-parameter operations to model temporal relationships [106–108], and modelling long and short-term temporal dynamics using separate network architectures [109,110].

CNN 's in WVAD —. It has been a common practice in WVAD to rely on I3D as the feature source, mainly for its public availability from works as [22,89], and the consistent improvements provided over the $3 \times 3 \times 3$ convolution-based C3D network. The I3D process relies heavily on pre-training, while C3D has few layers and is trained on small datasets, leading to a higher semantic feature content [21,84,88,89,93–98].

In that sense, there has been a lack of effective ablation studies per VAD community of latter architectures from the VU field to tackle efficiency and representativity. Although some works attend to that and provide results from different feature sources, they enable us to draw some conclusions.

Graph Convolutional Label Noise Cleaner (GCLNC) [84], Contrastive Attention (CAVAD) [91], and Weakly Supervised Anomaly Location (WSAL) [111] employ a BN-Inception [47] version of TSN [34,35] showing better results when compared to C3D/I3D.

WSAL goes further on comparing both I3D [33], R(2+1)D [104] and TSN, concluding that TSN surpasses the performance of CLAWS+ [92], with 3DResNet [36] showing better results compared to C3D.

3.1.2. Transformer temporal based models

Proposed initially for language translation [59], the Transformer architecture has revolutionized various fields, including Computer Vision (CV). Its ability to capture long-range dependencies and model global context has made it a powerful tool for video understanding tasks. However, adapting Transformers to the unique challenges of video data requires careful consideration of computational efficiency and incorporating inductive biases.

This section explores the application of transformer-based models to video understanding, highlighting their strengths and limitations in capturing temporal information. We will also discuss the emergence of Vision-Language Model (VLM)'s, which leverage the power of Transformers to learn joint representations of visual and textual data, opening up new possibilities for anomaly detection.

The Transformer consists of two distinct modules: encoder and decoder, each composed of several stacked Transformer layers. The encoder represents the source language sentence that is then attended by the decoder, which translates it into the target language. The Transformer excels at learning interactions of non-local contexts of the whole sequences at once, allowing for parallelization while removing the locality bias of traditional architectures like CNNs. However, the lack of inductive biases requires large amounts of data or several architectural modifications to accommodate the high redundancy of spatio-temporal information in videos. Furthermore, they scale quadratically with sequence length T (i.e., $O(T^2)$) due to the pair-wise affinity computation, which is exacerbated by the high dimensionality of the video.

The success of image Transformers has led to numerous transformer-based architectures for VU tasks. These works extend the design ideas of CNNs in the context of ViTs [100]. Some works perform spatial–temporal local self-attention [112–115], while [116] combine self-attention and convolution, employ 1D temporal attention [117], SqueezeTime [118] squeezes the temporal axis for a lightweight backbone targeting mobile video understanding.

Transformers in WVAD —. Despite the abundance of innovative transformer-based architectures for VU, their application in WVAD remains relatively limited. Works that have experimented with Video Swin Transformer (VSwin) [113], MSL [94] and MGFN [119], showed no relevant improvements over I3D, although further ablation studies are needed to inspect the anomalies-only subset.

3.1.3. Vision–language models

The appearance of massive web-scale multi-modal paired datasets, such as [120,121], with hundreds of millions of noisy image–text pairs, enabled foundational VLM like Contrastive Language-Image Pre-training (CLIP) [122], ALIGN [123], FLAVA [124], OpenCLIP [121], BLIP [125], EVA-CLIP [126,127] and FLIP [128]. These models employ a dual-encoder architecture that learns to align visual and language representations of image–text pairs using a contrastive loss and training (Fig. 6), producing strong and generalizable joint representation features that excel in zero-shot transfer learning.

Further works have focused on improving the dataset curation, training objective and exploring different encoder architectures. DFN [129] introduced data filtering, SigLIP [130] replaced softmax contrastive loss with sigmoid-based optimization, and SigLIP2 integrated caption-based pretraining [131] with DINOv2-style self-distillation [132]. UMG-CLIP [133] achieved multi-granular alignment (image/region/pixel).

MobileCLIP [134] optimized FastViT [83] for low-latency via multi-modal reinforcement leveraging knowledge transfer from an image captioning model, while ViTamin [135] re-evaluated vision encoders under CLIP, proposing hybrid architectures beyond vanilla ViT [67]. Temporal extensions like VideoCLIP [136] aligned overlapping video–text pairs, and ActionCLIP [137] reframed action recognition as video–text matching.

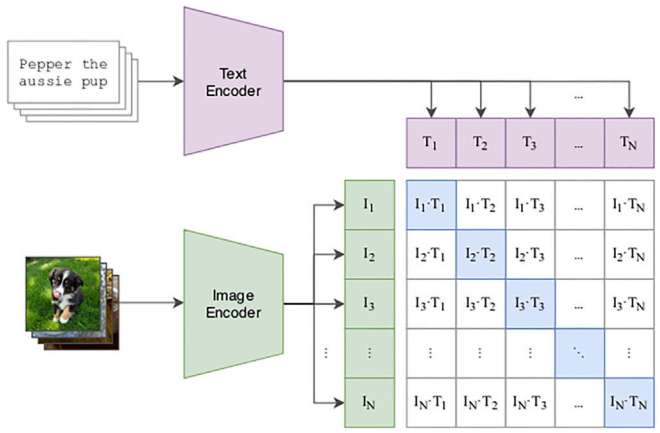


Fig. 6. Contrastive pre-training from [122].

LLMs: The evolution of LLMs has been driven by compact architectures and data-centric optimizations, with pioneering models like LLaMA [138–140], Vicuna [141], InternLM [142,143], demonstrating that smaller, well-designed models can rival larger counterparts. Efficiency breakthroughs emerged through diverse strategies: Phi [144–146] leveraged curated "textbook-quality" data for lean training, Gemma [147,148] distilled knowledge from larger models, and MiniCPM [149] redefined scaling laws for sub-7B parameter regimes. Pruning methods like LLM-Pruner [150] achieved parameter reduction without sacrificing capability, while Qwen [151–153] established versatile backbones for downstream VLM integration through architectural upgrades and expansive token training.

Efficient VLMs. Building on efficient LLMs, LLaVA [154,155] and LLaVA-NeXT [156–159] established parameter-efficient alignment recipes. MobileVLM [160,161] compressed visual tokens via lightweight adapters, while Phi-4 [162] and Gemma 3 [163] extended their LLM counterparts into multimodal domains. SF-LLaVA [164] combined LLaVA-NeXT 7B [156] with SlowFast [109] for training-free video understanding. FastVLM [165] introduces FastViTHD, a hybrid vision encoder that optimizes high-resolution VLM performance by scaling input images to balance token count, latency, and accuracy without token pruning, while LLaVAOneVision [166] unified image–text–video processing. VILA [167] focuses on pre-training techniques optimized for efficient edge deployment (Jetson Orin).

Following the Qwen series [168,169], Qwen2.5-VL [170] framework utilizes window attention and dynamic FPS sampling in the vision encoder to enable efficient, high-resolution video processing at native input resolution, unifying image and video representation by integrating multi-modal RoPE.

Training efficiency is advanced by Aquila-VL-2B's curated 40M dataset [171] (builds upon LLaVA-OneVision architecture [166], with Qwen-2.5-Instruct [172]/SigLIP [130] as the language/vision tower), and SAIL-VL(2B-8B)'s scalable Supervised Fine-Tuning (SFT) curriculum learning strategy [173].

Haplo-VL [174] fused multimodal embeddings early, Flash-VL 2B [175] merged SigLIP2, AIMv2, and Qwen-2.5-1.5B via tiling, token compression, and advanced data and training schemes. FluxViT [176] dynamically optimized tokens under compute budgets using UnMasked Teacher [177] pretraining.

NVILA [178] introduces an efficiency-oriented VLM framework built on VILA [167], employing a "scale-then-compress" strategy in space-and-time, integrates SigLIP [130] and Qwen2 [152] as vision and token encoders, optimized through dataset pruning (DeltaLoss) and FPS mixed-precision training, achieving lower latency and higher throughput than conventional VLMs like Qwen2-VL and LLaVA-OneVision.

Engineering for resource-efficient inference, SmolVLM [179] introduces a family of memory-efficient VLM (256M/500M/2.2B parameters, with single-image inference RAM usage of 0.8/1.2/4.9 GB), enabling adoption in various edge downstream applications. NanoVLMs [180] simplifies further in an educational VLM implementation.

More recently through train-free frameworks, LiveVLM [181] proposes a streaming-oriented KV cache to process video streams in real-time, while HoliTom [182] combines inner-LLM token merging and outer-LLM pruning strategies. Also, new benchmarks focusing on causality [183] and emergency detection [184] are of great interest for future studies in VAD.

The constant evolution of Vision–Language Models (VLMs), tracing developments from CLIP-style contrastive learning frameworks to architectures that integrate Large Language Models (LLMs). This progression encompasses advances in efficient backbone designs, hybrid architectures, and model compression strategies aimed at optimizing the trade-off between performance and computational efficiency. For detailed technical discussions, we refer to recent comprehensive surveys in the field [100,185–193].

VLMs in weakly-supervised VAD. CLIP's adoption in CLIP-TSA [194], UMIL [195], and AnomalyCLIP [196] demonstrated VLM features' superiority over ImageNet-trained models. CNN-ViT [98] fused CLIP with CNN features for anomaly subspace discrimination. BLIP-2 is employed in [8,197,198]. Nevertheless, there is still room to explore different lightweight encoders trained under other schemes and datasets, to compare how discriminative is the yielded space for anomalous events.

3.2. Audio feature representation

The acoustic landscape of our daily lives is rich with informative cues about the physical events unfolding in our surroundings. Environmental sounds, or everyday sounds, are ubiquitous in outdoor and indoor environments, generated by human and non-human activities. In real-world surveillance scenarios, anomalous events often have distinct audio signatures that can aid their identification, such as gunshots, screams, or breaking glass [199]. Beyond surveillance settings, sound plays a crucial role in various other environments present in online content. Within computational methods for analysing and understanding sound, Sound Event Recognition (SER) aims to automatically identify and categorize sounds occurring in our daily lives [200]. To that end, SER and Environmental Sound Classification (ESC) are tasked with a strong relevance to VAD.

3.2.1. Audio convolution-based features

CNN's have proven highly effective in extracting meaningful representations from audio data, enabling significant progress in tasks like environmental sound classification and speech recognition. This section explores the application of CNN's to audio feature extraction, highlighting key architectures and advancements in the field.

In the search for deeper embeddings, CNNs have been widely used in audio. There are two formats in which audio can be represented: the raw digital signal or its time–frequency version, a visual representation of the audio frequencies concerning time.

Learning from raw audio signals —. Early CNN-based approaches for audio processing focused on directly learning from raw audio waveforms. These methods employed 1D CNN architectures to capture temporal patterns and variations in the audio signal.

In the works of [201–204] a 1D CNN architectures are used to handle audio input without pre-processing, operating the raw input signal at different time scales, or with the input being initialized using gamma tone filter banks. In [205], a context-aware Active Learning framework for low-level feature selection and classification is proposed to achieve fast and accurate audio event annotation and classification on the UrbanSound8K dataset [206].

Spectrogram-based representations —. The spectrogram, a visual representation of audio frequencies over time, has become a dominant input format for audio processing.

A spectrogram can be generated using the Short-Time Fourier Transform (STFT) [207], which belongs to the family of Fourier-related transforms. The STFT is computed by splitting the input signal into overlapping frames, multiplying each frame by a window function, and applying the Fast Fourier Transform (FFT) to each frame. The resulting spectrogram can be further processed by applying Mel filter banks and a logarithmic operation to extract log-mel spectrograms. Mel-frequency cepstral coefficients (MFCC) can also be obtained from the spectrograms. The Mel scale is a perceptual scale of pitches [208], correlated to the human auditory system's frequency response.

With the image audio representation formed, well-known CNNs architectures from image domain tasks can be used as baselines. Firstly, [209] starts by operating on Mel-scaled spectrograms only. On [210], data augmentation techniques are evaluated. In the work [211], three networks operate on the raw audio, spectrograms, and the delta STFT coefficients. [212] employs a similar strategy using two networks with Mel-spectrograms and MFCC's as inputs.

Leveraging pre-trained models and transfer learning —. The availability of large-scale audio datasets, such as the general-purpose AudioSet [213], and pre-trained image recognition models from ImageNet has enabled the use of transfer learning for audio feature extraction. This approach involves initializing audio CNN's with weights learned from ImageNet or AudioSet, significantly improving performance and reducing training time.

One example is the work of [214] where the ImageNet pre-trained models (DenseNet [215], ResNet [50], and Inception-v1 [46]) are sued and fine-tuned them on different audio datasets, ESC-50 [209] and UrbanSound8K [206].

ESResNet [216] proposed a Siamese-like multichannel processing model using ResNet [50] as backbones with log-power spectrograms as input. In a subsequent work, [217] proposed a trainable time-frequency transformation based on frequency B-spline wavelets [218], using ResNeXt [51] as the backbone, setting the state-of-the-art on US8K. Both works obtained better results when the network was initialized on ImageNet and pre-trained on AudioSet.

PANNS [219] introduced a set of ResNets and MobileNets trained on raw AudioSet [213] recordings as versatile transfer learning systems. PSLA [220] proposed different training techniques, including ImageNet initialization, data augmentation, label enhancement, and an ensemble of various models using EfficientNet [55] as the backbone. Among these techniques, ImageNet pre-training seems to be the key ingredient, without which the performance decreases dramatically.

Addressing audio-specific challenges —. Researchers have also focused on addressing challenges specific to audio processing, such as shift-invariance and distortion introduced by subsampling operations in CNN's.

SINet [221] addressed the issue of subsampling operations in CNNs, which can lead to shift invariance and distortion problems [222]. Trainable low-pass filters are adopted by [223], and adaptive polyphase sampling is used by [224] in the max-pooling layers of a VGG model [45] variants while applying mix-up augmentation, achieving competitive results without pre-training on the FSD50k dataset [225].

DENet is proposed by [226] using a surveillance-oriented recurrent convolutional architecture that takes raw waveforms as input and learns the evolution of frequencies-of-interest over time, with experiments performed on the MIVIA datasets [199]. The proposed DENet architecture utilizes SincNet [227] as the backbone, and it introduces a denoising-enhancement (DE) layer that applies an attention map on the components of the band-filtered signal.

3.2.2. Audio transformer-based features

Transformer-based models, originally designed for NLP have also made significant inroads into the field of audio processing. This section explores the application of Transformers to audio feature extraction, highlighting their strengths and challenges in capturing temporal information and learning meaningful representations from audio data.

Purely attention-based models —. Audio Spectrogram Transformer (AST) [228] introduced the first purely attention-based audio classification model by treating Mel-spectrograms as a sequence of patches, incorporating pre-trained vision transformers ViT [67] and DeiT [68], achieving better results than CNNs in AudioSet tasks, although costly to train from scratch.

Transformers for raw audio —. Researchers have also explored the use of Transformers for processing raw audio waveforms, aiming to capture temporal dependencies directly from the signal without relying on spectrogram representations.

Audio Transformer [229] proposed learning upon raw signals without pre-training, evaluated on FSD50K [225], improving Transformer architectures with techniques such as pooling and multi-rate signal processing, and learning an adaptable time–frequency front-end representation for audio understanding.

Addressing computational complexity —. Patchout fast spectrogram Transformer (PaSST) [230] reduced the computation and memory complexity of Transformer training by applying Patchout to input sequences, improving generalization. Uses a modified version of both ViT and DeiT [67,68] with weights initialized from ImageNet and trained on AudioSet. Positional encoding is disentangled into time and frequency components, enabling inference on variable-length audio snippets. PaSST+ [231] further examined different approaches to extract general audio representations from PaSST under the HEAR [232] evaluation tool, concluding that mid-level features provide a more abstract representation beneficial for audio downstream tasks.

Multi-modal and contrastive learning —. Multi-modal and Contrastive Learning approaches have emerged as powerful techniques for learning robust audio representations by leveraging information from other modalities, such as text or video. Different modalities can accelerate compact learning in a single target modality by exploiting cross-modality structure.

The work of [233] contrastively induced audio representations from waveforms and log-mel spectrograms, allowing to learn better representations by maximizing the agreement between different augmented views of the same audio. It adopts PANNS [219] as the spectrogram encoder backbone and different unsupervised speech backbones as raw audio encoders. It is trained on AudioSet, with good generalization for ESC-50 [209] without fine-tuning. The work of [234] extends [233] to include correspondence with video frames, showing better learned audio representations. Both methods benefit from several Audio augmentations and larger batch sizes.

Knowledge distillation —. Because pre-training large-scale models requires large quantities of data and can be computationally expensive, another research direction has been distilling information from existing models trained on different modalities for which more data are available.

The work of AudioCLIP [235] combines ESResNeXt [217] with a pre-trained CLIP [122]. The trimodal hybrid architecture is trained on the AudioSet dataset using audio, video frames, and textual labels, resulting in minimum improvements over the baseline encoder in UrbanSound8K [206] and ESC-50 [209] datasets.

Wav2CLIP [236] used CLIP [122] in a two-step approach: first, an audio encoder is pre-trained on VGGSound by distilling CLIP image embeddings through videos, then fine-tuned to downstream tasks with frozen audio encoders. While no clear improvements were observed

Table 2
Summary of the identified works on audio feature extractor configurations.

	Input	Method	Image Init	Audio Pre-Train
CNN	R	CNN with RAW [201–205]		
	R+S	MS-CNN [211]		✗
	S	[209,210,212], SINet, DENet [221,226]		
		RCNN, PSLA [214,220]	✓	✗
		CNNAED, PANNs [219,240]	✗	✓
		ESResnet [216,217], EfficientAT [238,238]		✓
TRF	R	AT [229]		✗
	S	AST [228], PaSST [230,231]		
	S	AudioCLIP, Wav2CLIP [235,236]		✓
	R+S	RSCLIP, RSVCLIP [233,234]		

R: Raw audio signals; S: Spectrogram.

for downstream tasks, the embeddings derived from CLIP enable cross-modal text/audio-to-audio retrieval for the VAR task [237]. Moreover, in contrast to AudioCLIP [235], they did not learn the visual encoder but distilled CLIP into an audio model, resulting in one joint embedding space for three different modalities.

Efficient audio transformers —. Researchers have focused on developing efficient and high-performing audio Transformers, combining the strengths of CNN's and Transformers.

EfficientAT [238] is the ultimate example of that, with the focus on a low-complexity general-purpose audio embedding extractor based on a MobileNetV3 [54] architecture, complemented with Squeeze-and-Excitation layers. It is trained on AudioSet with initialization on ImageNet, using a Knowledge Distillation (KD) scheme with an ensemble of PaSST [230] models as the teacher. A range of models with varying complexities and spectrogram resolutions were introduced, from low-complexity models for edge devices to larger models, by scaling the model's width.

In [239], authors investigated the performance of EfficientAT with different levels of embeddings on downstream tasks using the HEAR [232] evaluation tool. Further refinements were made by replacing the original blocks with a Dynamic Inverted Residual Block [238], proving to be efficient, high-performing and easy-to-fine-tune audio models. It achieves state-of-the-art on AudioSet and competitive results in general-sounds-related tasks while balancing the efficiency of CNNs efficiency and the Transformer's superiority to scale up with large-scale datasets.

Audio feature extraction for WVAD —. Despite the advances in audio representation learning, the only audio embeddings available for the XD-Violence (XDV) dataset [22] are those provided by their authors. These embeddings were obtained using the VGG (configuration E) version of CNNAED [240], which pre-trains on a large-scale dataset of 70M YouTube videos (5.4M training hours). Recently, AVadCLIP [42] implemented Wav2CLIP as the audio FE.

Those works exploring distillation learning [236,238,238], attend on computational cost [230] and resort to signal processing enhancements of networks [217,221,229] showed promise as candidates to extract audio embeddings.

Table 2 summarizes the identified works and the different FE configurations.

3.3. Feature extractor summary

The evolution of FEs for VU has largely mirrored image classification, with adaptations to accommodate the additional temporal dimension. This trend is evident in both CNN and transformer-based networks. Early works attempted to extend modelling through the temporal axis, while later ones factorized spatial and temporal domains to achieve a better speed–accuracy tradeoff. This section has discussed various architectures, from the classic CNNs like AlexNet and VGG, to the more recent transformer-based models like ViT and Swin Transformer.

In the context of audio representation, SER and ESC are of great relevance to VAD. The spectrogram is the most common input representation for audio, although raw digital signals are also used. CNNs have been widely used for audio tasks, with architectures like [216,217,221] operating on spectrograms, while being lightweight. More recently, transformer-based architectures have been introduced for audio tasks, such as the [229] operating on raw signals and [230] on spectrograms. The work of [238] unifies both types of architecture, showing promise.

It is also important to note that real-world surveillance scenarios often have distinct audio signatures that can aid in identifying abnormal events. Works like [226] can integrate VAD in those situations. However, their integration with VAD requires a new UCFC-like benchmark containing audio signals in videos, for instance ECVA [241].

Transfer learning and pre-trained models have played a significant role in feature extraction for visual and audio data. The release of large-scale datasets like AudioSet and ImageNet has enabled the pre-training of models, which can then be fine-tuned for specific tasks. This approach has been particularly effective in audio tasks, with models like ESResNet and EfficientAT showing superior performance when initialized on ImageNet and pre-trained on AudioSet. Regarding the integration of VLM, works like Wav2CLIP [236], provide insights on how to integrate image, audio, and text data in a joint space.

In the context of WVAD, FEs for visual data have seen a similar shift from CNN-based architectures, C3D and I3D to transformer-based ones, VSwin and ViT. Only CNNAED [240] has been used to obtain audio embeddings for audio data. Pre-training and transfer learning have also been key to successfully applying these models. Integrating new proposals with WVAD, we can expect to see further advancements in the performance of VAD systems.

4. Datasets/benchmarks

Benchmark datasets are crucial in computer vision research, defining the problem scope and providing a fair comparison of different algorithms. In VAD, datasets reflect how the research community has interpreted and addressed real-world anomaly detection needs over time.

Table 3 summarizes the key characteristics of these datasets, highlighting their scene diversity, anomaly types, and annotation levels.

Early datasets often focused on public security applications with static cameras in restricted environments. This led to simple scenarios, artificial anomalies and limited scale, ultimately hindering the development of more generalizable VAD models.

Table 3
Summary of key characteristics of VAD datasets.

Dataset	Scene	Anomaly types	Annotation	Modality	Eval metric	#Videos/Length
Classical datasets						
Subway	Indoor	Wrong way, loitering	Temp	Visual	AUC	2/3h
UMN	Indoor/Outdoor	Staged panic	Temp	Visual	AUC	5
UCSD-PED	Outdoor (Campus)	Pedestrian (bike, etc.)	F/P	Visual	AUC	90/0.1h
CUHKAvenue	Outdoor (Campus)	Staged	F/P	Visual	AUC	35/0.5h
ShanghaiTech	Outdoor (Campus)	Staged (Pedestrian)	F/P	Visual	AUC	437/–
Recent Datasets						
UCF-Crime	Surveillance (YT)	13 real-world types	Weak: Video-Train Frame-Test	Visual	AUC	1900/128h
XD-Violence	In/Outdoor	Fight/Shoot/Riot Abuse/CarAcc/Explos	Weak	Vis/Audio	AP	4754/217h
UBnormal	Synthetic/29	22, Open-set, unseen	Pixel	Visual	AUC RBDC+TBDC	543/2.2h
NWPU	Surveillance/43	28 scene-dependent	Frame-Test	Visual	AUC Anticipation	547/16h
UCF/XDV-AR	Multiple	UCF+XDV	Video Caption	Vis/Text/Audio	R@K,MdR,SumR	–
HAWK	Diverse (7 sets)	Real-world	Video Captions +QA(Video Human-Centric)	Vis/Text	BLEU1-4 GPT-Guided	16 000/–
UCFA	Surveillance	UCF	Event Caption	Vis/Text	BLEU1-4,METEOR, ROUGE-L,CIDEr, IoU+AUC	1854/122h
MSAD	Diverse/500	35 human, 20 non-human	Weak	Visual	AUC	720/248h
HIVAU-70K	Multiple	UCFA+XDV	Frame +Video/Event/Clip Caption +QA(VEC Capt, Desc, Reason)	Vis/Text	AUC,AP+ BLEU1-4,METEOR ROUGE,CIDEr	5443/–
ECVA	Real-world/21 (YT)	100	Event Capt,Reason +Importance Curves	Vis/Text/Audio	AnomEval	2240/88h
M-VAE	Multi-scene	11 events, 14 scenes CUVA,	Frame+ QA(Event Quadruples)	Vis/Text	Macro-F1, T5/GPT-basedÜ mAP@tIoU+F2,FNRs	1000/32.50h
UCFVL	Surveillance	UCFA	QA(Detect+Classf +Temp Ground +MCQ+Event Desc)	Vis/Text	Acc+IOU+ GPT4o-BASED	1699/88.2h
VADD	In/Outdoor	18 (UCFC Throwing Action [242] Road-Accident)	Weak	Visual	AUC,AP	2591/–
UCFDVS	Surveillance	Motion-based-UCFC	Event-frames	Event	AUC,FAR	–
VANE	Real+Synthetic	5 types (e.g., pass-through, disappearance)	MC-Video QA	Vis/Text	VQA Acc	325/–
Surv.VQA-589K	Surveillance	18 (UCFA,MSAD MEVA,NWPU)	Frame+ Event Caption+ 589K QA(Det+Clas+Subj+ Desc+Cause+Result)	Vis/Text	VideoGPT+ based	3030/159h

R@K: Recall at K; **MdR**: Median Rank; **SumR**: Sum of all Recalls; **RBDC**: region-based detection criterion; **TBDC** track-based detection criterion;
BLEU: Bilingual Evaluation Understudy; **METEOR**: Metric for Evaluation of Translation with Explicit Ordering;
ROUGE: Recall Oriented Understudy of Gisting Evaluation; **CIDEr**: Consensus-based Image Description Evaluation.
AUC: Area Under the ROC Curve; **AP**: Average Precision; **METEOR**: Metric for Evaluation of Translation with Explicit ORdering; **IoU**: Intersection over Union.
FAR: False Alarm Rate; **mAP**: Mean Average Precision; **Acc**: Accuracy (overall correct predictions).

4.1. Early datasets

- **Subway** [243]: Captures events at entrance and exit gates of a subway station, focusing on anomalies like walking in the wrong direction and loitering. Its indoor setting and limited anomaly types restrict its generalizability.
- **UMN** [244]: Records staged crowd panic and escape events in indoor and outdoor scenes. Its artificial nature and lack of spatial annotations limit its relevance to real-world scenarios.
- **UCSD-PED** [245]: Focuses on pedestrian anomalies (e.g., biking, skateboarding) captured from two viewpoints on a university campus. Its simple scene and easily detectable anomalies have led to saturated model performance, hindering further development.
- **CUHKAvenue** [246]: Similar to UCSD-PED, it captures staged anomalies on a university campus, providing both frame-level and pixel-level annotations. However, its focus on a single scene limits its generalizability.

- **ShanghaiTech** [247]: introduce multi-view perspectives from 13 distinct surveillance cameras with varying angles across university campus locations, capturing over 270,000 training frames and 130 annotated abnormal events. It includes pixel-level ground truth for anomalies.

4.2. Recent datasets

The limitations of early datasets spurred the creation of more recent datasets that better reflect the open-set nature of real-world anomalies:

- **UCF-Crime** [21]: A large-scale real-world dataset with over 1900 surveillance and online videos spanning 13 anomaly categories *(Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, Vandalism, and Road Accident)*. It offers diverse indoor/outdoor scenes and includes temporal annotations for test videos.
- **XD-Violence** [22]: Introduces audio–visual modality for multimodal VAD. The dataset includes 2405 violent and 2349 non-violent videos, with 3954 for training and 800 for testing, totalling 4754 videos from diverse sources (e.g., movies, surveillance). Frame-level temporal annotations are provided for the test set.
- **UBnormal** [248]: A synthetic dataset for supervised open-set VAD, featuring virtual scenes with fine-grained pixel-level annotations. It enables controlled evaluation of models on unseen anomalies and includes a dedicated validation set for tuning. UBnormal facilitates open-set research, but introduces domain adaptation challenges when transferring to real-world data.
- **NWPU** [249]: It contains 43 scenes and 28 anomaly classes: single-person anomalies (e.g., climbing fence), interaction anomalies (e.g., stealing, snatching bag), group anomalies (e.g., protest, group conflict), scene-dependent anomalies (e.g., cycling on footpath, wrong turn, photographing in restricted area), location anomalies (e.g., car crossing square, crossing lawn), appearance anomalies (e.g., dogs, trucks) and trajectory anomalies (e.g., jaywalking, u-turn). Makes a total of 547 videos, 124 test videos with anomalies and 316 training videos. The first dataset for video anomaly anticipation (VAA).
- **VAR** [237]: Extends UCF-Crime and XD-Violence with text/audio annotations, enabling cross-modal video anomaly retrieval (VAR) via natural language or audio queries. 8 bilingual annotators (Chinese/English) reviewed videos to generate scene-specific anomaly/normal captions. For similar anomalies, two annotators per category detailed differences. Captions doubly quality-checked. For complex XD-Violence videos, annotations used audio descriptions due to dense content, leveraging audio–visual alignment, enabling cross-modal retrieval. It supports efficient offline search for specific anomalies in large-scale video archives
- **HAWK** [197]: A unified dataset integrating seven benchmarks (crime (UCF-Crime), campus (ShanghaiTech and CUHK Avenue), pedestrian walkways (UCSD Ped1 and Ped2), traffic (DoTA), and human behaviour (UBnormal)), totalling 8000 anomaly videos with GPT-4-generated descriptions and QA pairs. Annotations follow a two-step process: dense captioning via tools (e.g., InternVideo, GRiT), then GPT-4 synthesis for anomaly-specific descriptions and 5W2H-style QA pairs. Data is structured as `<VIDEO>: {DIS: <DESCRIPTION> | QA: <QUESTION> → <ANSWER>}`. Evaluation uses both text-level (BLEU, ROUGE) and GPT-guided metrics (Reasonability, Detail, Consistency), supporting context-aware, open-ended anomaly understanding across diverse scenarios. HAWK enhances practical applicability, enabling models to handle varied user inquiries while maintaining robust anomaly detection across complex, real-world scenes.
- **UCFA** [250]: UCFA extends UCF-Crime as the first large-scale multimodal dataset for Temporal-Specific Video Generation (TSGV), Video Captioning (VC), Dynamic Video Captioning (DVC), and Multimodal Anomaly Detection (MAD) tasks. It contains 1854 high-quality surveillance videos, annotated with over 23,000 sentence-level descriptions and 0.1s-precision timestamps by trained annotators under structured guidelines. It provides 110.7 h of data with concise, event-specific language. UCFA emphasizes fine-grained temporal grounding and domain relevance, supporting advanced tasks like captioning and query-based retrieval in dynamic, low-quality surveillance contexts. Dataset splits and ethical protocols are included to promote reproducible research. Ethical considerations and dataset splits (train/val/test) are detailed to support reproducible, responsible research in this emerging field.
- **Multi-Scenario AD (MSAD)** [12]: Addresses diverse surveillance challenges with broader scenarios/viewpoints. Contains 35 human-related anomalies (e.g., fighting, people falling, shop robberies) and 20 non-human-related anomalies (e.g., water leaks, factory fires, tree falls) across crime, pedestrian, and industrial domains. High-resolution fixed-camera footage (1920 × 1080, 30 fps), excluding low-quality clips, moving cameras, and Personally Identifiable Information (PII). The dataset addresses imbalanced anomaly distributions (e.g., "Fighting" at 4.2% in water incidents) and includes detailed statistics on frame numbers, video durations, and anomaly-type distributions (e.g., continuous vs. abrupt motion patterns). Evaluation protocols assess generalizability/adaptability (cross-scenario performance) and practical applicability (real-world robustness), with train/test splits tailored to these goals. Privacy and ethical considerations are central: face/vehicle blurring, restricted academic access via agreements, and pre-extracted features (I3D, SwinTransformer) to minimize data exposure. This design enables robust testing of VAD models in cross-scenario adaptation while balancing accuracy with privacy preservation.
- **VAD-Instruct50k** [251]: Built via a three-stage semi-automatic pipeline: (1) *Data Collection* gathers 5547 untrimmed videos from UCFC and XDV, filtered for quality; (2) *Annotation Enhancement* applies efficient single-frame labelling (2.35/video), generates pseudo-event clips (10–20s), and adds multimodal captions using Video-LLaVA; (3) *Instruction Construction* prompts LLaMA3-70B to synthesize 51,567 explainable QA pairs with natural language rationales (e.g., "Why is this abnormal?"), manually filtered for quality. The dataset offers fine-grained temporal supervision and rich semantic context, combining precision with scalable human-LLM collaboration.
- **HIVAU-70K** [13]: A hierarchical benchmark built from UCFC and XDV (a follow up work from VAD-Instruct50k) with 70,000+ annotations across three temporal levels: *clip-level* perception (short-term visual understanding), *event-level* reasoning (anomaly classification and description), and *video-level* analysis (causal explanations). It comprises: (1) *Video Decoupling* — 55,806 clips (5–20s) from 11,076 events in 5443 videos, manually annotated by five experts in 20 h ; (2) *Free-text Annotation* — clip-level captions via LLaVANext-Video and UCFA [250], and event/video-level summaries (Judgment, Description, Analysis) via recursive LLaMA3-70B prompting; (3) *Instruction Construction*—task-aligned prompts (Caption, Judgment, Description, Analysis) to support anomaly reasoning in VLMs. Covering short- to long-term context with balanced normal/anomalous content, HIVAU-70K enables interpretable, multi-scale anomaly detection and reasoning (see Figs. 7–10).
- **Exploring the Causation of Video Anomalies (ECVA)** [241]: An extension of CUVA [252], ECVA benchmark, a comprehensive dataset designed to advance causal reasoning in video anomaly understanding by addressing three dimensions: what (anomaly

type and description), why (underlying cause), and how (severity via temporally dynamic importance curves). ECVA comprises 2240 real-world videos spanning 21 categories and 100 subcategories, totalling 88.16 h, with meticulously curated annotations including free-text explanations, temporal event boundaries, and severity curves—a clever way to visualize anomaly severity over time, computed by (1) generating event descriptions, (2) ranking severity via GPT, (3) calculating frame-text similarity with CLIP, and (4) fusing scores. As raw curves can be noisy, a post-process improves precision using a voting mechanism over Video Captioning (VideoChat), Video Entailment (SEVILA), and Video Grounding (UniVTG), followed by wavelet smoothing. Evaluation is conducted via AnomEval, a metric assessing VLMs through:

  - Basic Reasoning: Coverage of key entities and logical coherence.
  - Consistency: Binary scoring using GPT to compare responses.
  - Hallucination: Robustness testing via edited videos.

  AnomEval achieves 82%–89% alignment with human judgment, surpassing traditional metrics like BLEU and ROUGE. ECVA supports fine-grained, causality-aware anomaly detection in long-form videos.
- **Multi-scene Video Abnormal Event extraction and localization (M-VAE)** [253]: is a instruction tuning dataset, derived from the CUVA benchmark (preliminary version of ECVA) [252] using its reason, result, and description tasks, built via a two-stage pipeline: a spatial understanding dataset by sampling 20K frames/images each from Ref-L4 [254], HumanML3D [255] (25K 1s clips), RSI-CB [256], and COCO [257], paired with handcrafted instructions to extract spatial cues (e.g., actions, objects, background); and generate abnormal event quadruples (subject, event type, object, scene) from 5 Q-A tasks via ChatGPT prompting and manual filtering.
  The dataset spans 1k videos (800K frames) with 1.68 abnormal events/80s duration per video on average, covering 11 event types *(Fighting, Animals, Water, Vandalism, Accidents, Robbery, Theft, Pedestrian, Fire, Violations, and Forbidden)* and 14 distinct scene categories *(School, Shop, Underwater, Street, Road, Boat, Wild, Forest, Residence, Bank, Commercial, Factory, Lawn and Other)*, validated with Kappa = 0.87. CUVA timestamps are used for localization (8 FPS, 800K frames, avg. 1.68 events/video). Evaluation includes Macro-F1 (element/pair/quadruple extraction), mAP@tIoU (0.1–0.3), F2 score, and FNRs, emphasizing low false-negative rates and statistical significance (t-tests). Sherlock enables interpretable, fine-grained abnormal event detection across diverse real-world scenes.
- **UCVL** [14]: is constructed by integrating the UCFC and UCFA datasets, refining 1829 videos (into 1699 total/1030 train/369 validation/300 test) with anomaly labels and segment-level descriptions. Leveraging Qwen2-72B [169] to generates QA pairs by summarizing UCFA annotations and formulating scenario-specific questions, while GPT-4o ensures answer accuracy through detailed prompting, the dataset generates 16,990 QA pairs across six task types: *anomaly detection (True or False, TF), anomaly classification (AC), anomaly temporal grounding (TG), multiple-choice questions (MCQ), event description (ED), and anomaly description (AD)*. Annotations combine human-labelled timestamps and event summaries from UCFA with LLM-synthesized questions. To mitigate bias and ensure fairness, the dataset excludes personally identifiable information and employs diverse scenario sampling. Evaluation benchmarks eight open-source MLLMs (0.5B–40B parameters) [166,169,258], assessing performance via a hybrid scoring system: pattern matching for MCQs and GPT-4o-based evaluations for open-ended responses (ED, AD). Finetuning protocols validate model adaptability (e.g., LLaVA-OV-7B increases scores by 6.2%), with results highlighting gaps in MLLMs' anomaly perception and guiding future research toward robust, interpretable surveillance systems.
- **VADD** [259]: *Video Anomaly Detection Dataset* is as an extension of UCF-Crime with 2591 videos (2202 train, 389 test) spanning 18 classes, including underrepresented anomalies like road accidents and dangerous throwing from UCFC, Throwing Action [242], and newly collected accident videos annotated, with video-level labels (train) and frame-level anomaly timestamps (test).
- **UCFDVS** [260]: First event-based VAD benchmark using Dynamic Vision Sensors (DVS). Captures asynchronous sparse event streams at high temporal resolution (1280 × 720, 242s/video). Unlike RGB, DVS encodes *ON/OFF* polarity changes, reducing redundancy and enhancing sensitivity to motion-related anomalies. Preserves temporal precision, ensures model compatibility. Novel benchmark for motion-centric anomaly detection in low-light/high-speed scenarios.
- **VANE** [261]: evaluates VLM in detecting and localizing anomalies in real-world and synthetic videos. It includes 325 video clips (197 synthetic, 128 real) and 559 QA pairs, covering five anomaly categories: unnatural transformations, appearances, pass-through, disappearance, and sudden appearance. Real-world samples derive from UCFC, UCSD Pedestrian, and Avenue; synthetic videos use closed/open-source text-to-video diffusion models (e.g., Open-Sora [262], VideoLCM [263]). QA pairs follow a multiple-choice video QA format, targeting subtle inconsistencies such as abrupt object removals or unnatural motion patterns. Built through a three-stage pipeline: (1) Annotators label subtle anomalies in synthetic and real-world videos, ensuring precise temporal alignment; (2) GPT-4o generates descriptive captions for videos, leveraging annotated frames to highlight anomalies; and (3) Custom prompts guide GPT-4o to produce QA pairs, balancing simplicity and complexity to test reasoning, localization, and contextual understanding.
  Benchmarked on nine Video-LMMs (e.g., LLaVA-NeXT, MiniGPT4-Video, Goldfish), it reveals poor open-source model accuracy (e.g., LLaVA-NeXT: 11.59% on SORA videos) and struggles with synthetic anomalies (e.g., 10.57% on VideoLCM). Closed-source models show marginally better stability. Human evaluation scores 19.44% on UCSD-Ped2, underscoring difficulty. VANE-Bench bridges gaps in evaluating LMMs for real-world/synthetic anomalies, with applications in misinformation detection and security systems. The hybrid design (real + synthetic) and MC-Video QA format enable robust testing, available under a non-commercial license.
- **SurveillanceVQA-589K** [15]: is the largest open-ended video question-answering (VQA) benchmark tailored to real-world surveillance scenarios, from UCFA [250] authors, containing 589,380 QA pairs across 12 cognitively diverse task types: temporal reasoning, causal inference, spatial understanding, anomaly interpretation, factual summarization, behaviour/spatial–temporal analysis, object interaction, motion dynamics, contextual description, event localization, safety-critical implications, and structured anomaly detection. Spanning 18 abnormal event categories (UCFC + Fire, Object/People Falling, Pursuit, Water Incidents), integrating four surveillance datasets (MEVA [264], NWPU [249], MSAD [12] and UCFA [250]), includes 3030 videos (159.18 h total) with 80%–20% train–test splits at the clip level.
  The dataset is constructed through a five-step pipeline integrating human annotators and large vision–language models (LVLMs):

  - Human-Labelled Annotations: Video clips segmented with manual spatiotemporal labels—precise timestamps, event descriptions (object interactions, motion patterns), and anomaly intervals.

- LVLM-Labelled Annotations: LLaVA-Video-7B-Qwen2 generates supplementary event narratives and scene summaries; validated against human labels for consistency.
- Integrated Annotations: Human/LVLM outputs fused via timestamp alignment and description merging; discrepancies resolved by majority voting/expert review. Output: synchronized timestamps (e.g., [35.1, 41.0]) + enriched descriptions.
- QA Generation: Qwen-Max-7B creates QA pairs (6 normal/6 abnormal tasks). Question types: *Temporal*: "When did the SUV leave?", *Causal*: "Why did the woman grab guns?", *Spatial*: "Where were individuals standing?". Answers video-grounded (e.g., "Two motorcycles left right").
- Evaluation: 8 LVLMs (0.5B–7B) tested on 4 metrics: Contextual Integration, Detail Orientation, Temporal Understanding, and Causal Reasoning.
- Key findings: Anomaly reasoning gaps & low scores on structured tasks. Fine-tuning boosts general understanding, not complex reasoning

Ethical considerations include blurring faces/license plates, fair compensation for annotators, and bias mitigation through diverse scenario inclusion (e.g., varying weather, lighting, demographics). This structured approach ensures high-quality, cognitively diverse QA pairs that challenge models to handle real-world surveillance complexity, from nuanced event descriptions to safety-critical anomaly detection.

The evolution of VAD datasets reflects the growing recognition of the complexity and diversity of real-world anomalies. Recent datasets, with their larger scale, more realistic scenarios, and richer annotations, provide a more challenging and relevant benchmark for evaluating and advancing VAD models.

5. A survey of WVAD methods

In the literature about WVAD, methods are characterized by a combination of several key components. Such components include FEs that process input video from single or multiple modalities using various backbone architectures and Feature Modulators. The latter transforms raw features from FE into more discriminative representations. The Segment Selection (SS) is responsible for optimizing segments of interest. This selection process can be based on different metrics and applied at different levels of the network, with or without the MIL scheme. The Loss function guides the learning process by defining the training objective to target the feature- and/or score-level, using only the video or pseudo-segment-level labels generated through an additional refinement stage.

The interactions between these elements, guided by the chosen AC, play a crucial role in shaping the overall anomaly detection system. These clusters e

The coming subsections will delve into a detailed exploration of these method clusters (Fig. 11), encompassing a wide spectrum of methods, focusing on specific datasets or input modalities to utilizing Vision–Language Model representations or incorporating multiple training stages. Some even define novel task settings within the broader VAD area. By examining each group's key aspects, strengths, and limitations, a comprehensive overview of the current state of WVAD research will follow.

5.1. Vanilla scores

Several studies investigate the **vanilla anomaly scores**, output from the regressor network, as confident cues to select and optimize. Some works adopt the top-K selection as an alternative to the maximum definition imposed by the ranking loss of [21]. Others go further to embed and/or modulate input features.

TCN-**IBL** [85] introduces a complementary inner bag loss to reduce intra-class distances and enlarge inter-class distances of instances simultaneously. This is achieved by iterating upon the ranking definition. A Temporal Convolutional Network (TCN) [265] is employed as a Feature Modulator (FM) to fuse information causally.

A Criss-Cross Attention [266] to aggregate the local **Spatial-Temporal Context (STA)** is introduced by [93]. Through the recurrence of this operation, global correlations are captured. A bidirectional recurrent network is used as the Score Head, and a mutual cosine embedding loss [267] is employed to centre normal representations and identify representations that deviate from the norm. The MIL ranking loss integrates the overall cost, operating at the score level.

A **Temporal Context Alignment Network (TCA)** is proposed by [95] as a new FM to capture temporal context at different scales, using a multiscale MHSA module. The method introduces a sparse continuous sampling strategy whereby input features are re-sampled in subsets of seven consecutive clips. It directly addresses the limitations of Segmentation. Furthermore, the ranking loss has been modified to consider each subset's top-K anomaly scores as the optimization unit. It adds the sparse constraint and temporal smoothness terms.

Temporal and Abnormal Information (TAI) is considered in [268], employing a FM composed of a combination of convolutional layers for short-term relations and a Global Context block [64] for long-term dependencies. This enables to achieve a balanced choice compared to those utilizing Non-local (NL) blocks [31]. An N-pair loss [269] is proposed to address the anomalous inner bag variance and provide a more extensive range of segments during optimization at both feature and score levels. It also provides insights about Deep Metric Learning (DML).

Table 4 summarizes the identified works.

For reference, metrics detailed throughout the next tables in this section will always be *Area Under ROC Curve (AUC-ROC) for UCFC*, and *Average Precision (AP) for XDV*, mentioned in original works. If more than one value is present, it will be referent to evaluation over the *Overall Set/Abnormal Only Set / False Alarm Rate (FAR)*, respectively.

Key Points of Vanilla Scores Methods: These methods focus on utilizing the raw anomaly scores output by the regressor network as the primary indicator for anomaly detection.

- **IBL:** Introduces an inner bag loss to refine the ranking objective and minimize intra-class distances while maximizing inter-class distances.
- **STA:** Employs criss-cross attention to capture spatial–temporal context and uses a mutual cosine embedding loss to centre normal representations.
- **TCA:** Develops a multiscale MHSA module for temporal context aggregation and introduces a sparse continuous sampling strategy.
- **TAI:** Combines convolutional layers and a Global Context block for feature modulation and utilizes an N-pair loss to address inner bag variance.

While these methods offer valuable insights into the limitations of MIL and propose novel loss functions, different Segmentation processes, or top-K selections, their reliance on regressor scores solely trained on the video-level label as the primary AC to select relevant segments may hinder their ability to generalize. While incorporating more sophisticated techniques, many subsequent works still base their SS on this basic criterion. This highlights the need to explore alternative AC and refine the supervision process to improve the robustness and generalizability of VAD methods.

5.2. Temporal dynamics

Another line of research investigates the **temporal dynamics** for anomaly detection, guided by the hypothesis that anomalies that

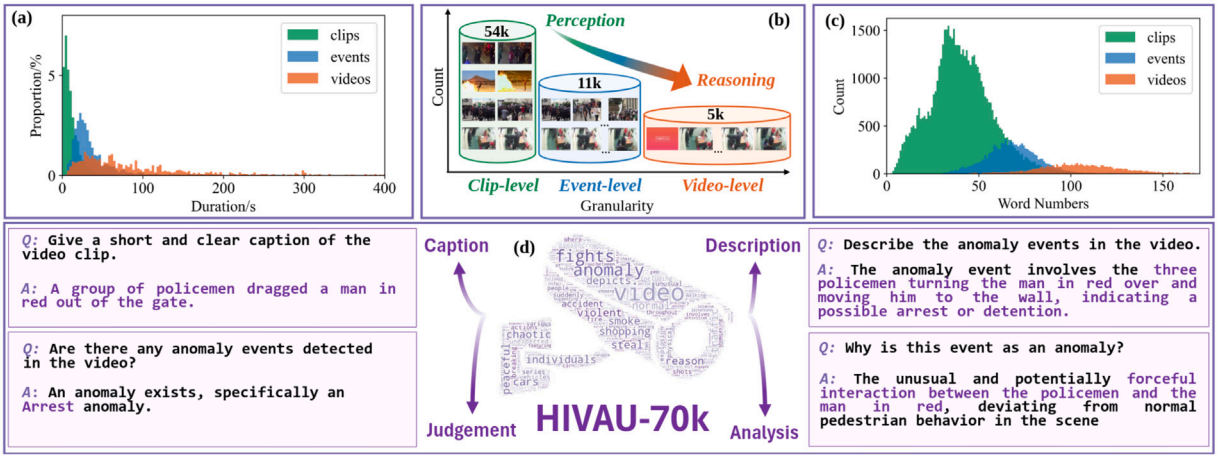


Figure 3. **HIVAU-70k dataset**. (**a**) Duration distributions for clips, events, and full videos, showing dominance of short clips. (**b**) Hierarchical data organization from clip-level to video-level, enabling perception-to-reasoning insights. (**c**) Word count variations across annotation levels, with more detailed descriptions at the video level. (**d**) Sample annotations capturing captioning, judgment, description, and anomaly analysis, highlighting nuanced understanding of anomaly events in complex scenes.

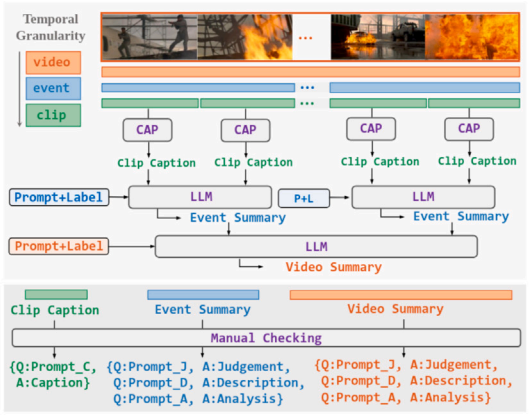


Figure 2. **Data Engine**. We present a structured workflow for generating hierarchical annotations across video, event, and clip levels. Clips are first captioned, then processed through a large language model (LLM) with prompts for event summarization. The outputs include clip captions, event summaries, and video summaries, followed by manual checking and refinement. This multi-step approach enriches the dataset with detailed judgments, descriptions, and analyses of anomalies, enabling robust contextual understanding at varying granularities.

Figure B. **An example of clip-level instruction data.**

Figure C. **An example of event-level instruction data.**

Figure D. **An example of video-level instruction data.**

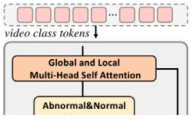


Fig. 7. Image source from [13].

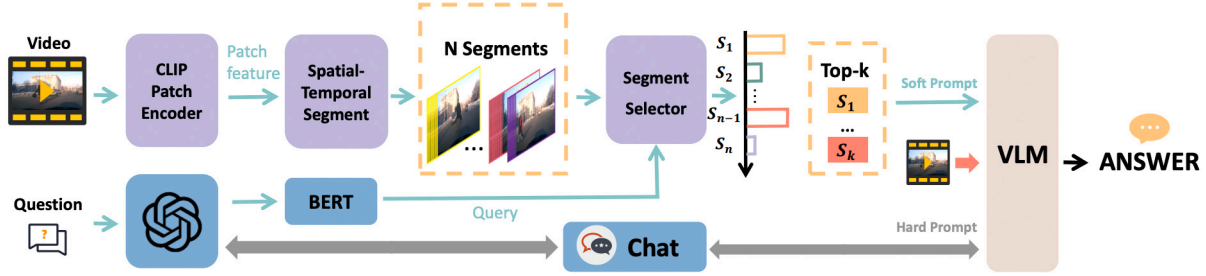
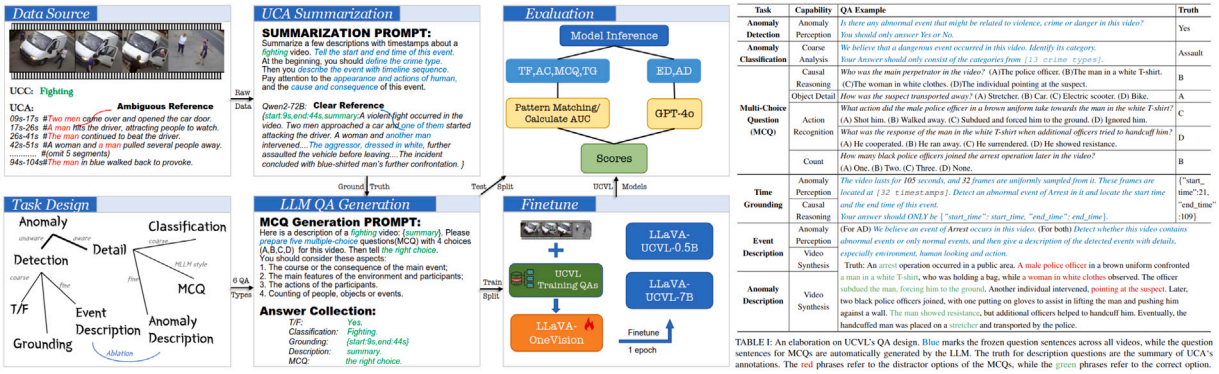


Fig. 8. Image source from [252].

occur infrequently among normal patterns result in notable changes in the time domain. Consequently, dynamic variations may serve as an indicator of an anomaly.

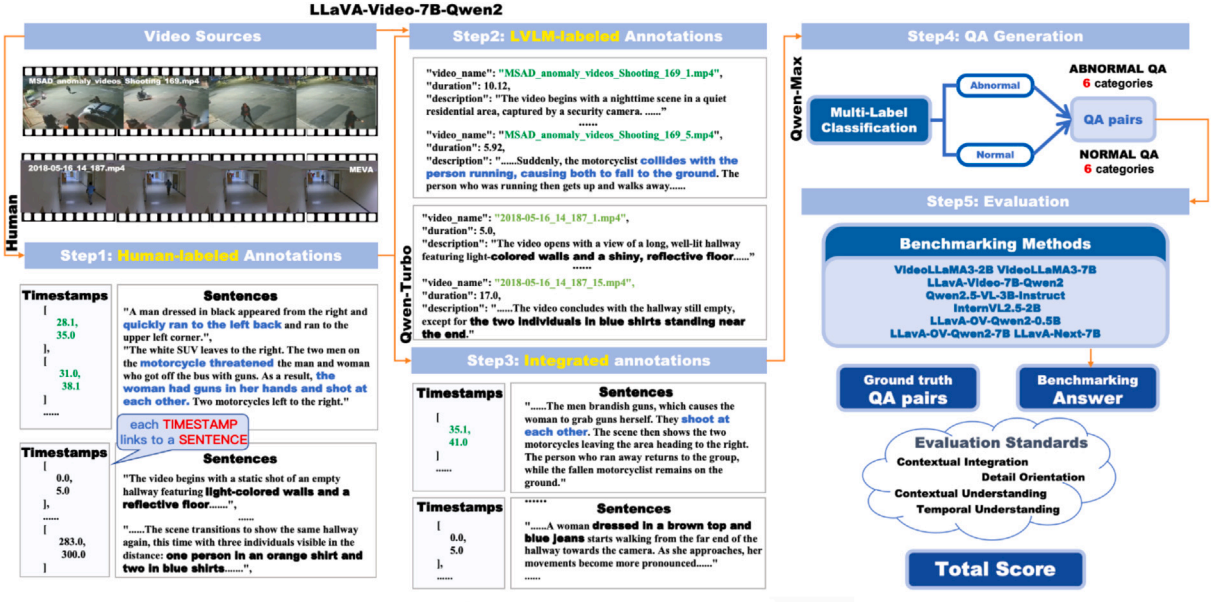
WSAL [111], a representative work for the surveillance task, explores temporal dynamics at different levels. The FE, TSN [34,35], accounts for the long-range temporal structure of the video through sparse sampling. The High-Order Context Encoding module first embeds feature through a convolutional layer and then generates scores from intermediate features' immediate semantic and dynamic variation cues. Training is conducted through a normal MIL configuration, whereby the margin between max-selected abnormal and normal instances is augmented. WSAL is the first work to provide results for



Task	Capability	QA Example	Truth
Anomaly Detection	Anomaly Perception	*Is there any abnormal event that might be related to violence, crime or danger in this video? You should only answer Yes or No.*	Yes
Anomaly Classification	Coarse Analysis	*We believe that a dangerous event occurred in this video. Identify its category. Your Answer should only consist of the categories from [13 crime types].*	Assault
Multi-Choice Question (MCQ)	Causal Reasoning	*Who was the main perpetrator in the video?* (A)The police officer. (B)The man in a white T-shirt. (C)The woman in white clothes. (D)The individual pointing at the suspect.	B
	Object Detail	*How was the suspect transported away?* (A) Stretcher. (B) Car. (C) Electric scooter. (D) Bike.	A
	Action Recognition	*What action did the male police officer in a brown uniform take towards the man in the white T-shirt?* (A) Shot him. (B) Walked away. (C) Subdued and forced him to the ground. (D) Ignored him.	C
		What was the response of the man in the white T-shirt when additional officers tried to handcuff him? (A) He cooperated. (B) He ran away. (C) He surrendered. (D) He showed resistance.	D
	Count	*How many black police officers joined the arrest operation later in the video?* (A) One. (B) Two. (C) Three. (D) None.	B
Time Grounding	Anomaly Perception	*The video lasts for 105 seconds, and 32 frames are uniformly sampled from it. These frames are located at [32 timestamps]. Detect an abnormal event of Arrest in it and locate the start time and the end time of this event. Your answer should ONLY be {"start_time": start_time, "end_time": end_time}.*	{"start_time":21, "end_time":109}
	Causal Reasoning		
Event Description	Anomaly Perception	(For AD) *We believe an event of Arrest occurs in this video.* (For both) *Detect whether this video contains abnormal events or only normal events, and then give a description of the detected events with details, especially environment, human looking and action.*	
	Video Synthesis	Truth: An arrest operation occurred in a public area. A male police officer in a brown uniform confronted a man in a white T-shirt, who was holding a bag, while a woman in white clothes observed. The officer subdued the man, forcing him to the ground. Another individual intervened, pointing at the suspect. Later, two black police officers joined, with one putting on gloves to assist in lifting the man and pushing him against a wall. The man showed resistance, but additional officers helped to handcuff him. Eventually, the handcuffed man was placed on a stretcher and transported by the police.	
Anomaly Description	Video Synthesis		

TABLE I: An elaboration on UCVL's QA design. Blue marks the frozen question sentences across all videos, while the question sentences for MCQs are automatically generated by the LLM. The truth for description questions are the summary of UCA's annotations. The red phrases refer to the distractor options of the MCQs, while the green phrases refer to the correct option.

Fig. 9. Image source from [14].



Dataset	Number of Videos	Total Video Duration	Number of Text Annotations	Number of Segments		QA Pairs
				Normal	Abnormal	
MSAD	201	4.23h	1783	1417	366	34290
MEVA	720	16.76h	2057	2044	13	37104
NWPU	255	16.29h	4166	4121	45	75258
UCA	1854	121.9h	23542	20380	3162	442728
Total	**3030**	**159.18h**	**31548**	**27962**	**3586**	**589380**

Fig. 10. Image source from [15].

the anomalous subset in WVAD, paramount for comprehending the approach's influence on detecting genuine anomalies.

Dissimilarity Attention Module (DAM) [270] employs a Dissimilarity Attention Module as the FM to capture abrupt changes in both channel and temporal dynamics. More specifically, dissimilarity among two consecutive clips, calculated either by Manhattan distance or cross-covariance matrix, obtains both channel-wise and temporal-wise attention maps, which are further processed by a many-to-one vanilla LSTM network to encode those channels which significantly changed over the temporal scale. During the training phase, a weighted ranking loss (wRL) assures the clip attention weights from the DAM and scores are aligned, including temporal smoothing and sparsity constraints [21]. As DAM captures local temporal dissimilarity among consecutive clips, it neglects global enhancement, thus making it effective for live applications.

Locality-aware Attention Network (LAN) [271] achieves causal consistency over the temporal dimension by mining temporal dynamics in model architecture and loss formulation. The Locality-aware

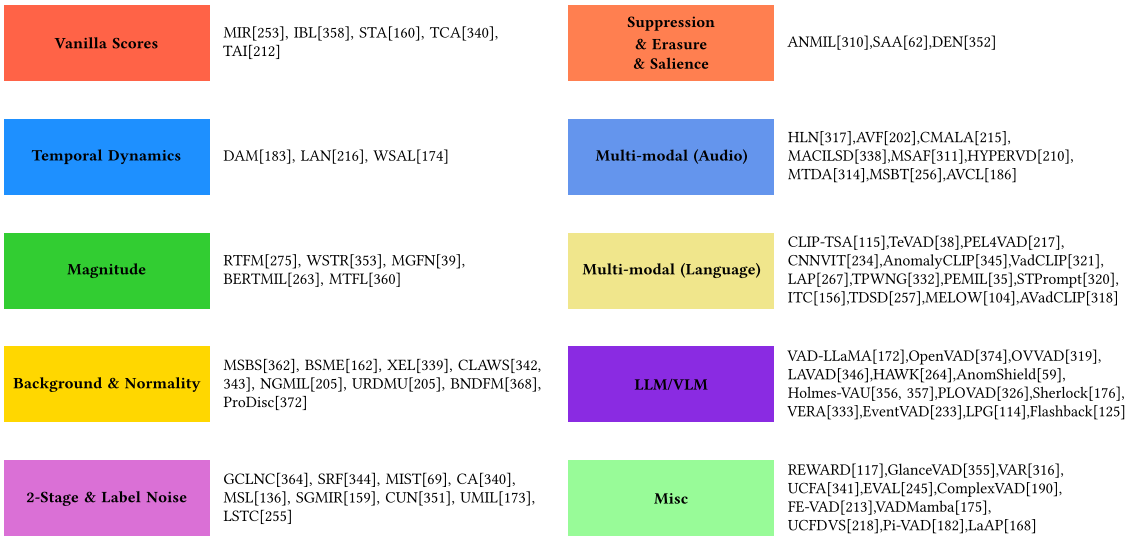


Fig. 11. Works across taxonomy clusters.

Table 4
Summary of works using Vanilla approaches.

Method	FE	FM	MIL	SS	LT		Metrics	
					FL	SL	UCFC	XDV
MIR	C3D I3D	✗	✓	max	✗	✓	75.41 77.92	✗
IBL	C3D	Conv Layer	✓	max/min	✗	✓	78.66	✗
STA	C3D I3D	Recurrent Criss-Cross Attention	✓	max/min	✓	✓	81.60 83.00	✗ ✗
TCA	C3D I3D	Multi-Scale MHSA	✓	✓	Top-K	✓	82.08/–/0.11 83.75/–/0.05	✗ ✗
TAI	I3D	Conv layer + GC	✓	multi-Top-K	✓	✓	85.73	✗

FE: Feature Extractor; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Frame-Level; **SL**: Score-Level.
Metrics: UCFC-AUCROC, XDV-AP, Overall/Abnormal Only/FAR.

Table 5
Summary of works that employ temporal dynamics.

Method	FE	FM	MIL	SS	LT		Metrics	
					FL	SL	UCFC	XDV
DAM	I3D	Att+LSTM	✓	max	✗	✓	82.67/–/0.3	✗
LAN	I3D	Att	✓	Top-Kwink	✓	✓	85.12/–/–	80.72
WSAL	TSN	Conv	✓	wink	✓	✓	85.38/67.38/–	✗

FE: Feature Extractor; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

Attention Network (LA-Net) models long-range dependencies and recalibrates the locality preference of adjacent snippets using a self-attention mechanism and Gaussian-like location prior as a bias term. A dense MLP extracts the final enhanced feature representation, and a causal convolution provides anomaly scores. LAN adopts Discriminative Dynamics Learning, using a dynamics ranking loss to amplify the variation of anomaly score magnitude between positive and negative bags and a dynamics alignment loss (KL divergence) to coordinate feature dynamics with score dynamics within each bag, in addition to the top-K MIL loss function. Note that this alignment's impact varies across datasets, with a more significant impact in UCFC, highlighting the importance of dataset-specific design considerations.

Table 5 summarizes the Temporal dynamics works.

Key Points of Temporal Dynamics Methods:

These methods focus on leveraging temporal variations and dynamic changes in video features as indicators of anomalous events.

- **WSAL:** Emphasizes temporal dynamics at multiple levels, using TSN for feature extraction and a High-Order Context Encoding module to capture dynamic variations.
- **DAM:** Employs a Dissimilarity Attention Module to capture abrupt changes in channel and temporal dynamics, making it suitable for live applications.
- **LAN:** Achieves causal consistency over the temporal dimension through a Locality-aware Attention Network and a Discriminative Dynamics Learning approach, with a significant impact on the UCFC dataset.

Exploring temporal dynamics has led to diverse approaches, including capturing dissimilarity between consecutive clips, modelling long-range dependencies, and aligning feature and score dynamics. These methods highlight the importance of considering temporal context and dynamic variations for effective anomaly detection. However, the challenge of adapting these strategies to dataset-specific characteristics remains a key area for further research.

5.3. Magnitude

Another approach uses the feature magnitude as the AC, and selects confident features and scores based on their L2 Norm, also known

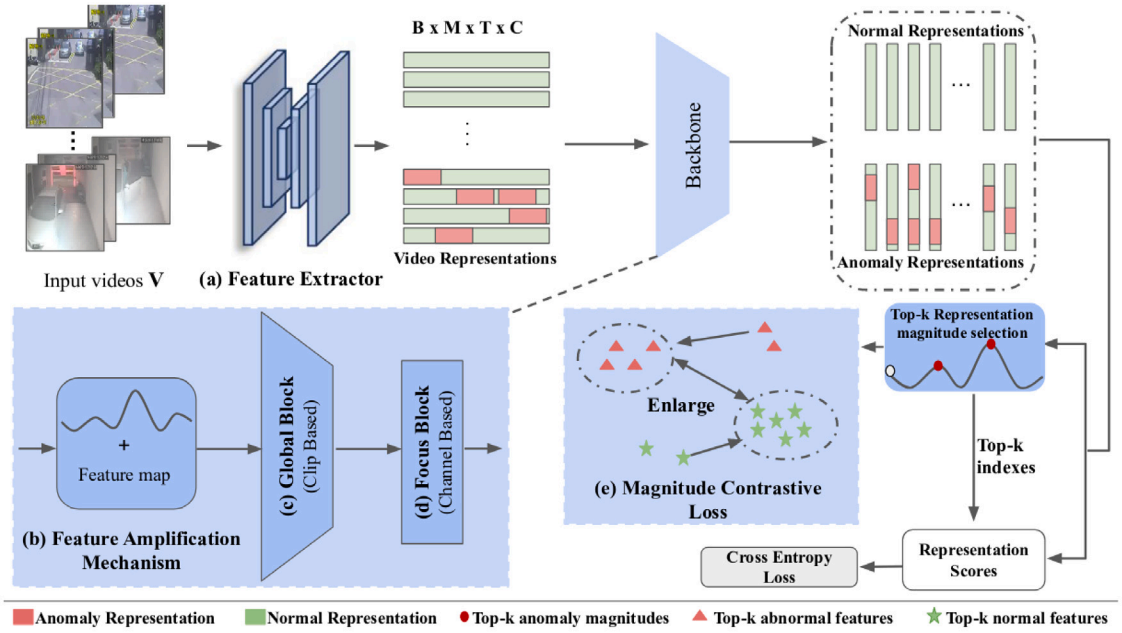


Fig. 12. MGFN [119] proposed method.

as Euclidean distance. This concept was introduced by **Robust Temporal Feature Magnitude Learning (RTFM)** [89]. The underlying assumption is that the mean feature magnitude of abnormal snippets is more significant than that of normal snippets. The proposal RTFM loss function enforces large margins between the top-K segments, with the largest magnitudes from abnormal and normal videos at both feature and score levels. It also employs a combination of Pyramid of Dilated Convolutions (PDC) [272] and a NL module (NLNet) [31] to capture both global and local correlations of vanilla encoded features. Subsequent works have adapted various aspects of RTFMs approach, either architectural or following the same AC.

Transformer-Enabled Temporal Relation Learning (WSTETR) [273] explores long-term temporal relations using a Transformer-enabled temporal relation encoder. The input video features are encoded with sine and cosine functions of different frequencies and further processed by a stack of Transformer encoder formed of multi-head relation aggregation (MHRA) layer and feed-forward layer from the original Transformer architecture [59]. This allows for learning multiple types of temporal relations from different representation subspaces. WSTETR uses the same cost functions as RTFM [89].

MGFN [119] (Fig. 12) builds upon RTFMs assumptions regarding loss formulation but concludes that pushing abnormal feature magnitudes to be larger than normal ones is only valid for similar scenes, as other elements besides anomalies also influence the magnitude. The proposed FM consists of three modules: a Feature Amplification Mechanism (FAM), Glance Block (GB), and Focus Block (FB). FAM amplifies input features by adding the 1D-Convolution modulated feature norm. The Glance block uses MHSA with a Feed-Forward Network (FFN) to provide the network global correlations for each space through time. The Focus block employs self-attention convolution with a fixed kernel size and an FFN to learn local channel-wise. A Magnitude Contrastive loss uses the pairwise distance between top feature magnitudes as the metric to guide the training, ensuring that the selected feature representations of normal and abnormal videos are at least margin units apart. Similar representations of the selected top-K are divided into two bags, minimizing the distance between feature magnitudes. This approach allows for flexibility in feature magnitudes, not forcing all abnormal features in different scenes to be larger than normal ones while allowing a regular video with substantial movement to have larger feature magnitudes than an abnormal one.

BERTMIL-RTFM [274] extends RTFM to incorporate video-level classification as a correction term for the final anomaly score, effectively improving both ranking and feature magnitude loss functions. This method demonstrates the potential of leveraging additional video-level information to enhance anomaly detection, particularly on datasets like XDV.

Multi-Timescale Feature Learning (MTFL) [259] is designed to capture anomalies of varying temporal durations by integrating features from short, medium, and long temporal tubelets (8, 32, and 64 frames, respectively). It employs a VSwin [113] to extract spatio-temporal features at these scales, which are then fused via a multi-stage process: pairwise cross-attention (PFL) merges features across scales, 1D convolutions (LTL) model local temporal dependencies, and self-attention (GTL) captures global snippet correlations. A classifier generates anomaly scores using the same magnitude-based optimization [89] (binary cross-entropy, feature magnitude maximization, and temporal smoothness constraints). The method leverages pre-trained VST models (e.g., Kinetics-400 or VADD-augmented variants) to enhance feature discriminability, achieving state-of-the-art results on UCFC (89.78% AUC) and XDV (84.57% AP). To address data scarcity in anomaly detection, the authors propose VADD (Video Anomaly Detection Dataset), an extension of UCFC with 2591 videos (2202 train, 389 test) spanning 18 classes, including underrepresented anomalies like road accidents and dangerous throwing. VADD incorporates data from UCFC, Throwing Action [], and newly collected accident videos annotated with video-level labels (train) and frame-level anomaly timestamps (test). This expansion enables robust evaluation of complex, real-world anomalies, with MTFL achieving 88.42% AUC on VADD, outperforming adapted baselines like RTFM*. The dataset's diversity and scale highlight MTFL's ability to generalize across subtle contextual anomalies (e.g., burglary) and rapid motion events (e.g., throwing).

Table 6 summarizes the works focusing on Magnitude.

Key Points of Magnitude-based Methods:

- **RTFM:** Enforces large feature magnitude margins between normal and abnormal segments using PDC and NL blocks.
- **WSTETR:** Captures long-term temporal relations using a transformer-based encoder while maintaining the magnitude-based loss formulation of RTFM.

Table 6
Summary of works that focus on magnitude.

Method	FE	FM	MIL	SS	LT		Metrics	
					FL	SL	UCFC	XDV
RTFM	C3D I3D	PCD+NL	✓	mean Top-K	✓	✓	83.28 84.30	75.89 77.81
WSTR	I3D	MHRA	✓	mean Top-K	✓	✓	83.17	✗
MGFN	I3D VSwin	Conv & 2*ConvMHSA	✓	multi Top-K	✓	✓	86.98 86.67	79.19 80.11
BERTMIL-RTFM	I3D	PCD+NL	✓	mean Top-K	✓	✓	82.10	✗
MTFL	VSwin	LTL	✓	mean Top-K	✓	✓	89.78	85.57

FE: Feature Extractor; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

- **MGFN:** Introduces a more flexible approach to magnitude, allowing for variations in feature magnitudes across different scenes and using a contrastive loss to separate normal and abnormal feature representations.
- **BERTMIL-RTFM:** Extends RTFM by incorporating video-level classification as a correction term to refine anomaly scores. This dual supervision improves both ranking and feature magnitude losses, effectively leveraging global video context.
- **MTFL:** Captures multi-timescale anomalies by extracting spatio-temporal features over short, medium, and long tubelets (8, 32, 64 frames) using a Video Swin Transformer. Features are fused through cross-attention, temporal convolutions, and self-attention, producing anomaly scores using magnitude-based objectives.

These methods leverage the assumption that abnormal events exhibit higher feature magnitudes. While this assumption holds in some cases, it can be limiting, especially in diverse datasets like XDV, where other factors besides anomalies can influence feature magnitude.

5.4. Background & normality

Another approach in WVAD leverages the prevalence of normal events. This is achieved by constructing a reliable normality reference or capitalizing on the background's influence on decision-making. These two branches interconnect, as the background in anomalous videos often equates to spatial information devoid of anomalous events. While some methods focus solely on noise-free normal videos, others exploit the abundant normal events present, even in videos labelled as anomalous.

Background —. Those addressing the background-bias problem [30] treat the task as a two-class problem: foreground and background. MS-BS [275] and BSME [276] address the background bias problem by treating WVAD as a two-class problem: foreground and background. These methods employ a two-branch background suppression approach, incorporating a Multi-scale Temporal Convolution Module to handle anomaly events of various lengths.

Multi Scale Background Suppression (MSBS) [275] proposes a two-branch background suppression method based on [277], an approach for the Temporal Action Localization task. It incorporates the Multi-scale Temporal Convolution Module from RTFM [89] as the FM. The base branch assumes that the background is present in every input feature, thus forcing the top-K mean of foreground scores to be close to the video-level label and the background to be close to 1. In the suppression branch, the filter module weights the input features to the temporal convolution module. The foreground class behaves the same as in the base branch, but the top-K mean background scores are now moved close to 0. This forces the filter to modulate features and suppress background information as input to the FM. An additional normalization loss is applied to the filter module to constrain its polarized weights. Moreover, the ranking loss from MIR [21] is used.

Background Suppression Motion Enhanced (BSME) [276] further explores the background bias and proposes a background-suppressed and motion-enhanced network. It adopts the same components as RTFM [89], as the global and local temporal FM, to generate 3 attention maps, weighting the anomaly scores into 3 different attention-based sequences. More specifically, a background-suppressed sequence captures background information, a discriminative features sequence accounts only for semantically ambiguous segments in the foreground and background, and an enhanced motion focuses on salient motions in the foreground. The training is performed under the MIL scheme, and the cross-entropy loss targeting the scores is applied over the average of the top scores for each of the three branches. To further align feature representations of FM, the modified feature magnitude ranking loss from [89] is applied over the average score-selected top-K features for all 3 branches.

Normality —. While some methods directly address the background bias problem, exposing the background as a new class, others focus on establishing a robust representation of normality.

CLAWS [87] (Fig. 13) exploits the noise-free nature of normal videos by designing an alternative training data pipeline and self-attention network to beneficially suppress all the normality, both feature- and score-wise. More precisely, input clip-level features are the basic unit to form a batch, inheriting video labels and are randomly selected as train input. Doing so neglects the need for additional Segmentation and possible detour of information. It also augments the number of normal exemplars to train on, complementing the self-attention suppression mechanism.

The proposed network comprises an MLP with 2 dense layers, with 2 additional connections responsible for weighting intermediate features. It proposed Clustering as Supervision to guide the train by grouping the intermediate feature representations into two clusters using K-means clustering. A clustering loss brings the clip's features in a normal batch closer to the centre and far apart in an abnormal batch. The mean square error with sparsity and temporal smoothness as MIR [21] forms the loss parts that guide the prediction. The result is a model that learns to minimize highlights in the attention map, so normal information in features is suppressed so that low scores are produced, which, in return, high values are considered anomalies. CLAWS+ [92] improves upon this approach by using a cluster compactness loss to reduce intra-cluster variation and 3DResNet [36] as an additional FE.

Cross Evaluation Learning (XEL) [90] addresses the imbalanced data issue in WVAD datasets by using a Hard Instance Bank (HIB) to collect the segments with the highest scores from normal videos and updating the HIB by reiterating all normal features through the network at each epoch end. Two loss functions are applied: a validation loss to penalize HIB scores and a dynamic margin loss to augment the gap between the maximum segment-level scores of anomalous bags and HIB scores.

Normality Guided MIL (NGMIL) [96] leverages a memory bank of compact normality prototypes, computed using all normal segments across normal videos, to generate an attention map based on the cosine similarity between each segment and the prototypes. By applying an inverse version of this map over the scores, high scores are attributed to abnormal temporal instances, refining the anomalous scores. The

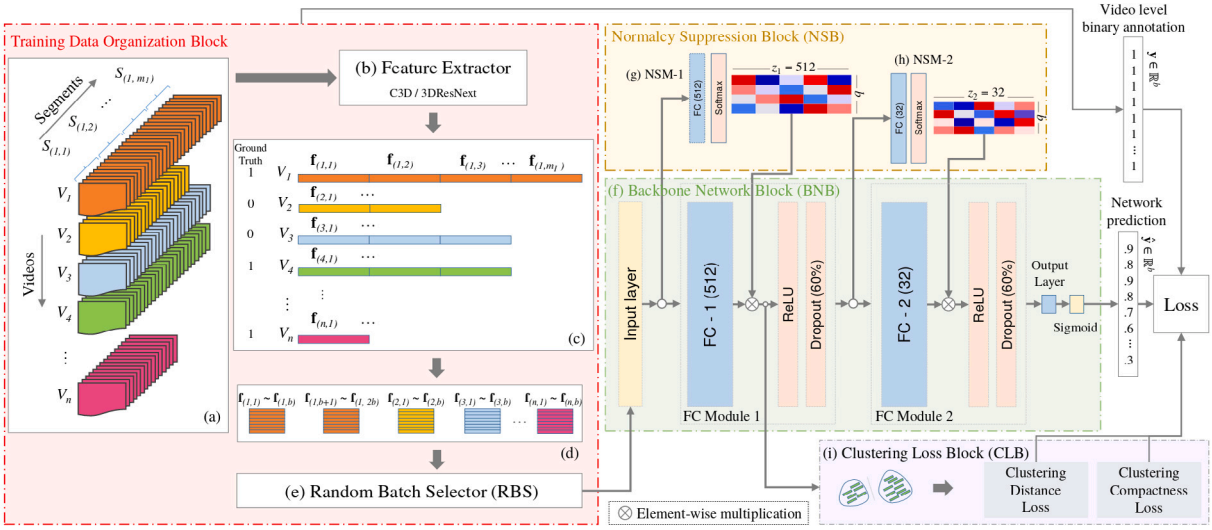


Fig. 13. CLAWS [87,92] proposed method.

training uses a Normality Clustering Loss, which brings each instance in the negative bag closer to the normal prototypes, and a triplet loss on the top-K and bottom-K representations based on refined scores to enhance positive-bag inter-class separability.

URDMU [278] (Fig. 14) extends the concept of memory banks to modulate both normal and abnormal events. As the FM, URDMU adopts MHSA with a temporal mask and extends memory banks to dual units (normal/
abnormal), where attention-driven combinations explicitly regulate uncertainty: normal inputs anchor certainty in the normal bank while suppressing abnormal prototypes; anomalous inputs conversely activate abnormal prototypes (label 1 for topk) and partially disrupting normal patterns (label 1 for inverted topk). A triplet loss further sharpens this divergence. Taking knowledge from [89], that normal and abnormal features are in opposite levels of magnitude, it is further observed that normal features exhibit fluctuations due to noises coming from camera switching, subject changing, etc. [119]. To this intent, it adapted a Normal data Uncertainty Learning (NUL), by assuming normal patterns follow a Gaussian distribution, learning both mean feature and uncertainty variance to generate a latent space of normal variance, helping with unknown normal cases, as well as augmenting abnormal robustness by enlarging the magnitude distance of its mean in the normal latent space.

BNDFM [279] (Fig. 15) leverages the mean vector of the BatchNorm layer as a strong statistical reference of normality to introduce the Divergence of Feature from Mean (DFM) as a novel anomaly metric. DFM is calculated as the Mahalanobis distance [280] between features of hidden space and the normality prototype. As most of the segments within a batch are normal, the mean vector inevitably captures the feature distribution of normality. As a result, characteristics from abnormal segments will diverge from the mean vector, enabling the formulation of DFM as an efficient selection metric. The notion of the batch is further explored by adopting Sample-Batch Selection (SBS), batch-wise for abnormal segments and video-wise for normal instances, with the number of selected segments regulated according to the abnormal length distributions of each dataset. This approach enlarges the abnormal selection pool to contain all abnormal videos within a batch while maintaining modularity and coherence with the dataset. The same structure as URDMU [278] is used for the global and temporal FM, and the BatchNorm incorporates the Conv-based regressor in its intermediate layers. During optimization, the mean vector is an anchor for a triplet loss, encouraging the divergence of the top-K-selected abnormal segments and pulling the top-K normal segments closer to their anchor. Since the Score Head trains on certain normal data, it eliminates label noise from abnormal videos. It enables segment-level score supervision by minimizing the magnitude of all normal scores within each batch. During inference, the Score Head predictions are enhanced with the DFM metric to form the anomaly scores, which have been proved to exhibit increased robustness to label noise since they are acquired in dense feature space. Considering both test sets, a detailed ablation study proved the complementary nature of the DFM metric and the triplet loss in yielding a discriminative hidden space. Audio features were concatenated for the XDV dataset, leaving room for further study.

Pro Discriminability VAD (ProDisc-VAD) [281] in an exceptional parameter efficiency method combines two novel components for CLIP's feature discriminability. The Prototype Interaction Layer (PIL) models normality using a small set of learnable prototypes, which interact with video features via attention to establish a robust normality baseline while preventing dominance by normal data. The Pseudo-Instance Discriminative Enhancement (PIDE) loss enhances feature separability by applying supervised contrastive learning exclusively to extreme-scoring segments (rgmax/argmin anomalies/normal), leveraging reliable pseudo-labels to mitigate noise. The total loss also integrates the video-level classification MIL loss for weak supervision, optimizing discriminative feature learning. Remarkably, ProDisc-VAD achieves competitive results on ShanghaiTech (ST) and UCFC using only 0.4M parameters — over 800× fewer than ViT-based methods like VadCLIP — while maintaining a lightweight model size (1.7 MB) and fast inference (0.0009s per video). This efficiency stems from its focused design, avoiding complex architectures while synergizing prototype-driven normality modelling and targeted contrastive learning, making it a practical solution for real-world surveillance applications.

Table 7 summarizes the works exploring Background and Normality.

Key Points of Background and Normality Methods: These methods highlight the importance of addressing background bias and modelling normality in VAD.

- **MSBS:** Employs a two-branch architecture with a filter module to suppress background information.
- **BSME:** Utilizes a three-branch network to separate background, discriminative features, and motion information.
- **CLAWS and CLAWS+:** Introduce Random Batch Selector as a novel training pipeline and clustering-based supervision to suppress normal features and scores.
- **XEL:** Employs a hard instance bank to refine normal event representation with a Validation loss and uses a dynamic margin loss to separate normal and abnormal scores.

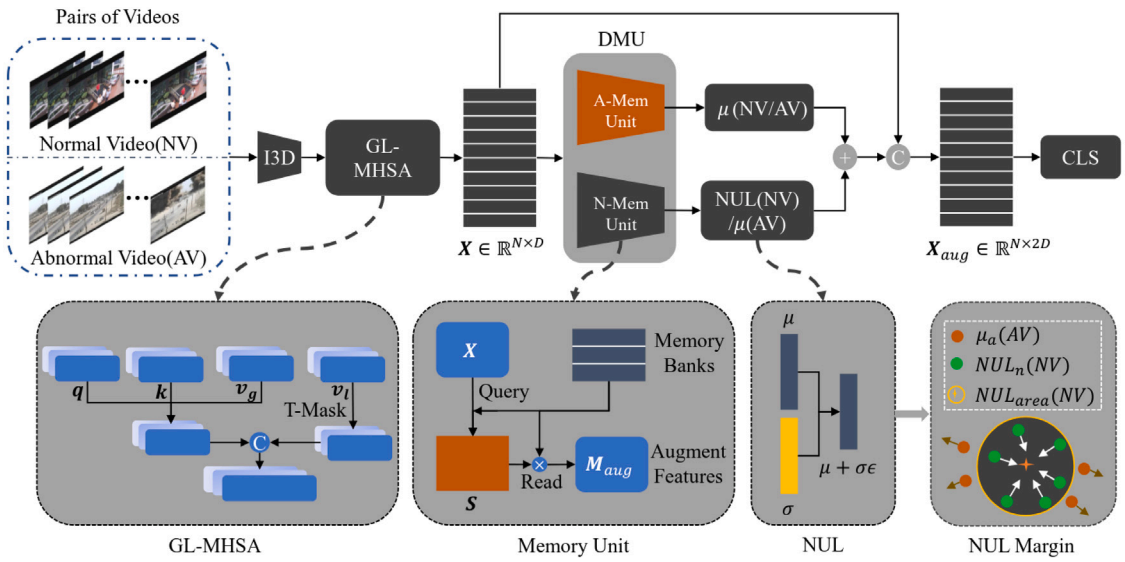


Fig. 14. URDMU [278] proposed method.

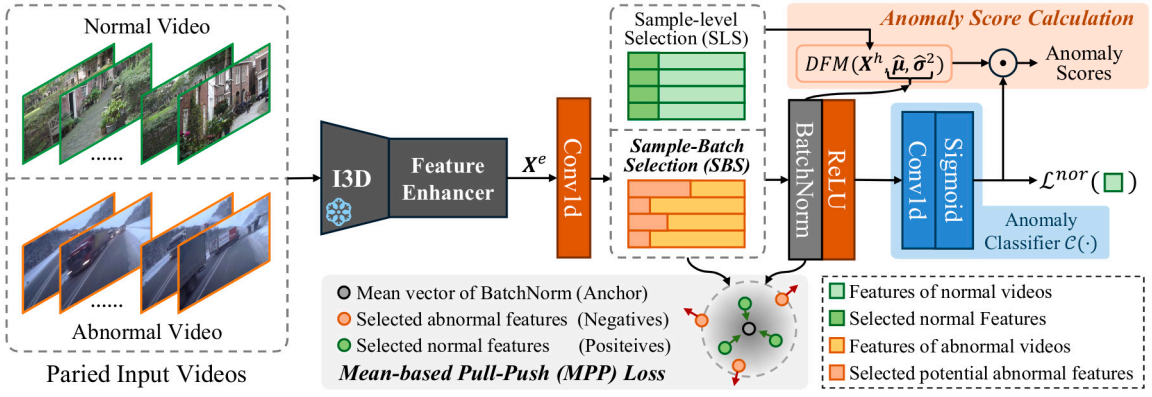


Fig. 15. BNDFM [279] proposed method.

Table 7
Summary of methods that explore Background and Normality.

Method	FE	FM	MIL	SS	LT		Metrics	
					FL	SL	UCFC	XDV
MSBS	I3D	PCD+NL& Filter Module	✓	Top-K Scores	✓	✓	83.53	✗
BSME	I3D	PCD+NL & Conv1D	✓	Top-K Scores	✓	✓	83.63	✗
XEL	C3D	✗	✓	✗	✓	✓	82.60	✗
CLAWS	C3D 3DRN	✗	✗	Batch Scores	✓	✓	80.94 81.27	✗
CLAWS+	C3D 3DRN	✗	✗	K-Means	✓	✓	83.37/–/0.11 84.16/–/0.09	✗
NGMIL	C3D I3D	✗	✓	Top-K Scores	✓	✓	83.43 85.63	75.91 78.51
URDMU	I3D (+VGG)	MHSA	✓	Top-K Scores & Mem. Scores	✓	✓	86.97/–/1.05	81.66/–/0.65 81.77
BNDFM	I3D (+VGG)	Conv1D & MHSA	✓	Top-K DFM	✓	✓	87.24/71.71/–	84.93/85.45/– 85.26
ProDisc-VAD	CLIP	Prototype Att	✓	Max&Min	✓	✓	87.12	✗

FE: Feature Extractor; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

- **NGMIL:** Leverages a memory bank of normality prototypes and a normality-guided Top-K SS strategy. Employs Clustering, Triplet and Ranking Loss, targeting both scores and features.
- **URDMU:** Utilizes memory modules to modulate both normal and abnormal events and incorporates normal data uncertainty learning.
- **BNDFM:** Introduces the DFM metric based on the BatchNorm layer's mean vector and employs dynamic SS guided by a triplet loss.
- **ProDisc-VAD:** Combines a Prototype Interaction Layer (PIL) and a Pseudo-Instance Discriminative Enhancement (PIDE) loss to enhance feature space discriminability of both Score Head and Feature Modulator.

The exploration of background bias and normality modelling has led to diverse approaches, ranging from background suppression to memory-based normality representation. Although memory modules are expensive and prone to weak generalizability.

On the other hand, those who shine leverage inner network embeddings and batch-level processing to model normality and detect anomalies. BNDFM introduces the novel DFM anomaly metric based on BatchNorm statistics, while CLAWS learns to suppress normal information within the feature embedding space. Both methods employ pair-based loss functions and highlight the potential of batch-level analysis for anomaly detection. Notably, BNDFM's SBS strategy overcomes limitations of traditional static SS methods.

5.5. 2-Stage & label noise

Another group of methods employs a two-stage training scheme to directly address the limitations of weak supervision and the unreliability of scores to guide selection in early epochs. These methods generate confident pseudo labels in the first stage, enabling training with reduced label noise in the second stage. Despite explicitly employing pseudo-label generation, these methods can still mine various forms of AC.

GCLNC [84] treats WVAD as a supervised learning task under one-sided label noise and proposes a two-branch Graph Convolutional Network (GCN), leveraging feature similarity and temporal consistency of input features to clean label noise for iterative model training. Both C3D [24] architecture, pre-trained on the Sports-1M [25], and Temporal Segment Network (TSN) [34,35], pre-trained on Kinetics-400 [33] are used a FE. However, training both a GCN and MIL is computationally expensive and progresses slowly.

A **Self-Reasoning Framework (SRF)** is proposed by [86], followed the noisy labelling stated in [84] and further using a K-means clustering algorithm to generate binary pseudo-labels to aid the training of a regression network, supervised by clustering distance and cross-entropy losses.

Multiple Instance Self-Training (MIST) [88] adopts the Segmentation strategy of TCA [95] and trains a similar model as MIR [21] in the first stage. Then, under the supervision of generated anomalous pseudo-labels, a self-attention module and an additional Score Head are trained to minimize the cross-entropy loss function.

CAVAD [91] leverages the dominance of normalcy in WVAD and assumes the classifier's reliability in learning normal features. In the first stage, a temporal Graph Convolution Network captures the temporal context of input features. It generates a fused attention-weighted feature to update parameters in both the CA module and classifier with only video-level labels. In the second stage, similar to GANs, the attention block is fine-tuned to recognize more anomalous segments using video-level features based on the inverted attention map, tricking the classifier into predicting normal videos. Attention consistency loss aligns attention weights with classifier segment scores, removing misclassified normal segments. The two loss functions are executed alternately using two optimizers to update different parameters in each training iteration, first training the whole network, including the contrastive attention module and classifier with video-level features and labels, and then using the converted video-level features to refine the attention module through the classifier.

In a **Self-Training MSL**, [94] (Fig. 16) introduced a sequence selection strategy by changing the [21] ranking objective to choose the sequence with the highest mean of anomaly scores.

In the first stage, video-level labels are used as initial snippet-level pseudo-labels to select sequences, and the model is optimized through the hinge-based MSL ranking loss. In the second stage, scores are directly used to select sequences, and the model is further optimized. The two-stage self-training strategy starts from a sequence length equal to the number of input segments. It gradually reduces the length by half to progressively improve its ability to detect anomalies at a finer granularity. MSL proposes a MHSA with a DepthWise Separable 1D Convolution (DW Conv1D) [52] as the linear projection, serving as the building blocks for both the FM and score head. As the FE is used as the transformer-based backbone, VSwin [113] is pre-trained on Kinetics-400 [33].

The work of [282] proposes a **self-guiding MIR**-based method (SG-MIR), clustering abnormal bags into two using the K-means algorithm. The cluster module generates pseudo-labels based on their relative distance to the cluster centre, guiding the training of a bidirectional Recurrent Neural Network (RNN) regression module. A clustering loss is introduced to optimize the feature module, encouraging negative bag clustering centres to be close while keeping the boundary between clusters in positive bags distinct.

In a **Completeness and Uncertainty Network (CUN)** [283], it is explored the pseudo-labels generation process of MIST and MSL, to conclude it ignores both completeness of abnormal events, as positive bag may contain multiple abnormal clips, and the uncertainty of generated pseudo labels in the second stage, leading to a gradually deviating self-training process guided by noisy pseudo labels. To this end, in the first stage to tackle completeness, a parallel multi-regression head network is trained under MIL scheme with ranking loss with a ranking loss and a diversity loss, which minimizes the cosine similarity of the distribution between any two heads to enforce the predicted score to be distinct from each other. In the second stage, an uncertainty estimation leveraging Monte Carlo Dropout [284] is used to discard low-confident samples as pseudo-segment-level labels in optimizing the regressor.

An **Unbiased-MIL (UMIL)** framework is proposed in [195], where VAD is streamlined through an end-to-end process that involves fine-tuning the FE and regressor training. Leveraging a random augmented sample selection method, UMIL deals with videos in their raw frame format, which allows for the incorporation of self-training via data augmentation from FixMatch [285]. In the first stage, the classifier builds on the fine-tuned X-CLIP-B/32 model [286] from Kinectics-400 [33], is pre-trained using only video-level labels under the MIL scheme. A prediction history for each snippet informs the formation of a confidence set. Any remaining snippets form an ambiguous set, further divided into two clusters using a pairwise-trained cluster head. These clusters guide the final classifier in optimizing anomaly prediction. Extensive ablation studies were conducted on both UCFC and TAD datasets, focusing on abnormal class-wise metrics as introduced by the authors in [111].

In a **Long-Short Temporal Co-Teaching (LSTC)** framework, [97] employs two tubeless-based spatiotemporal Transformer networks to learn from short and long-term clips, mining both fine-grained spatial features and global temporal dependencies. Tubelets, or patch clips, are embedded as tokens, passed through multiple Transformer layers, and a regression layer to predict anomaly scores for input clips. The co-teaching strategy [287] is introduced to enhance both networks' training while being more robust to label noise compared to self-training strategies [84,88,94]. In the first round, the short-term network takes short clip sequence features to generate pseudo labels by training under the supervision of the MIL ranking loss. The long-term

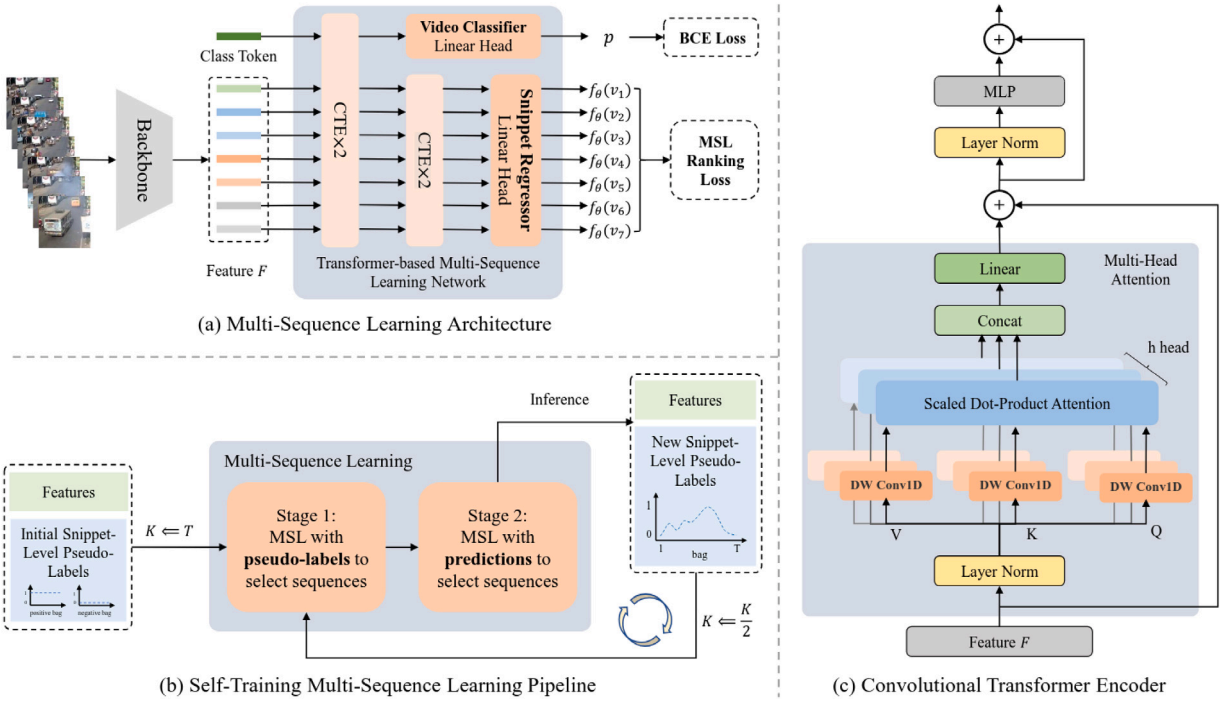


Fig. 16. Self-Training MSL [94] proposed method.

Table 8
A summary of works that employ a 2-stage approach.

Method	FE	FM	MIL	LT		Metrics	
				FL	SL	UCFC	XDV
GCLNC	TSN	Graph Conv	✓	✗	✓	82.12	✗
SRF	C3D	✗	✓	✓	✓	79.54/–/0.13	✗
MIST	C3D I3D	SA	✓	✗	✓	81.40/–/2.19 82.30/–/0.13	✗
CA	I3D	Graph Conv	✗	✗	✓	84.62	79.60
MSL	I3D VSWIN	Conv MHSA	✓	✗	✓	85.30 85.62	78.28 78.59
SGMIR	I3D	Dense MLP	✓	✓	✓	81.70	✗
CUN	I3D (+VGGish)	✗	✓	✗	✓	86.22	78.74 81.43
UMIL	X-CLIP	✗	✓	✓	✓	86.75/68.68/–	✗
LSTC	C3D I3D	MHSA	✓	✗	✓	83.47 85.88	✗

FE: Feature Extractor; **FM**: Feature Modulator; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

network takes the pseudo labels for supervision, adding a cross-entropy loss into the backward pass adjustment process.

Table 8 summarizes works that employ a 2-stage approach.

Key Points of Two-Stage and Label Noise Mitigation Methods: By employing two-stage training schemes, these methods address the limitations of weak supervision and label noise in WVAD.

- **GCLNC:** Uses a graph convolutional network to refine noisy pseudo-labels.
- **SRF:** Generates binary pseudo-labels via clustering.
- **MIST:** Fine-tunes a self-attention module using pseudo-labels.
- **CAVAD:** Employs a two-stage adversarial approach to emphasize anomalous attention regions.
- **MSL:** Introduces a self-training strategy on sequences with a convolutional Transformer encoder.
- **SGMIR:** Extends MIL ranking with a self-guiding clustering loss.
- **CUN:** Tackles incompleteness and uncertainty in pseudo-labels using multi-head prediction and Monte Carlo dropout.
- **UMIL:** Develops an unbiased end-to-end approach with random augmented sampling and cluster-based self-training.
- **LSTC:** Uses two Transformer networks for short and long-range clips with a co-teaching strategy for robustness to label noise.

These methods showcase various techniques for generating and refining pseudo-labels, including clustering, self-attention, adversarial training, and uncertainty estimation. The two-stage paradigm, with its initial pseudo-labelling followed by semi-supervised training, has proven effective in mitigating label noise and improving anomaly detection performance. Although, the overall process requires additional and often complex steps in the training scheme.

5.6. Anomaly suppression & erasure & salience

Another line of thought emphasizes extracting and analysing salient features to improve anomaly detection. These methods often incorporate an erasure process, which paradoxically involves removing the most anomalous segments or features identified during the initial selection process. The rationale behind this approach is that the network must focus on less conspicuous regions of the input features by deleting the most obvious anomalies. This may reveal subtler anomalies initially overshadowed by more prominent ones, leading to a more comprehensive understanding of the video content.

Suppression —. The *Segment Anomaly Attention (SAA)* [288] shifts attention from obvious to subtle anomalies, blending components from RTFM [89] for feature modulation with a convolutional MLP to generate a soft, segment-level attention sequence. During the SS process, the attention sequence is leveraged to provide a threshold to suppress the most attentive parts, both raw scores and attention-weighted versions. The approach employs Multi-Branch Supervision (MBS), which uses extended video labels to supervise the original and suppressed versions. It implements constraints on video-level supervision (using Binary Cross Entropy (BCE) loss overall score versions) and the attention mechanism. For the latter, the distribution of the anomalous attention is encouraged to align with the final anomalous scores. The attention sequence is minimized for normal videos, while a threshold step adjusts the similarity between the attention and score sequences for anomalous videos. After reaching the threshold, the similarity is calculated against a binary version of the scores. This adjustment allows the model to gradually shift its attention from obvious anomalies to less obvious ones throughout training. A normalization loss is introduced to account for the sparsity of video anomalies. This drives the model to pay more attention to the few anomalous segments in a video, rather than spreading its attention evenly across all segments.

SAA stands out for its significant impact on the performance of the XDV dataset compared to UCFC. It handles long anomalous events and fast movements/actions, outperforming URDMU [278]. Its method of co-interacting segment attention and scores during inference aligns with the conclusions of [94,274,279]. Despite its reliance on the accuracy of regressor predictions and its primary focus on scores in its loss function, SAA still offers room for potential improvements in supervising the embedding space. The method also leverages audio signals, enhancing performance in scenes without visual cues for anomaly detection. However, careful consideration is needed due to the potential for noise introduction.

Erasure —. In a **DEN** [289] (Fig. 17) presents a dynamic erasing process to mitigate potential biases in selecting anomalous segments in MIL-based methods, accounting for large variability within abnormal videos. The method utilizes a multiscale temporal modelling module (MSTM) as their FM to identify events of varying durations. This module, inspired by RTFM [89,290], uses multiple 1D convolutional layers with different strides and kernel sizes, along with a positional matrix, to learn multiscale global-aware local representations.

DEN's dynamic erasing strategy is guided by a proposed completeness metric. This metric measures the completeness of detection based on the segment similarity of an abnormal video and the sum of all normal videos in a batch. The segment-level features with the highest and lowest anomaly scores are selected to calculate the similarity in each video. If the calculated completeness is larger than zero, the video is considered to contain no further abnormal segments, and the erasure operation is not performed. Otherwise, prominent abnormal segments are erased. The erased features are treated as augmented abnormal video features and re-integrated into the network, compelling the model to extend its anomaly detection beyond the most prominent segments.

The loss function of DEN includes the hinge-based MIL ranking loss to encourage higher anomaly scores in anomalous video segments than in normal ones. Additionally, a local variation term is included to favour a larger feature-wise variation in abnormal videos compared to normal ones. The overall loss function balances the unerased and erased loss with hyperparameters. Compared to SAA, whose suppress mechanism targets scores, the erasure is applied directly to features and processed again by the model.

While DEN's dynamic erasing strategy improves the completeness of detected anomalies, selecting features for calculating similarity is based on regressor scores, which may not be entirely accurate, and the erasure threshold is a constant. Nevertheless, the method achieves competitive performance on both UCFC and XDV datasets while proving to be a more robust suppression mechanism than SAA. The proposed FM showed better performance than the widely used PDC and NL from RTFM, with no parameters count or gflops/macs given.

It also incorporates audio representations, yielding a more significant impact than SAA. However, as the audio representations are merely concatenated, the method may not fully leverage the potential benefits of this additional modality.

Salience —. Another particular work focuses on extracting salient features efficiently, relying solely on the video labels to guide anomaly detection through a mechanism of ANMIL [291]. It investigated the impact of the FM on capturing the temporal relationships of features by randomly rearranging the temporal indexes of input features, using both MIR [21] and RTFM [89] methods as a baseline. The results indicated that capturing the temporal order between segments does not contribute to the accuracy of the anomaly detector. The observation led the authors to hypothesize that networks have a mechanism that can extract salient features over the entire feature map.

Building on this insight, ANMIL extends the network design from CLAWS [87,92] by incorporating a lightweight network with a self-attention mechanism, vastly reducing the parameter count to 1.3% compared to RTFM. This mechanism aggregates spatial features over the temporal dimension, enabling it to handle variable-length inputs and making it apt for real-time applications. In contrast to CLAWS, which is trained to suppress features belonging to normal data, ANMIL's attention maps are trained to produce strong video representations by emphasizing salient features over the entire feature map. Fig. 18 depicts the proposed network.

For the XDV dataset, ANMIL employs a bidirectional LSTM (bd-LSTM) to obtain compact input features since the original network underperformed for both modalities. The results obtained with this framework provide insights into the role of temporal modulation in various approaches, particularly regarding the benefit of global modelling across the whole segmented video. Even if the loss only targets video classification during optimization, good results were attained, especially on XDV, hinting once more for its power [274].

Comparatively, both SAA and ANMIL proposed networks can be viewed as extensions of the self-attention mechanism introduced in CLAWS, but applied to different anomaly criteria. While CLAWS suppresses normality, SAA shifts attention from the most obvious anomalies to less conspicuous ones, and ANMIL enhances the salience of video representations. Both CLAWS and ANMIL cleverly utilize only the video labels during training. CLAWS expands normal exemplars via its Random Batch Selector, while ANMIL focuses on creating strong video-level representations. However, CLAWS could benefit from integrating an erasure process similar to that in DEN to suppress normal information on abnormal videos better.

Table 9 summarizes the identified works.

Key Points of Anomaly Suppression, Erasure, and Salience Methods: These methods emphasize the importance of extracting and analysing salient features for anomaly detection.

- **SAA:** Employs a suppression strategy based on a learned attention sequence, gradually shifting attention from obvious to less conspicuous anomalies.

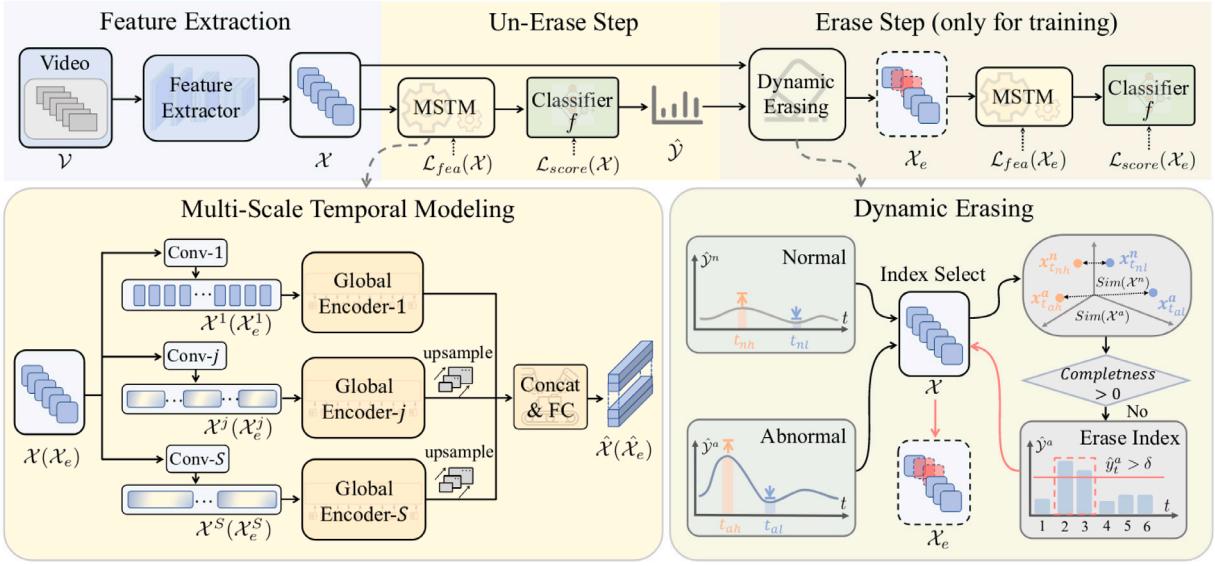


Fig. 17. DEN [289] proposed framework.

Table 9
Summary of works that focus on Anomaly Suppression, Erasure, and Salience.

Method	FE	FM	MIL	SS	LT		Metrics	
					FL	SL	UCFC	XDV
ANMIL	I3D	✗	✗	✗	✗	✗	82.99	84.91
SAA	I3D (+VGGish)	PDC+NL	✓	✗	✗	✓	86.19/68.77	83.59/84.19 84.23
DEN	I3D (+VGGish)	MSTM	✓	✗	✓	✓	86.33	81.66 83.13

FE: Feature Extractor; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

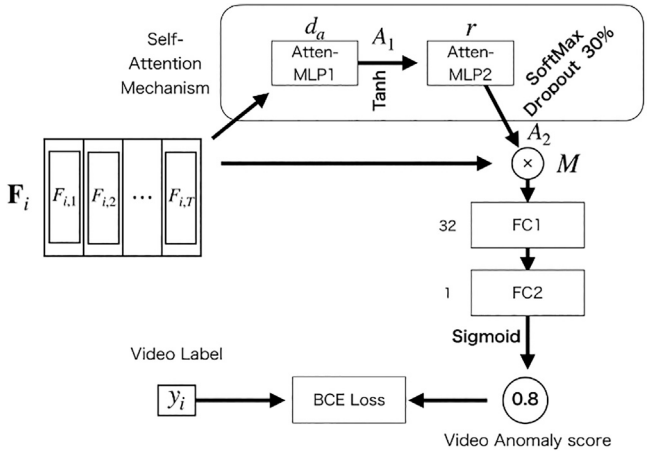


Fig. 18. ANMIL [291] proposed network.

- **DEN:** Introduces a dynamic erasing process guided by a completeness metric, forcing the model to attend to subtler anomalies.
- **ANMIL:** Leverages a self-attention mechanism to enhance the salience of video representations, relying solely on video labels for training.

These methods offer practical and easy-to-implement solutions, highlighting the potential of erasure and salience as anomaly criteria. They provide unique insights into the role of attention mechanisms and feature modulation in detecting prominent and subtle anomalies. Notably, SAA and DEN demonstrate the effectiveness of incorporating audio information, while ANMIL showcases the power of lightweight, self-attention-based architectures for efficient anomaly detection. Further research in this area could explore more sophisticated erasure and suppression strategies, the integration of multi-modal information, and the alignment of hidden feature representations with output anomaly scores.

5.7. Multi-modal

In a Multi-modal setup, different approaches explored ways of combining visual, audio, and sometimes language data to optimize detection performance. Some methods focus solely on multi-modal data, while others focus on a specific dataset, such as the XDV dataset, with little attention given to fusion methods.

5.7.1. Vision–audio

Holistic-Localized Network (HLN) [22] proposes a three-branch graph convolutional network (GCN) formed by a holistic branch capturing long-range dependencies, a localized branch modelling short-range interactions, and a dynamic score branch. A Holistic and Localized Cue (HLC) approximator is introduced to enable online inference. The dynamic score branch computes a weighted sum of features based on predicted scores, while the HLC approximator uses previous video segments to generate predictions guided by HLN. The concatenated representations are projected to the label space, and the average of top-K selected scores is input to the BCE function, denoted as a video classification loss. The overall cost includes a KD loss to encourage the HLC to approximate the HLN output. It also introduced the XDV dataset, whose audio features were extracted using a pre-trained VGG, and video features with I3D. All the following works rely on those as

input features. The authors also released I3D features for UCFC, which enable standardized ablation studies.

Audio-Visual Fusion (AVF) [292] explores three ways to combine audio and visual signals. An attention module comprises co-attention and two self-attention mechanisms to extract multi-modal and uni-modal dependencies. These dependencies are further concatenated to form enhanced visual and audio features. A fusion module is proposed, which applies an element-wise product between enhanced features, followed by power normalization. It is noted that using the L2-norm destroys the correlation between generated features at each time step, thus not recommended. An additional mutual learning module performs self-attention only over input visual features. An MSE loss is applied over top-K scores with video labels to optimize Fusion and Mutual Learning modules. To align the output from both modules, the whole sequence is used into an MSE loss, thus prioritizing the visual information.

Cross-Modal Awareness-Local Arousal (CMALA) [293] proposes a multi-modal fusion mechanism to enhance visual information by computing cross-attention regarding audio. Modalities are temporally aligned by adding a self-adaptive position before the global attention map without affecting the original feature distribution. Such a mechanism enables the capture of local context while suppressing channel redundancy due to the properties of a Gaussian-like kernel function. It is both lightweight and flexible for variable-length videos, taking inspiration from [294], an audiovisual synchronization work. A temporal causal convolution layer with a kernel size equal to 7 is used as the Regressor network. During training, video labels and top-K scores are used to calculate the loss of the BCE.

Modality-Aware Contrastive Instance Learning with Self-Distillation (MACILSD) [295] (Fig. 19) proposes a novel approach to tackle the challenge of audio–visual asynchrony and semantic mismatch in multi-modal data. By leveraging a two-stream architecture, the model processes visual features through a self-attention block, while cross-modality attention handles audio–visual features. An additional dense layer on both streams functions as the regressor. The SS creates positive and negative pairs of embeddings for both modalities to be input into the InfoNCE loss function [296]. The underlying strategy clusters each input into violent or normal background representations, basing its selection upon visual and audio scores, with bottom-K representing the background present at all times. In contrast, the video-level score moderates if a certain video in a batch will be regarded as abnormal or normal. The video-level score is the average sum of top audio/visual logits. While the abnormal selection is represented by the average of the selected K features, the negative elements (normal and background) encompass the full selected K features. This approach aligns with the argument that audio and visual abnormal instances with diverse positions could be semantically mismatched, such as expressing the beginning and ending of an abnormal event. By conducting average pooling only on the abnormal embeddings, MACILSD ensures that audio and visual representations express event-level semantics, alleviating the noise issue. The overall function is optimized to bring abnormal pairs closer together and push normal/background pairs further apart. Furthermore, having both stream networks with similar architectures enables the infusing of parameters from the visual into the audiovisual network via an exponential moving average strategy [297]. Self-distillation transfers knowledge from the visual module to the audio–visual module, thereby ensuring modality noise reduction and robust modality-agnostic knowledge. The lightweight solution has 0.347M (two-stream net) and 0.678M (MACIL and SD) parameters.

Multimodal Supervise-Attention Enhanced Fusion (MSAF) [298] proposes a MIL-based regressor [21] to aid the train of the multi-modal attention module, which is supervised by a segment-level BCE loss. Unimodal features undergo self-attention and cross-attention before being combined with their concatenated version through a Hadamard product. The 3 tensors are concatenated and then scored using a cascade of MLPs to predict anomaly scores. During training, a binary cross-entropy calculate the error between pseudo-segment-level labels and predictions. The MLP supervisor is trained to enlarge, for both modalities, the margin between top-K scores in different bags, while anomaly features from different modalities are constrained to converge. On top of this, the cross-attended multi-modal feature calculates attention loss with pseudo-clip-level labels directly. An over-designed and unreliable supervision scheme with shallow performance.

HyperVD [299] (Fig. 20) addresses the limitation in existing graph representation-based methods [22,84] which learn in the Euclidean space, despite graph-like data's propensity for a non-Euclidean latent structure [300,301].

It proposes projecting features directly into hyperbolic geometry via the Lorentz model [302], to then reuse a two-branch graph convolutional network as FM, akin to HLN [22], for learning temporal relations and feature similarity. This approach better captures subtle semantic nuances often overlooked in the Euclidean space. Further, HyperVD explores various fusion methods to assert the modality's dimensions, named detour fusion, which employs a 1D convolution MLP to match the dimension of RGB features with audio, inherently prioritizing visual information contribution to the network.

Multi Temporal Dynamic Aggregation (MTDA) [303] models global and local temporal relations using Multi-Head Attention (MHA), two inflated 1D convolutions with different expansion rates PDC, and channel-wise shifting [107] in a three-branch method. Concatenated features are added through a residual connection over original audio-visual concatenated features, and an MLP regressor is supervised using BCE over top-K mean scores at the video level. The combination of various temporal enhancements is shown to complement each other. Notably, the combination of MHA and PDC is highlighted as a key component in the network's performance. It is also noted that 2 heads in MHA is enough to modulate global relationships.

Multi-scale Bottleneck Transformer (MSBT) in [304], inspired by multimodal bottleneck transformer (MBT) [305], integrates three modalities (RGB, optical-flow, audio), where each independently encoded modality shares transformers to generate contextualized embeddings. The MSBT addresses information redundancy and modality imbalance via hierarchical cross-modal fusion: bottleneck tokens iteratively condense information across layers, transmitting features between modality pairs through cross-attention. This asymmetric fusion ensures focused interaction, while a weighting mechanism prioritizes informative fused features using learned bottleneck token relevance. To resolve temporal asynchrony and to take inspiration from [295], a Temporal Consistency Contrast (TCC) loss, based cosine similarity, aligns fused features from different modalities at corresponding timesteps, contrasting them with temporally mismatched pairs. The training combines video-level classification loss with the TCC loss for joint optimization—a global transformer aggregates multimodal context before the final prediction.

In a **Audio–Visual Collaborative Learning (AVCL)** [306] addresses ambiguity by combining audio mutations and visual robustness through two core modules, The Audio–Visual Hard-case Separation (AVHS) module identifies ambiguous boundary regions by detecting abrupt audio score variations (via thresholded temporal differences) and refines them using visual consistency checks to filter false positives. The Multi-modal Mutual Learning (MML) module establishes bidirectional knowledge transfer via Kullback–Leibler divergence between single-modal (visual/audio) and fused multi-modal branches, mitigating modality-specific noise. The architecture employs Graph Convolutional Networks (GCN) to model intra-/inter-modality relationships, constructing cross-modality graphs where nodes represent visual/audio snippets and edges encode local/global dependencies. Optimization combines video classification loss, AVHS margin loss, and MML divergence loss, balanced by trade-off weights, to maximize inter-class separation. When integrated as a plug-in module with MACIL-Self Distillation [295], AVCL enhances anomaly localization by resolving

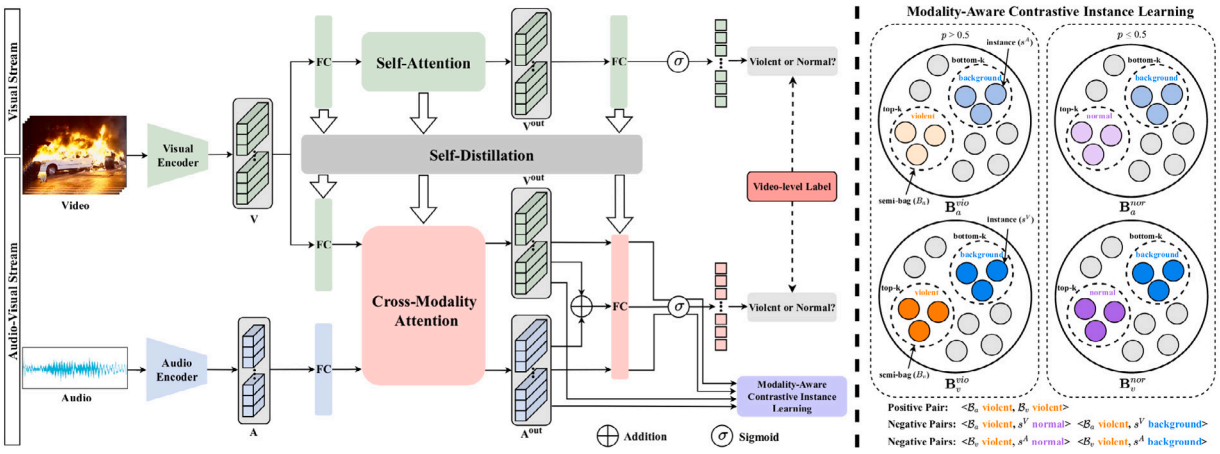


Fig. 19. MACILSD [295] proposed method.

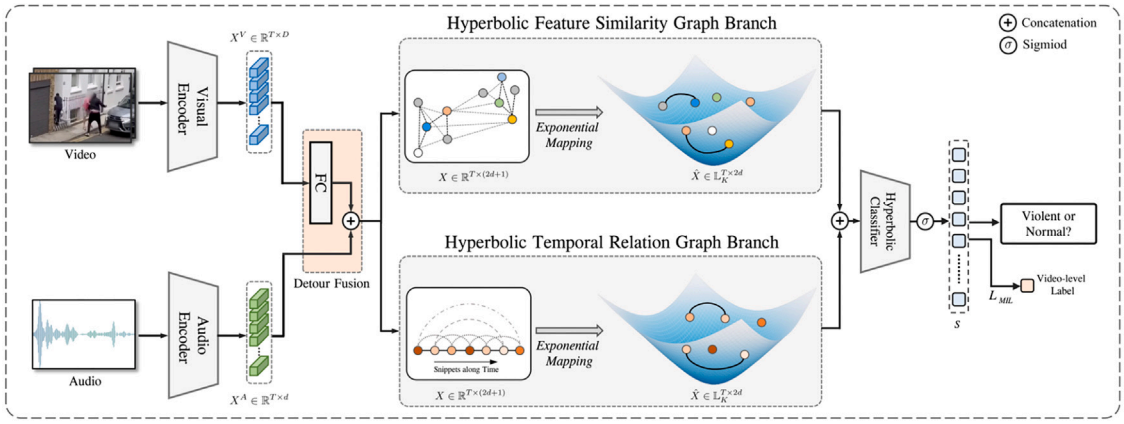


Fig. 20. HyperVD [299] proposed method.

Table 10
Summary of audio–visual works.

Method	FUSE			FM	MIL	SS	LT		Metrics	
	E	M	L				FL	SL	UCFC	XDV
HLN	Concat	✗	✗	Graph Conv	✗	Mean Top-K	✗	✓	82.44	78.64
AVF	✗	Concat & Product	✗	Aud/Vis SAtt & XMod Att & Vis SAtt	✗	Mean Top-K	✗	✓	✗	81.69
CMALA	✗	Att	✗	XMod Att	✗	Mean Top-K	✗	✓	✗	83.54
MACILSD	✗	Att	Concat	Vis Att+ XMod Att	✓	Aud/Vis Bot-K/Top-K	✓	✓	✗	83.40
MSAF	Concat	Att Concat	✗	Aud/Vis SAtt & XMod Att	✓	Aud/Vis Top-K	✗	✓	✗	80.51
HyperVD	Concat	✗	✗	Hyperbolic GraphConv	✗	Mean Top-K	✗	✓	✗	85.67
MTDA	Concat	Sum	✗	MHSA +PDC +TempShift	✗	Mean Top-K	✗	✓	✗	84.44
MSBT	✗	Att & SA Concat	SA	MBT	✓	Mean Top-K	✓	✓	✗	82.54 84.32 (+flow)
AVCL	✗	Graph	✗	GCN	✓	Max&Min	✓	✓	✗	81.11 83.98 (+SD)

FE: Feature Extractor; **FUSE**: **E**arly **M**id Late; Concatenation Summation Attention; **SAtt**: Self-Attention; **XMod**: Cross-modality; **FM**: Feature Modulator; **SS**: Segment Selection; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level.

transient edge cases, synergizing audio–visual cues, showing promises in scenarios with subtle boundary transitions.

Table 10 summarizes the various approaches to audio–visual fusion and anomaly detection.

Key Points of Audio–Visual Methods: These methods explore the integration of audio and visual information for anomaly detection, highlighting the importance of effective fusion strategies and temporal modelling.

- **HLN:** Introduces a three-branch GCN architecture and a KD approach for online inference. Despite its instance-level learning, it can be viewed as a deviation from traditional MIL due to its use of average top-K scores and a video-level BCE loss.
- **AVF:** Explores various fusion methods, including element-wise product and power normalization, and employs a mutual learning module to enhance visual features.
- **CMALA:** Proposes a lightweight cross-attention mechanism to fuse audio and visual information, emphasizing local context and temporal alignment.
- **MACILSD:** Addresses audio–visual asynchrony and semantic mismatch using contrastive learning with the InfoNCE loss and incorporates self-distillation for knowledge transfer.
- **MSAF:** Utilizes a multi-modal attention module supervised by segment-level BCE loss and a MIL-based regressor to guide training.
- **HyperVD:** Projects features into hyperbolic geometry to capture subtle semantic nuances, and explores detour fusion to align modality dimensions.
- **MTDA:** Models global and local temporal relations using a combination of multi-head attention, dilated convolutions, and temporal shifting.
- **MSBT:** Uses a multimodal transformer integrating RGB, optical flow, and audio via shared encoders and hierarchical bottleneck-based cross-modal fusion to reduce redundancy and address balanced modality contributions, while enforcing temporal alignment of features across modalities via contrastive loss (TCC).
- **AVCL:** pioneers audio-guided boundary refinement, leveraging abrupt audio shifts to pinpoint ambiguous anomaly transitions and cross-validating them with visual consistency, while bidirectional modality distillation (via mutual KL divergence) resolves noise-induced conflicts.

The choice of fusion strategy, temporal modelling techniques, and loss formulation significantly impacts performance. MACILSD stands out for effectively handling asynchrony and semantic mismatch, while CMALA offers a lightweight and efficient approach. However, the field faces challenges in evaluating the full XDV dataset, potentially obscuring performance on specific anomaly types. Future research should prioritize exploring alternative audio backbones, refining fusion strategies, and evaluating performance on anomaly-only subsets to gain a more in-depth understanding of audio–visual anomaly detection capabilities.

5.7.2. Vision-Language

Integrating Vision–Language models into WVAD opens up exciting avenues for research. These models enrich the discriminatory feature space by exploiting semantic relationships between vision and language, enabling a more in-depth understanding of video content. Several innovative methods have emerged, each presenting unique strategies and mechanisms. Those using CLIP-based methods as feature sources always leveraged the ViT-B/16 backbone from [122].

CLIP Temporal Self-Attention (CLIP-TSA) [194] introduces the CLIP model [122] with a ViT backbone, into WVAD. This FE captures visual features from each clip's middle frame. Both temporal FM and optimization functions are the same, used per RTFM [89]. The key innovation is the Temporal Self-Attention (TSA), inspired by [307], which comprises three modules: the temporal scorer network, the top-K score nominator, and the fusion process. First, the temporal scorer network, a 2-layer MLP, transforms input features into a score sequence that is further cloned M times and perturbed through Gaussian noise. This process enables the creation of an empirical mean, forming the basis for the later calculation of the expectation error. Next, the top-K selected indices from each clone perturbed score are one-hot encoded and further averaged to produce a stack of K soft vectors. Finally, from the element-wise multiplication over the vanilla feature results, a stack of K perturbed feature vectors is independently summed up to form the input to FM. The gradient expectation is calculated during the backwards pass, and the network is encouraged to focus on segments with the highest scores. Even if the attention mask over input features effectively discards a small portion of information and guides the network towards a more accurate selection, it does so based on the video-level feature magnitude criterion.

Text-empowered VAD (TeVAD) [308] enhances the RTFM [89] FM to incorporate textual information, capturing both long and short-range temporal dependencies between visual and text features. Unlike CLIP-TSA and UMIL, which primarily focus on visual features, TeVAD explicitly leverages the semantic relationships between vision and language. It utilizes SwinBERT [309], a video captioning model pre-trained on the VATEX [310] dataset, to generate textual descriptions for each video clip. TeVAD aims to improve anomaly detection by capturing a richer understanding of the video content by processing visual and text features through a similar temporal modelling pipeline. During training, the video classification loss from RTFM is applied to both the enhanced visual and textual features, as well as the final anomaly scores.

Prompt-Enhanced Learning for VAD (PEL4VAD) [39] presents an innovative three-component method consisting of a Temporal Context Aggregation module, Prompt-Enhanced Learning (PEL), and a score smoothing process during inference. The TCA module improves upon previous authors' work [293] from audio–visual by using a masking window to obtain locally calibrated features. The two global and local reweighed features are further fused in a balanced way by a learnable parameter. The resultant enhanced feature undergoes dimensionality reduction through a two-layer MLP, while a causal convolution acts as regressor layer. It is optimized by the BCE loss of average top-K scores for abnormal videos and max score for normal videos.

For PEL training, prompt representations are encoded through a pre-trained CLIP model over a specific dataset constructed concept dictionary through ConceptNet [311]. During optimization, anomaly scores activate the inner layer MLP to obtain video-level foreground and background representations. These are then aligned to their corresponding knowledge-based prompt feature. For an abnormal video, background/normal and foreground/abnormal alignments are made, while for a normal video, only the foreground is aligned with the normal prompt. This division strategy is similar to works on Background [275,276], Vanilla [268], Normality [282], where an anomaly prototype is constructed, and particularly the [295] on an audio–vision. However, it extends binary labels to rich context text features, adding a layer of interpretability by feeding anomaly priors through external knowledge. During inference, a sliding window mechanism is implemented to smooth possible peaks in the anomaly scores from the camera jitter and other potential environmental perturbances.

CNN-ViT [98] proposes a blend of backbone structures, leveraging CNN-based (like C3D and I3D), and ViT-encoded visual features from CLIP to close the domain gap of existing pre-trained FE. Similarly to CLIP-TSA, it employs a Temporal Self-Attention module (TSA) to process features before modulation.

Although, does it differently in two ways. First CNN features have dimensions asserted by a low-variance-filter. Then, the proposed top-K nominator bases its feature selection on attention scores calculated by the Mahalanobis distances of vanilla features. The remaining steps to obtain the selected feature map are the same as the original [307]: the generation of top stacked attention masks and its fusion with raw features. To handle the simultaneous presence of CNN and CLIP

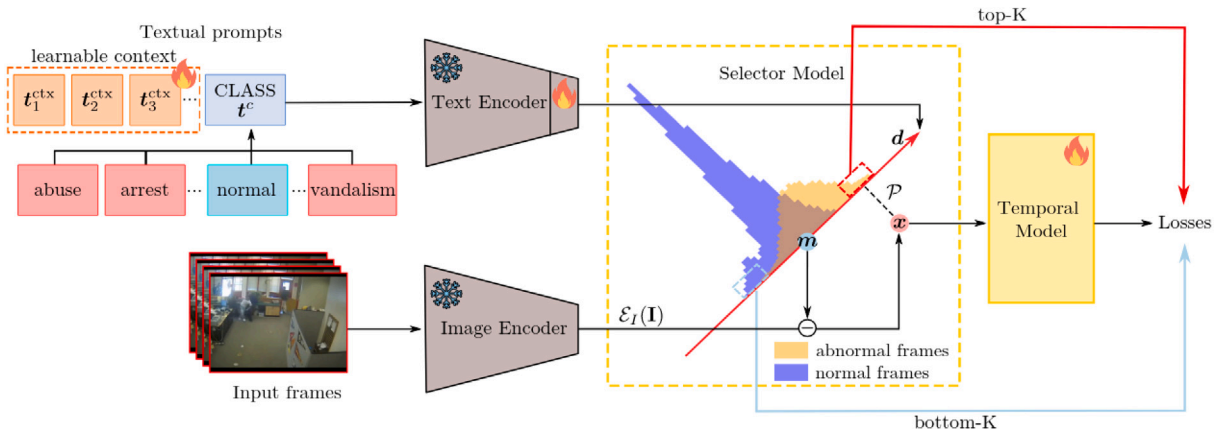


Fig. 21. AnomalyCLIP [196] proposed method.

features, CNN-ViT selects each independently before merging them via addition. This fusion leads to the final enhanced feature map, which undergoes further processing by the PDC and NL network, mirroring the RTFM method. The training is guided by the same magnitude of video classification loss as CLIP-TSA, targeting scores and features.

AnomalyCLIP [196] (Fig. 21) jointly address both VAD and Video Anomaly Recognition (VAR) tasks. An end-to-end approach exploits the vision–text alignment in the CLIP feature space, employing a CoOp prompt learning strategy [312]. This strategy learns a set of directions in the CLIP latent space, calculated by determining the difference between each class's textual prompt embedding and a normality prototype in the proposed Selector module. A sequence of token embeddings and learnable context vectors denotes each class. At the same time, the prototype is pre-computed as the global average of all CLIP encoded frames within regular videos in the dataset. Projected features into each class's learned directions serve as class activations at the frame level. The magnitude of these activations is then summed within a segment to determine the segment-level anomalous class score. To mitigate the unbalanced feature magnitude within different anomalous events, batch normalization is used instead of simple projection, ensuring that regular features remain close to the origin of the usual prototype. The training process follows the top-K MIL schema, guided by the selector's segment-level scores. AnomalyCLIP operates end-to-end, segmenting each video into a fixed number of segments and randomly selecting a set of 16 consecutive frames to form each segment representation. While this setup effectively captures all video information, it necessitates offline operation. To guide training, AnomalyCLIP maximizes each anomalous class's top-K segment scores for both module outputs, simultaneously maximizing the bottom-K average segment scores from the Temporal module and all standard frame scores from the Selector module. The Temporal module, equipped with an Axial Transformer [313], captures both short-term and long-term temporal dependencies within and between video segments, focusing on computational efficiency and robustness against overfitting. The final frame anomaly score for each class is the combination of results from the Selector module and the score from the Temporal module, yielding both abnormal and normal frame scores. Ablation studies conducted with AnomalyCLIP introduced the standard class-wise metrics of video AR into the anomaly detection area. The method showed clear improvement over inspiration methods [136,137] in recognizing anomalies in both UCFC and XDV, while maintaining competitive results in detection only in UCFC and a discrete improvement in XDV compared to even uni-modal methods [119].

Learn suspected Anomalies from event Prompts (LAP) [314] leverage text–visual fusion exploiting prompts by a dynamic dictionary and adaptive thresholds, improving TeVAD's fixed captions, enable broader adaptability in novel scenarios.

The proposed method is composed by three modules: (1) the Feature Synthesis combines visual features from CLIP/I3D with textual descriptions generated by SwinBERT's visual-to-text encoder, aligning them via SimCSE to form enriched spatio-temporal representations, producing corresponding semantic features; (2) the Multi-Prompt Learning (MPL) module aligns synthetic features with prompt-derived anomaly representations, by constructing a sentence-based prompt dictionary (e.g., "A man is shooting"), extracting semantic embeddings through SimCSE and measuring snippet-prompt similarity via an anomaly matrix, then aggregated into an anomaly vector; and (3) the Pseudo Anomaly Labelling (PAL) module assigns pseudo labels using dynamic thresholds, refining localization through semantic-textual alignment. The total loss combines VL MIL, MPL's Triplet loss to separate normal (anchor), low-score abnormal (positive), and high-score (negative) features across videos, enhancing cross-context discrimination, and PAL consistency, improving fine-grained anomaly detection.

LAP achieves good results with plug-and-play compatibility, boosting existing methods, while additionally providing class-wise, cross-dataset and open-set evaluations. Its efficiency stems from prompt-tuning-based adaptation, avoiding full model retraining while leveraging semantic priors for open-set anomaly generalization. Compared to TeVAD, which relies on fixed SwinBERT captions, text as auxiliary input and temporal dependency modelling between RTFM features, LAP proactively integrates semantic prompts, aligns features globally via anomaly vectors, and generates pseudo labels using dynamic thresholding, reducing ambiguity in weakly supervised settings. This approach improves open-set generalization — LAP's AUC drops 0.6% on UCF-Crime compared to TeVAD's 1.2% — and avoids full-model retraining through its modular design, unlike TeVAD's end-to-end pipeline.

VadCLIP [40] (Fig. 22) introduces a dual-branch method that directly utilizes visual features for binary classification while using both visual and textual features for language–image alignment, as opposed to previous CLIP-based methods [194,195]. The Classification Branch (C-Branch) employs a TopK-based binary cross-entropy loss to distill temporal anomaly scores by averaging the highest frame-level predictions. The Alignment Branch (A-Branch) aligns frame features with learnable text prompts (context tokens prepended to class labels, e.g., `learnable context "Fighting"`) and visual prompts (anomaly-focused feature aggregation via segment-level scores) to refine CLIP's text embeddings. Similarities between frame features and prompts are optimized via MIL-Align loss, which selects TopK frame-text matches per class and applies temperature-scaled cross-entropy. A contrastive loss further diversifies prompt embeddings to avoid semantic collapse. As the FM for temporal modelling, the Local–Global Temporal Adapter (LGT-Adapter) combines a windowed transformer (local dependencies) and a lightweight GCN (global dependencies). By employing a dual-branch structure: a classification branch for anomaly scoring and an alignment branch for enhancing text embeddings with

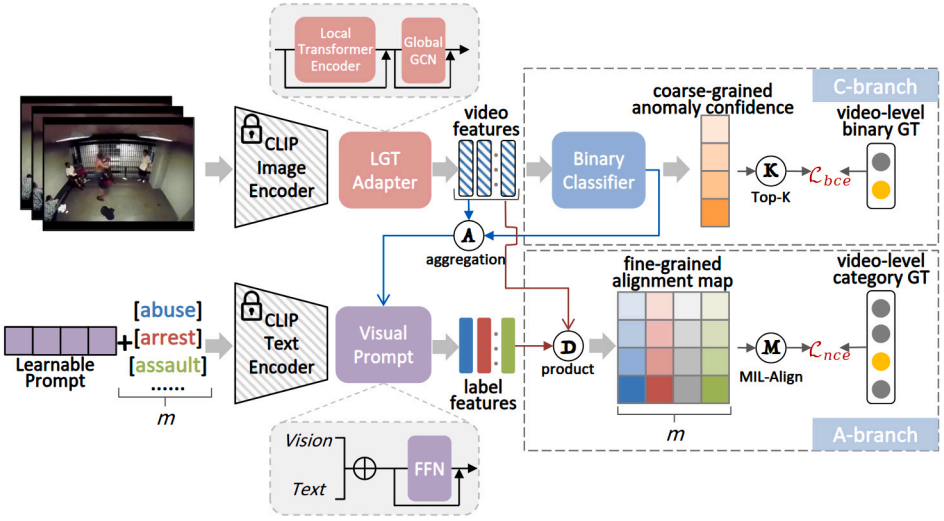


Fig. 22. VadCLIP [40] proposed method.

multimodal prompts, it allows CLIP's representations to be effectively transferred to the VAD task and dynamically refine text embeddings with key multimodal features, strengthening the semantic alignment between video content and corresponding textual labels. For fine-grained evaluation — a pioneering shift from frame-level regression to categorization-aware temporal localization — VadCLIP adopts mean Average Precision (mAP) across IoU thresholds (0.1–0.5), measuring both multi-category classification accuracy (e.g., distinguishing "fighting" vs. "shoplifting") and temporal segment continuity (overlap between predicted and ground-truth anomaly intervals). This metric reflects how well localized anomalies align with their textual semantics, with VadCLIP achieving 24.70% average mAP on XDV, surpassing prior works (e.g., +4.49% over AVVD).

In a **Spatio-Temporal Prompt (STPrompt)** [315] improves upon dual-branch prompt learning strategy by VadCLIP [40] proposing a two-stream architecture: a temporal anomaly detection branch and a spatial anomaly localization branch, for both anomaly detection and localization (WSVADL).

For temporal detection, the method incorporates a spatial attention aggregation (SA^2) mechanism, which calculates motion magnitude by applying L2 normalization to the temporal difference between a patch in the current frame and its adjacent frames, emphasizing abrupt changes in localized regions while suppressing static background noise. This motion-aware weighting prioritizes patches with significant temporal deviations, enabling the model to aggregate spatial features by selectively attending to dynamic regions most likely associated with anomalies. As the FM, a lightweight transformer enhances the global context by modelling relative temporal distances instead of feature similarity, following OVVAD [6]. For joint optimization, STPrompt adopts the VL classification loss, a temperature-scaled MIL-Align loss for fine-grained frame-text matching, and a contrastive loss, the same as VadCLIP.

For spatial localization, the method utilizes LLM-generated text prompts to describe normal (e.g., common background objects) and abnormal (e.g., augmented category phrases) regions, enabling training-free patch-level retrieval through CLIP-based similarity matching. This dual-branch design decouples spatio-temporal modelling, avoiding complex annotations or auxiliary detectors. Evaluations on UCFC, ST, and UBNormal (UB) benchmarks on tasks, achieving high interpretability by localizing anomalies via heatmaps generated from text-patch alignments, demonstrating robustness and efficiency in weakly supervised settings.

Text Prompt with Normality Guidance (TPWNG) [316] integrates learnable text prompts and visual–language alignment to enhance pseudo-label generation. First, the Text and Normality Visual Prompt (NVP) fine-tunes CLIP's [122] text encoder with ranking [21] and distributional inconsistency losses by prepending learnable context vectors to event category labels, enabling domain-specific alignment between textual descriptions and video frames. It also aggregates visual&text prompt features from normal videos using similarity-weighted pooling to create a normality reference, helping suppress interference from normal frames in abnormal videos during text–video alignment. The Pseudo-Label Generation (PLG) module computes and fuses abnormal/normal text-frame similarity scores with a guidance weight, applying a threshold to prioritize regions where abnormal text aligns strongly. To train the classifier, Temporal Context Self-Adaptive Learning (TCSAL) replaces standard Transformer attention with a soft-masked mechanism, dynamically adjusting span lengths via a learnable parameter to focus on relevant time windows (e.g., short spans for abrupt anomalies, longer spans for gradual events). At the same time, a piecewise mask function weights attention scores, ensuring adaptive focus on critical temporal dependencies while ignoring irrelevant frames.

A **Prompt-Enhanced MIL (PE-MIL)** [41] (Fig. 23) processes multimodal (video/audio) features through a Temporal Feature Fusion (TFF) module, which models long- and short-range dependencies via self-attention and dynamic position encoding, following TCA module from PEL4VAD. A Scale-Aware Prediction Head generates multi-scale anomaly scores using causal convolutions and GELU activations. Abnormal-Aware Prompt Learning (APL) integrates semantic priors into visual features to address the insufficiency of binary labels in capturing diverse anomalies. First, event-context separation isolates anomaly-related snippets from their surrounding context using scaled anomaly scores, amplifying high-confidence regions. Next, abnormal-aware prompts are constructed by augmenting textual class annotations (e.g., "explosion") with learnable prompt vectors, which are optimized via a prompt constraint loss to maintain semantic consistency with the original labels. These prompts are encoded into text features using a frozen language model (e.g., CLIP's text encoder). A dynamic cross-modal alignment module then computes fine-grained relevance scores between the isolated event/context visual features and the text prompts, guided by an event relevance reasoning mechanism that enforces semantic coherence (e.g., aligning "fire" with "explosion" contexts). This alignment is optimized via a KL divergence loss, ensuring visual features absorb semantic cues to distinguish diverse anomalies. APL injects class-specific semantics into the model by iteratively refining prompts and their alignment, enabling it to detect varied anomalies without explicit frame-level supervision.

Simultaneously, the Normal Context Prompt (NCP) — learned via a two-stage training strategy — summarizes normal patterns to decouple

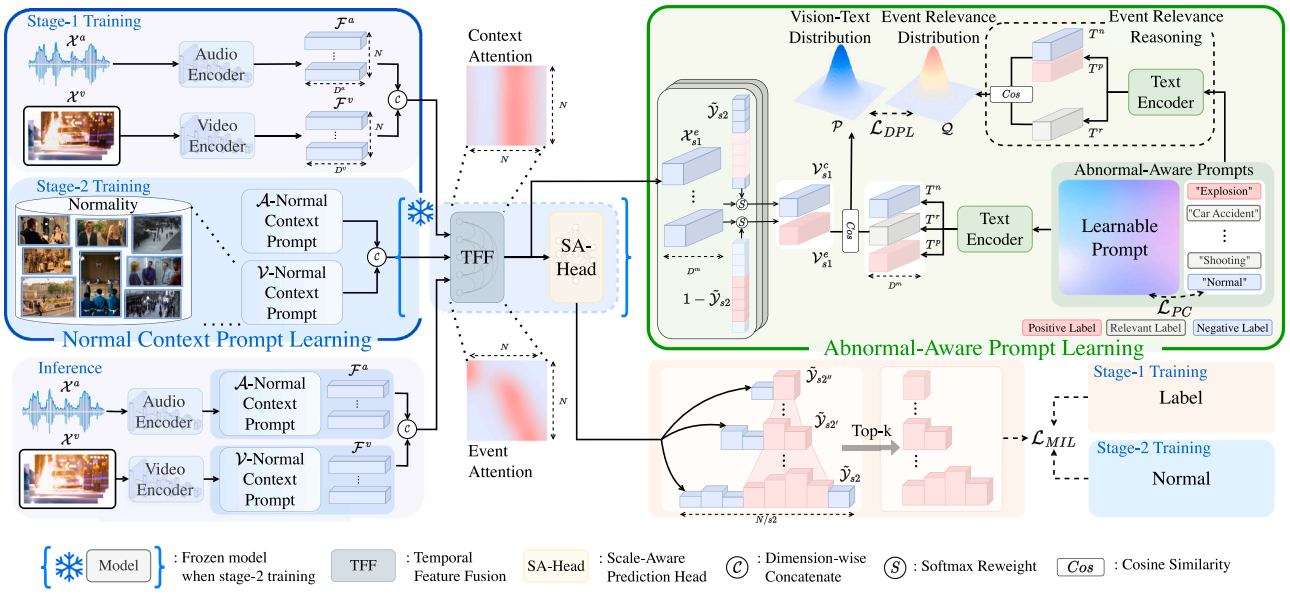


Fig. 23. PE-MIL [41] proposed method.

anomalies from their context, enhancing boundary clarity by enriching ambiguous context features during inference. The training combines video-level MIL loss (binary cross-entropy on top-K scores), APL loss (KL divergence for cross-modal alignment), and prompt constraint loss (semantic consistency), followed by NCP fine-tuning with mean square error. Evaluated on UCFC, ST, and XDV, PE-MIL achieves good performance by leveraging semantic prompts and context refinement to detect diverse anomalies with precise temporal boundaries.

Injecting Text Clues (ITC) [317] (Fig. 24) addresses false alarms and incomplete localization through a dual-branch architecture that integrates text clues for enhanced cross-modal learning. The Text-Guided Anomaly Discovering (TAG) branch leverages a hierarchical matching scheme, combining text–video matching (TVM) to align video-level embeddings (based on threshold segment-level scores) with anomaly-category text queries and text-snippet matching (TSM) to directly measure local snippet-text similarities (weighted trough segment-level scores), enhanced by learnable prompts and CLIP-based embeddings. The segment-level scores (pp) act as a bridge between TSM and TVM, as both strategies leverage the same anomaly-aware scores to filter irrelevant contexts and amplify critical cues, ensuring consistency across granularities.

Complementing this, the Anomaly-Conditioned Text Completion (ATC) branch employs a generative task where masked anomaly descriptions are reconstructed using Transformer blocks with anomaly-conditioned attention, forcing the model to gather event semantics from relevant snippets. A dual-branch mutual learning strategy enforces consistency between the anomaly scores of both branches via MSE loss, ensuring complementary knowledge transfer.

The framework is jointly optimized through a combination of MIL loss, cross-modal matching losses (TVM, TSM), text reconstruction loss (ATR), and mutual learning loss. By integrating discriminative (TAG) and generative (ATC) objectives, ITC reduces false alarms through hierarchical text-visual alignment and improves localization completeness via semantic reconstruction. Evaluations on UCF-Crime and XD-Violence benchmarks demonstrate notable performance, with significant improvements in AUC (e.g., +1.46%) and average mAP (e.g., +2.13%), validating its effectiveness in capturing discriminative anomalies while maintaining temporal integrity.

The **Text-Driven Scene-Decoupled (TDSD)** [318] (Fig. 25) framework addresses weakly supervised video anomaly detection by integrating vision–language models (e.g., CLIP) to extract scene and object semantics, enabling scene-dependent anomaly identification. At its core, TDSD employs a Text-Driven Scene-Decoupled Module (TDSDM), which decomposes scene understanding into two orthogonal components: Context Semantic Injection (CSI) and Object Semantic Injection (OSI). CSI leverages CLIP's text encoder to generate contextual scene descriptions (e.g., "shopping mall corridor") by cross-attending to a predefined set of scene categories (e.g., Places365). These contextual embeddings are fused with spatiotemporal features (extracted via I3D) to enrich scene-aware representations. OSI extends this by generating object-level semantic features (e.g., "motor scooter") using CLIP's zero-shot capability, cross-attending to 1000 object categories from ImageNet-1K. This dual injection ensures the model captures both global scene context and fine-grained object interactions, critical for distinguishing normal vs. anomalous events in scene-specific settings (e.g., detecting a thief in a store). To enhance feature discriminability, TDSD introduces a Fine-Grained Visual Augmentation (FVA) module, which applies spatial–temporal attention to amplify subtle anomalies in local regions. The framework is trained weakly supervised, relying solely on video-level labels (normal/abnormal) without frame-level annotations. Design choices prioritize computational efficiency by reusing pre-trained CLIP weights and avoiding task-specific fine-tuning, while cross-attention mechanisms dynamically align visual and textual features to minimize modality gaps.

MELOW extends VadCLIP's dual-branch framework by integrating Multi-scale Temporal Visual Modelling (MTVM) and specially Multi-modal Evidential Collaborative Learning (MECL), addressing to the Open-World setting. The C-Branch uses CLIP's visual features with MTVM — a Local-Global Temporal Adapter combining windowed transformers (local dependencies) and GCN (global similarity/position relations) — to capture multi-scale temporal contexts, while the A-Branch aligns MTVM-enhanced visual features with CLIP-derived textual embeddings.

The MECL module, under Subjective Logic Theory (SLT) [319], introduces a multimodal Evidential Deep Learning approach to address ambiguity in open-world detection by modelling uncertainty across visual, textual, and joint modalities. For each modality, non-negative evidence scores are computed for anomaly categories, converted into concentration parameters to parameterize a Dirichlet distribution. This distribution quantifies belief masses and uncertainty (reflects the lack of evidence). MECL dynamically prioritizes the most confident modality (visual/textual) per category and fuses joint-modal evidence via a weighted loss, penalizing high uncertainty and aligning predictions

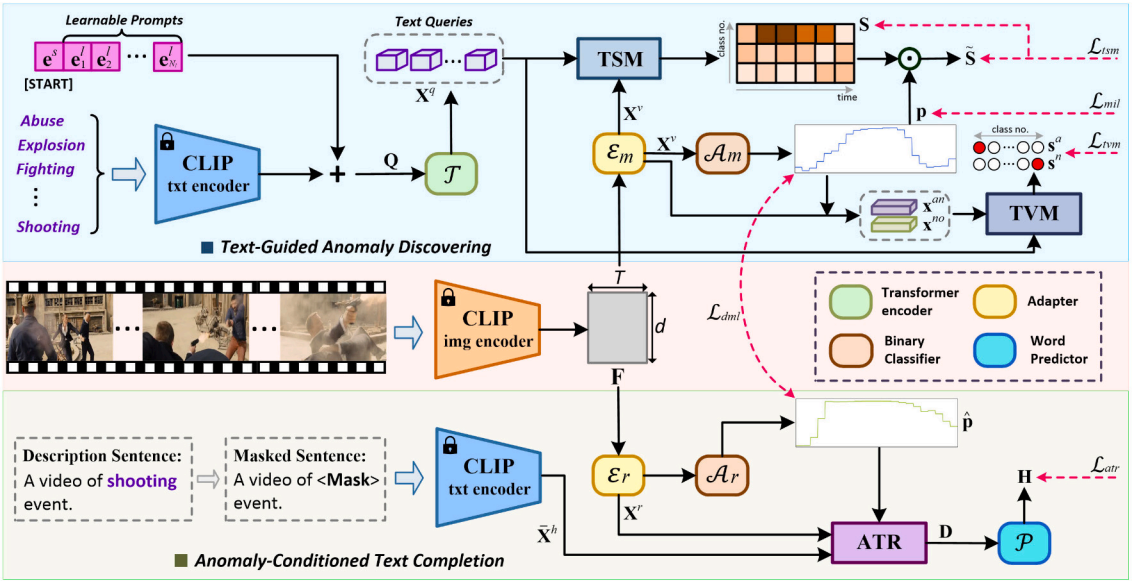


Fig. 24. ITC [317] proposed method.

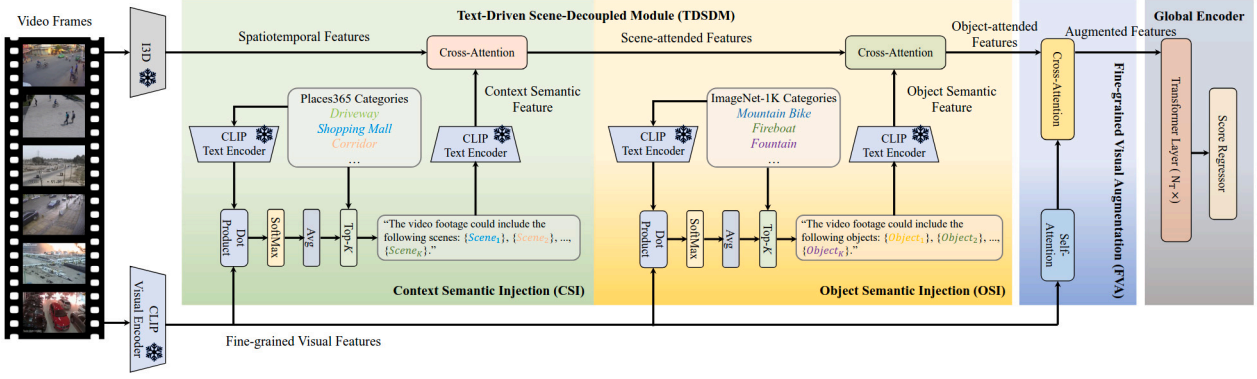


Fig. 25. TDSD [318] proposed method.

with reliable modalities. By treating predictions as subjective opinions rather than deterministic outputs, MECL reduces overconfidence in ambiguous or unseen scenarios, enabling adaptive calibration of anomaly boundaries. This contrasts with prior single-modal EDL methods OpenVAD [4] Section 5.8, as MECL leverages cross-modal correlations (e.g., textual descriptions refine visual uncertainty) to improve robustness in open-world settings.

For optimization, a joint-modality loss dynamically calibrates anomaly boundaries by weighting evidence confidence, alongside VadCLIP's TopK video-level classification loss and MIL-Align. This framework achieves open-world detection by generalizing to unseen anomalies via uncertainty estimation and multi-scale temporal fusion.

Audio–Visual Anomaly Detection with CLIP (AVadCLIP) [42] (Fig. 26) leverages CLIP's cross-modal alignment capabilities in an audio–visual collaboration, building upon VadCLIP dual-branch framework, incorporates three components: (1) an adaptive audio–visual fusion module that dynamically combines visual (CLIP ViT-B/16) and audio (Wav2CLIP) features via lightweight projection networks, prioritizing visual dominance while enhancing with audio cues; (2) an audio–visual prompt mechanism that enriches text embeddings with multimodal context by aligning global video features with class labels, improving semantic specificity; (3) and an Uncertainty-driven KD (UKD) module that models feature uncertainty to synthesize robust unimodal representations when audio is missing. The UKD treats audio–visual fusion features as noisy observations of enhanced visual features from the student model, with additive Gaussian noise. A Gaussian likelihood loss quantifies feature alignment uncertainty, while a three-layer CNN branch predicts variance to weight the distillation loss. This uncertainty-aware formulation minimizes an adaptive MSE term, downweighting high-uncertainty features, avoiding overfitting and improving generalization. It also employs the VadCLIP [40] two-branch framework (video-level classification and categorization alignment branches). Evaluated on XDV and CCTV-Fights, AVadCLIP achieves good performance while maintaining robustness in unimodal scenarios.

Table 11 summarizes the works employing Vision Language models.

Key Points of Vision–Language Methods: These methods leverage the power of Vision–Language models to enhance anomaly detection by incorporating semantic information from text.

- **CLIP-TSA:** Introduces a Temporal Self-Attention (TSA) module guided by feature magnitude to refine feature representations from CLIP.
- **TeVAD:** Augments the RTFM FM to incorporate text features, capturing temporal dependencies across both visual and textual modalities.
- **PEL4VAD:** Improves previous work [293] global/local temporal modelling and incorporates Prompt-Enhanced Learning (PEL) to leverage knowledge from a concept dictionary. To note the simple self-attention mechanism that provides lightweight global/local temporal modelling, contrasting to approaches like HLN, RTFM, URDMU and HyperVD.

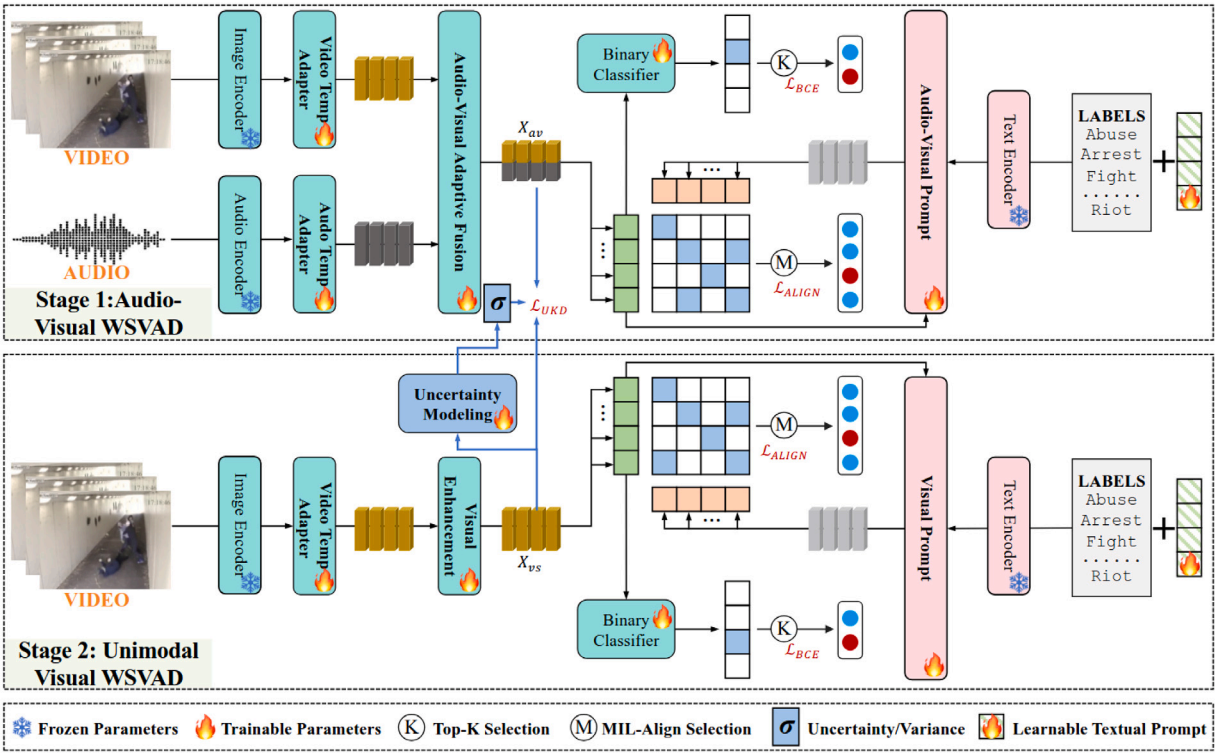


Fig. 26. AVadCLIP [42] proposed method.

- **CNN-ViT:** Combines CNN-based and ViT-encoded visual features from CLIP to bridge the domain gap, and utilizes a TSA module for feature refinement.
- **AnomalyCLIP:** Exploits the vision–text alignment in CLIP's latent space using a CoOp prompt learning strategy, enabling both anomaly detection and retrieval.
- **VadCLIP:** Introduces a dual-branch method with learnable and visual prompts, utilizing visual features for binary classification and aligning visual and textual features for language–image understanding.
- **LAP:** Employs prompt-based learning using CLIP and pseudo-labelling (PAL) guided by video-caption similarity. Combines visual-semantic synthesis and multi-prompt contrastive learning using an prompt dictionary to improve fine-grained anomaly detection and open-set generalization.
- **TPWNG:** Fine-tunes CLIP with Normality Visual Prompts (NVP) and learnable text prompts, generating pseudo-labels guided by contrasting normal/abnormal textual similarity scores to train an anomaly classifier. Introduces Temporal Context Self-Adaptive Learning (TCSAL) for dynamic attention over video timelines, enhancing detection of gradual or abrupt anomalies.
- **PE-MIL:** Fuses audio and video features using temporal attention and scale-aware anomaly scores, enhanced by Abnormal-Aware Prompt Learning (APL) to inject semantic class priors. It refines event-context alignment via CLIP-based text prompts and Normal Context Prompt to fit the captured normal distribution.
- **STPrompt:** A two-stream framework combining temporal detection via motion-prior spatial attention (SA^2) to suppress static backgrounds and a lightweight transformer for global context, and spatial localization using LLM-generated text prompts for training-free patch-level retrieval via CLIP similarity.
- **ITC:** Injecting Text Clues (ITC) through a dual-branch framework combines a discriminative Text-Guided Anomaly Discovering (TAG) branch — which aligns video segments with textual anomaly queries via hierarchical CLIP-based matching — and a generative Anomaly-Conditioned Text Completion (ATC) branch, which reconstructs masked anomaly descriptions to gather semantic cues.
- **TDSD:** introduces a text-driven, scene-decoupled framework for weakly supervised video anomaly detection, combining Context Semantic Injection (CSI) (CLIP-based scene category fusion) and Object Semantic Injection (OSI) (zero-shot object feature extraction) with Fine-Grained Visual Augmentation (FVA) to enhance discriminability.
- **MELOW:** Extends VadCLIP's dual-branch framework by integrating multi-scale temporal modelling and multimodal evidential learning (Dirichlet-based uncertainty fusion of visual–textual evidence), enabling robust detection of unseen anomalies in open-world settings.
- **AVadCLIP:** Builds on CLIP and VadCLIP by integrating adaptive audio–visual fusion, multimodal-enriched prompts, and an uncertainty-aware knowledge distillation module to handle missing or noisy modalities.

Vision–language models have significantly advanced the field of VAD by incorporating semantic information and enabling a more in-depth understanding of video content. The TSA module has proven effective in refining feature representations through its top-K nominator mechanism. The prompt-based learning strategies, like those in VadCLIP, AnomalyCLIP, TPWNG, STPrompt, AVadCLIP, and especially, PEL4VAD, PE-MIL and ITC, demonstrate the potential of leveraging external knowledge for improved anomaly detection and classification. The foundational dual-branch approach of VadCLIP for detection (coarse) and classification (fine) sets the direction for STPrompt's additional train-free spatial anomaly localization, ITC's hierarchical text–video alignment & anomaly-conditioned text completion, MELOW's multimodal evidence learning and AVadCLIP's probabilistic uncertainty distillation strategy.

AnomalyCLIP's success in jointly addressing anomaly detection and retrieval highlights the power of exploring CLIP's latent space for multi-modal understanding. A key takeaway from AnomalyCLIP's contribution is that the latent CLIP space becomes discriminative when

Table 11
Summary of visual Language models.

Method	FE	FUSE			FM	MIL	SS	LT		Metrics	
		E	M	L				FL	SL	UCFC	XDV
CLIP-TSA	I3D CLIP	✗	✗	✗	TSA PDC+NL	✓	magn Top-K	✓	✓	84.66 87.58	78.19 82.19
TeVAD	I3D+ SwinBERT	✗	✓	✗	PDC+NL	✗	Magn Top-K	✓	✓	84.90	79.80
PEL4VAD	I3D	✗	✗	✗	SAtt	✗	Top-K Scores	✓	✓	86.76/72.24/0.43	85.59/70.26/0.57
CNNVIT	C3D (+CLIP) I3D (+CLIP) CLIP	✓	✗	✗	TSA PDC+NL	✓	Magn Top-K	✓	✓	85.78 86.50 87.63 88.02 88.97	✗
Anomaly-CLIP	CLIP	✓	✗	✗	Axial Transf	✓		✗	✓	86.36	78.51
VadCLIP	CLIP (img&txt)	✓	✓	✗	Local TE Global GC	✓	Top-K Scores	✓	✓	88.02/70.23	84.51
LAP	CLIP (img&txt)	✗	✓	✓	PAL	✓	AUC Top-K	✓	✓	87.7	82.6
TPWNG	CLIP (img&txt)	✓	✓	✗	✗	✗	Max&Min	✓	✓	87.79	83.68
PE-MIL	I3D (+VGG)	✓	✓	✓	TFF	✓	Top-K Scores	✓	✓	86.83	88.05
STPrompt	CLIP (img&txt)	✗	✓	✗	Spatial Att	✓	Top-K Scores	✓	✓	88.08	✗
ITC	CLIP (img&txt)	✓	✓	✓	TC	✓	Scores	✗	✓	89.04	85.45
TDSD	I3D+CLIP (img&txt)	✓	✓	✗	SA	✓	Top-K Categ.&Scores	✗	✓	85.49 (UCF-SHT)	84.69
MELOW	CLIP (img&txt)	✓	✓	✓	LGT MSVM	✓	Top-K Scores	✓	✓	87.80x	85.13
AVadCLIP	CLIP (img&txt) (+Wav2CLIP)	✗	✓	✗	Temp Transf GCN	✓	Top-K Scores	✓	✓	✗	85.53 86.04

FE: Feature Extractor; **FUSE**: **E**arly **M**id **L**ate; **FM**: Feature Modulator; **SS**: Segment Selection; Magnitude; **LT**: Loss Target; **FL**: Feature-Level; **SL**: Score-Level; **TC**: Temporal Convolution.

realigned with a pre-constructed standard prototype, also explored in NGMIL [96], main idea of relying on a divergence measure between abnormal and normal global representations (Section 5.4). This finding aligns well with BN-DFM [279], which bases its SS on an embedding space statistically rooted around a prototype.

However, further research is needed to explore more diverse prompt engineering techniques and backbones other than CLIP-ViT, solutions for dependency in vanilla segment-level scores (Section 5.1) and evaluate the generalizability of these methods across different datasets and anomaly types, to provide a more nuanced interpretation of video content and pave the way for further innovations towards explainable VAD.

5.8. Large language model and VLM for VAD

Integrating Large Language Models (LLMS) into VAD marks a paradigm shift from conventional pattern-based approaches toward semantically guided reasoning, enabling models to interpret better, localize, and generalize abnormal events in unstructured video data.

The survey **Quo Vadis, Anomaly Detection?** [3] details the integration of Large Language Model (LLM) and VLM in VAD, highlighting their role in tackling core challenges. These include semantic interpretability (e.g., VAD-LLAMA [320], LAVAD [8], HAWK [197]), temporal modelling (e.g., OVVAD [6], Holmes-VAD [251]), few-/zero-shot detection (e.g.,VERA [321], Flashback [10]), and open-world generalization (e.g., OVVAD [6], Holmes-VAU [13], LAVAD [8], PLOVAD [7]). In this work, we complement the survey findings with a deeper analysis of recent advancements and broaden the scope by incorporating additional LLM-based approaches relevant to VAD problematic.

Integrating LLM's has begun to be explored by **VAD-LLaMA** [320] using Video-LLaMA [322] to provide textual explanations for detected anomalies, addressing the struggle with manual threshold selection by traditional anomaly-scoring methods. Existing video-based large language models (VLLMs) further face challenges in long-range context modelling and require extensive labelled data for domain-specific tuning. To bridge these gaps, the authors propose VAD-LLaMA, a framework integrating VLLMs into VAD to enable threshold-free anomaly detection with textual explanations. The core novelty lies in two components: (1) a Long-Term Context (LTC) module that dynamically aggregates normal/abnormal clip features via cross-attention to enhance long-range context modelling, and (2) a three-phase training strategy that minimizes VAD data requirements by co-training the LTC-enhanced detector (VADor) [] with frozen VLLM components and generating pseudo-instructions from anomaly scores. This approach eliminates manual thresholding while improving interpretability through natural language explanations. Experiments on UCFC (1900 videos, video-level weak labels) and TrafficAD (TAD) achieved AUC improvements of up to +3.86 p.p and +4.96 p.p, respectively, with significant gains in anomaly-specific metrics (AUCAA). The LTC module proves critical, boosting performance on context-dependent anomalies like "Arson" by leveraging long-term dependencies. It establishes a precedent for efficiently adapting general-purpose VLLMs to specialized tasks with limited labelled data, broadening their applicability in multimodal reasoning domains.

Open-world. **Open-World video anomaly detection** addresses the critical challenge of identifying unforeseen anomalies in real-world deployments, where models must contend with both known and novel abnormal events post-training. Traditional closed-set approaches, trained

on fixed anomaly categories, falter under semantic distribution shifts, leading to high false positives or missed detections. This open-set risk underscores the need for methods that balance discriminative power on known anomalies with robust uncertainty awareness to reject unknowns.

OpenVAD [4] addresses VAD in an open-world setting by unifying MIL with Evidential Deep Learning (EDL) and normalizing flows (NFs) under weak supervision. The proposed method leverages a dual graph convolutional network (GCN) to model temporal and feature-wise dependencies, enhanced by triplet loss for discriminative embedding. EDL employs a Beta distribution prior to predict per-clip evidence, enabling uncertainty-aware selection of high-confidence anomalies (via thresholds on confidence and evidence) to mitigate label noise in MIL training. Concurrently, NFs learn the latent distribution of normal clips, synthesizing pseudo-anomalies by sampling low-density regions in the feature space, thereby bounding open-set risk. A multi-stage training protocol first optimizes feature embeddings and EDL classifier using MIL loss (Type II maximum likelihood) and triplet loss, then freezes the encoder to train NFs, and finally fine-tunes with selected clean anomalies and pseudo-anomalies. Evaluated on UCFC, XDV, and ST demonstrates robustness under varying ratios of known anomalies and limited supervision.

Open-Vocabulary VAD (OVVAD) [6] (Fig. 27) pushes the boundaries of VAD by utilizing pre-trained vision–language models (e.g. [122]-ViT) to detect and categorize both seen and unseen anomalies, taking advantage of their powerful zero-shot generalization ability. The framework decouples OVVAD into two synergistic tasks: class-agnostic detection (identifying anomalies) and class-specific classification (labelling anomaly types). A lightweight graph convolutional network models the temporal dependencies by constructing an adjacency matrix based on frame proximity. Through Semantic Knowledge Injection (SKI), a cross-modal alignment mechanism fuses textual embeddings of normal/abnormal scenarios generated via LLM prompts and encoded by CLIP's text encoder with CLIP visual features via sigmoid-weighted similarity scores. A FFN is the regressor network trained under a top-K video-level classification loss. For categorization, a learnable prompt is employed, similar to the work of [39], and the similarity between aggregated video-level features and textual category embeddings IS used in a video-level cross-entropy loss. In a fine-tuning stage for unseen categories, a Novel Anomaly Synthesis (NAS) module is proposed by generating pseudo-anomalous videos via LLM-guided prompts and AIGC models (e.g., DALL-E). This approach holds promise, especially in a real-world UCFC, while suboptimal in scenarios requiring temporal modelling XDV, leaving the investigation room open-world settings by investigating class-wise performance.

LAnguage-based VAD (LAVAD) [8] introduces a training-free anomaly detection framework that synergizes pretrained VLM's and LLM. The method first generates frame-level captions using an ensemble of BLIP-2 variants [323] (including Flan-T5-XXL and OPT-6.7B) to maximize caption diversity, then refines them by selecting the most semantically aligned caption per frame via ImageBind's [324] cross-modal similarity between visual and textual embeddings. Temporal context is modelled through sliding 10-s windows, sampling 1 frame/sec, where Llama-2-13b [139] aggregates captions into scene summaries via task-specific prompts (e.g., "Summarize events in this surveillance context..."). Initial anomaly scores from LLM reasoning are refined using video–text alignment, where scores from frames with semantically similar summaries (via ImageBind's video/text encoders) are weighted by their cosine similarity. This approach uniquely addresses spatial–temporal ambiguity through cross-modal noise suppression, LLM-guided temporal aggregation, and video–language correspondence, establishing a new paradigm for zero-shot VAD.

HAWK [197] (Fig. 28) proposes a vision–language framework for open-world video anomaly understanding by explicitly incorporating motion cues, recognizing their significance in distinguishing anomalous events. The architecture, based on Video-LLaMA [322], adopts a dual-branch design — one for appearance and another for motion — based on frozen video encoders from BLIP-2 [323] (EVA-CLIP [126] and a pre-trained Video Q-Former). Motion features are extracted via optical flow (Farneback algorithm [325]). To enhance anomaly focus, two auxiliary losses are introduced: (i) a video-motion consistency loss that aligns appearance and motion features, and (ii) a motion-language matching loss that associates motion features with parsed linguistic entities and verbs. The model undergoes a three-stage training process: (1) pre-training on WebVid [326], (2) fine-tuning on the proposed curated anomaly dataset (8000 videos from seven benchmarks) with GPT-4-generated descriptions and QA pairs, and (3) task-specific evaluation (using metrics as Text-Level benchmarks (e.g., BLEU, ROUGE) and GPT-Guided metrics). HAWK enables interactive tasks such as anomaly description and open-ended QA, showing one more promise across domains like surveillance and traffic. Its multimodal integration establishes a new paradigm for explainable and generalizable VAD, despite scalability challenges in real-world deployment.

Anomaly Shield (AnomShield) is proposed for Exploring the Causation of Video Anomalies (ECVA) [241,252] benchmark (2240 real-world videos with 100 types of anomalies) to address 3 sub-tasks: the What (anomaly type/description), Why (cause), and How (severity via importance curves) of an anomaly.

The proposed AnomShield VLM integrates a novel pipeline for causal reasoning. It employs chain-of-thought prompting to identify anomaly-critical segments: (1) coarse sampling extracts uniform frames, (2) captioning generates segment descriptions via a VLM, (3) retrieval leverages GPT to align descriptions with user queries, and (4) dense resampling focuses on key segments. These segments are processed by a bidirectional Mamba-based connector, which combines CLIP-L/14 visual features with Mistral-7B's [327] language capabilities, enhanced with spatiotemporal position embeddings. Training occurs in three stages: alignment of image–text features by freezing the Base LLM and visual encoder while pre-training the connector with the low-quality short-text data, adaptation of the connector to long-text narratives, and video-specific fine-tuning using Low-Rank Adaptation (LoRA) [328] on multiple datasets [329–331], ensuring robust multimodal integration.

Evaluation is conducted via the proposed AnomEval, a metric assessing VLMs through basic reasoning (coverage of key entities and logical coherence), consistency (GPT-based binary scoring) and hallucination (robustness to edited videos). AnomEval achieves 82%–89% alignment with human judgment, surpassing traditional metrics like BLEU and ROUGE. Experiments demonstrate excellence in causal reasoning (e.g., linking traffic accidents to red-light violations). At the same time, ablation studies confirm the necessity of both hard and soft prompts (for anomaly focus) (for spatiotemporal modelling). This work bridges video–language alignment with causal reasoning, offering a foundational framework for real-world anomaly understanding. ECVA, AnomShield, and AnomEval collectively advance the field by prioritizing interpretability, temporal granularity, and robustness, setting a new standard for future research.

Holmes-VAD [251], an interpretable framework integrating multimodal Large Language Model (LLM) towards unbiased and explainable VAD. The authors first construct VAD-Instruct50k, a novel multimodal VAD benchmark (VAD-Instruct50k) developed via a semi-automatic labelling pipeline, combining sparse single-frame temporal annotations (applied to untrimmed videos from UCFC and XDV) with LLM-synthesized explanations. The framework employs a three-stage architecture: a frozen video encoder from LanguageBind [332] to extract frame-level features, a lightweight temporal sampler to select high-response frames using pseudo-label supervision from sparse annotations, reducing computational overhead, and a VLM fine-tuned with LoRA adapters to generate natural language explanations for identified anomalies. Key innovations include the efficient annotation protocol balancing labelling costs with supervision quality, and the temporal

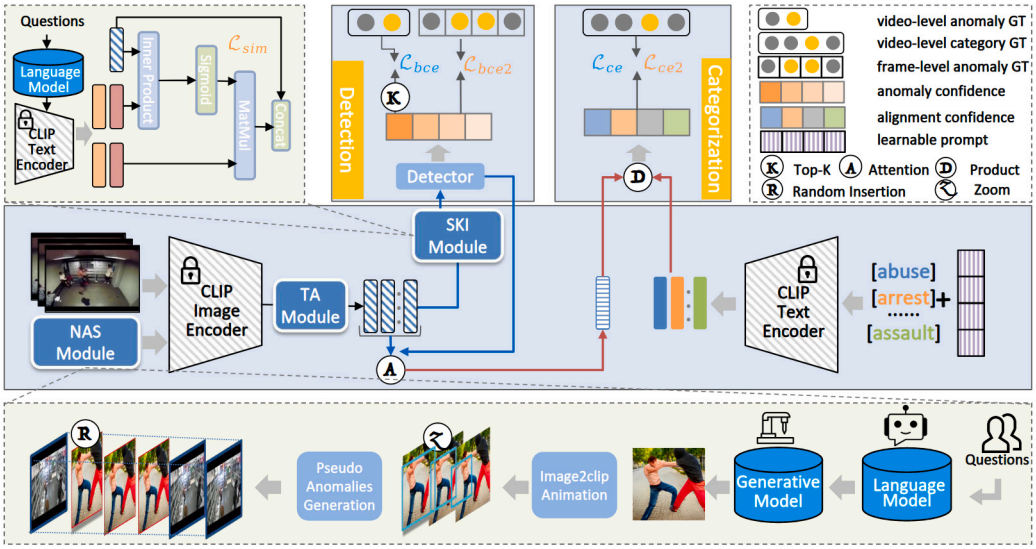


Fig. 27. OVVAD [6] proposed method.

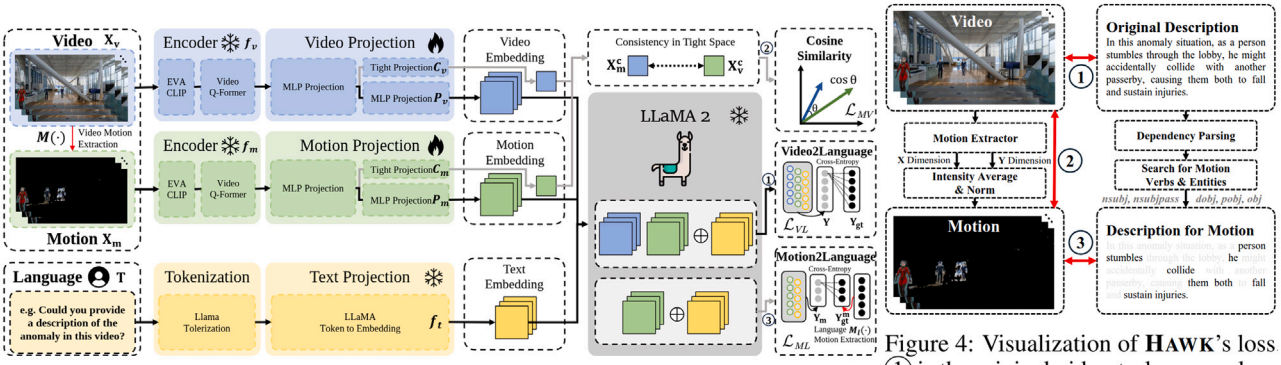


Figure 3: Overview of **HAWK**. During training (**Black** and Gray path), we aim to optimize for video-language matching loss, along with Video-Motion Consistency and Motion-Language Matching. During inference (only Gray path), we generate language descriptions using video, motion, and text.

Figure 4: Visualization of **HAWK**'s loss. ① is the original video-to-language loss. ② is the cosine similarity loss for motion modality adaptation. ③ is the motion-to-language loss.

Fig. 28. HAWK [197] proposed method.

sampler's dual role in enabling hour-long video processing while feeding salient segments to the LLM. Experimental validation demonstrates state-of-the-art performance (90.67% AP on XDV, 89.51% AUC on UCFC) alongside improved interpretability through human evaluations, establishing a new paradigm for explainable anomalies in long-form video analysis.

Holmes-VAU [13] (Fig. 29) upgrades the previous contribution of Holmes-VAD and VAD-Instruct50k [251] for hierarchical video anomaly understanding (VAU), consisting in classical temporal anomaly detection and anomaly explainability, model's ability to provide an anomaly-related response attending both visual perception (recognizing main entities in the video) and anomaly reasoning (model's judgment and analysis of the anomaly content). The proposed framework has at its core an Anomaly-focused Temporal Sampler (ATS), which dynamically prioritizes anomaly-dense segments in long videos using a two-stage process: the anomaly scorer (UR-DMU [278] architecture) generates frame-level anomaly probabilities. In contrast, a density-aware sampler converts these scores into a cumulative distribution for non-uniform frame sampling. This sampling ensures computational resources focus on critical regions (e.g., sudden collisions) while reducing redundancy in normal footage. The selected frames are processed by a multimodal LLM (InternVL2-2B [333]) fine-tuned with LoRA adapters on hierarchical instruction data, enabling natural language explanations of anomalies. The anomaly scorer is first optimized during training using the annotated frame-level labels. At the same time, the LLM undergoes instruction fine-tuning through LoRA [328], both using the proposed updated benchmark HIVAU-70k's multi-granular data (clip captions, event summaries, video analyses). The ATS's adaptive sampling reduces inference latency by 4–8× compared to uniform sampling while maintaining detection accuracy and efficiently analysing hour-long surveillance footage. Evaluation for anomaly detection is measured with AUC/AP on UCFC and XDV, while BLEU [334], CIDEr [335], METEOR [336] and ROUGE [337] for the quality of the reasoning text output by the model. By integrating temporal anomaly localization with hierarchical semantic reasoning, the framework provides both "when" (temporal detection) and "why" (textual explanation) insights, bridging low-level visual patterns and high-level contextual understanding.

Prompting VLMs for Open Vocabulary (PLOVAD) [7] (Fig. 30) proposes CLIP to detect and categorize both seen and unseen anomalies in videos, addresses challenges of data scarcity and open-world generalization by leveraging prompt tuning and temporal modelling. The framework comprises two core modules: (1) a Prompting Module that generates domain-specific prompts (learnable vectors capturing task-specific knowledge) and anomaly-specific prompts (LLM-generated textual descriptions of anomalies, e.g., "fighting involves aggressive physical contact"), enriching semantic understanding of diverse anomalies; and (2) a Temporal Module using a graph attention network (GAT)

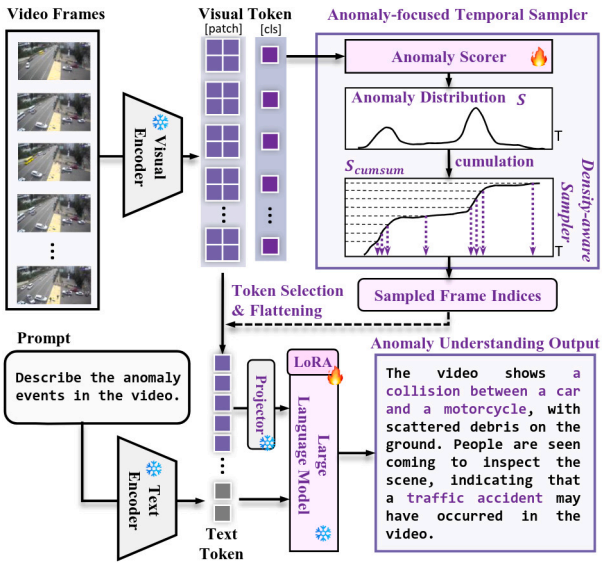


Fig. 29. Holmes-VAU [13] proposed method.

to model spatiotemporal dependencies across video frames, bridging the gap between static image features and dynamic video contexts. During training, PLOVAD aligns visual features (enhanced by temporal reasoning) with text embeddings from prompts via cross-modal loss, enabling the detection and categorization of anomalies without requiring labelled data for unseen classes. By freezing the VLM backbone and tuning only lightweight components (prompts and GAT), PLOVAD achieves scalability, outperforming traditional weakly supervised methods on benchmarks like UCFC and ST in both open-vocabulary and weakly supervised settings.

The **Sherlock** model, a Global–local Spatial-sensitive Large Language Model [253], proposes the M-VAE (Multi-scene Video Abnormal Event Extraction and Localization) task by extracting structured event quadruples (subject, event type, object, scene) with corresponding timestamps. The primary challenges — modelling global and local spatial information and addressing spatial imbalance — are tackled through a Global–local Spatial-enhanced Mixture of Experts (MoE) module, featuring four spatial experts (Action, Object Relation, Background, and Global Scene), and a Spatial Imbalance Regulator (SIR) with a Gated Spatial Balancing Loss (GSB) to mitigate data imbalance among experts. The model is trained on a custom M-VAE instruction dataset derived from the CUVA benchmark [252], comprising two stages: (1) pre-tunes Video-LLaVA [338] on high-quality datasets like Ref-L4 [254], HumanML3D [255], RSI-CB [256] and COCO [257] to enhance spatial understanding, while (2) fine-tunes the model on the M-VAE dataset to localize events and extract quadruples. Sherlock is benchmarked against advanced Video-LLMS (e.g., Video Chatgpt, Video-LLaVA) and performs better, achieving a 10.85% improvement in event extraction and 11.42% better localization (mAP@tIoU), with significantly lower false negative rates (FNRs)/higher F2 across all scenes.

Verbalized Learning Framework (VERA) [321] proposes a verbalized learning (VL) framework to support VAD without the need for manually crafted guiding questions for frozen visual–language models (VLMS). Instead, it introduces a data-driven learning objective that treats guiding questions as learnable parameters. These are optimized through feedback from a VLM using a binary video classification subtask, allowing the model to discover practical anomaly characterization questions from coarsely labelled datasets. This Characterization enables efficient capture of video-specific temporal characteristics while avoiding detailed instance-level annotations. In the inference phase, VERA follows a coarse-to-fine procedure: it first computes segment-level anomaly scores by prompting the VLM with the learned guiding question, then refines those scores through softmax-weighted ensembling based on scene similarity. Finally, it produces e-level scores by applying Gaussian smoothing and position-based temporal weighting, enabling accurate localization and interpretable VAD. The model is evaluated on the UCFC and XDV dataset with an average video length of 1.62 min.

EventVAD [9] acknowledges limitations of previous train-free in localizing anomalies with high temporal precision and combines event-aware segmentation and VLM understanding. The proposed method combines four core components: an Event-Aware Dynamic Graph Construction builds spatiotemporal graphs by fusing RAFT [339]-based optical flow (capturing motion dynamics) and EVA-CLIP's ViT-B/16 [126] semantic features (encoding visual context), with edges weighted by cross-modal similarity and temporal decay to prioritize short-term event correlations, segmenting videos into event units to enhance temporal consistency. Second, Graph Attention Propagation (GANP) refines node features through orthogonal feature projection and iterative message passing, amplifying inter-frame differences at event boundaries. Third, Statistical Boundary Detection identifies event transitions by combining abrupt feature-space jumps and directional changes in the manifold trough dissimilarity scores, with Savitzky–Golay filtering for signal smoothing and median absolute deviation (MAD) thresholding, adaptively segmenting videos into semantically coherent event units to avoid fragmented VLM inputs. Finally, in Event-Centric Anomaly Scoring, a VideoLLaMA2.1-7B-16F [340] analyses segmented events using hierarchical prompts, to first provide video descriptions and then generate frame-level anomaly scores, addressing long-tail fragmentation and improving interpretability, showing better performance compared to the 13B-parameter LAVAD.

Local Patterns Generalize Better for Novel Anomalies (LPG) [198] (Fig. 31) proposes a vision–language framework for unsupervised VAD that prioritizes semantically consistent local patterns to address open-set generalization. The pipeline begins by cropping regions of interest (via YOLOv7 [341] or Qwen-VL [168]) and processes them through a two-stage approach: first, the Image–Text Alignment Module (ITAM) leverages BLIP-2's frozen backbones [323] (EVA-CLIP and Q-Former) to extract text-informative spatial patterns (e.g., body joints, object parts) from cropped regions, encoding them as image tokens aligned with generic textual descriptions (e.g., "a person moving arms"). These tokens capture domain-agnostic semantics, enabling recombination of known components to represent unseen anomalies. Next, the Cross-Modality Attention Module (CMAM) refines these patterns by weighting image tokens based on their similarity to temporally coherent captions generated by the Temporal Sentence Generation Module (TSGM). TSGM integrates a State Machine Module (SMM) to propagate high-resolution textual context from prior frames (e.g., "a man pushing a stroller") into low-resolution or occluded observations, ensuring caption consistency across temporal variations. To model dynamics, motion vectors from H.265(HEVC)-encoded videos [342] are extracted and fused with spatial patterns in a Reconstruction Module (RM), which is trained exclusively on normal data to detect anomalies via reconstruction error, jointly optimizing spatial–temporal feature fidelity. Evaluated on ST, UB, NWPU Campus (NWPU), UCFC, and XDV, shows a framework trained using normal data is capable of detecting unseen anomalies, by exploiting vision–language alignment for robust open-set detection, where cross-modal attention suppresses noisy backgrounds while motion estimation flags irregular dynamics.

Flashback [10] introduces a two-stage paradigm for zero-shot, real-time video anomaly detection by decoupling offline pseudo-scene memory construction and online caption-retrieval inference. In the offline phase, a frozen LLM (GPT-4o) generates 1M+ normal and anomalous scene captions using structured prompts, which are embedded via a cross-modal encoder (e.g., ImageBind [324] and PerceptionEncoder [343]) and stored in memory. To enhance separation between

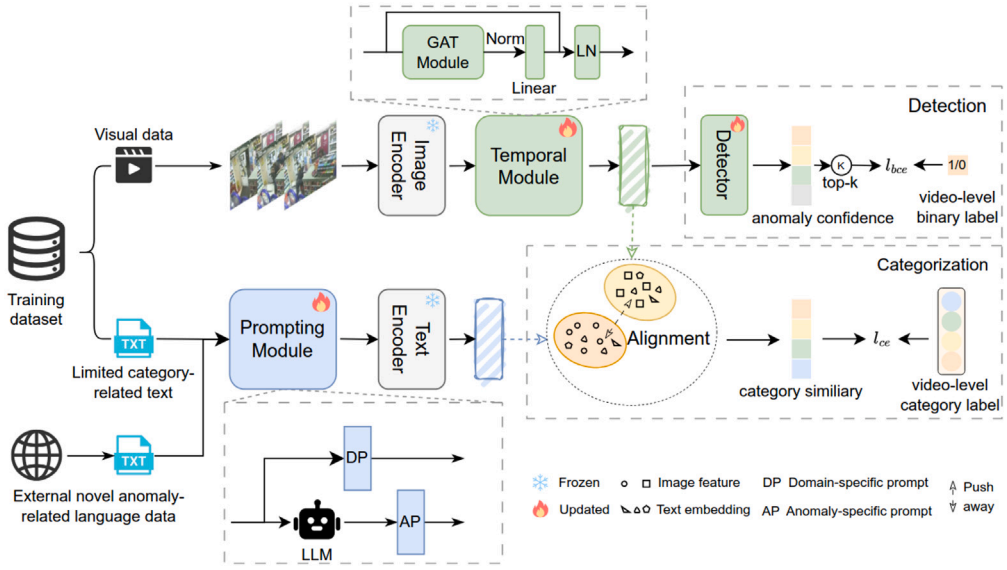


Fig. 2. Overview of the proposed framework (PLOVAD). PLOVAD comprises two primary modules: the Temporal Module and the Prompting Module, catering to two sub-tasks: the detection sub-task and the categorization sub-task. The Temporal Module integrates temporal information using GAT stacking atop frame-wise visual features to address the transition from static images to videos. The Prompting Module is employed to formulate a domain-specific prompt (DP) to capture domain-specific knowledge and an anomaly-specific prompt (AP) crafted by a LLM to capture semantic nuances and enhance generalization.

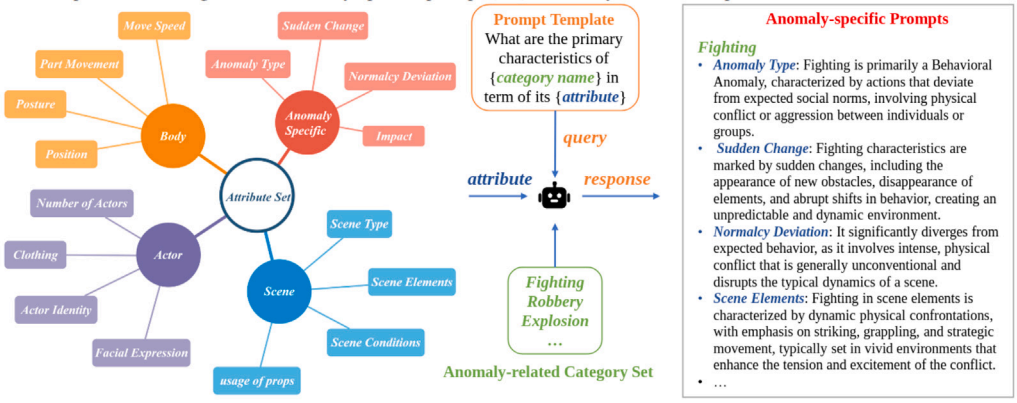


Fig. 3. Illustration of Anomaly-specific Prompt(AP) generation workflow. On the left, the process of defining the attribute set is visualized. The middle section depicts the querying process with LLMs, transforming anomaly-related categories and attributes into APs. On the right, we present sample snippets of the prompts generated.

Fig. 30. PLOVAD [7] proposed method.

normal and anomalous embeddings, repulsive prompting is applied: normal captions are prefixed with "Normal" and anomalous ones with "Anomalous", while lightweight templates further widen their feature-space distance. Additionally, scaled anomaly penalization attenuates the magnitudes of anomalous embeddings to mitigate residual bias during retrieval.

During online inference, incoming video segments are encoded into embeddings and matched against the memory using similarity search (dot product), with top-K retrieved captions aggregated into segment-level anomaly scores. These scores are smoothed via Gaussian filtering to yield frame-level predictions. By eliminating LLM calls at inference — replacing autoregressive captioning with efficient retrieval — Flashback achieves 42.06 FPS on an RTX 3090, 34× faster than prior methods, while attaining 87.29% AUC on UCF-Crime and 90.54%AUC/75.13% AP on XD-Violence. Key design choices, such as repulsive prompting (increasing normal/anomalous centroid angles from 8.12° to 33.29°) and memory size optimization (1M captions vs. 10k), ensure robustness and scalability, balancing computational efficiency with strong zero-shot performance.

Table 12 summarizes the works employing Large Language Model and Vision Language models. Methods designed for Zero-Shoot/Train-Free are inherently Open-World, for others it means works attend to that setting specifically.

Key Points of LLM and Vision–Language VAD Methods: These methods leverage the multimodal reasoning capabilities of large Vision–Language Models (VLMs) and Large Language Models (LLMs) to enhance anomaly detection with rich semantic understanding, enabling threshold-free detection, textual explanation, and causal reasoning.

- **VAD-LLaMA:** Integrates VLLMs with a Long-Term Context (LTC) module and a three-phase training strategy to enable threshold-free anomaly detection with natural language explanations. Uses frozen VLLM components and pseudo-instruction generation to reduce labelled data requirements while enhancing interpretability.
- **OpenVAD:** Tackles open-world VAD by unifying MIL with evidential deep learning (EDL) and normalizing flows (NFs) under weak supervision. Uses dual GCNs for temporal and feature modelling, EDL for uncertainty-aware anomaly selection, and NFs to synthesize pseudo-anomalies from low-density regions. A three-stage

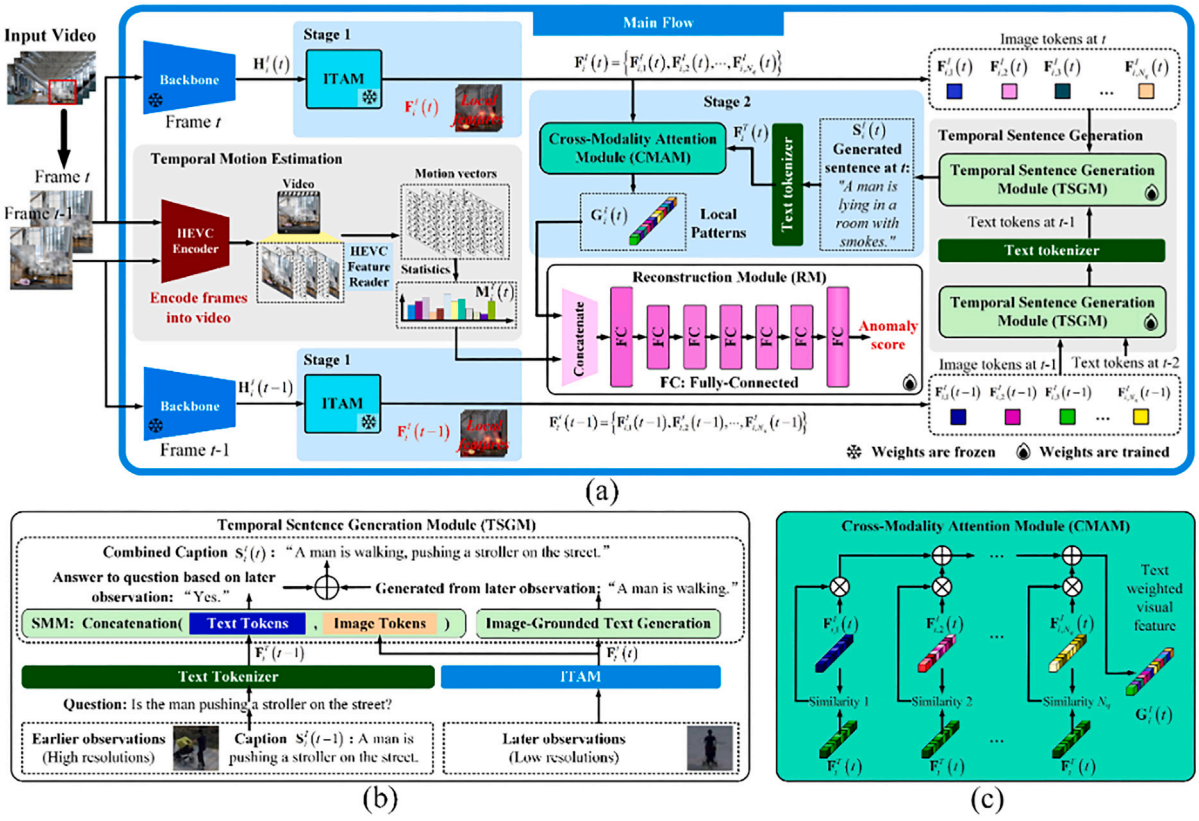


Figure 2: Structure of the method. (a) Main flow: Visual features are extracted using backbone and Image-Text Alignment Module (ITAM) which identifies caption-informative local features as image tokens. Conditioned on image tokens, the Temporal Sentence Generation Module (TSGM) generates sentences which are then combined with image tokens using the Cross-Modality Attention Module (CMAM) to highlight key patterns. HEVC Encoder estimates inter-frame motion through video compression. Spatial local patterns and motion are jointly analyzed to detect anomalies through reconstruction. (b) TSGM uses the State Machine Module (SMM) for sentence generation based on image tokens and earlier sentences. (c) CMAM is implemented based on image-text similarity.

Fig. 31. LPG [198] proposed method.

Table 12
Summary of large language models and visual language models.

Method	OW	Sup.	FE	FUSE	FM	LT	UCFC	XDV	Description
VAD-LLaMA	–	Weak	Video-LLaMA	E, M, L	LTC, pseudo-instr., frozen VLM	SL	✓	–	Threshold-free, NLE
OpenVAD	✓	Weak	MIL, EDL, NF	E, M	Dual-GCN, ER, PS	FL	–	–	Weakly supervised, open-set
OVVAD	✓	Weak (VLM)	CLIP + ViT	E, L	SKI fusion, PL, FFN	SL	–	✓	Zero-shot prompt-based
LAVAD	✓	Train-Free (L+VLM)	BLIP-2, LLaMA-2, ImageBind	E, M, L	Caption, Sim score	–	–	✓	Training-free zero-shot
HAWK	–	Descrip. & Answer (L+VLM)	Video-LLaMA, EVA-CLIP	E, M	Dual-branch, align.	SL	–	✓	GPT-based QA on anomaly
AnomShield	–	Descrip. & Answer (L+VLM)	CLIP-L/14, Mistral-7B	E, M	CoT, causal analysis	SL	✓	✓	AnomEval: what–why–how
Holmes-VAU	–	Fully(frame-level) & Descript. L+VLM	InternVL2-2B, UR-DMU	E, M, L	ATS, LoRA, hierarchy	FL, SL	–	✓	Efficient "when+why" explain.
PLOVAD	✓	Weak	CLIP + GAT	E, L	Prompt + GAT temporal	FL	✓	–	Scalable, temporal context
Sherlock	–	Instruction Tuning	Video-LLaVA, MoE	E, M	Event extraction, SIR	FL, SL	–	✓	Structured anomaly scene event quadruples & timestamps
VERA	–	Multi (L+VLM)	CLIP + prompts	E, L	Verbal prompts, ens.	FL	✓	✓	Prompt-based question mining
EventVAD	✓	Train-Free (L+VLM)	EVA-CLIP, VideoLLaMA2.1	E, M, L	Graph Attention	–	✓	✓	Precise, long-tail aware
LPG	–	Weak (VLM)	BLIP-2, Q-Former	E, M	Tokens, temp. caption	FL	–	✓	Local recomposition gen.
Flashback	✓	Train-Free	ImageBind, PerceptionEncoder	E, M	RepulsivePprompting	–	–	✓	RT trough Pseudo-Scene Memory

OW: Open-World, **Sup.**: Supervision, **FE**: Feature Extractor, **FUSE**: Fusion (E = Early, M = Mid, L = Late), **FM**: Feature Modulator, **SS**: Segment Selection, **LT**: Loss Target (FL = Feature, SL = Score), **UCFC/XDV**: Benchmark used, PL = Prompt Learning, ER = Evidential Reasoning, PS = Pseudo-synthesis, CoT = Chain-of-Thought, SIR = Spatial Imbalance Reg.

training scheme enhances robustness under limited supervision and open-set conditions.

- **OVVAD:** Uses CLIP-based vision–language alignment for open-vocabulary anomaly detection and classification. Employs Semantic Knowledge Injection (SKI), prompt learning, and a lightweight GCN for temporal modelling. Introduces NAS for unseen class generation via LLM prompts and AIGC.
- **LAVAD:** Training-free, it fuses BLIP-2-based captioning with LLM reasoning (Llama-2) and ImageBind similarity scoring. Uses LLMs for contextual aggregation and video–text alignment for zero-shot anomaly detection with temporal reasoning.
- **HAWK:** Open-world framework combining appearance and motion branches. Integrates optical flow and motion-language matching to enhance spatiotemporal reasoning. Trained using

WebVid pretraining and anomaly-specific fine-tuning with GPT-generated descriptions.

- **ECVA + AnomShield:** Introduces a causality-focused benchmark and pipeline combining VLM-based captioning, GPT-based retrieval, and Mamba-based bidirectional reasoning for What–Why–How anomaly understanding. AnomEval assesses reasoning and hallucination alignment with human judgment.
- **Holmes-VAD:** Introduces VAD-Instruct50k benchmark with LLM-synthesized labels. Uses LanguageBind features and a temporal sampler to efficiently feed salient segments to a LoRA-tuned LLM for interpretable anomaly reasoning in long videos.
- **Holmes-VAU:** Enhances Holmes-VAD with hierarchical reasoning and temporal sampling. Anomaly-focused Temporal Sampler (ATS) and InternVL2 LLM enable fine-grained "when" and "why" insights using the proposed multi-granular HIVAU-70k benchmark data.
- **PLOVAD:** Uses CLIP with a prompting module and graph-based temporal modelling to detect and categorize both seen and unseen anomalies without labelled data for novel classes. A GAT module aligns visual and textual features across time for open-vocabulary, weakly supervised VAD.
- **Sherlock:** Tackles multi-scene video anomaly detection by extracting structured event quadruples using a global-local spatial mixture-of-experts and a spatial imbalance regulator.
- **VERA:** Introduces a verbalized learning approach where anomaly characterizations are treated as learnable guiding questions optimized via binary video classification.
- **EventVAD:** Combines event-aware graph segmentation and vision–language understanding using optical flow, CLIP features, and dynamic graphs to detect temporally coherent events.
- **LPG:** A VLM-based unsupervised VAD pipeline that learns local, semantically aligned spatial patterns and combines them with motion features for open-set generalization.
- **Flashback:** A zero-shot, real-time framework by decoupling LLM-based pseudo-scene memory construction (offline) from caption-retrieval inference (online), leveraging repulsive prompting (to widen normal/anomalous embedding separation) and scaled anomaly penalization (to mitigate residual bias), achieving 87.29 AUC on UCFC at 42.06 fps on consumer GPUs.

The evolution of VLMs has enabled transformative advances in WVAD, shifting from computationally intensive architectures (e.g., 3D CNNs like C3D/I3D) to lightweight, efficient designs. LLM-integrated VAD systems combine vision–language representations and natural language reasoning to support zero-shot detection, interpretability, and domain generalization, while enabling the curation of new and multimodal benchmarks. These frameworks surpass manual thresholding and binary scoring, addressing both causality ("Why"), classification ("What"), and temporality ("When") in anomaly understanding. Key challenges remain in scalability, fine-tuning costs, and the robustness of cross-modal alignment, especially in long-duration or ambiguous surveillance footage.

Future directions toward deployable, generalizable WVAD systems include integrating lightweight backbones (e.g., FastViTHD, HiLoViT)t, expanding datasets for open-world learning, and refining semantic alignment with motion cues and temporal reasoning. Unified benchmarks with evaluation metrics for both class-wise detection and explanation quality (e.g., BLEU, CIDEr, AnomEval) are essential to quantify progress in explainable and causally-aware VAD.

5.9. Misc

This section highlights several noteworthy WVAD approaches that either do not neatly fall into the previous categories or represent promising new directions for future research.

REWARD [344] proposes an end-to-end WVAD framework that transforms video-level labels into frame-level pseudo-labels for binary classification, similar to ANMIL [291] and 2Stage-based methods (Section 5.5). The method consists of a three-stage self-supervised pipeline designed to bypass memory limitations of metric-learning approaches like MIR [21]. Initially, a kkNN classifier compares features from anomalous videos to a nominal feature bank, identifying coarse pseudo-labels with temporal smoothing. These are refined via an MLP classifier trained on high-confidence segments. Final pseudo-labels enable joint fine-tuning of a transformer-based feature extractor (Uniformer-32 [116]) and a binary classifier using BCE loss. By removing the need for feature aggregation and optimizing both components together, REWARD achieves real-time inference (32 frames @ 5 fps, 6.4 s window) while surpassing RTFM-based methods (Section 5.3). Notably, the authors exclude the first and last 20% of segments in UCF-Crime training videos to eliminate noise from static banners.

GlanceVAD [43] (Fig. 32) presents a paradigm shift in annotation efficiency with its "Glance Supervision" approach. It requires annotators to mark only a single frame representing the anomaly, significantly reducing annotation effort and subjectivity. Adapting Gaussian Splatting from 3D scene representation, they utilize Gaussian as the core anomaly representation, focusing on temporal kernel optimization. The resulting 1D anomaly score serves as smoother pseudo-labels from sparse glance annotations. GlanceVAD seamlessly integrates with existing MIL-based weakly supervised methods, demonstrating significant performance improvements over state-of-the-art VAD techniques. Achieving 91.96% AUC on UCFC and 89.40% AP on XDV, it showcases a strong balance between annotation cost and performance. While acknowledging potential limitations in baseline method selection (RTFM and UR-DMU), the authors highlight their established performance in weakly supervised settings.

Video Anomaly Retrieval (VAR) [237]: This work introduces the task of VAR, shifting the focus from anomaly detection to retrieving relevant videos containing specific anomalous events based on cross-modal queries, such as text descriptions. This approach addresses the practical need for efficiently searching large video archives for particular anomalies.

UCF-Crime Annotation (UCFA) [250]: Recognizing the limitations of traditional VAD methods in semantic understanding, a richly annotated version of the UCFC dataset is proposed, including 23,542 fine-grained sentence descriptions and temporal annotations. The authors demonstrate the dataset's utility through comprehensive benchmarking of multiple video–language understanding tasks, including temporal sentence grounding (TSGV), video captioning (VC), dense video captioning (DVC), and multimodal anomaly detection (MAD). For MAD specifically, they propose an enhanced TEVAD [308] framework that incorporates general and surveillance-specific Swinbert models for caption generation. Experiments show that state-of-the-art models trained on standard video datasets underperform on surveillance footage. However, domain-specific caption features from UCFA-trained models complement visual cues to boost anomaly detection, enabling multimodal learning for richer semantic understanding in surveillance contexts.

Explainable Video Anomaly Localization (EVAL) [345] is a transparent VAD framework that detects and explains anomalies using exemplar-based nearest-neighbour matching and interpretable attributes such as motion, appearance, and spatial context. It leverages pretrained models (e.g., MS-COCO) to extract semantic features and selects exemplar instances from normal videos for comparison. Anomalies are identified when test instances deviate from these exemplars in attribute space, with explanations tied to specific mismatches (e.g., a stationary car in traffic). While EVAL is effective and interpretable in simple scenes (e.g., Street Scene, UCSD Ped2), it has limited evaluation on complex interaction anomalies.

ComplexVAD [346] detects interaction-based anomalies using scene graphs to model object relationships. Each frame is represented as

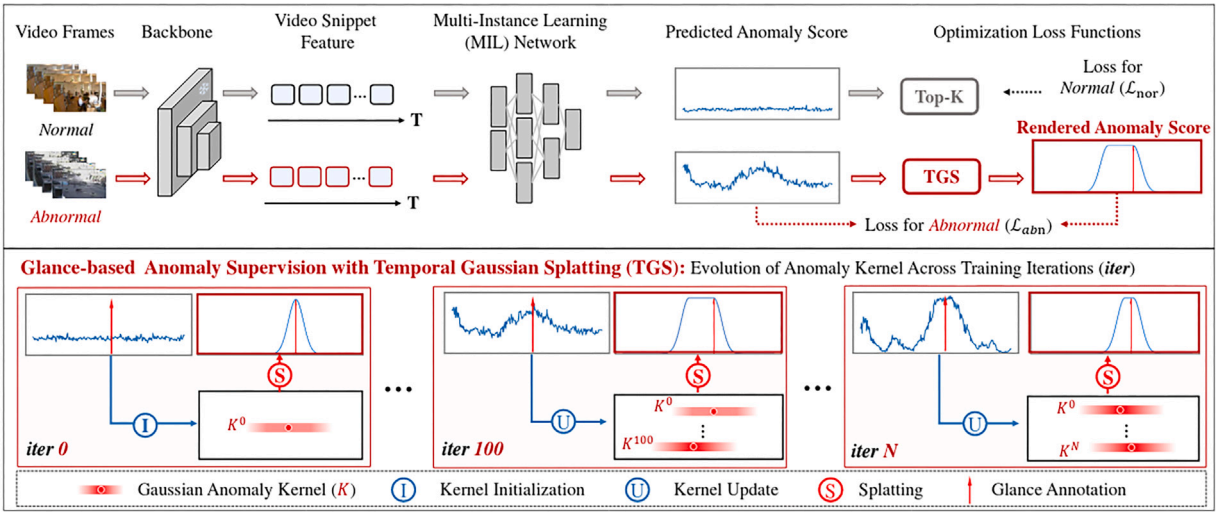


Fig. 32. GlanceVAD [43] proposed method.

a graph with nodes (objects) and edges (spatial–temporal relations), enriched with attributes like trajectory and pose (using Detectron2 for detection and pose, ByteTrack for tracking). Normal interactions are captured via exemplar node pairs and isolated nodes from nominal videos, filtered by distance metrics. Anomalies are flagged when test instances deviate from these exemplars, emphasizing interaction irregularities over single-object behaviour. Evaluated on a custom dataset of 217 videos, the method uses frame-level AUC, RBDC, and TBDC for performance assessment.

A **Frequency Enhanced (FE-VAD)** [347], addresses the limitation of existing methods that rely solely on spatio-temporal features and integrates temporal strengthening and a novel frequency-domain analysis: (1) A Temporal Strengthening Network (TSN) employs masked self-attention to model global dependencies and local 1D convolutions to refine temporal features, prioritizing future anomaly prediction by suppressing past influences. (2) A High–Low Frequency Enhancement Network (HLFN) decomposes features via Fourier transforms into high-frequency (detail-sensitive) and low-frequency (global-context) components, enhanced through Gaussian filtering and adaptive convolutions in both temporal and spatial domains. These complementary frequency features are fused and optimized via a High–Low MIL Loss, which differentially weights snippet selection. During inference, a video-specific scaling and smoothing strategy dynamically adjusts anomaly scores using pseudo-labels derived from training data, refining frame-level predictions. Evaluated on ST, UCFC, and XDV, FE-VAD demonstrates that frequency analysis effectively complements spatio-temporal modelling to detect subtle/local and persistent/global anomalies. The work pioneers frequency-domain integration in WSVAD, offering a robust solution for real-world surveillance systems requiring adaptability to diverse anomaly types.

VADMamba [348] pioneers the use of state space models (SSMs), specifically Mamba [349], for fast VAD. It combines frame prediction and optical flow reconstruction in a hybrid framework. The proposed VQ-Mamba Unet (VQ-Mau) compresses normal features via vector quantization (VQ) and uses Non-negative Vision State Space (NVSS) blocks with pre-activation (Relu → Conv → BN) to accelerate convergence. A clip-level fusion strategy dynamically selects optimal anomaly scores (frame or flow-based) per video segment. Ablations confirm the efficacy of VQ and NVSS in balancing speed and performance, with Mamba's linear scalability enabling efficient long-range modelling. Although further evaluations on richer datasets are needed.

The **UCFDVS** [260] dataset represents the first event-based benchmark for VAD, utilizing Dynamic Vision Sensors (DVS) to capture asynchronous, sparse event streams with high temporal resolution (1280 × 720, 242s/video). Unlike RGB data, DVS encodes *ON/OFF* polarity changes, reducing redundancy and enhancing motion sensitivity, particularly useful for dynamic anomalies. To leverage this modality, the authors introduce a Multi-Scale Spiking Fusion Network (MSF) based on Spiking Neural Networks (SNNs), which are well-suited for processing discrete event data. MSF incorporates pyramidal dilated convolutions to extract multi-scale spiking features, a Spiking Graph Convolutional Network (GCN) to model global temporal dependencies, and a Temporal Interaction Module (TIM) to fuse historical and current spike states. The work establishes a new direction for efficient, temporally-aware anomaly detection, showcasing the promise of event cameras and SNNs for real-time, motion-centric VAD.

The **Poly-modal Induced VAD (Pi-VAD)** [350] (Fig. 33) framework enhances WVAD by integrating five modalities — pose, depth, panoptic masks, optical flow, and text — into RGB analysis through a efficient two-module architecture, the Poly-modal Inductor.

At its core, the Pseudo Modality Generation (PMG) module synthesizes modality-specific embeddings directly from RGB features using an encoder–decoder structure. The encoder, a 1D convolutional layer, projects RGB snippet features into a shared latent space, while five parallel decoder (linear layer + 1D convolution) generate pseudo-modality embeddings by translating latent representations. This design eliminates reliance on external modality backbones during inference, relying instead on MSE reconstruction loss during training to align pseudo-modalities with ground-truth embeddings extracted from pre-trained models (pose: YOLOV7-posewangYOLOv72022, depth: DepthAnythingV2yangDepth2024, panoptic mask: SAMkirillovSegment2023, optical-flow: RAFTteedRAFT2020, text modality ground-truths: VifiCLIPrasheedViFiCLIP2023). The shared encoder ensures cross-modal consistency, while decoders distil task-specific cues, reducing redundancy and noise inherent in raw multimodal data.

The Cross Modal Induction (CMI) module aligns pseudo-modalities with RGB through a two-stage process: (1) snippet-level bi-directional InfoNCE loss ($\mathcal{L}_{align}$), enforces fine-grained alignment by maximizing cosine similarity for same-snippet pairs while repelling dissimilar pairs, and (2) distillation loss ($\mathcal{L}_{distill}$) leverages a frozen teacher model — pre-trained on the WSVAD task — to guide the student by minimizing the MSE between the student's refined multi-modal features ($\mathcal{F}_M^*$) and the teacher's RGB features ($\mathcal{F}_{teach}$).

The Poly-modal Inductor processes the output representation from Student's Block-i and injects the refined multi-modal feature into Student's Block i+1, regardless of the student's specific block. It can be integrated throughout the teacher–student architecture (authors placed in initial/final blocks to capture low/high level multi-modal features).

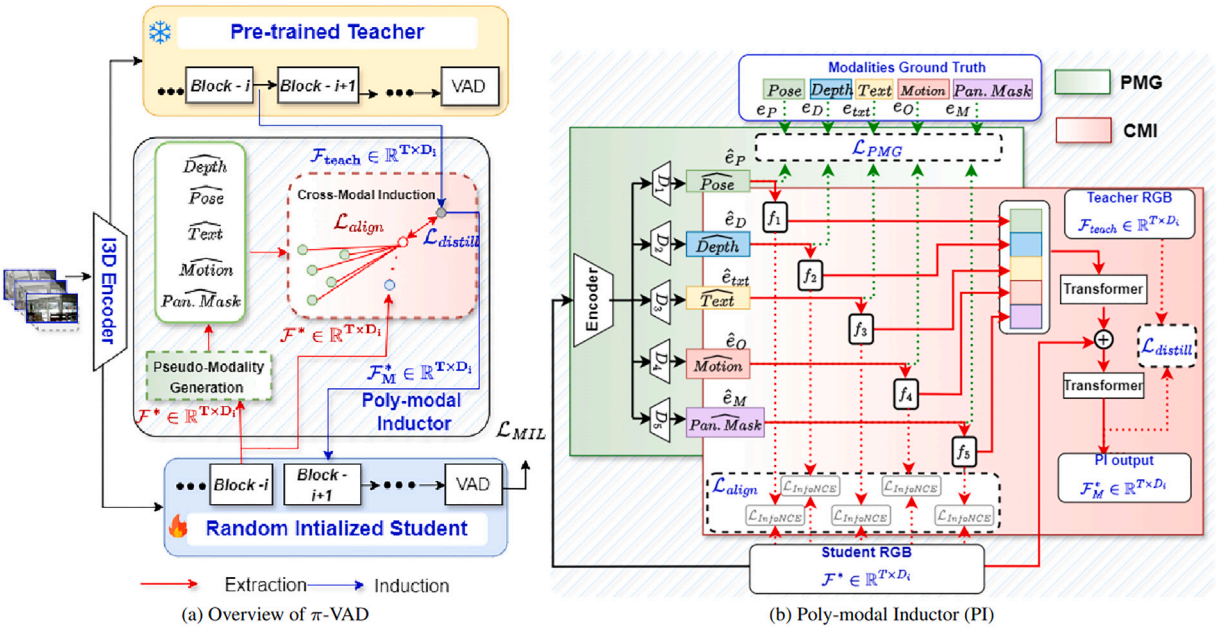


Fig. 33. Pi-VAD [350] proposed method.

Training uses a teacher–student setup: the teacher (pre-trained WS-VAD model) remains frozen, while the student undergoes two phases: warm-up ($\mathcal{L}_{PMG} + \mathcal{L}_{align} + \mathcal{L}_{distill}$) initializes modality synthesis/alignment, and task training combines MIL loss ($\mathcal{L}_{MIL}$) with auxiliary losses. At inference, only the Student and PI's run, generating pseudo-modalities solely from RGB (19.88 GFLOPs vs. 2561 GFLOPs for raw modalities) at 30 FPS, making it practical for CCTV and surveillance systems. PI-VAD achieves 90.33% AUC on UCF-Crime (+2.75% over RGB baselines) and 85.37% AP on XD-Violence, with depth (spatial context) as key contributor. Ablations show PMG+CMI jointly boost AUC by 5.67%, while distillation adds 1.45%. Efficiency and robustness in real-world anomalies (e.g., explosions, shoplifting) validate its applicability.

The **Latency-aware Average Precision (LaAP)** [11] metric redefines VAD evaluation by prioritizing timely detection, crucial for real-world safety systems. Traditional metrics (AUC/AP), while effective for frame-wise classification, neglect the temporal nature of anomalies, failing to distinguish between early and late detections within an event.

LaAP integrates a time-decaying weighting mechanism into the precision–recall framework, penalizes delayed predictions and rewarding models that identify anomalies closer to their onset through adaptive scoring and sparse sampling. LaAP assumes that only one abnormal event exists in a video due to the sparsity of anomalies (not ready for XDV). Alongside LaAP, the authors tackle annotation bias through multi-round averaged AUC/AP, mitigating inconsistencies arising from subjective labelling practices. To further address dataset limitations, they introduce synthetic benchmarks (UCF-HN, MSAD-HN) generated via diffusion models, which simulate normal behaviour within anomaly-prone scenes to rigorously test scene overfitting.

Ablation studies showed:

- Overfitting: Models like CLIP-TSA [194] and VadCLIP [40] showed 44%–100% FAR on synthetic benchmarks, despite low FAR on original data, revealing heavy reliance on scene biases.
- Latency Matters: LaAP uncovered performance gaps invisible to AUC/AP; e.g., MGFN [119] and URDMU [278] lagged in timeliness despite competitive comparable AUC/AP scores.
- Supervision Gap: Methods with extra supervision (e.g., Holmes-VAU [13]) resisted overfitting, while weakly supervised models faltered, highlighting the need for robust training paradigms.
- Annotation Bias Mitigation: Models like RTFM [89] and PEL [39] showed significant ranking shifts under averaged AUC/AP, validating the necessity of this approach for reliable evaluation.

By harmonizing temporally aware metrics with synthetically augmented data, this work pioneers a holistic framework that bridges data quality and metric reliability, advancing the development of robust, generalizable VAD systems capable of nuanced temporal reasoning and scene-agnostic performance.

In sum. , these diverse approaches highlight the ongoing evolution of WVAD, expanding beyond traditional paradigms to encompass new tasks, annotation strategies, integrating powerful language models, new modalities and evaluation metrics. Exploring these emerging directions holds significant potential for advancing the field towards more efficient, interpretable, and adaptable anomaly detection systems.

6. Edge deployability considerations in weakly supervised VAD

Despite the increasing sophistication of Weakly Supervised VAD methods, their deployability at the edge is an essential dimension often overlooked in current reviews. Real-world surveillance systems rely on edge devices with constrained computational resources, memory, and power budgets [351]. Therefore, the practical utility of WSVAD techniques is tightly coupled with their ability to operate efficiently under such limitations.

Edge deployment imposes several critical requirements: low-latency inference, minimal memory footprint, model compression, and the ability to run without reliance on high-throughput cloud infrastructure [352]. Many state-of-the-art WSVAD models leverage deep architectures with millions of parameters, often requiring GPUS and substantial bandwidth for video processing, making direct edge deployment difficult [353,354].

Edge VAD challenges in [355] highlights the increasing importance of edge devices for real-time video VAD in smart cities. It contrasts traditional, centralized approaches with edge-based solutions that offer greater efficiency, privacy, and scalability. The paper categorizes existing methods, emphasizing the context-sensitive and time-critical nature of VAD, while also noting key challenges of edge deployment, such as limited computational resources, energy, and memory constraints.

Benchmarking is the focus of [352], discussing an end-to-end VAD detection system implemented on various NVIDIA Jetson edge devices, focusing on performance optimization and real-time analysis for surveillance applications. It highlights the efficiency and effectiveness of these devices in real-time surveillance applications, employing a weakly supervised model previously identified, the RTFM [89] for benchmarking comparison.

Lightweight and robust framework for anomaly detection in large-scale surveillance video data is presented in [356]. The approach combines 2D-CNNS for video feature extraction, autoencoders for representation learning, and Echo State Networks (ESNS) for sequence modelling and anomaly detection. It was specifically designed for deployment on edge devices, the framework supports secure and intelligent surveillance. When evaluated on challenging surveillance datasets, the method outperformed several existing approaches in terms of performance and efficiency.

Vision Transformer Anomaly Recognition (ViT-ARN) framework [357] enables anomaly detection and classification in smart city surveillance videos through a two-stage process: a lightweight one-class neural network performs online anomaly detection, followed by classification of detected events. To ensure edge compatibility, the model is compressed using geometric median-based filter pruning, and refined features are analysed via a multi-reservoir Echo State Network for recognizing complex anomalies like vandalism and traffic incidents.

Online Active Learning (OAL) is explored by [358], introducing a framework for deep neural networks on edge devices, using Singular Value Decomposition (SVD) to assess model quality and trigger retraining without ground truth or teacher labels. It supports efficient teacher–student knowledge distillation and intelligent frame selection to maintain real-time performance and reduce overfitting. Evaluated on human pose estimation and object detection tasks using models like YOLO and ResNet on NVIDIA Jetson NX, the approach significantly reduces unnecessary retraining.

Fast-DAVAD [359] framework addresses VAD on AIoT-enabled edge devices by tackling domain shift through multi-level adversarial domain adaptation in an unsupervised, lightweight manner. Designed for resource-constrained environments, it uses a Residual U-Net and a memory module to ensure low latency and maintain accuracy without relying on labelled real-world data. Experiments on public datasets and platforms like edge NVIDIA Jetson devices confirm its efficiency and performance compared to state-of-the-art methods.

EdANo-Vision [360] VAD system using edge computing for real-time surveillance analysis, supported on NVIDIA's Jetson Orin Nano 8 GB. The edge device supports the I3D [33] model, which detects anomalies, providing probabilities and confidence percentages. The model processes video frames at 240 × 230 pixels and 30 FPS, and the system's performance is evaluated using the UCFC dataset. The Integrated edge-based anomaly detector effectively communicates detected anomalies to users.

Table 10 summarizes the various approaches focused on edge VAD (see Table 13).

Key Points of Edge VAD Methods: These methods leverage the power of small embedded devices to enhance anomaly detection in the edge devices, minimizing information flow and bandwidth, while maintaining VAD accuracy.

- **Lightweight:** Employs the RTFM model for anomaly detection, optimized for edge deployment on Jetson Nano with low power consumption (41.7 W). Prioritizes efficiency and real-time inference while maintaining competitive AUC (84.39).
- **OAL:** Proposes a teacher–student framework incorporating Singular Value Decomposition (SVD) and Echo State Networks. Focuses on unsupervised representation learning without explicit edge deployment, emphasizing low-rank modelling and temporal abstraction.
- **Fast-DAVAD:** Utilizes a residual U-Net with multi-level adversarial domain adaptation. Integrates an autoencoder reconstruction scheme and is deployed on Jetson hardware, balancing inference efficiency (7.14 W) with robust domain transfer capabilities.
- **EdANo-Vision:** Builds on I3D and Inflated 3D CNNs, enhanced with depth-wise convolution and 8-bit compression for hardware-aware deployment on Jetson. Demonstrates a trade-off between accuracy (AUC 81.0) and power efficiency (15 W), targeting efficient high-dimensional spatiotemporal modelling.

Edge-oriented VAD systems prioritize lightweight architectures such as MobileNet or pruned models to meet the constraints of low-power hardware like Raspberry Pi or Jetson Nano. By processing data locally, they enhance privacy, reduce latency, and minimize bandwidth usage by transmitting only alerts or metadata. These energy-efficient systems support scalable, decentralized surveillance networks but face challenges in maintaining detection accuracy under computational constraints and in updating models on-device.

Emerging techniques—such as compression, feature distillation, and federated learning aim to improve adaptability and performance at the edge. Without assessing deployability, conclusions about a method's real-world applicability remain incomplete. Future reviews and benchmarks should incorporate edge-aware evaluations, including runtime efficiency, memory footprint, and performance on resource-constrained hardware.

7. Emerging privacy challenges in Weakly Supervised Video Anomaly Detection (WSVAD)

As VAD becomes increasingly prevalent in domains such as surveillance, healthcare, and smart cities, the adoption of weak supervision has facilitated more scalable and cost-effective model development. Nevertheless, this approach introduces significant privacy concerns that are frequently neglected. Weakly supervised VAD typically relies on large volumes of minimally annotated video data, which raises serious issues regarding data governance and personal identity protection [361, 362], something addressed in more recent benchmarks [12,15,261].

A key privacy risk arises from the nature of real-world video footage, which, unlike anonymized datasets, may contain personally identifiable information such as faces, gait patterns, vehicle license plates, and contextual details of private behaviour. Since weak supervision reduces annotation requirements, it often involves using raw, unfiltered video, thereby amplifying privacy risks [363]. Even in cases where raw footage is not directly shared, model parameters or gradient updates may leak sensitive information via model inversion or membership inference attacks [364], or transfer attacks [365]. This is particularly concerning in high-stakes environments like hospitals or public transport, where trust and data protection must be prioritized [366].

Another emerging concern is the lack of standardized frameworks for privacy-preserving weak supervision [367,368]. While techniques such as differential privacy, homomorphic encryption, and synthetic data generation are being explored in supervised contexts, their adaptation to weakly labelled, sequential video data remains an open research area [368]. These methods must also balance the trade-off between privacy and anomaly detection performance, as over-sanitization of the data may obscure subtle but important behavioural deviations that are relevant to be detected [369].

Privacy-Preserving Video Anomaly Detection (P2VAD) survey [368] identified those challenges addressing the growing concerns over personal privacy in surveillance-based VAD, which often captures identifiable information. Traditional VAD research lacks transparency and interpretability, hindering its real-world adoption. This article offers the first comprehensive review of P2VAD, defining its scope, categorizing existing approaches, and analysing their assumptions and effectiveness.

Table 13
Summary of Edge-Oriented VAD Works.

Method	Power		FM	MIL	SC	Edge	Metrics	
	Watts	GFLOPs					UCFC	XDV
Lightweight	✗	41.733	Detection RTFM model	✗	AUC	Jetson Nano	84.39	✗
OAL	✗	✗	Teacher-Student & Singular Value Decomposition (SVD) & Echo State Network	✗	AUC	✗	✗	✗
Fast-DAVAD	7.142	✗	Residual U-Net & Multi-level adversarial domain	✗	AUC	Autoencoder Reconstruct & Jetson	✗	✗
EdANo-Vision	15	✗	I3D & Inflated & 3D CNN	✗	AUC	Depth-wise convolution & 8-bit Compression & Jetson	81.0	✗

Power: Approximate consumption; **FM**: Feature Modulator; **MIL**: Multiple Instance Learning; **SC**: Score Metric; **Edge**: Uses edge supervision or strategy used for deploy on edge hardware; **SAtt**: Self-Attention; **XMod**: Cross-modality; **UCFC/XDV**: Benchmark datasets.

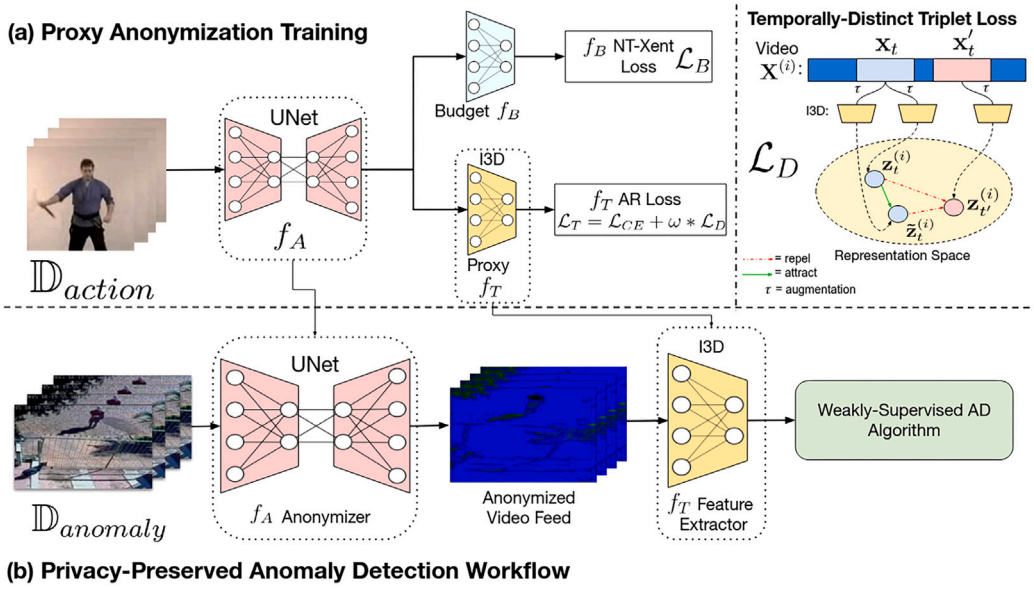


Figure 2: Full TeD-SPAD framework consisting of the proxy anonymization training followed by the privacy-preserved anomaly detection. (a) shows this proxy training, where UNet is used to anonymize frames in such a way that reduces mutual information between frames while maintaining utility performance. We complement the standard cross-entropy loss with our proposed temporally-distinct triplet loss, which enforces a difference in clip features at distinct timesteps. After training the anonymizer and feature extractor, (b) shows the privacy-preserved workflow, where the anomaly dataset videos are passed through the proxy-trained f_A, f_T, then into any WSAD algorithm.

Fig. 34. Ted-SPAD [370] proposed method.

Ted-SPAD [370] (Fig. 34) proposes a privacy-aware framework that anonymizes visual data in a self-supervised way while preserving anomaly detection accuracy. By introducing a temporally distinct triplet loss, the method enables the anonymization training of the model. Tested on three VAD benchmark datasets, ST, UCFC, and XDV [22], and [371] to evaluate the leakage of privacy attributes. Ted-SPAD achieves strong privacy-utility trade-offs, reducing private attribute prediction with minimal impact on anomaly detection performance.

Moving towards **human-centric VAD**, [372,373] a novel dataset prioritizing privacy in human activity analysis, by enforcing the use of features, such as pose, trajectory, and optical flow, to protect individual privacy and prevent discrimination against minority groups. These features are alternatives to raw pixel data, potentially improving model generalizability and performance. Alongside UCAL (Unsupervised Continual Anomaly Learning)—an innovative framework that empowers models to evolve continuously through incremental learning. This approach facilitates a seamless transition from offline training to dynamic real-world implementation, addressing the critical gap between static model development and adaptive deployment scenarios.

Recent efforts have turned toward systematically studying privacy across data, features, and system levels. However, existing work remains fragmented and mainly focused on RGB-based methods, neglecting privacy leakage and appearance bias. Although many works try to address the privacy problem in VAD, future research must prioritize the development of privacy-aware weakly supervised learning methods that incorporate anonymization, access control, and ethical auditing directly into the training pipeline. Furthermore, regulatory frameworks

like the GDPR and HIPAA underscore the need for transparency, consent management, and risk mitigation strategies in any system that processes personal video data, regardless of the supervision level.

8. Overall conclusions and limitations

This survey has undertaken a comprehensive exploration of VAD, examining both the theoretical underpinnings and practical challenges of this rapidly evolving field. Tracing the trajectory of the paradigm of VAU, we reveal a dynamic convergence of methodologies—where once-disparate components and abnormality criteria mostly in a hybrid setting, and recently increasing the multimodal systems that incorporates textual cues, causal reasoning, and cross-modal alignment.

The experimental taxonomy proposed in this work yields valuable insights into how data curation, feature representations, sampling strategies, and optimization techniques collectively influence performance across all stages of VAU. A defining trend is blending feature extraction and modulation with novel additions — such as language-grounded encoders or multimodal vision–language models (VLMs) — augmenting traditional pipelines. This crisscross of plug-and-play modules signals the field's maturation into a holistic ecosystem. Notably, recent efforts explore dynamic sampling strategies, like the SBS (Sample-Batch Selection) [279], which adaptively samples more relevant video segments, especially when combined with powerful image–text pre-trained models or cutting-edge VLM encoders Section 3.1.3.

Despite these advances, several core challenges persist. Chief among them is the absence of unified, modular experimentation platforms. We advocate for a standardized, open-source framework to integrate plug-and-play modules, network-agnostic methods, and supervision strategies (e.g., weakly supervised, open-world, and LLM-driven). Such a framework would support ablation studies and modular benchmarking while promoting cross-dataset evaluation to better assess model robustness across heterogeneous surveillance domains.

A significant obstacle to practical deployment remains the high False Alarm Rate (FAR), particularly prevalent in weakly supervised methods that rely on imprecise video-level labels, most of the time forgotten [39,86–88,95,270,278]. Models often struggle to localize anomalies without accurate temporal annotations, leading to unreliable detections and alert fatigue in real-world applications. Moreover, most current works prioritize coarse metrics like AUC or AP, overlooking semantic interpretability, cross-scenario generalization, and anomaly-specific reasoning. Some works have begun incorporating class-wise metrics [39,40,43,111,195,196,279,288,320,350] and fine-grained mAP [40,42,196,317] to address inter-class variability, and spatial anomaly location (Temporal Intersection-over-Union (TIoU)) [30,253,315], yet inconsistencies persist due to differing AP calculation methodologies. The method of calculating Average Precision (AP) — either via trapezoidal interpolation, which risks overestimation by assuming linear precision–recall (PR) segments, or the non-interpolated approach that weights precision at each recall threshold — critically impacts performance estimates in imbalanced datasets, with the latter offering a more conservative and reliable measure.

Emerging metrics like AnomEval [241], M-VAE [253] and LaAP [11] build on these insights by integrating causal reasoning and temporal sensitivity, addressing gaps left by conventional metrics. LaAP explicitly rewards early anomaly detection through time-decayed scoring, while AnomEval introduces causal localization metrics to disentangle spurious correlations. These advancements highlight the need for enriched evaluation protocols that unify fine-grained mAP classification, GPT-4o-guided coherence analysis, and anomaly-specific reasoning.

Benchmarks also evolve to reflect real-world complexity, as seen by UCFA [250], MSAD [12], HIVAU-70K [13], M-VAE [253], UCFVL [14] and SurveillanceVQA-589K [15], diversify anomaly types, contexts and annotations/instructions, synthetic benchmarks (e.g., VANE [261], UCF-HN/MSAD-HN [11]) expose latent biases like scene overfitting through diffusion-generated normal videos. Incorporating multimodal signals, particularly audio–language embeddings Section 3.2 into VAD pipelines [42,295], further promises to enhance detection fidelity, as seen in ECVA [241], and UCFDVS [260].

Existing frameworks based on Multiple Instance Learning (MIL) [21] remain influential, yet their reliance on discrete, non-differentiable operations (e.g., top-k selection) using noisy selection criteria (e.g. scores, feature magnitude) often undermines stability and generalization. These strategies typically neglect temporal context and smooth decision boundaries, making them problematic in noisy settings. Recent advances like REWARD [344] address these limitations by adopting end-to-end training with frame-level pseudo-labels, akin of 2-stage works (Section 5.5), bypassing the memory bottlenecks of metric learning losses [95,268,295,304,314] while enabling joint optimization of feature extractors/classifiers. A deeper analysis is necessary, along with the investigation of different alternatives that better capture temporal coherence and anomaly semantics (e.g. DEN's dynamic erasing [289], AnomCLIP feature space vision–text alignment [196], batch-based DFM anomaly metric [279], MELOW's Multimodal Evidential Collaborative Learning [5] and PEMIL's Abnormal-Aware Prompt Learning [41]).

Furthermore, real-world deployment demands emphasize the need for efficient inference—recent works [10,344,350] highlight trade-offs between detection accuracy, computational throughput (FPS), Params, and MACS underscoring the importance of lightweight architectures and hardware-aware optimizations for edge/cloud deployment and real-time inference.

Looking forward, we propose an extension of our current work with UWS4VAD—a modular, extensible platform to advance the state of video anomaly understanding through:

- Unified Modular Architecture:
  - Network-agnostic design supporting plug-and-play components (feature extractors/modulators, loss functions, distillation layers [42,295], teacher–student [350]).
  - Multimodal data handling (RGB, audio, text, event) with hierarchical configuration for reproducibility.
  - Flexible interplay between datasets, supervision paradigms (weak, glance, instruction, train-free), and optimization strategies.
- Comprehensive Benchmark Suite:
  - Standardized splits, cross-dataset protocols and open-world settings.
  - Multi-granular evaluation: detection (AUC/AP/LaAP/F2/FNR), classification (mAP), reasoning (AnomEval), scene overfitting (FAR), computational throughput (GFLOPs).
- Collaborative Research Hub:
  - Centralized tracking of model performance with systematic results and cross-model comparison (plots, metrics, visualizations, att/embeddings maps).
  - Version-controlled modules for inference/training with deployment mechanisms addressing efficiency (edge/cloud).
  - Documentation portal for WVAD/VAU advancements, integrating community contributions.

The framework advocates a configuration-driven, reusable engine for end-to-end experimentation, bridging data, models, and evaluation into a single reproducible workflow.

This initiative invites the community to join development and contributions of new methods, datasets, evaluation protocols, and real-world applications. By fostering reproducibility, comparability, and interdisciplinary collaboration, we aim to shift VAD from a fragmented landscape of isolated methods into a robust, interpretable, and deployable discipline.

CRediT authorship contribution statement

Rui Z. Barbosa: Writing – original draft. **Hugo S. Oliveira:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **João Manuel R.S. Tavares:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hugo S. Oliveira reports financial support was provided by Foundation for Science and Technology. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financed by the Portuguese Foundation of Science and Technology — FCT under the Ph.D grant “2021.06275.BD”. We thank Junxi Chen and Yujiang Pu for their valuable feedback on late drafts of this survey.

Data availability

No data was used for the research described in the article.

References

[1] Jing Liu, Yang Liu, Jieyu Lin, Jielin Li, Peng Sun, Bo Hu, Liang Song, Azzedine Boukerche, Victor C.M. Leung, NSVAD: Networking systems for video anomaly detection: a tutorial and survey, 2024, arXiv:2405.10347 http://arxiv.org/abs/2405.10347 [cs].

[2] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Liang Song, GVAD: generalized video anomaly event detection: systematic taxonomy and comparison of deep models, 2023, arXiv:2302.05087 http://arxiv.org/abs/2302.05087 [cs].

[3] Xi Ding, Lei Wang, SURV-vllm-VAD: Quo vadis, anomaly detection? LLMs and VLMs in the spotlight, 2024, http://dx.doi.org/10.48550/arXiv.2412.18298, arXiv:2412.18298, http://arxiv.org/abs/2412.18298 [cs].

[4] Yuansheng Zhu, Wentao Bao, Qi Yu, OpenVAD: towards open set video anomaly detection, 2022, http://dx.doi.org/10.48550/arXiv.2208.11113, arXiv:2208.11113, http://arxiv.org/abs/2208.11113 [cs].

[5] Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, Bob Zhang, MELOWVAD: Multimodal evidential learning for open-world weakly-supervised video anomaly detection, IEEE Trans. Multimed. (2025) 1–12, http://dx.doi.org/10.1109/TMM.2025.3557682, https://ieeexplore.ieee.org/document/10948323/.

[6] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, Yanning Zhang, OVVAD: Open-vocabulary video anomaly detection, 2024, http://dx.doi.org/10.48550/arXiv.2311.07042, arXiv:2311.07042, http://arxiv.org/abs/2311.07042 [cs].

[7] Chenting Xu, Ke Xu, Xinghao Jiang, Tanfeng Sun, PLOVAD: prompting vision-language models for open vocabulary video anomaly detection, IEEE Trans. Circuits Syst. Video Technol. (2025) http://dx.doi.org/10.1109/TCSVT.2025.3528108, 1–1, https://ieeexplore.ieee.org/document/10836858/?arnumber=10836858.

[8] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, Elisa Ricci, LAVAD: Harnessing large language models for training-free video anomaly detection, 2024, http://dx.doi.org/10.48550/arXiv.2404.01014, arXiv:2404.01014, http://arxiv.org/abs/2404.01014 [cs].

[9] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, Xiaochen Wang, Hao Tang, Yan Wang, Shuyan Li, EventVAD: Training-free event-aware video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2504.13092, arXiv:2504.13092, http://arxiv.org/abs/2504.13092 [cs].

[10] Hyogun Lee, Haksub Kim, Ig-Jae Kim, Yonghun Choi, Flashback: Memory-driven zero-shot, real-time video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2505.15205, arXiv:2505.15205, http://arxiv.org/abs/2505.15205 [cs].

[11] Zihao Liu, Xiaoyu Wu, Wenna Li, Linlin Yang, LaAP: rethinking metrics and benchmarks of video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2505.19022, arXiv:2505.19022, http://arxiv.org/abs/2505.19022 [cs].

[12] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, Chen Chen, MSAD: Advancing video anomaly detection: a concise review and a new dataset, 2024, http://dx.doi.org/10.48550/arXiv.2402.04857, arXiv:2402.04857, http://arxiv.org/abs/2402.04857 [cs].

[13] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, Nong Sang, Holmes-VAU: towards long-term video anomaly understanding at any granularity, 2024, http://dx.doi.org/10.48550/arXiv.2412.06171, arXiv:2412.06171, http://arxiv.org/abs/2412.06171 [cs].

[14] Haoran Chen, Dong Yi, Moyan Cao, Chensen Huang, Guibo Zhu, Jinqiao Wang, UCVL: a benchmark for crime surveillance video analysis with large models, 2025, http://dx.doi.org/10.48550/arXiv.2502.09325, arXiv:2502.09325, http://arxiv.org/abs/2502.09325 [cs].

[15] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, Tongtong Yuan, Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models, 2025, http://dx.doi.org/10.48550/arXiv.2505.12589, arXiv:2505.12589, http://arxiv.org/abs/2505.12589 [cs].

[16] Bharathkumar Ramachandra, Michael J. Jones, Ranga Raju Vatsavai, A survey of single-scene video anomaly detection, 2020, arXiv:2004.05993 http://arxiv.org/abs/2004.05993 [cs].

[17] Guansong Pang, Chunhua Shen, Longbing Cao, Anton van den Hengel, Deep learning for anomaly detection: a review, ACM Comput. Surv. 54 (2) (2021) 1–38, http://dx.doi.org/10.1145/3439950, arXiv:2007.02500 http://arxiv.org/abs/2007.02500.

[18] Rashmiranjan Nayak, Umesh Chandra Pati, Santos Kumar Das, A comprehensive review on deep learning-based methods for video anomaly detection, Image Vis. Comput. 106 (2021) 104078, http://dx.doi.org/10.1016/j.imavis.2020.104078, https://www.sciencedirect.com/science/article/pii/S0262885620302109.

[19] Nomica Choudhry, Jemal Abawajy, Shamsul Huda, Imran Rao, A comprehensive survey of machine learning methods for surveillance videos anomaly detection, IEEE Access 11 (2023) 114680–114713, http://dx.doi.org/10.1109/ACCESS.2023.3321800, https://ieeexplore.ieee.org/document/10271300/.

[20] Minqi Jiang, Chaochuan Hou, Ao Zheng, Xiyang Hu, Songqiao Han, Hailiang Huang, Xiangnan He, Philip S. Yu, Yue Zhao, Weakly supervised anomaly detection: A Survey, 2023, arXiv:2302.04549 http://arxiv.org/abs/2302.04549 [cs].

[21] Waqas Sultani, Chen Chen, Mubarak Shah, MIR: Real-world anomaly detection in surveillance videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 6479–6488, http://dx.doi.org/10.1109/CVPR.2018.00678, https://ieeexplore.ieee.org/document/8578776/.

[22] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, Zhiwei Yang, XDV-HLNET: Not only look, but also listen: learning multimodal violence detection under weak supervision, 2020, arXiv:2007.04687 http://arxiv.org/abs/2007.04687 [cs].

[23] Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, Sarah Vluymans, Multiple instance learning, in: Francisco Herrera, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, Sarah Vluymans (Eds.), Multiple Instance Learning: Foundations and Algorithms, Springer International Publishing, Cham, 2016, pp. 17–33, http://dx.doi.org/10.1007/978-3-319-47759-6_2.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, C3D: Learning spatiotemporal features with 3D convolutional networks, 2015, arXiv:1412.0767 http://arxiv.org/abs/1412.0767 [cs].

[25] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei, Sports-1M: Large-scale video classification with convolutional neural networks, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732, http://dx.doi.org/10.1109/CVPR.2014.223, https://ieeexplore.ieee.org/document/6909619.

[26] Hakan Bilen, Andrea Vedaldi, Weakly supervised deep detection networks, 2016, http://dx.doi.org/10.48550/arXiv.1511.02853, arXiv:1511.02853, http://arxiv.org/abs/1511.02853 [cs].

[27] Peng Tang, Xinggang Wang, Xiang Bai, Wenyu Liu, Multiple instance detection network with online instance classifier refinement, 2017, http://dx.doi.org/10.48550/arXiv.1704.00138, arXiv:1704.00138, http://arxiv.org/abs/1704.00138 [cs].

[28] Sujoy Paul, Sourya Roy, Amit K. Roy-Chowdhury, W-TALC: weakly-supervised temporal activity localization and classification, 2018, http://dx.doi.org/10.48550/arXiv.1807.10418, arXiv:1807.10418, http://arxiv.org/abs/1807.10418 [cs].

[29] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, Shih-Fu Chang, AutoLoc: weakly-supervised temporal action localization, 2018, http://dx.doi.org/10.48550/arXiv.1807.08333, arXiv:1807.08333, http://arxiv.org/abs/1807.08333 [cs].

[30] Kun Liu, Huadong Ma, BackBias: Exploring background-bias for anomaly detection in surveillance videos, in: Proceedings of the 27th ACM International Conference on Multimedia, in: MM ’19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1490–1499, http://dx.doi.org/10.1145/3343031.3350998, https://dl.acm.org/doi/10.1145/3343031.3350998.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, 2017, http://dx.doi.org/10.48550/arXiv.1711.07971, arXiv:1711.07971, http://arxiv.org/abs/1711.07971 [cs].

[32] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, Huadong Ma, T-C3D: Temporal convolutional 3D network for real-time action recognition, Proc. the AAAI Conf. Artif. Intell. 32 (1) (2018) http://dx.doi.org/10.1609/aaai.v32i1.12333, https://ojs.aaai.org/index.php/AAAI/article/view/12333.

[33] João Carreira, Andrew Zisserman, I3D: quo vadis, action recognition? a new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724–4733, http://dx.doi.org/10.1109/CVPR.2017.502, https://ieeexplore.ieee.org/document/8099985.

[34] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool, TSN: Temporal segment networks: towards good practices for deep action recognition, 2016, http://dx.doi.org/10.48550/arXiv.1608.00859, arXiv:1608.00859, http://arxiv.org/abs/1608.00859 [cs].

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool, TSN: Temporal segment networks for action recognition in videos, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2019) 2740–2755, http://dx.doi.org/10.1109/TPAMI.2018.2868668, https://ieeexplore.ieee.org/document/8454294.

[36] Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh, 3DResNet: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?, 2018, http://dx.doi.org/10.48550/arXiv.1711.09577, arXiv:1711.09577, http://arxiv.org/abs/1711.09577 [cs].

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, Learning deep features for discriminative localization, 2015, http://dx.doi.org/10.48550/arXiv.1512.04150, arXiv:1512.04150, http://arxiv.org/abs/1512.04150 [cs].

[38] Rui Z. Barbosa, Hugo S. Oliveira, A unified approach to video anomaly detection: advancements in feature extraction, weak supervision, and strategies for class imbalance, IEEE Access 13 (2025) 60969–60986, http://dx.doi.org/10.1109/ACCESS.2025.3557948, https://ieeexplore.ieee.org/document/10949172/.

[39] Yujiang Pu, Xiaoyu Wu, Shengjin Wang, PEL4vad: Learning prompt-enhanced context features for weakly-supervised video anomaly detection, 2023, arXiv:2306.14451 http://arxiv.org/abs/2306.14451 [cs].

[40] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, Yanning Zhang, VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection, 2023, arXiv:2308.11681 http://arxiv.org/abs/2308.11681 [cs].

[41] Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, Qingming Huang, PEMIL: prompt-enhanced multiple instance learning for weakly supervised video anomaly detection, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 18319–18329, http://dx.doi.org/10.1109/CVPR52733.2024.01734, https://ieeexplore.ieee.org/document/10657732/.

[42] Peng Wu, Wanshun Su, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, Yanning Zhang, AVadCLIP: Audio-visual collaboration for robust video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2504.04495, arXiv:2504.04495, http://arxiv.org/abs/2504.04495 [cs].

[43] Huaxin Zhang, Xiang Wang, Xiaohao Xu, Xiaonan Huang, Chuchu Han, Yuehuan Wang, Changxin Gao, Shanjun Zhang, Nong Sang, GlanceVAD: exploring glance supervision for label-efficient video anomaly detection, 2024, http://dx.doi.org/10.48550/arXiv.2403.06154, arXiv:2403.06154, http://arxiv.org/abs/2403.06154 [cs].

[44] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, AlexNet: ImageNet classification with deep convolutional neural networks, Neural Inf. Process. Syst. 25 (2012) http://dx.doi.org/10.1145/3065386.

[45] Karen Simonyan, Andrew Zisserman, VGG: Very deep convolutional networks for large-scale image recognition, 2015, http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv:1409.1556, http://arxiv.org/abs/1409.1556 [cs].

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Inception: going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9, http://dx.doi.org/10.1109/CVPR.2015.7298594, https://ieeexplore.ieee.org/document/7298594.

[47] Sergey Ioffe, Christian Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, http://dx.doi.org/10.48550/arXiv.1502.03167, arXiv:1502.03167, http://arxiv.org/abs/1502.03167 [cs].

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, InceptionV2V3: Rethinking the inception architecture for computer vision, 2015, arXiv:1512.00567 http://arxiv.org/abs/1512.00567 [cs].

[49] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi, Inception-v4, inception-ResNet and the impact of residual connections on learning, 2016, arXiv:1602.07261 http://arxiv.org/abs/1602.07261 [cs].

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, ResNet: Deep residual learning for image recognition, 2015, http://dx.doi.org/10.48550/arXiv.1512.03385, arXiv:1512.03385, http://arxiv.org/abs/1512.03385 [cs].

[51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, ResNeXt: Aggregated residual transformations for deep neural networks, 2017, arXiv:1611.05431 http://arxiv.org/abs/1611.05431 [cs].

[52] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, MobileNets: efficient convolutional neural networks for mobile vision applications, 2017, arXiv:1704.04861 http://arxiv.org/abs/1704.04861 [cs].

[53] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, MobileNetV2: inverted residuals and linear bottlenecks, 2019, arXiv:1801.04381 http://arxiv.org/abs/1801.04381 [cs].

[54] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, Searching for MobileNetV3, 2019, arXiv:1905.02244 http://arxiv.org/abs/1905.02244 [cs].

[55] Mingxing Tan, Quoc V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, 2020, arXiv:1905.11946 http://arxiv.org/abs/1905.11946 [cs, stat].

[56] Mingxing Tan, Quoc V. Le, EfficientNetV2: Smaller models and faster training, 2021, arXiv:2104.00298 http://arxiv.org/abs/2104.00298 [cs].

[57] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei, Deformable convolutional networks, 2017, http://dx.doi.org/10.48550/arXiv.1703.06211, arXiv:1703.06211, http://arxiv.org/abs/1703.06211 [cs].

[58] Xizhou Zhu, Han Hu, Stephen Lin, Jifeng Dai, Deformable ConvNets v2: more deformable, better results, 2018, http://dx.doi.org/10.48550/arXiv.1811.11168, arXiv:1811.11168, http://arxiv.org/abs/1811.11168 [cs].

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, in: NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.

[60] Han Hu, Zheng Zhang, Zhenda Xie, Stephen Lin, LRNet: Local relation networks for image recognition, 2019, http://dx.doi.org/10.48550/arXiv.1904.11491, arXiv:1904.11491, http://arxiv.org/abs/1904.11491 [cs].

[61] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens, Stand-alone self-attention in vision models, 2019, ArXiv, https://www.semanticscholar.org/paper/Stand-Alone-Self-Attention-in-Vision-Models-Ramachandran-Parmar/d6dccb5d71fbb6f5765f89633ba3a8e6809a720d.

[62] Hengshuang Zhao, Jiaya Jia, Vladlen Koltun, Exploring self-attention for image recognition, 2020, http://dx.doi.org/10.48550/arXiv.2004.13621, arXiv:2004.13621, http://arxiv.org/abs/2004.13621 [cs].

[63] Jie Hu, Li Shen, Gang Sun, Squeeze-and-Excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745, https://ieeexplore.ieee.org/document/8578843.

[64] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, Han Hu, Gcnet: non-local networks meet squeeze-excitation networks and beyond, 2019, http://dx.doi.org/10.48550/arXiv.1904.11492, arXiv:1904.11492, http://arxiv.org/abs/1904.11492 [cs].

[65] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, Han Hu, Disentangled non-local neural networks, 2020, http://dx.doi.org/10.48550/arXiv.2006.06668, arXiv:2006.06668, http://arxiv.org/abs/2006.06668 [cs].

[66] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, Quoc V. Le, Attention augmented convolutional networks, 2020, http://dx.doi.org/10.48550/arXiv.1904.09925, arXiv:1904.09925, http://arxiv.org/abs/1904.09925 [cs].

[67] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, Vit: An Image is worth 16x16 words: transformers for image recognition at scale, 2020, http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv:2010.11929, http://arxiv.org/abs/2010.11929 [cs].

[68] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou, Deit: training data-efficient image transformers & distillation through attention, 2021, http://dx.doi.org/10.48550/arXiv.2012.12877, arXiv:2012.12877, http://arxiv.org/abs/2012.12877 [cs].

[69] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: hierarchical vision transformer using shifted windows, 2021, http://dx.doi.org/10.48550/arXiv.2103.14030, arXiv:2103.14030, http://arxiv.org/abs/2103.14030 [cs].

[70] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, Baining Guo, Swin transformer V2: scaling up capacity and resolution, 2022, http://dx.doi.org/10.48550/arXiv.2111.09883, arXiv:2111.09883, http://arxiv.org/abs/2111.09883 [cs].

[71] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, Jonathon Shlens, Scaling local self-attention for parameter efficient visual backbones, 2021, arXiv:2103.12731 http://arxiv.org/abs/2103.12731 [cs].

[72] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, Levent Sagun, ConViT: Improving vision transformers with soft convolutional inductive biases, J. Stat. Mech. Theory Exp. 2022 (11) (2022) 114005, http://dx.doi.org/10.1088/1742-5468/ac9830, arXiv:2103.10697 http://arxiv.org/abs/2103.10697 [cs].

[73] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang, CvT: Introducing convolutions to vision transformers, 2021, arXiv:2103.15808 http://arxiv.org/abs/2103.15808 [cs].

[74] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, Chang Xu, CMT: Convolutional neural networks meet vision transformers, 2022, arXiv:2107.06263 http://arxiv.org/abs/2107.06263 [cs].

[75] Zihang Dai, Hanxiao Liu, Quoc V. Le, Mingxing Tan, CoAtNet: Marrying convolution and attention for all data sizes, 2021, ArXiv.Org, https://arxiv.org/abs/2106.04803v2.

[76] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, Shuicheng Yan, Inception transformer, 2022, arXiv:2205.12956 http://arxiv.org/abs/2205.12956 [cs].

[77] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, ConvNeXt: a ConvNet for the 2020s, 2022, arXiv:2201.03545 http://arxiv.org/abs/2201.03545 [cs].

[78] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, Cho-Jui Hsieh, DynamicViT: efficient vision transformers with dynamic token sparsification, 2021, http://dx.doi.org/10.48550/arXiv.2106.02034, arXiv:2106.02034, http://arxiv.org/abs/2106.02034 [cs].

[79] Sachin Mehta, Mohammad Rastegari, MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022, http://dx.doi.org/10.48550/arXiv.2110.02178, arXiv:2110.02178, http://arxiv.org/abs/2110.02178 [cs].

[80] Haoran You, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peizhao Zhang, Haoqi Fan, Peter Vajda, Yingyan Celine Lin, Castling-ViT: Compressing self-attention via switching towards linear-angular attention at vision transformer inference, 2022, http://dx.doi.org/10.48550/arXiv.2211.10526, arXiv:2211.10526, http://arxiv.org/abs/2211.10526 [cs].

[81] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, Yixuan Yuan, EfficientViT: memory efficient vision transformer with cascaded group attention, 2023, arXiv:2305.07027 http://arxiv.org/abs/2305.07027 [cs].

[82] Zizheng Pan, Jianfei Cai, Bohan Zhuang, HiLoViT: Fast vision transformers with hilo attention, 2023, arXiv:2205.13213 http://arxiv.org/abs/2205.13213 [cs].

[83] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, Anurag Ranjan, FastViT: A fast hybrid vision transformer using structural reparameterization, 2023, http://dx.doi.org/10.48550/arXiv.2303.14189, arXiv:2303.14189, http://arxiv.org/abs/2303.14189 [cs].

[84] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, Ge Li, GCNVAD: Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1237–1246, http://dx.doi.org/10.1109/CVPR.2019.00133, https://ieeexplore.ieee.org/abstract/document/8953791.

[85] Jiangong Zhang, Laiyun Qing, Jun Miao, TCN-IBL: temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, Taipei, Taiwan, 2019, pp. 4030–4034, http://dx.doi.org/10.1109/ICIP.2019.8803657, https://ieeexplore.ieee.org/document/8803657/.

[86] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, Seung-Ik Lee, SRF: A self-reasoning framework for anomaly detection using video-level labels, IEEE Signal Process. Lett. 27 (2020) 1705–1709, http://dx.doi.org/10.1109/LSP.2020.3025688, https://ieeexplore.ieee.org/document/9204830.

[87] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, Seung-Ik Lee, CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection, 2020, http://dx.doi.org/10.48550/arXiv.2011.12077, arXiv:2011.12077, http://arxiv.org/abs/2011.12077 [cs].

[88] Jia-Chang Feng, Fa-Ting Hong, Wei-Shi Zheng, MIST: Multiple instance self-training framework for video anomaly detection, 2021, arXiv:2104.01633 cs.

[89] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, Gustavo Carneiro, RTFM: weakly-supervised video anomaly detection with robust temporal feature magnitude learning, 2021, arXiv:2101.10030 http://arxiv.org/abs/2101.10030 [cs].

[90] Shenghao Yu, Chong Wang, Qiaomei Mao, Yuqi Li, Jiafei Wu, XEL: Cross-epoch learning for weakly supervised anomaly detection in surveillance videos, IEEE Signal Process. Lett. 28 (2021) 2137–2141, http://dx.doi.org/10.1109/LSP.2021.3117737, https://ieeexplore.ieee.org/document/9560033/.

[91] Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, Zhiying Zhou, CA: contrastive attention for video anomaly detection, IEEE Trans. Multimed. 24 (2021) 4067–4076, http://dx.doi.org/10.1109/TMM.2021.3112814, https://ieeexplore.ieee.org/document/9540293/.

[92] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, Seung-Ik Lee, CLAWS+: Clustering aided weakly supervised training to detect anomalous events in surveillance videos, 2022, arXiv:2203.13704 http://arxiv.org/abs/2203.13704 [cs].

[93] Yang Liu, Jing Liu, Xiaoguang Zhu, Donglai Wei, Xiaohong Huang, Liang Song, STA: learning task-specific representation for video anomaly detection with spatial-temporal attention, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Singapore, Singapore, 2022, pp. 2190–2194, http://dx.doi.org/10.1109/ICASSP43922.2022.9746822, https://ieeexplore.ieee.org/document/9746822/.

[94] Shuo Li, Fang Liu, Licheng Jiao, MSL: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection, Proc. the AAAI Conf. Artif. Intell. 36 (2) (2022) 1395–1403, http://dx.doi.org/10.1609/aaai.v36i2.20028, https://ojs.aaai.org/index.php/AAAI/article/view/20028.

[95] Shenghao Yu, Chong Wang, Lehong Xiang, Jiafei Wu, TCA-VAD: Temporal context alignment network for weakly supervised video anomly detection, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1–6, http://dx.doi.org/10.1109/ICME52920.2022.9859607, https://ieeexplore.ieee.org/document/9859607.

[96] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, Kwanghoon Sohn, NGMIL: Normality guided multiple instance learning for weakly supervised video anomaly detection, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2023, pp. 2664–2673, http://dx.doi.org/10.1109/WACV56688.2023.00269, https://ieeexplore.ieee.org/document/10030221/.

[97] Shengyang Sun, Xiaojin Gong, LSTC: Long-short temporal co-teaching for weakly supervised video anomaly detection, 2023, arXiv:2303.18044 http://arxiv.org/abs/2303.18044 [cs].

[98] Md Haidar Sharif, Lei Jiao, Christian W. Omlin, CNN-vit supported weakly-supervised video segment level anomaly detection, Sensors 23 (18) (2023) 7734, http://dx.doi.org/10.3390/s23187734, https://www.mdpi.com/1424-8220/23/18/7734.

[99] Karen Simonyan, Andrew Zisserman, Two-stream convolutional networks for action recognition in videos, 2014, http://dx.doi.org/10.48550/arXiv.1406.2199, arXiv:1406.2199, http://arxiv.org/abs/1406.2199 [cs].

[100] Yulin Wang, Yizeng Han, Chaofei Wang, Shiji Song, Qi Tian, Gao Huang, Computation-efficient deep learning for computer vision: a survey, 2023, arXiv:2308.13998 http://arxiv.org/abs/2308.13998 [cs].

[101] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, Kevin Murphy, S3D: rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification, 2018, arXiv:1712.04851 http://arxiv.org/abs/1712.04851 [cs].

[102] Du Tran, Heng Wang, Lorenzo Torresani, Matt Feiszli, CSN: video classification with channel-separated convolutional networks, 2019, arXiv:1904.02811 http://arxiv.org/abs/1904.02811 [cs].

[103] Christoph Feichtenhofer, X3D: Expanding architectures for efficient video recognition, 2020, http://dx.doi.org/10.48550/arXiv.2004.04730, arXiv:2004.04730, http://arxiv.org/abs/2004.04730 [cs].

[104] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri, R(2+1)D: a closer look at spatiotemporal convolutions for action recognition, 2018, arXiv:1711.11248 http://arxiv.org/abs/1711.11248 [cs].

[105] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, Tong Lu, TAM: Temporal adaptive module for video recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 13688–13698, http://dx.doi.org/10.1109/ICCV48922.2021.01345, https://ieeexplore.ieee.org/document/9710203/.

[106] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, Junjie Yan, STM: SpatioTemporal and motion encoding for action recognition, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 2000–2009, http://dx.doi.org/10.1109/ICCV.2019.00209, https://ieeexplore.ieee.org/document/9010925/.

[107] Ji Lin, Chuang Gan, Song Han, TSM: Temporal shift module for efficient video understanding, 2019, arXiv:1811.08383 http://arxiv.org/abs/1811.08383 [cs].

[108] Swathikiran Sudhakaran, Sergio Escalera, Oswald Lanz, GSN: gate-shift networks for video action recognition, 2020, arXiv:1912.00381 http://arxiv.org/abs/1912.00381 [cs].

[109] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He, SlowFast networks for video recognition, 2019, http://dx.doi.org/10.48550/arXiv.1812.03982, arXiv:1812.03982, http://arxiv.org/abs/1812.03982 [cs].

[110] Limin Wang, Zhan Tong, Bin Ji, Gangshan Wu, TDN: Temporal difference networks for efficient action recognition, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 1895–1904, http://dx.doi.org/10.1109/CVPR46437.2021.00193, https://ieeexplore.ieee.org/document/9577569/.

[111] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, Jian Yang, WSAL: Localizing anomalies from weakly-labeled videos, IEEE Trans. Image Process. 30 (2020) 4505–4515, http://dx.doi.org/10.1109/TIP.2021.3072863, https://ieeexplore.ieee.org/abstract/document/9408419.

[112] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid, Vivit: a video vision transformer, 2021, http://dx.doi.org/10.48550/arXiv.2103.15691, arXiv:2103.15691, http://arxiv.org/abs/2103.15691 [cs].

[113] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu, VSwin: video swin transformer, 2021, http://dx.doi.org/10.48550/arXiv.2106.13230, arXiv:2106.13230, http://arxiv.org/abs/2106.13230 [cs].

[114] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, Christoph Feichtenhofer, Mvit: Multiscale vision transformers, 2021, http://dx.doi.org/10.48550/arXiv.2104.11227, arXiv:2104.11227, http://arxiv.org/abs/2104.11227 [cs].

[115] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, Christoph Feichtenhofer, MViTv2: Improved multiscale vision transformers for classification and detection, 2022, http://dx.doi.org/10.48550/arXiv.2112.01526, arXiv:2112.01526, http://arxiv.org/abs/2112.01526 [cs].

[116] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, Yu Qiao, UniFormer: Unified transformer for efficient spatiotemporal representation learning, 2022, http://dx.doi.org/10.48550/arXiv.2201.04676, arXiv:2201.04676, http://arxiv.org/abs/2201.04676 [cs].

[117] Gedas Bertasius, Heng Wang, Lorenzo Torresani, TimeSformer: is space-time attention all you need for video understanding?, 2021, arXiv:2102.05095 http://arxiv.org/abs/2102.05095 [cs].

[118] Yingjie Zhai, Wenshuo Li, Yehui Tang, Xinghao Chen, Yunhe Wang, SqueezeTime: No Time to waste: squeeze time into channel for mobile video understanding, 2024, http://dx.doi.org/10.48550/arXiv.2405.08344, arXiv:2405.08344, http://arxiv.org/abs/2405.08344 [cs].

[119] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, Yik-Chung Wu, MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection, 2022, arXiv:2211.15098 http://arxiv.org/abs/2211.15098 [cs].

[120] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, Aran Komatsuzaki, LAION-400m: open dataset of CLIP-filtered 400 million image-text pairs, 2021, http://dx.doi.org/10.48550/arXiv.2111.02114, arXiv:2111.02114, http://arxiv.org/abs/2111.02114 [cs].

[121] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, Ludwig Schmidt, DataComp: In search of the next generation of multimodal datasets, 2023, arXiv:2304.14108 http://arxiv.org/abs/2304.14108 [cs].

[122] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, I. Sutskever, CLIP: learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, https://www.semanticscholar.org/paper/Learning-Transferable-Visual-Models-From-Natural-Radford-Kim/6f870f7f02a8c59c3e23f407f3ef00dd1dcf8fc4.

[123] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, Tom Duerig, ALIGN: Scaling up visual and vision-language representation learning with noisy text supervision, 2021, arXiv:2102.05918 http://arxiv.org/abs/2102.05918 [cs].

[124] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela, FLAVA: A foundational language and vision alignment model, 2022, http://dx.doi.org/10.48550/arXiv.2112.04482, arXiv:2112.04482, http://arxiv.org/abs/2112.04482 [cs].

[125] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022, http://dx.doi.org/10.48550/arXiv.2201.12086, arXiv:2201.12086, http://arxiv.org/abs/2201.12086 [cs].

[126] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, Yue Cao, EVA-CLIP: Exploring the limits of masked visual representation learning at scale, 2022, http://dx.doi.org/10.48550/arXiv.2211.07636, arXiv:2211.07636, http://arxiv.org/abs/2211.07636 [cs].

[127] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, Yue Cao, EVA-CLIP: Improved training techniques for CLIP at scale, 2023, http://dx.doi.org/10.48550/arXiv.2303.15389, arXiv:2303.15389, http://arxiv.org/abs/2303.15389 [cs].

[128] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, Kaiming He, FLIP: Scaling language-image pre-training via masking, 2023, http://dx.doi.org/10.48550/arXiv.2212.00794, arXiv:2212.00794, http://arxiv.org/abs/2212.00794 [cs].

[129] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, Vaishaal Shankar, DFN-CLIP: Data filtering networks, 2023, http://dx.doi.org/10.48550/arXiv.2309.17425, arXiv:2309.17425, http://arxiv.org/abs/2309.17425 [cs].

[130] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, Lucas Beyer, Siglip: Sigmoid Loss for language image pre-training, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, 2023, pp. 11941–11952, http://dx.doi.org/10.1109/ICCV51070.2023.01100, https://ieeexplore.ieee.org/document/10377550/.

[131] Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, Xiaohua Zhai, LocCa: Visual pretraining with location-aware captioners, 2024, http://dx.doi.org/10.48550/arXiv.2403.19596, arXiv:2403.19596, http://arxiv.org/abs/2403.19596 [cs].

[132] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, Piotr Bojanowski, DINOv2: Learning robust visual features without supervision, 2024, http://dx.doi.org/10.48550/arXiv.2304.07193, arXiv:2304.07193, http://arxiv.org/abs/2304.07193 [cs].

[133] Bowen Shi, Peisen Zhao, Zichen Wang, Yuhang Zhang, Yaoming Wang, Jin Li, Wenrui Dai, Junni Zou, Hongkai Xiong, Qi Tian, Xiaopeng Zhang, UMG-CLIP: A unified multi-granularity vision generalist for open-world understanding, 2024, http://dx.doi.org/10.48550/arXiv.2401.06397, arXiv:2401.06397, http://arxiv.org/abs/2401.06397 [cs].

[134] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, Oncel Tuzel, MobileCLIP: Fast image-text models through multi-modal reinforced training, 2024, arXiv:2311.17049 http://arxiv.org/abs/2311.17049 [cs].

[135] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, Liang-Chieh Chen, Vitamin: designing scalable vision models in the vision-language era, 2024, ArXiv.Org https://arxiv.org/abs/2404.02132v2.

[136] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, Christoph Feichtenhofer, VideoCLIP: Contrastive pre-training for zero-shot video-text understanding, 2021, http://dx.doi.org/10.48550/arXiv.2109.14084, arXiv:2109.14084, http://arxiv.org/abs/2109.14084 [cs].

[137] Mengmeng Wang, Jiazheng Xing, Yong Liu, ActionCLIP: A new paradigm for video action recognition, 2021, http://dx.doi.org/10.48550/arXiv.2109.08472, arXiv:2109.08472, http://arxiv.org/abs/2109.08472 [cs].

[138] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, Llama: open and efficient foundation language models, 2023, http://dx.doi.org/10.48550/arXiv.2302.13971, arXiv:2302.13971, http://arxiv.org/abs/2302.13971 [cs].

[139] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom, Llama 2: open foundation and fine-tuned chat models, 2023, http://dx.doi.org/10.48550/arXiv.2307.09288, arXiv:2307.09288, http://arxiv.org/abs/2307.09288 [cs].

[140] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira

Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, Zhiyu Ma, The Llama 3 herd of models, 2024, http://dx.doi.org/10.48550/arXiv.2407.21783, arXiv:2407.21783 http://arxiv.org/abs/2407.21783 [cs].

[141] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, Eric P. Xing, Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023, https://lmsys.org/blog/2023-03-30-vicuna/.

[142] InternLM Team, Internlm: A multilingual language model with progressively enhanced capabilities, 2023, https://github.com/InternLM/InternLM-techreport.

[143] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, Dahua Lin, InternLM2 technical report, 2024, http://dx.doi.org/10.48550/arXiv.2403.17297, arXiv:2403.17297, http://arxiv.org/abs/2403.17297 [cs].

[144] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, Yuanzhi Li, Phi-1: Textbooks are all you need, 2023, http://dx.doi.org/10.48550/arXiv.2306.11644, arXiv:2306.11644, http://arxiv.org/abs/2306.11644 [cs].

[145] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, Yin Tat Lee, Textbooks are all you need II: Phi-1.5 technical report, 2023, http://dx.doi.org/10.48550/arXiv.2309.05463, arXiv:2309.05463, http://arxiv.org/abs/2309.05463 [cs].

[146] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, Xiren Zhou, Phi-3 technical report: A highly capable language model locally on your phone, 2024, http://dx.doi.org/10.48550/arXiv.2404.14219, arXiv:2404.14219 http://arxiv.org/abs/2404.14219 [cs].

[147] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree

Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, Kathleen Kenealy, Gemma: open models based on gemini research and technology, 2024, http://dx.doi.org/10.48550/arXiv.2403.08295, arXiv:2403.08295, http://arxiv.org/abs/2403.08295 [cs].

[148] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M.R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, Alek Andreev, Gemma2: improving open language models at a practical size, 2024, http://dx.doi.org/10.48550/arXiv.2408.00118, arXiv:2408.00118, http://arxiv.org/abs/2408.00118 [cs].

[149] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, Maosong Sun, Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024, http://dx.doi.org/10.48550/arXiv.2404.06395, arXiv:2404.06395, http://arxiv.org/abs/2404.06395 [cs].

[150] Xinyin Ma, Gongfan Fang, Xinchao Wang, LLM-Pruner: On the structural pruning of large language models, 2023, http://dx.doi.org/10.48550/arXiv.2305.11627, arXiv:2305.11627, http://arxiv.org/abs/2305.11627 [cs].

[151] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, Tianhang Zhu, Qwen technical report, 2023, http://dx.doi.org/10.48550/arXiv.2309.16609, arXiv:2309.16609, http://arxiv.org/abs/2309.16609 [cs].

[152] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, Zhihao Fan, Qwen2 technical report, 2024, http://dx.doi.org/10.48550/arXiv.2407.10671, arXiv:2407.10671, http://arxiv.org/abs/2407.10671 [cs].

[153] Qwen Team, Qwen2.5: A party of foundation models, 2024, https://qwenlm.github.io/blog/qwen2.5/.

[154] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee, LLaVA: Visual instruction tuning, 2023, http://dx.doi.org/10.48550/arXiv.2304.08485, arXiv:2304.08485 http://arxiv.org/abs/2304.08485 [cs].

[155] Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee, Llava: Improved baselines with visual instruction tuning, 2024, http://dx.doi.org/10.48550/arXiv.2310.03744, arXiv:2310.03744 http://arxiv.org/abs/2310.03744 [cs].

[156] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, Yong Jae Lee, Llava-next: Improved reasoning, OCR, and world knowledge, 2024, https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[157] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, Chunyuan Li, LLaVA-NeXT: Stronger llms supercharge multimodal capabilities in the wild, 2024, https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

[158] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, Chunyuan Li, Llava-next: What else influences visual instruction tuning beyond data?, 2024, https://llava-vl.github.io/blog/2024-05-25-llava-next-ablations/.

[159] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, Chunyuan Li, LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models, 2024, http://dx.doi.org/10.48550/arXiv.2407.07895, arXiv:2407.07895, http://arxiv.org/abs/2407.07895 [cs].

[160] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, Chunhua Shen, Mobilevlm : a fast, strong and open vision language assistant for mobile devices, 2023, http://dx.doi.org/10.48550/arXiv.2312.16886, arXiv:2312.16886, http://arxiv.org/abs/2312.16886 [cs].

[161] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, Chunhua Shen, Mobilevlm-V2: faster and stronger baseline for vision language model, 2024, http://dx.doi.org/10.48550/arXiv.2402.03766, arXiv:2402.03766, http://arxiv.org/abs/2402.03766 [cs].

[162] Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, Xiren Zhou, Phi-4-mini technical report: compact yet powerful multimodal language models via mixture-of-loras, 2025, http://dx.doi.org/10.48550/arXiv.2503.01743, arXiv:2503.01743, http://arxiv.org/abs/2503.01743 [cs].

[163] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C.J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan,

Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Gemma3 technical report, 2025, http://dx.doi.org/10.48550/arXiv.2503.19786, arXiv:2503.19786, http://arxiv.org/abs/2503.19786 [cs].

[164] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, Afshin Dehghan, SlowFast-LLaVA: A strong training-free baseline for video large language models, 2024, http://dx.doi.org/10.48550/arXiv.2407.15841, arXiv:2407.15841, http://arxiv.org/abs/2407.15841 [cs].

[165] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, Hadi Pouransari, FastVLM: Efficient vision encoding for vision language models, 2025, http://dx.doi.org/10.48550/arXiv.2412.13303, arXiv:2412.13303, http://arxiv.org/abs/2412.13303 [cs].

[166] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, Chunyuan Li, LLaVA-OneVision: Easy visual task transfer, 2024, http://dx.doi.org/10.48550/arXiv.2408.03326, arXiv:2408.03326 http://arxiv.org/abs/2408.03326 [cs].

[167] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, Song Han, VILA: On pre-training for visual language models, 2024, http://dx.doi.org/10.48550/arXiv.2312.07533, arXiv:2312.07533, http://arxiv.org/abs/2312.07533 [cs].

[168] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, Jingren Zhou, Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond, 2023, http://dx.doi.org/10.48550/arXiv.2308.12966, arXiv:2308.12966, http://arxiv.org/abs/2308.12966 [cs].

[169] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, Junyang Lin, Qwen2-VL: enhancing vision-language model's perception of the world at any resolution, 2024, http://dx.doi.org/10.48550/arXiv.2409.12191, arXiv:2409.12191 http://arxiv.org/abs/2409.12191 [cs].

[170] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin, Qwen2.5-VL technical report, 2025, http://dx.doi.org/10.48550/arXiv.2502.13923, arXiv:2502.13923, http://arxiv.org/abs/2502.13923 [cs].

[171] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, Zhenchong Hu, Bo-Wen Zhang, Jijie Li, Dong Liang, Yingli Zhao, Songjing Wang, Yulong Ao, Yiming Ju, Huanhuan Ma, Xiaotong Li, Haiwen Diao, Yufeng Cui, Xinlong Wang, Yaoqi Liu, Fangxiang Feng, Guang Liu, Aquila-VL_Infinity-MM: scaling multimodal performance with large-scale and high-quality instruction data, 2025, http://dx.doi.org/10.48550/arXiv.2410.18558, arXiv:2410.18558, http://arxiv.org/abs/2410.18558 [cs].

[172] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Qwen2.5 technical report, 2025, http://dx.doi.org/10.48550/arXiv.2412.15115, arXiv:2412.15115, http://arxiv.org/abs/2412.15115 [cs].

[173] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, Jiao Ran, SAIL-VL: Scalable vision language model training via high quality data curation, 2025, http://dx.doi.org/10.48550/arXiv.2501.05952, arXiv:2501.05952, http://arxiv.org/abs/2501.05952 [cs].

[174] Rui Yang, Lin Song, Yicheng Xiao, Runhui Huang, Yixiao Ge, Ying Shan, Hengshuang Zhao, HaploVL: A single-transformer baseline for multi-modal understanding, 2025, http://dx.doi.org/10.48550/arXiv.2503.14694, arXiv:2503.14694, http://arxiv.org/abs/2503.14694 [cs].

[175] Bo Zhang, Shuo Li, Runhe Tian, Yang Yang, Jixin Tang, Jinhao Zhou, Lin Ma, Flash-VL 2B: optimizing vision-language model performance for ultra-low latency and high throughput, 2025, http://dx.doi.org/10.48550/arXiv.2505.09498, arXiv:2505.09498, http://arxiv.org/abs/2505.09498 [cs].

[176] Chenting Wang, Kunchang Li, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, Limin Wang, FluxUMTViT: Make your training flexible: towards deployment-efficient video models, 2025, http://dx.doi.org/10.48550/arXiv.2503.14237, arXiv:2503.14237, http://arxiv.org/abs/2503.14237 [cs].

[177] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, Yu Qiao, UnMaskedTeacher: Towards training-efficient video foundation models, 2024, http://dx.doi.org/10.48550/arXiv.2303.16058, arXiv:2303.16058, http://arxiv.org/abs/2303.16058 [cs].

[178] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, Yao Lu, NVILA: efficient frontier visual language models, 2025, http://dx.doi.org/10.48550/arXiv.2412.04468, arXiv:2412.04468, http://arxiv.org/abs/2412.04468 [cs].

[179] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, Thomas Wolf, SmolVLM: Redefining small and efficient multimodal models, 2025, http://dx.doi.org/10.48550/arXiv.2504.05299, arXiv:2504.05299, http://arxiv.org/abs/2504.05299 [cs].

[180] Luis Wiedmann, Aritra Roy Gosthipaty, Andrés Marafioti, Nanovlm, 2025, https://github.com/huggingface/nanoVLM.

[181] Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, Jieru Zhao, LiveVLM: Efficient online video understanding via streaming-oriented KV cache and retrieval, 2025, http://dx.doi.org/10.48550/arXiv.2505.15269, arXiv:2505.15269, http://arxiv.org/abs/2505.15269 [cs].

[182] Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, Huan Wang, HoliTom: holistic token merging for fast video large language models, 2025, http://dx.doi.org/10.48550/arXiv.2505.21334, arXiv:2505.21334, http://arxiv.org/abs/2505.21334 [cs].

[183] Zeqing Wang, Shiyuan Zhang, Chengpei Tang, Keze Wang, TimeCausality: Evaluating the causal ability in time dimension for vision language models, 2025, http://dx.doi.org/10.48550/arXiv.2505.15435, arXiv:2505.15435, http://arxiv.org/abs/2505.15435 [cs].

[184] Dasol Choi, Seunghyun Lee, Youngsook Song, Better safe than sorry? overreaction problem of vision language models in visual emergency recognition, 2025, http://dx.doi.org/10.48550/arXiv.2505.15367, arXiv:2505.15367 http://arxiv.org/abs/2505.15367 [cs].

[185] Hugo Laurençon, Andrés Marafioti, Victor Sanh, Léo Tronchon, Building and better understanding vision-language models: Insights and future directions, 2024, http://dx.doi.org/10.48550/arXiv.2408.12637, arXiv:2408.12637, http://arxiv.org/abs/2408.12637 [cs].

[186] Shakti N. Wadekar, Abhishek Chaurasia, Aman Chadha, Eugenio Culurciello, The evolution of multimodal model architectures, 2024, http://dx.doi.org/10.48550/arXiv.2405.17927, arXiv:2405.17927 http://arxiv.org/abs/2405.17927 [cs].

[187] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, Chenliang Xu, Video understanding with large language models: a survey, 2024, http://dx.doi.org/10.48550/arXiv.2312.17432, arXiv:2312.17432, http://arxiv.org/abs/2312.17432 [cs].

[188] Lorenzo Papa, Paolo Russo, Irene Amerini, Luping Zhou, A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 7682–7700, http://dx.doi.org/10.1109/TPAMI.2024.3392941, arXiv:2309.02031 http://arxiv.org/abs/2309.02031 [cs].

[189] Ahmed Sharshar, Latif U. Khan, Waseem Ullah, Mohsen Guizani, Vision-language models for edge networks: A comprehensive survey, 2025, http://dx.doi.org/10.48550/arXiv.2502.07855, arXiv:2502.07855 http://arxiv.org/abs/2502.07855 [cs].

[190] Nitesh Patnaik, Navdeep Nayak, Himani Bansal Agrawal, Moinak Chinmoy Khamaru, Gourav Bal, Saishree Smaranika Panda, Rishi Raj, Vishal Meena, Kartheek Vadlamani, Small vision-language models: A Survey on compact architectures and techniques, 2025, http://dx.doi.org/10.48550/arXiv.2503.10665, arXiv:2503.10665 http://arxiv.org/abs/2503.10665 [cs].

[191] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, Guangyao Shi, A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges, 2025, http://dx.doi.org/10.48550/arXiv.2501.02189, arXiv:2501.02189 http://arxiv.org/abs/2501.02189 [cs].

[192] Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, Nirmalya Roy, A survey on efficient vision-language models, 2025, http://dx.doi.org/10.48550/arXiv.2504.09724, arXiv:2504.09724, http://arxiv.org/abs/2504.09724 [cs].

[193] Shaibal Saha, Lanyu Xu, Vision transformers on the edge: a comprehensive survey of model compression and acceleration strategies, 2025, http://dx.doi.org/10.48550/arXiv.2503.02891, arXiv:2503.02891, http://arxiv.org/abs/2503.02891 [cs].

[194] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, Ngan Le, CLIP-TSA: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection, in: 2023 IEEE International Conference on Image Processing (ICIP), 2022, pp. 3230–3234, http://dx.doi.org/10.1109/ICIP49359.2023.10222289, https://ieeexplore.ieee.org/document/10222289.

[195] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, Hanwang Zhang, UMIL: unbiased multiple instance learning for weakly supervised video anomaly detection, 2023, arXiv:2303.12369 http://arxiv.org/abs/2303.12369 [cs].

[196] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, Elisa Ricci, AnomalyClip: Delving into CLIP latent space for video anomaly recognition, 2023, arXiv:2310.02835 http://arxiv.org/abs/2310.02835 [cs].

[197] Jiaqi Tang, Hao Lu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, Ying-Cong Chen, Hawk: learning to understand open-world video anomalies, 2024, http://dx.doi.org/10.48550/arXiv.2405.16886, arXiv:2405.16886, http://arxiv.org/abs/2405.16886 [cs].

[198] Yalong Jiang, LPG: local patterns generalize better for novel anomalies, 2025.

[199] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, Mario Vento, MiviaAudio events: Reliable detection of audio events in highly noisy environments, Pattern Recognit. Lett. 65 (2015) 22–28, http://dx.doi.org/10.1016/j.patrec.2015.06.026, https://www.sciencedirect.com/science/article/pii/S0167865515001981.

[200] Eduardo Fonseca, Fonseca: Training sound event classifiers using different types of supervision [n.d.]..

[201] Yuji Tokozume, Tatsuya Harada, Learning environmental sounds with end-to-end convolutional neural network, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, New Orleans, LA, 2017, pp. 2721–2725, http://dx.doi.org/10.1109/ICASSP.2017.7952651, http://ieeexplore.ieee.org/document/7952651/.

[202] Yuji Tokozume, Yoshitaka Ushiku, Tatsuya Harada, Learning from between-class examples for deep sound recognition, 2017, http://dx.doi.org/10.48550/arXiv.1711.10282, http://arxiv.org/abs/1711.10282 arXiv:1711.10282 [cs, eess, stat].

[203] Boqing Zhu, Kele Xu, Dezhi Wang, Lilun Zhang, Bo Li, Yuxing Peng, Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features, 2018, http://dx.doi.org/10.48550/arXiv.1805.09752, arXiv:1805.09752 http://arxiv.org/abs/1805.09752 [cs, eess].

[204] Sajjad Abdoli, Patrick Cardinal, Alessandro Lameiras Koerich, End-to-End environmental sound classification using a 1D convolutional neural network, 2019, http://dx.doi.org/10.48550/arXiv.1904.08990, arXiv:1904.08990 http://arxiv.org/abs/1904.08990 [cs, stat].

[205] M. Mehrdad Morsali, Hoda Mohammadzade, Saeed Bagheri Shouraki, Face : Fast, accurate and context-aware audio annotation and classification, 2023, arXiv:2303.03666 http://arxiv.org/abs/2303.03666 [cs, eess].

[206] Justin Salamon, Christopher Jacoby, Juan Pablo Bello, UrbanSound8K: a dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, Orlando Florida USA, 2014, pp. 1041–1044, http://dx.doi.org/10.1145/2647868.2655045, https://dl.acm.org/doi/10.1145/2647868.2655045.

[207] J. Allen, Short term spectral analysis, synthesis, and modification by discrete Fourier transform, IEEE Trans. Acoust. Speech Signal Process. 25 (3) (1977) 235–238, http://dx.doi.org/10.1109/TASSP.1977.1162950, https://ieeexplore.ieee.org/document/1162950.

[208] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, J. Acoust. Soc. Am. 8 (3) (1937) 185–190, http://dx.doi.org/10.1121/1.1915893.

[209] Karol J. Piczak, Environmental sound classification with convolutional neural networks, in: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, Boston, MA, USA, 2015, pp. 1–6, http://dx.doi.org/10.1109/MLSP.2015.7324337, http://ieeexplore.ieee.org/document/7324337/.

[210] Justin Salamon, Juan Pablo Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal Process. Lett. 24 (3) (2017) 279–283, http://dx.doi.org/10.1109/LSP.2017.2657381, arXiv:1608.04363 http://arxiv.org/abs/1608.04363 [cs].

[211] Xinyu Li, Venkata Chebiyyam, Katrin Kirchhoff, Multi-stream network with temporal attention for environmental sound classification, 2019, http://dx.doi.org/10.48550/arXiv.1901.08608, arXiv:1901.08608 http://arxiv.org/abs/1901.08608 [cs, eess].

[212] Yu Su, Ke Zhang, Jingyu Wang, Kurosh Madani, Environment sound classification using a two-stream CNN based on decision-level fusion, Sensors 19 (7) (2019) 1733, http://dx.doi.org/10.3390/s19071733, https://www.mdpi.com/1424-8220/19/7/1733.

[213] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, Marvin Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780, http://dx.doi.org/10.1109/ICASSP.2017.7952261.

[214] Kamalesh Palanisamy, Dipika Singhania, Angela Yao, Rethinking CNN models for audio classification, 2020, arXiv:2007.11154 http://arxiv.org/abs/2007.11154 [cs, eess].

[215] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, DenseNet: Densely connected convolutional networks, 2018, http://dx.doi.org/10.48550/arXiv.1608.06993, arXiv:1608.06993, http://arxiv.org/abs/1608.06993 [cs].

[216] Andrey Guzhov, Federico Raue, Jörn Hees, Andreas Dengel, EsResNet: Environmental sound classification based on visual domain models, 2020, http://dx.doi.org/10.48550/arXiv.2004.07301, arXiv:2004.07301 http://arxiv.org/abs/2004.07301 [cs, eess].

[217] Andrey Guzhov, Federico Raue, Jörn Hees, Andreas Dengel, ESResNe(X)t-fbsp: Learning robust time-frequency transformation of audio, 2021, arXiv:2104.11587 cs,eess.

[218] Anthony Teolis, Computational Signal Processing with Wavelets, Modern Birkhäuser Classics, Springer International Publishing, Cham, 2017, http://dx.doi.org/10.1007/978-3-319-65747-9, http://link.springer.com/10.1007/978-3-319-65747-9.

[219] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, Mark D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, 2020, arXiv:1912.10211 http://arxiv.org/abs/1912.10211 [cs, eess].

[220] Yuan Gong, Yu-An Chung, James Glass, PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 29 (2021) 3292–3306, http://dx.doi.org/10.1109/TASLP.2021.3120633, arXiv:2102.01243 http://arxiv.org/abs/2102.01243 [cs, eess].

[221] Eduardo Fonseca, Andres Ferraro, Xavier Serra, Sinet : Improving sound event classification by increasing shift invariance in convolutional neural networks, 2021, arXiv:2107.00623 http://arxiv.org/abs/2107.00623 [cs, eess].

[222] Aharon Azulay, Yair Weiss, Why do deep convolutional networks generalize so poorly to small image transformations?, 2019, arXiv:1805.12177 http://arxiv.org/abs/1805.12177 [cs].

[223] Richard Zhang, TLPF: making convolutional networks shift-invariant again, 2019, http://dx.doi.org/10.48550/arXiv.1904.11486, arXiv:1904.11486, http://arxiv.org/abs/1904.11486 [cs].

[224] Anadi Chaman, Ivan Dokmanić, APS: truly shift-invariant convolutional neural networks, 2021, http://dx.doi.org/10.48550/arXiv.2011.14214, arXiv:2011.14214, http://arxiv.org/abs/2011.14214 [cs].

[225] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, Xavier Serra, FSD50k: An open dataset of human-labeled sound events, 2022, arXiv:2010.00475 http://arxiv.org/abs/2010.00475 [cs, eess, stat].

[226] Antonio Greco, Antonio Roberto, Alessia Saggese, Mario Vento, Denet: A deep architecture for audio surveillance applications, Neural Comput. Appl. 33 (17) (2021) 11273–11284, http://dx.doi.org/10.1007/s00521-020-05572-5, https://doi.org/10.1007/s00521-020-05572-5.

[227] Mirco Ravanelli, Yoshua Bengio, SincNet: speaker recognition from raw waveform with SincNet, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 1021–1028, http://dx.doi.org/10.1109/SLT.2018.8639585, https://ieeexplore.ieee.org/document/8639585.

[228] Yuan Gong, Yu-An Chung, James Glass, AST: Audio spectrogram transformer, in: Interspeech 2021, ISCA, 2021, pp. 571–575, http://dx.doi.org/10.21437/Interspeech.2021-698, https://www.isca-archive.org/interspeech_2021/gong21b_interspeech.html.

[229] Prateek Verma, Jonathan Berger, AudioTransformers: Transformer architectures for large scale audio understanding. adieu convolutions, 2021, arXiv:2105.00335 http://arxiv.org/abs/2105.00335 [cs, eess].

[230] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, Gerhard Widmer, Passt: Efficient training of audio transformers with patchout, 2021, ArXiv.Org, https://arxiv.org/abs/2110.05069v1.

[231] Khaled Koutini, Shahed Masoudian, Florian Schmid, Hamid Eghbal-zadeh, Jan Schlüter, Gerhard Widmer, PaSST+: Learning general audio representations with large-scale training of patchout audio transformers, 2023, arXiv:2211.13956 http://arxiv.org/abs/2211.13956 [cs, eess].

[232] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, Yonatan Bisk, HEAR: Holistic evaluation of audio representations, 2022, arXiv:2203.03022 http://arxiv.org/abs/2203.03022 [cs, eess, stat].

[233] Luyu Wang, Aaron van den Oord, Multi-format contrastive learning of audio representations, 2021, arXiv:2103.06508 http://arxiv.org/abs/2103.06508 [cs, eess].

[234] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, Aaron van den Oord, Multimodal self-supervised learning of general audio representations, 2021, ArXiv.Org, https://arxiv.org/abs/2104.12807v2.

[235] Andrey Guzhov, Federico Raue, Jörn Hees, Andreas Dengel, Audioclip: Extending CLIP to image, text and audio, 2021, arXiv:2106.13043http://arxiv.org/abs/2106.13043 [cs, eess].

[236] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, Juan Pablo Bello, Wav2CLIP: Learning robust audio representations from CLIP, 2022, arXiv:2110.11499 http://arxiv.org/abs/2110.11499 [cs, eess].

[237] Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, Yanning Zhang, VAR: Towards video anomaly retrieval from video anomaly detection: new benchmarks and model, 2023, arXiv:2307.12545 http://arxiv.org/abs/2307.12545 [cs].

[238] Florian Schmid, Khaled Koutini, Gerhard Widmer, Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation, 2023, arXiv:2211.04772 http://arxiv.org/abs/2211.04772 [cs, eess].

[239] Florian Schmid, Khaled Koutini, Gerhard Widmer, GPAE: low-complexity audio embedding extractors, 2023, arXiv:2303.01879 http://arxiv.org/abs/2303.01879 [cs, eess].

[240] Shawn Hershey, Sourish Chaudhuri, Daniel P.W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson, CNN architectures for Large-scale audio classification, 2017, arXiv:1609.09430 http://arxiv.org/abs/1609.09430 [cs, stat].

[241] Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, Jun Liu, ECVA: Exploring What Why and how: a multifaceted benchmark for causation understanding of video anomaly, 2024, http://dx.doi.org/10.48550/arXiv.2412.07183, arXiv:2412.07183, http://arxiv.org/abs/2412.07183 [cs].

[242] Ivo P.C. Kersten, Erkut Akdag, Egor Bondarev, Peter H. De With, ThrowingAction: detection of object throwing behavior in surveillance videos, Electron. Imaging 35 (9) (2023) 286–1–286–9, http://dx.doi.org/10.2352/EI.2023.35.9.IPAS-286, https://library.imaging.org/ei/articles/35/9/IPAS-286.

[243] Amit Adam, Ehud Rivlin, Ilan Shimshoni, Daviv Reinitz, SUBWAY: robust real-time unusual event detection using multiple fixed-location monitors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (3) (2008) 555–560, http://dx.doi.org/10.1109/TPAMI.2007.70825.

[244] Xinyi Cui, Qingshan Liu, Mingchen Gao, Dimitris Metaxas, UMN: Abnormal detection using interaction energy potentials, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 3161–3167, http://dx.doi.org/10.1109/CVPR.2011.5995558.

[245] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, Nuno Vasconcelos, UCSD_PED: Anomaly Detection in crowded scenes, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981, http://dx.doi.org/10.1109/CVPR.2010.5539872.

[246] Cewu Lu, Jianping Shi, Jiaya Jia, AVENUE: Abnormal event detection at 150 FPS in MATLAB, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 2720–2727, http://dx.doi.org/10.1109/ICCV.2013.338.

[247] Weixin Luo, Wen Liu, Shenghua Gao, ShanghaiTechCampus: A revisit of sparse coding based anomaly detection in stacked RNN framework, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, 2017, pp. 341–349, http://dx.doi.org/10.1109/ICCV.2017.45, http://ieeexplore.ieee.org/document/8237307/.

[248] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, Mubarak Shah, UBnormal: New benchmark for supervised open-set video anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20143–20153, https://openaccess.thecvf.com/content/CVPR2022/html/Acsintoae_UBnormal_New_Benchmark_for_Supervised_Open-Set_Video_Anomaly_Detection_CVPR_2022_paper.html.

[249] Congqi Cao, Yue Lu, Peng Wang, Yanning Zhang, NWPU campus: a new comprehensive benchmark for semi-supervised video anomaly detection and anticipation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, 2023, pp. 20392–20401, http://dx.doi.org/10.1109/CVPR52729.2023.01953, https://ieeexplore.ieee.org/document/10203814/.

[250] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, Zhenzhen Jiao, UCFA: Towards surveillance video-and-language understanding: new dataset, baselines, and challenges, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2024, pp. 22052–22061, http://dx.doi.org/10.1109/CVPR52733.2024.02082, https://ieeexplore.ieee.org/document/10656129/.

[251] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, Nong Sang, Holmes-VAD: towards unbiased and explainable video anomaly detection via multi-modal LLM, 2024, http://dx.doi.org/10.48550/arXiv.2406.12235, arXiv:2406.12235, http://arxiv.org/abs/2406.12235 [cs].

[252] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, Xiaofeng Tao, CUVA: Uncovering What, Why and How: A comprehensive benchmark for causation understanding of video anomaly, 2024, http://dx.doi.org/10.48550/arXiv.2405.00181, arXiv:2405.00181, http://arxiv.org/abs/2405.00181 [cs].

[253] Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, Guodong Zhou, Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive LLM, 2025, http://dx.doi.org/10.48550/arXiv.2502.18863, arXiv:2502.18863, http://arxiv.org/abs/2502.18863 [cs].

[254] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, Hongyang Zhang, Ref-L4: revisiting referring expression comprehension evaluation in the era of large multimodal models, 2024, http://dx.doi.org/10.48550/arXiv.2406.16866, arXiv:2406.16866, http://arxiv.org/abs/2406.16866 [cs].

[255] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, Li Cheng, HumanML3D: Generating Diverse and Natural 3D Human Motions From Text[n.d.].

[256] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, Ling Zhao, RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data, Sensors 20 (6) (2020) 1594, http://dx.doi.org/10.3390/s20061594, https://www.mdpi.com/1424-8220/20/6/1594.

[257] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, COCO: common objects in context, in: David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[258] InternVL2 https://internvl.github.io/blog/2024-07-02-InternVL-2.0/ [n.d].

[259] Yiling Zhang, Erkut Akdag, Egor Bondarev, Peter H.N. De With, MTFL: Multi-timescale feature learning for eeakly-supervised anomaly detection in surveillance videos, in: Wolfgang Osten (Ed.), Seventeenth International Conference on Machine Vision (ICMV 2024), SPIE, Edinburg, United Kingdom, 2025, p. 14, http://dx.doi.org/10.1117/12.3055069, https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13517/3055069/MTFL–multi-timescale-feature-learning-for-weakly-supervised-anomaly/10.1117/12.3055069.full.

[260] Yuanbin Qian, Shuhan Ye, Chong Wang, Xiaojie Cai, Jiangbo Qian, Jiafei Wu, UCF-crime-DVS: a novel event-based dataset for video anomaly detection with spiking neural networks, 2025, http://dx.doi.org/10.48550/arXiv.2503.12905, arXiv:2503.12905, http://arxiv.org/abs/2503.12905 [cs].

[261] Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, Salman Khan, VANE-bench: video anomaly evaluation benchmark for conversational LMMs, 2025, http://dx.doi.org/10.48550/arXiv.2406.10326, arXiv:2406.10326, http://arxiv.org/abs/2406.10326 [cs].

[262] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, Yang You, Open-sora: Democratizing efficient video production for all, 2024, arXiv preprint arXiv:2412.20404 arXiv:2412.20404.

[263] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, Nong Sang, VideoLCM: Video latent consistency model, 2023, http://dx.doi.org/10.48550/arXiv.2312.09109, arXiv:2312.09109, http://arxiv.org/abs/2312.09109 [cs].

[264] Kellie Corona, Katie Osterdahl, Roderic Collins, Anthony Hoogs, MEVA: a large-scale multiview, multimodal video dataset for activity detection, in: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2021, pp. 1059–1067, http://dx.doi.org/10.1109/wacv48630.2021.00110, https://ieeexplore.ieee.org/document/9423413/.

[265] Shaojie Bai, J. Zico Kolter, Vladlen Koltun, TCN: an empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018, http://dx.doi.org/10.48550/arXiv.1803.01271, arXiv:1803.01271, http://arxiv.org/abs/1803.01271 [cs].

[266] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, Thomas S. Huang, Ccnet: Criss-cross attention for semantic segmentation, 2020, http://dx.doi.org/10.48550/arXiv.1811.11721, arXiv:1811.11721, http://arxiv.org/abs/1811.11721 [cs].

[267] Yuli Wu, Long Chen, Dorit Merhof, CosEmbdLoss: Improving pixel embedding learning through intermediate distance regression supervision for instance segmentation, in: Adrien Bartoli, Andrea Fusiello (Eds.), Computer Vision – ECCV 2020 Workshops, vol. 12540, Springer International Publishing, Cham, 2020, pp. 213–227, http://dx.doi.org/10.1007/978-3-030-65414-6_16, https://link.springer.com/10.1007/978-3-030-65414-6_16.

[268] Ruoyan Pi, Xiangteng He, Yuxin Peng, TAI: weakly supervised video anomaly detection with temporal and abnormal information, in: Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 594–608, http://dx.doi.org/10.1007/978-3-031-18913-5_46.

[269] Kihyuk Sohn, Npair: Improved deep metric learning with multi-class N-pair loss objective, in: Advances in Neural Information Processing Systems, 29, Curran Associates, Inc., 2016, https://papers.nips.cc/paper_files/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

[270] Snehashis Majhi, Srijan Das, Francois Bremond, DAM: Dissimilarity attention module for weakly-supervised video anomaly detection, in: 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, Washington, DC, USA, 2021, pp. 1–8, http://dx.doi.org/10.1109/AVSS52988.2021.9663810, https://ieeexplore.ieee.org/document/9663810/.

[271] Yujiang Pu, Xiaoyu Wu, LAN: locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection, 2022, arXiv:2208.05636 http://arxiv.org/abs/2208.05636 [cs].

[272] Fisher Yu, Vladlen Koltun, Multi-scale context aggregation by dilated convolutions, 2016, http://dx.doi.org/10.48550/arXiv.1511.07122, arXiv:1511.07122, http://arxiv.org/abs/1511.07122 [cs].

[273] Dasheng Zhang, Chao Huang, Chengliang Liu, Yong Xu, WSTR: Weakly supervised video anomaly detection via transformer-enabled temporal relation learning, IEEE Signal Process. Lett. 29 (2022) 1197–1201, http://dx.doi.org/10.1109/LSP.2022.3175092, https://ieeexplore.ieee.org/abstract/document/9774889.

[274] Weijun Tan, Qi Yao, Jingfeng Liu, BERT-MILRTFM : Overlooked video classification in weakly supervised video anomaly detection, 2023, arXiv:2210.06688 http://arxiv.org/abs/2210.06688 [cs].

[275] Yang Zhen, Yuanfang Guo, Jinjie Wei, Xiuguo Bao, Di Huang, MS-BS: Multi-scale background suppression anomaly detection in surveillance videos, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, Anchorage, AK, USA, 2021, pp. 1114–1118, http://dx.doi.org/10.1109/ICIP42928.2021.9506580, https://ieeexplore.ieee.org/document/9506580/.

[276] Yang Liu, Wanxiao Yang, Hangyou Yu, Lin Feng, Yuqiu Kong, Shenglan Liu, BS-me: background suppressed and motion enhanced network for weakly supervised video anomaly detection, in: Shiqi Yu, Zhaoxiang Zhang, Pong C. Yuen, Junwei Han, Tieniu Tan, Yike Guo, Jianhuang Lai, Jianguo Zhang (Eds.), Pattern Recognition and Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 678–690, http://dx.doi.org/10.1007/978-3-031-18913-5_52.

[277] Pilhyeon Lee, Youngjung Uh, Hyeran Byun, BS-WSTAL: Background suppression network for weakly-supervised temporal action localization, 2019, http://dx.doi.org/10.48550/arXiv.1911.09963, arXiv:1911.09963, http://arxiv.org/abs/1911.09963 [cs].

[278] Hang Zhou, Junqing Yu, Wei Yang, URDMU: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection, 2023, arXiv:2302.05160 http://arxiv.org/abs/2302.05160 [cs].

[279] Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, Hengtao Shen, BN-DFM: BatchNorm-based weakly supervised video anomaly detection, 2023, arXiv:2311.15367 http://arxiv.org/abs/2311.15367 [cs].

[280] Hamid Ghorbani, Mahalanobis DISTANCE AND its application FOR DETECTING multivariate OUTLIERS, Facta Univ. Ser.: Math. Informatics (2019) 583, http://dx.doi.org/10.22190/FUMI1903583G, http://casopisi.junis.ni.ac.rs/index.php/FUMathInf/article/view/5028.

[281] Tao Zhu, Qi Yu, Xinru Dong, Shiyu Li, Yue Liu, Jinlong Jiang, Lei Shu, ProDisc-VAD: An Efficient System for weakly-supervised anomaly detection in video surveillance applications, 2025, http://dx.doi.org/10.48550/arXiv.2505.02179, arXiv:2505.02179, http://arxiv.org/abs/2505.02179 [cs].

[282] Yang Liu, Jing Liu, Wei Ni, Liang Song, SG-MIR: abnormal event detection with self-guiding multi-instance ranking framework, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 01–07, http://dx.doi.org/10.1109/IJCNN55064.2022.9892231, https://ieeexplore.ieee.org/document/9892231.

[283] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, Ming-Hsuan Yang, CUN: exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection, 2022, arXiv:2212.04090 http://arxiv.org/abs/2212.04090 [cs].

[284] Yarin Gal, Zoubin Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, 2016, http://dx.doi.org/10.48550/arXiv.1506.02142, arXiv:1506.02142 http://arxiv.org/abs/1506.02142 [stat].

[285] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, Colin Raffel, FixMatch: Simplifying semi-supervised learning with consistency and confidence, 2020, http://dx.doi.org/10.48550/arXiv.2001.07685, arXiv:2001.07685, http://arxiv.org/abs/2001.07685 [cs].

[286] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, Haibin Ling, X-CLIP: Expanding language-image pretrained models for general video recognition, 2022, http://dx.doi.org/10.48550/arXiv.2208.02816, arXiv:2208.02816, http://arxiv.org/abs/2208.02816 [cs].

[287] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, Masashi Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, 2018, http://dx.doi.org/10.48550/arXiv.1804.06872, arXiv:1804.06872, http://arxiv.org/abs/1804.06872 [cs].

[288] Yidan Fan, Yongxin Yu, Wenhuan Lu, Yahong Han, SAA: Weakly-supervised video anomaly detection with snippet anomalous attention, 2023, arXiv:2309.16309 http://arxiv.org/abs/2309.16309 [cs].

[289] Chen Zhang, Guorong Li, Yuankai Qi, Hanhua Ye, Laiyun Qing, Ming-Hsuan Yang, Qingming Huang, DEN: Dynamic erasing network based on multi-scale temporal features for weakly supervised video anomaly detection, 2023, arXiv:2312.01764 http://arxiv.org/abs/2312.01764 [cs].

[290] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S. Ryoo, Francois Bremond, MS-TCT: Multi-scale temporal ConvTransformer for action detection, 2022, http://dx.doi.org/10.48550/arXiv.2112.03902, arXiv:2112.03902, http://arxiv.org/abs/2112.03902 [cs].

[291] Yudai Watanabe, Makoto Okabe, Yasunori Harada, Naoji Kashima, ANMIL: Real-world video anomaly detection by extracting salient features in videos, 2022, arXiv:2209.06435 http://arxiv.org/abs/2209.06435 [cs].

[292] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, Yan-Xiong Li, AVF: Violence detection in videos based on fusing visual and audio information, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2260–2264, http://dx.doi.org/10.1109/ICASSP39728.2021.9413686.

[293] Yujiang Pu, Xiaoyu Wu, CMALA: Audio-guided attention network for weakly supervised violence detection, in: 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2022, pp. 219–223, http://dx.doi.org/10.1109/ICCECE54139.2022.9712793.

[294] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, Yuejie Zhang, Look, listen, and attend: Co-Attention Network for self-supervised audio-visual representation learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3884–3892, http://dx.doi.org/10.1145/3394171.3413869, arXiv:2008.05789 http://arxiv.org/abs/2008.05789 [cs].

[295] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, Yuejie Zhang, MACIL-SD: Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection, 2022, arXiv:2207.05500 http://arxiv.org/abs/2207.05500 [cs].

[296] Aaron van den Oord, Yazhe Li, Oriol Vinyals, InfoNCE: Representation learning with contrastive predictive coding, 2019, arXiv:1807.03748 http://arxiv.org/abs/1807.03748 [cs, stat].

[297] Everette S. Gardner, Exponential smoothing: The state of the art—part II, Int. J. Forecast. 22 (4) (2006) 637–666, http://dx.doi.org/10.1016/j.ijforecast.2006.03.005, https://linkinghub.elsevier.com/retrieve/pii/S0169207006000392.

[298] Donglai Wei, Yang Liu, Xiaoguang Zhu, Jing Liu, Xinhua Zeng, MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection, IEEE Signal Process. Lett. 29 (2022) 2178–2182, http://dx.doi.org/10.1109/LSP.2022.3216500, https://ieeexplore.ieee.org/document/9926192/.

[299] Xiaogang Peng, Hao Wen, Yikai Luo, Xiao Zhou, Keyang Yu, Yigang Wang, Zizhao Wu, Hypervd: learning weakly supervised audio-visual violence detection in hyperbolic space, 2023, arXiv:2305.18797 http://arxiv.org/abs/2305.18797 [cs].

[300] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, Pierre Vandergheynst, Geometric deep learning: Going beyond euclidean data, IEEE Signal Process. Mag. 34 (4) (2017) 18–42, http://dx.doi.org/10.1109/MSP.2017.2693418, arXiv:1611.08097 http://arxiv.org/abs/1611.08097 [cs].

[301] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, Jure Leskovec, Hierarchical graph representation learning with differentiable pooling, 2019, http://dx.doi.org/10.48550/arXiv.1806.08804, arXiv:1806.08804 http://arxiv.org/abs/1806.08804 [cs].

[302] Maximilian Nickel, Douwe Kiela, Lorentz: Learning continuous hierarchies in the Lorentz model of hyperbolic geometry, 2018, http://dx.doi.org/10.48550/arXiv.1806.03417, arXiv:1806.03417, http://arxiv.org/abs/1806.03417 [cs].

[303] Huixin Wu, Mengfan Yang, Fupeng Wei, Ge Shi, Wei Jiang, Yaqiong Qiao, Hangcheng Dong, MTDA: Weakly-supervised video anomaly detection with MTDA-net, Electron. 12 (22) (2023) 4623, http://dx.doi.org/10.3390/electronics12224623, https://www.mdpi.com/2079-9292/12/22/4623.

[304] Shengyang Sun, Xiaojin Gong, MSBT: Multi-scale bottleneck transformer for weakly supervised multimodal violence detection, 2024, arXiv:2405.05130 http://arxiv.org/abs/2405.05130 [cs].

[305] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, Chen Sun, Attention bottlenecks for multimodal fusion, 2022, http://dx.doi.org/10.48550/arXiv.2107.00135, arXiv:2107.00135, http://arxiv.org/abs/2107.00135 [cs].

[306] Jingke Meng, Huilin Tian, Ge Lin, Jian-Fang Hu, Wei-Shi Zheng, AVCL: Audio-visual collaborative learning for weakly supervised video anomaly detection, IEEE Trans. Multimed. (2025) 1–12, http://dx.doi.org/10.1109/TMM.2025.3535377, https://ieeexplore.ieee.org/document/10855604/.

[307] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, Thomas Unterthiner, Differentiable patch selection for image recognition, 2021, http://dx.doi.org/10.48550/arXiv.2104.03059, arXiv:2104.03059 http://arxiv.org/abs/2104.03059 [cs].

[308] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, David Aik-Aun Khoo, TEVAD: Improved video anomaly detection with captions, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Vancouver, BC, Canada, 2023, pp. 5549–5559, http://dx.doi.org/10.1109/CVPRW59228.2023.00587, https://ieeexplore.ieee.org/document/10208872/.

[309] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, Lijuan Wang, Swinbert: End-to-end transformers with sparse attention for video captioning, 2022, http://dx.doi.org/10.48550/arXiv.2111.13196, arXiv:2111.13196, http://arxiv.org/abs/2111.13196 [cs].

[310] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, William Yang Wang, VATEX: A Large-Scale, high-quality multilingual dataset for video-and-language research, 2020, http://dx.doi.org/10.48550/arXiv.1904.03493, arXiv:1904.03493, http://arxiv.org/abs/1904.03493 [cs].

[311] Robyn Speer, Joshua Chin, Catherine Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, 2018, http://dx.doi.org/10.48550/arXiv.1612.03975, arXiv:1612.03975, http://arxiv.org/abs/1612.03975 [cs].

[312] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu, CoOp: Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348, http://dx.doi.org/10.1007/s11263-022-01653-1, arXiv:2109.01134 http://arxiv.org/abs/2109.01134 [cs].

[313] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, Tim Salimans, Axial attention in Multidimensional transformers, 2019, arXiv:1912.12180 http://arxiv.org/abs/1912.12180 [cs].

[314] Chenchen Tao, Chong Wang, Yuexian Zou, Xiaohao Peng, Jiafei Wu, Jiangbo Qian, LAP: Learn suspected anomalies from event prompts for video anomaly detection, 2024, arXiv:2403.01169 http://arxiv.org/abs/2403.01169 [cs].

[315] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, Yanning Zhang, STPrompt: Weakly supervised video anomaly detection and localization with spatio-temporal prompts, 2024, http://dx.doi.org/10.48550/arXiv.2408.05905, arXiv:2408.05905, http://arxiv.org/abs/2408.05905 [cs].

[316] Zhiwei Yang, Jing Liu, Peng Wu, TPWNG: Text Prompt with normality guidance for weakly supervised video anomaly detection, 2024, http://dx.doi.org/10.48550/arXiv.2404.08531, arXiv:2404.08531, http://arxiv.org/abs/2404.08531 [cs].

[317] Tianshan Liu, Kin-Man Lam, Bing-Kun Bao, ITC: injecting text clues for improving anomalous event detection from weakly labeled videos, IEEE Trans. Image Process. 33 (2024) 5907–5920, http://dx.doi.org/10.1109/TIP.2024.3477351, https://ieeexplore.ieee.org/document/10719608.

[318] Shengyang Sun, Jiashen Hua, Junyi Feng, Dongxu Wei, Baisheng Lai, Xiaojin Gong, TDSD: Text-driven scene-decoupled weakly supervised video anomaly detection, in: Proceedings of the 32nd ACM International Conference on Multimedia, ACM, Melbourne VIC Australia, 2024, pp. 5055–5064, http://dx.doi.org/10.1145/3664647.3680934, https://dl.acm.org/doi/10.1145/3664647.3680934.

[319] Audun Jøsang, Subjective Logic, Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing, Cham, 2016, http://dx.doi.org/10.1007/978-3-319-42337-1, http://link.springer.com/10.1007/978-3-319-42337-1.

[320] Hui Lv, Qianru Sun, VAD-LLaMA: video anomaly detection and explanation via large language models, 2024, http://dx.doi.org/10.48550/arXiv.2401.05702, arXiv:2401.05702, http://arxiv.org/abs/2401.05702 [cs].

[321] Muchao Ye, Weiyang Liu, Pan He, VERA: Explainable video anomaly detection via verbalized learning of vision-language models, 2025, http://dx.doi.org/10.48550/arXiv.2412.01095, arXiv:2412.01095, http://arxiv.org/abs/2412.01095 [cs].

[322] Hang Zhang, Xin Li, Lidong Bing, Video-LLaMA: an instruction-tuned audio-visual language model for video understanding, 2023, http://dx.doi.org/10.48550/arXiv.2306.02858, arXiv:2306.02858 http://arxiv.org/abs/2306.02858 [cs].

[323] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023, http://dx.doi.org/10.48550/arXiv.2301.12597, arXiv:2301.12597, http://arxiv.org/abs/2301.12597 [cs].

[324] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra, ImageBind: one embedding space to bind them all, 2023, http://dx.doi.org/10.48550/arXiv.2305.05665, arXiv:2305.05665 http://arxiv.org/abs/2305.05665 [cs].

[325] G. Farneback, Farneback: Fast and accurate motion estimation using orientation tensors and parametric motion models, in: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, 1, 2000, pp. 135–139 vol.1, http://dx.doi.org/10.1109/ICPR.2000.905291, https://ieeexplore.ieee.org/document/905291/.

[326] Max Bain, Arsha Nagrani, Gul Varol, Andrew Zisserman, WebVid-2M: Frozen in time: A joint video and image encoder for end-to-end retrieval, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 1708–1718, http://dx.doi.org/10.1109/ICCV48922.2021.00175, https://ieeexplore.ieee.org/document/9711165/.

[327] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed, Mistral 7B, 2023, http://dx.doi.org/10.48550/arXiv.2310.06825, arXiv:2310.06825, http://arxiv.org/abs/2310.06825 [cs].

[328] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, LoRA: Low-rank adaptation of large language models, 2021, http://dx.doi.org/10.48550/arXiv.2106.09685, arXiv:2106.09685, http://arxiv.org/abs/2106.09685 [cs].

[329] Li Xu, He Huang, Jun Liu, SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events, 2021, http://dx.doi.org/10.48550/arXiv.2103.15538, arXiv:2103.15538, http://arxiv.org/abs/2103.15538 [cs].

[330] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Video-ChatGPT: Towards detailed video understanding via large vision and language models, 2024, http://dx.doi.org/10.48550/arXiv.2306.05424, arXiv:2306.05424, http://arxiv.org/abs/2306.05424 [cs].

[331] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Khan, VideoGPT+: Integrating Image and video encoders for enhanced video understanding, 2024, http://dx.doi.org/10.48550/arXiv.2406.09418, arXiv:2406.09418, http://arxiv.org/abs/2406.09418 [cs].

[332] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, Li Yuan, LanguageBind: extending video-language pretraining to n-modality by language-based semantic alignment, 2024, http://dx.doi.org/10.48550/arXiv.2310.01852, arXiv:2310.01852, http://arxiv.org/abs/2310.01852 [cs].

[333] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, Wenhai Wang, InternVL1.5: how far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites, 2024, http://dx.doi.org/10.48550/arXiv.2404.16821, arXiv:2404.16821 http://arxiv.org/abs/2404.16821 [cs].

[334] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: A method for automatic evaluation of machine translation, in: Pierre Isabelle, Eugene Charniak, Dekang Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318, http://dx.doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/.

[335] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh, CIDEr: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575, http://dx.doi.org/10.1109/CVPR.2015.7299087, https://ieeexplore.ieee.org/document/7299087.

[336] Satanjeev Banerjee, Alon Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Jade Goldstein, Alon Lavie, Chin-Yew Lin, Clare Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/Or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72, https://aclanthology.org/W05-0909/.

[337] Chin-Yew Lin, ROUGE: A Package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81, https://aclanthology.org/W04-1013/.

[338] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, Li Yuan, Video-LLaVA: learning united visual representation by alignment before projection, 2024, http://dx.doi.org/10.48550/arXiv.2311.10122, arXiv:2311.10122, http://arxiv.org/abs/2311.10122 [cs].

[339] Zachary Teed, Jia Deng, RAFT: recurrent all-pairs field transforms for optical flow, in: Andrea Vedaldi, Horst Bischof, Thomas Brox, Jan-Michael Frahm (Eds.), Computer Vision – ECCV 2020, vol. 12347, Springer International Publishing, Cham, 2020, pp. 402–419, http://dx.doi.org/10.1007/978-3-030-58536-5_24, https://link.springer.com/10.1007/978-3-030-58536-5_24.

[340] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, Lidong Bing, VideoLLaMA2: advancing spatial-temporal modeling and audio understanding in video-LLMs, 2024, http://dx.doi.org/10.48550/arXiv.2406.07476, arXiv:2406.07476, http://arxiv.org/abs/2406.07476 [cs].

[341] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022, http://dx.doi.org/10.48550/arXiv.2207.02696, arXiv:2207.02696, http://arxiv.org/abs/2207.02696 [cs].

[342] Hao Zeng, Zhiyong Zhang, Lulin Shi, HEVC: Research and implementation of video codec based on ffmpeg, in: 2016 International Conference on Network and Information Systems for Computers (ICNISC), 2016, pp. 184–188, http://dx.doi.org/10.1109/ICNISC.2016.049, https://ieeexplore.ieee.org/document/7945976.

[343] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, Christoph Feichtenhofer, Perception encoder: the best visual embeddings are not at the output of the network, 2025, http://dx.doi.org/10.48550/arXiv.2504.13181, arXiv:2504.13181, http://arxiv.org/abs/2504.13181 [cs].

[344] Hamza Karim, Keval Doshi, Yasin Yilmaz, REWARD: Real-time weakly supervised video anomaly detection, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2024, pp. 6834–6842, http://dx.doi.org/10.1109/WACV57701.2024.00670, https://ieeexplore.ieee.org/document/10483693/.

[345] Ashish Singh, Michael J. Jones, Erik G. Learned-Miller, EVAL: explainable video anomaly localization, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada, 2023, pp. 18717–18726, http://dx.doi.org/10.1109/CVPR52729.2023.01795, https://ieeexplore.ieee.org/document/10205367/.

[346] Furkan Mumcu, Michael J. Jones, Yasin Yilmaz, Anoop Cherian, ComplexVAD: Detecting interaction anomalies in video, 2025, http://dx.doi.org/10.48550/arXiv.2501.09733, arXiv:2501.09733, http://arxiv.org/abs/2501.09733 [cs].

[347] Ruoyan Pi, Jinglin Xu, Yuxin Peng, FE-VAD: high-low frequency enhanced weakly supervised video anomaly detection, in: 2024 IEEE International Conference on Multimedia and Expo (ICME), 2024, pp. 1–6, http://dx.doi.org/10.1109/ICME57554.2024.10688326, https://ieeexplore.ieee.org/document/10688326/.

[348] Jiahao Lyu, Minghua Zhao, Jing Hu, Xuewen Huang, Yifei Chen, Shuangli Du, VADMamba: Exploring state space models for fast video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2503.21169, arXiv:2503.21169, http://arxiv.org/abs/2503.21169 [cs].

[349] Xiao Liu, Chenxu Zhang, Lei Zhang, Vision mamba: a comprehensive survey and taxonomy, 2024, http://dx.doi.org/10.48550/arXiv.2405.04404, arXiv:2405.04404 http://arxiv.org/abs/2405.04404 [cs].

[350] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, Francois Bremond, PiVAD: just dance with {π}! a poly-modal inductor for weakly-supervised video anomaly detection, 2025, http://dx.doi.org/10.48550/arXiv.2505.13123, arXiv:2505.13123, http://arxiv.org/abs/2505.13123 [cs].

[351] Manuel Barusco, Francesco Borsatti, Davide Dalle Pezze, Francesco Paissan, Elisabetta Farella, Gian Antonio Susto, Paste: improving the efficiency of visual anomaly detection at the edge, 2024, http://dx.doi.org/10.48550/arXiv.2410.11591, arXiv:2410.11591, http://arxiv.org/abs/2410.11591 [cs].

[352] Hoang Viet Pham, Thinh Gia Tran, Chuong Dinh Le, An Dinh Le, Hien Bich Vo, Jetsonvad: bnchmarking jetson edge devices with an end-to-end video-based anomaly detection system, vol. 920, 2024, pp. 358–374, http://dx.doi.org/10.1007/978-3-031-53963-3_25, arXiv:2307.16834 http://arxiv.org/abs/2307.16834 [cs].

[353] Arianna Stropeni, Francesco Borsatti, Manuel Barusco, Davide Dalle Pezze, Marco Fabris, Gian Antonio Susto, Towards scalable IoT deployment for visual anomaly detection via efficient compression, 2025, http://dx.doi.org/10.48550/arXiv.2505.07119, arXiv:2505.07119, http://arxiv.org/abs/2505.07119 [cs].

[354] Musrrat Ali, Lakshay Goyal, Chandra Mani Sharma, Sanoj Kumar, Edge-computing-enabled abnormal activity recognition for visual surveillance, Electron. 13 (2) (2024) 251, http://dx.doi.org/10.3390/electronics13020251, https://www.mdpi.com/2079-9292/13/2/251.

[355] Devashree R. Patrikar, Mayur Rajaram Parate, EdgeVAD: anomaly detection using edge computing in video surveillance system: Review, Int. J. Multimed. Inf. Retr. 11 (2) (2022) 85–110, http://dx.doi.org/10.1007/s13735-022-00227-8, https://doi.org/10.1007/s13735-022-00227-8.

[356] Muhammad Islam, Abdulsalam S. Dukyil, Saleh Alyahya, Shabana Habib, An IoT enable anomaly detection system for smart city surveillance, Sensors 23 (4) (2023) 2358, http://dx.doi.org/10.3390/s23042358, https://www.mdpi.com/1424-8220/23/4/2358.

[357] Waseem Ullah, Tanveer Hussain, Sung Wook Baik, Vision transformer attention with multi-reservoir echo state network for anomaly recognition, Inf. Process. Manage. 60 (3) (2023) 103289, http://dx.doi.org/10.1016/j.ipm.2023.103289, https://www.sciencedirect.com/science/article/pii/S0306457323000262.

[358] Michele Boldo, Mirco De Marchi, Enrico Martini, Stefano Aldegheri, Nicola Bombieri, Domain-adaptive online active learning for real-time intelligent video analytics on edge devices, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 43 (11) (2024) 4105–4116, http://dx.doi.org/10.1109/TCAD.2024.3453188, https://ieeexplore.ieee.org/abstract/document/10745828.

[359] Haocheng Shen, Bin Guo, Yasan Ding, Jie Xiao, Mingze Lv, Zhiwen Yu, Fast-DAVAD : Domain adaptation for fast video anomaly detection on resource-constrained edge devices, in: 2023 IEEE Smart World Congress (SWC), 2023, pp. 1–8, http://dx.doi.org/10.1109/SWC57546.2023.10448698, https://ieeexplore.ieee.org/document/10448698.

[360] Kahlil Muchtar, Adhiguna Mahendra, Muhammad Rizky Munggaran, Maya Fitria, Al Bahri, Fitri Arnia, Chih-Yang Lin, EdANo-Vision: An edge AI-powered anomaly detector using flask web-app framework, in: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2024, pp. 1–6, http://dx.doi.org/10.1109/AVSS61716.2024.10672600, https://ieeexplore.ieee.org/document/10672600.

[361] Dave Maass, Cooper Quintin, New ALPR vulnerabilities prove mass surveillance is a public safety threat, Electron. Front. Found. (2024) https://www.eff.org/deeplinks/2024/06/new-alpr-vulnerabilities-prove-mass-surveillance-public-safety-threat.

[362] Nazia Aslam, Kamal Nasrollahi, Balancing privacy and action performance: A penalty-driven approach to image anonymization, 2025, http://dx.doi.org/10.48550/arXiv.2504.14301, arXiv:2504.14301, http://arxiv.org/abs/2504.14301 [cs].

[363] BCDVideo, HIPAA compliance in video surveillance for health care settings, 2024, https://www.bcdvideo.com/blog/hipaa-compliance-in-video-surveillance-for-health-care-settings/,

[364] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov, Membership inference attacks against machine learning models, 2017, http://dx.doi.org/10.48550/arXiv.1610.05820, arXiv:1610.05820, http://arxiv.org/abs/1610.05820 [cs].

[365] Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, James Bailey, X-transfer attacks: Towards super transferable adversarial attacks on CLIP, 2025, http://dx.doi.org/10.48550/arXiv.2505.05528, arXiv:2505.05528 http://arxiv.org/abs/2505.05528 [cs].

[366] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, Li Fei-Fei, Privacy-preserving action recognition for smart hospitals using low-resolution depth images, 2018, http://dx.doi.org/10.48550/arXiv.1811.09950, arXiv:1811.09950, http://arxiv.org/abs/1811.09950 [cs].

[367] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, Zhangyang Wang, Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset, 2021, http://dx.doi.org/10.48550/arXiv.1906.05675, arXiv:1906.05675, http://arxiv.org/abs/1906.05675 [cs].

[368] Jing Liu, Yang Liu, Xiaoguang Zhu, PP2VAD: Privacy-preserving video anomaly detection: a survey, 2024, http://dx.doi.org/10.48550/arXiv.2411.14565, arXiv:2411.14565, http://arxiv.org/abs/2411.14565 [cs].

[369] Anas Al-lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, Karthik Nandakumar, Collaborative learning of anomalies with privacy (CLAP) for unsupervised video anomaly detection: a new baseline, 2024, http://dx.doi.org/10.48550/arXiv.2404.00847, arXiv:2404.00847, http://arxiv.org/abs/2404.00847 [cs].

[370] Joseph Fioresi, Ishan Rajendrakumar Dave, Mubarak Shah, Ted-SPAD: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection, 2023, http://dx.doi.org/10.48550/arXiv.2308.11072, arXiv:2308.11072, http://arxiv.org/abs/2308.11072 [cs].

[371] Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz, VISPR: Towards a visual privacy advisor: understanding and predicting privacy risks in images, 2017, http://dx.doi.org/10.48550/arXiv.1703.10660, arXiv:1703.10660 http://arxiv.org/abs/1703.10660 [cs].

[372] Ghazal Alinezhad Noghre, HuVAD1: Privacy-preserving real-world video anomaly detection, in: 2023 IEEE International Conference on Smart Computing (SMARTCOMP), 2023, pp. 235–254, http://dx.doi.org/10.1109/SMARTCOMP58114.2023.00067, https://ieeexplore.ieee.org/document/10207609/.

[373] Armin Danesh Pazho, Shanle Yao, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Vinit Katariya, Hamed Tabkhi, HuVAD2: towards adaptive human-centric video anomaly detection: a comprehensive framework and a new benchmark, 2025, http://dx.doi.org/10.48550/arXiv.2408.14329, arXiv:2408.14329, http://arxiv.org/abs/2408.14329 [cs].