Video-Analysis-as-a-Service with Foundation Models for Real-Time Multi-Event Detection

Vincenzo Carletti[1][0000−0002−9130−5533], Antonio Greco[1][0000−0002−5495−2432], Andrea Vincenzo Ricciardi[1][0009−0004−9476−8420], Alessia Saggese[1][0000−0003−4687−7994], Bruno Vento[2][0009−0006−7687−4929], and Antonio Vitale[1][0009−0003−7735−7964]

[1] Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Italy {vcarletti, agreco, anricciardi, asaggese, antovitale}@unisa.it

[2] Consorzio Interuniversitario Nazionale per l'Informatica (CINI), Via Ariosto 25, 00185 Rome, Italy brunovento.it@gmail.com

Abstract. The increasing deployment of camera networks in smart city environments has increased the need of video surveillance systems able to detect heterogeneous events in real-time while operating under resource-constrained conditions. Traditional video analytics approaches usually rely on task-specific models, which require dedicated training data and architectural redesign. In this work, we propose a unified framework for real-time multi-event detection based on foundation models for video-text matching. The proposed approach represents short video segments and textual event descriptions within a shared embedding space, enabling event detection through cosine similarity between video and text embeddings. The framework is implemented within a Video-Analysis-as-a-Service architecture, where the Open Vocabulary Event Detection (OVED) service acts as a modular semantic engine invoked on demand for different surveillance tasks. The proposed method is evaluated across three surveillance scenarios, including fall detection, fight recognition, and illegal waste dumping detection. Experimental results show that the framework achieves competitive or superior performance compared with specialized task-specific methods, while requiring only the modification of textual queries to address new events. Moreover, the proposed approach remains feasible on constrained hardware, confirming its suitability for scalable edge and cloud deployments.

Keywords: Video surveillance · Foundation models · Video-text matching · Multi-event detection · Smart cities

1 Introduction

The rapid proliferation of camera networks in urban environments has made video surveillance a central component of smart city infrastructures [15]. These systems are expected to operate continuously, detect a wide range of heterogeneous events, and scale across distributed and resource-constrained deployments. Traditional approaches for video analysis typically rely on task-specific

models trained for narrowly defined scenarios (e.g., action recognition, anomaly detection, or object tracking) [9, 4]. While effective in controlled settings, such approaches often lack generalization capabilities, require extensive labeled data, and impose significant maintenance costs when extended to new tasks or environments.

Recent advances in foundation models have opened new perspectives for visual understanding by enabling a unified representation space across modalities [11]. In particular, video-language models trained on large-scale multimodal data have demonstrated strong zero-shot and few-shot capabilities, allowing semantic alignment between visual content and natural language descriptions. This paradigm offers a promising alternative to conventional supervised pipelines by shifting the focus from task-specific training to flexible query-driven inference.

In this work, we propose a novel video surveillance framework based on a foundation model for video-text matching, designed to detect events of interest in live video streams. The core idea is to represent both short video segments and textual event descriptions within a shared embedding space, and to perform event detection through cosine similarity comparison. Incoming video streams are processed as temporal chunks (e.g., 20 seconds), enabling near real-time analysis while preserving temporal context. This formulation allows the system to detect diverse events without retraining, but by modifying the textual query.

We demonstrate the effectiveness and generality of the proposed approach across multiple surveillance scenarios, including illegal waste dumping, fight detection and fall detection. Despite the variability of these tasks in terms of visual appearance and temporal dynamics, the proposed method achieves competitive performance on challenging benchmark datasets, highlighting the robustness of foundation model-based representations in real-world conditions.

Beyond the methodological contribution, we address key practical constraints of operational surveillance systems. We introduce a video-analysis-as-a-service architecture in which the foundation model is invoked on demand by modular plugins associated with different event detectors. This design enables efficient resource utilization by avoiding continuous processing when no relevant activity is present, and supports deployment across heterogeneous computing environments, from edge devices to centralized servers. The modularity of the system further facilitates extensibility, allowing new event types to be integrated with minimal effort.

In summary, the contributions of this paper are threefold:

- A unified framework for event detection in video surveillance based on video-text similarity using foundation models.
- A scalable and resource-aware video-analysis-as-a-service architecture tailored to real-world smart city deployments.
- An evaluation across diverse application domains demonstrating the flexibility and the accuracy of the approach.

These results suggest that foundation models can be successfully adopted in the next generation of adaptive and intelligent video surveillance systems.

2 Related work

The analysis of anomalous and complex events in video streams has been extensively studied in computer vision, particularly in the context of intelligent surveillance systems for smart cities [15]. Existing approaches can be broadly categorized into task-specific detection methods and general-purpose foundation models for video understanding leveraging multimodal representations.

Among the task-specific detection methods, we focus on the models dedicated to fall detection, fight recognition, and illegal waste dumping detection. While these approaches achieve strong performance within their target domains, they are typically tailored to a single task and require task-specific datasets and training procedures.

The fall detection task has been extensively analyzed in literature, evolving from early heuristic-based approaches to semantic reasoning models. Initial works relied on handcrafted features and geometric heuristics. Rougier et al. [30] proposed a method based on body aspect ratio variation and orientation, modeled through Gaussian Mixture Models (GMMs) to identify abnormal posture transitions associated with falls. While computationally efficient, these methods were highly sensitive to environmental conditions, limiting their applicability in real-world scenarios. To mitigate these limitations, Anderson et al. [4] introduced a three dimensional human body representation, named voxel person, constructed from multi-camera silhouettes. However, this system required complex multi-camera setups and remained difficult to scale in real world deployments. Early deep architectures typically combined 2D Convolutional Neural Networks (CNNs) for spatial representation with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units to capture temporal dynamics [23, 25, 13]. More advanced architectures introduced 3D CNNs [2] and Three-Stream [34] networks, where RGB frames and optical flow were processed separately to disentangle appearance and motion cues. Despite their effectiveness in controlled environments, these approaches often suffered from domain shift. More recently, research has focused on Transformer-based architectures to better capture long-range spatio-temporal dependencies. Sanjalawe et al. [31] integrate YOLOv8 for human detection, followed by temporal modeling through Time-Space Transformers applied to capture fall trajectories. Additionally, Kibet et al. [18] combine pose estimation with constrained Generative Adversarial Networks (cGANs) augmented by Transformer layers to enhance robustness under occlusions and noisy observations. However, the high computational cost of attention mechanisms poses significant challenges for real-time deployment.

Fight detection in videos is a well-established research topic within the broader field of action recognition, focusing on the identification of aggressive human interactions in surveillance and unconstrained environments [27, 29]. Early approaches relied on handcrafted features through video preprocessing, feature extraction, and classification pipelines [26]. Within this paradigm, Datta et al. [9] introduced a person-on-person violence detection method leveraging motion trajectories and bodies orientations. Subsequently, Nguyen et al. [28] employed a Hidden Markov Model (HMM) combined with a particle filter, allowing real-time

activity recognition. Following these approaches, Li et al. [20] proposed a framework for anomaly detection in crowded scenarios based on the joint modeling of appearance and dynamics, where normal behavior is represented as a dynamic texture and anomalies are identified as deviations in spatial and temporal patterns. Handcrafted feature-based methods suffer from limited representation capacity and environment sensitivity real-world scenarios [39]. One of the first approaches that used Deep Learning was introduced by Ding et al. [12], employing 3D CNNs to automatically learn spatio-temporal representations from sequences of frames. Following this, hybrid approaches were applied. For instance, Mabrouk [24] extract spatio-temporal optical flow features by modeling motion magnitude and orientation distributions, which are encoded into descriptors and classified using an SVM. Sudhakaran [35] introduced a more advanced strategy, who aggregate frame-level features using Convolutional LSTM (ConvLSTM), allowing spatial filtering and temporal modeling through convolutional recurrent units. Although deep learning-based approaches significantly improve detection accuracy, they often involve high computational complexity and large numbers of parameters, limiting their deployment in real-time surveillance systems [39]. A recent research direction is represented by the adoption of transformer-based architectures. Vision Transformers (ViT), as introduced by Dosovitskiy et al. [11], leverage self-attention mechanisms to model global dependencies by processing video frames as sequences of patches. Based on this paradigm, Bertasius et al. [6] proposed the TimeSformer architecture, which employs a divided attention mechanism to separately model spatial and temporal dependencies in video sequences.

Illegal Waste Dumping Detection (IWDD) is an emerging research topic, focusing on the recognition of human-object interactions that lead to the unauthorized abandonment of waste. The task is particularly challenging, as it requires modeling the temporal evolution of actions rather than simply detecting waste presence, and has a strong impact on public health and environmental sustainability, motivating the need for automated real-time monitoring systems [7]. Different approaches in IWDD have been based on deep learning architectures designed for spatio-temporal modeling. For instance, Zbuce et al. [42] proposed SUDOPI, a hybrid pipeline combining YOLOv8 for object detection, BoT-SORT for multi-object tracking, and a dual-timescale MOG2 background model to identify newly deposited waste through temporal consistency constraints. Similarly, ALEXKNIGHTS [3] adopts a lightweight X3D-M backbone based on depthwise separable 3D convolutions and channel-wise attention to efficiently model motion patterns. Multi-stage architectures have also been explored, such as UNICASS [10], which integrates a ResNet-18 for frame-level feature extraction with a bidirectional GRU to capture temporal dependencies across fixed-length frame sequences. Other approaches shift towards Transformer-based and multimodal paradigms. In this direction, GARBERUS [32] employs a TimeSformer backbone within a multi-task learning framework to perform event detection, temporal localization, and disposal-type classification. In parallel, in ARRAY [17] formulates the task as a multimodal temporal classification problem by combin-

ing vision-language captioning with Sentence-BERT embeddings for temporal decision making.

These methods are special-purpose, with few data available to recognize complex and specific event patterns. There is the lack of a unified framework capable of handling multiple heterogeneous events without task-specific redesign and retraining. To overcome the limitations of task-specific approaches, research has moved toward learning general-purpose video representations. Early attempts relied on large-scale supervised training of deep architectures, including Transformer-based models, to capture spatio-temporal patterns across diverse datasets [37]. While these models improve generalization, they typically still require fine-tuning for downstream tasks and remain dependent on labeled data. Recent advances in foundation models have introduced a new paradigm based on vision-language models, which learn a shared embedding space where visual content and textual descriptions can be directly compared [33]. This enables zero-shot and query-driven inference, significantly reducing the need for task-specific supervision. Extensions of these models to the video domain have demonstrated promising results in tasks such as video retrieval, captioning, and action recognition [21, 16, 40]. By aligning video segments with natural language descriptions, these approaches allow flexible semantic querying of visual content. However, their application to continuous video surveillance and real-time event detection remains largely unexplored. In particular, existing works rarely address the challenges of streaming data, temporal chunking, and operational constraints typical of real-world surveillance systems.

To this aim, significant efforts have been devoted to design scalable and resource-efficient video surveillance architectures. These include edge computing solutions [36], event-triggered processing pipelines [14], and distributed inference frameworks aimed at reducing latency and computational cost [22]. Despite these advances, most systems are tightly coupled with task-specific models and lack the flexibility to dynamically adapt to new event types. The integration of foundation models into such architectures remains an open challenge, particularly in scenarios requiring real-time responsiveness and efficient resource utilization.

In this paper, we address the limitations of existing approaches by introducing a unified and flexible framework for multi-event detection in video surveillance. Unlike prior work, our method:

- leverages a foundation model for video-text matching, enabling query-driven detection without task-specific retraining.
- supports heterogeneous event types within a single framework.
- is deployed through a video-analysis-as-a-service architecture, allowing on-demand inference and efficient resource usage.

This combination of multimodal foundation models and service-oriented system design distinguishes our approach from both traditional task-specific pipelines and existing general-purpose foundation models for video understanding leveraging multimodal representations.

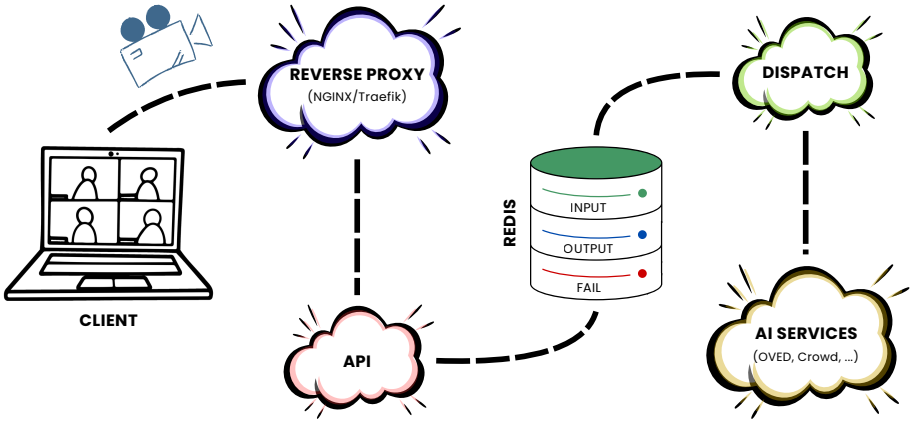

Fig. 1. End to end communication pipeline of the AI-VAS platform, showing the interaction between all the components.

3 Proposed method

The proposed methodology introduces a unified framework for multi-event detection in urban surveillance that shifts away from traditional task-specific supervised models toward a more scalable and resource-efficient video-analysis-as-a-service paradigm. This architecture is structured into two primary components: the AI-VAS platform, which functions as the foundational communication infrastructure, and the Open Vocabulary Event Detection (OVED) service, a modular analytical plugin that serves as the core semantic engine for the specific events examined in this study.

3.1 AI-VAS

The AI-VAS [8] (Video-Analysis-as-a-Service) platform uses microservices for scalable on-demand video and image analysis. This modular design supports containerization and orchestration via Kubernetes across cloud or edge environments. As illustrated in Fig. 1, the processing of a client request traverses a well-defined communication pipeline comprising independent yet closely cooperating components.

The entry point of this pipeline is managed by a reverse proxy, typically implemented via Nginx or Traefik depending on the deployment environment. This component is responsible for Transport Layer Security (TLS) termination and mutual authentication (mTLS), ensuring that only authorized clients with internal Certification Authority (CA) certificates access the system. Validated requests reach the API service, which maintains persistent WebSocket connections for streaming. This service validates payloads via Pydantic models and bundles media streams into structured packages.

To decouple data reception from computation, the platform uses Redis as an asynchronous message broker. Requests are enqueued into task-specific channels and retrieved by the dispatch service. Using priority mechanisms like min-heaps, the dispatcher balances system load before forwarding packages to computational nodes via asynchronous HTTP POST requests.

The actual semantic inference is executed by the AI services, which form the computational core of the framework. These microservices encapsulate the operational logic of the various analytical services and expose dedicated HTTP endpoints. To maximize throughput and hardware utilization, incoming frames are dynamically aggregated into mini-batches based on predefined size thresholds or time constraints (patience time), enabling the underlying models to process multiple vectors simultaneously. Once inference is complete, results return through the pipeline: the dispatch service routes output to a Redis queue, where an asynchronous thread in the API Service retrieves and transmits the final detection to the client via the active socket.

3.2 The OVED service

The Open Vocabulary Event Detection service represents the specific video analytics evaluated in this work. Operating on top of the broader AI-VAS infrastructure, this service leverages a zero-shot multimodal alignment strategy that bridges the gap between visual surveillance data and natural language. By projecting both dynamic video sequences and descriptive textual queries into a shared high-dimensional embedding space, the OVED service enables the detection of heterogeneous events without the need for domain-specific fine-tuning or extensive labeled datasets. This service is tailored to handle specific security concerns, namely fall detection, fight recognition, and illegal waste dumping detection.

The foundation model. The technical foundation of the proposed system is the Qwen3-VL-Embedding model [19], a specialized bi-encoder architecture derived from the Qwen3-VL multimodal series [5]. For the purposes of this study, the variant with 2B parameters was selected to balance high-precision semantic retrieval with computational efficiency. To further optimize resource utilization, the model was quantized to 8 bits. This 8-bit quantization process effectively halves both the model's storage footprint and the required VRAM, all without incurring any loss in overall performance. The architecture is engineered to project diverse modalities, including text, images, and video sequences, into a unified high-dimensional manifold where semantic relevance is measured by spatial proximity (see Fig. 2). At its core, the model integrates a SigLIP-2 vision encoder [38], which is optimized for dynamic resolution handling, allowing the system to process frames in their native aspect ratios without the spatial distortions typically introduced by static resizing. To maintain temporal and spatial coherence in video data, the model employs Interleaved Multimodal Rotary Positional Embedding (MROPE). This mechanism enables the simultaneous

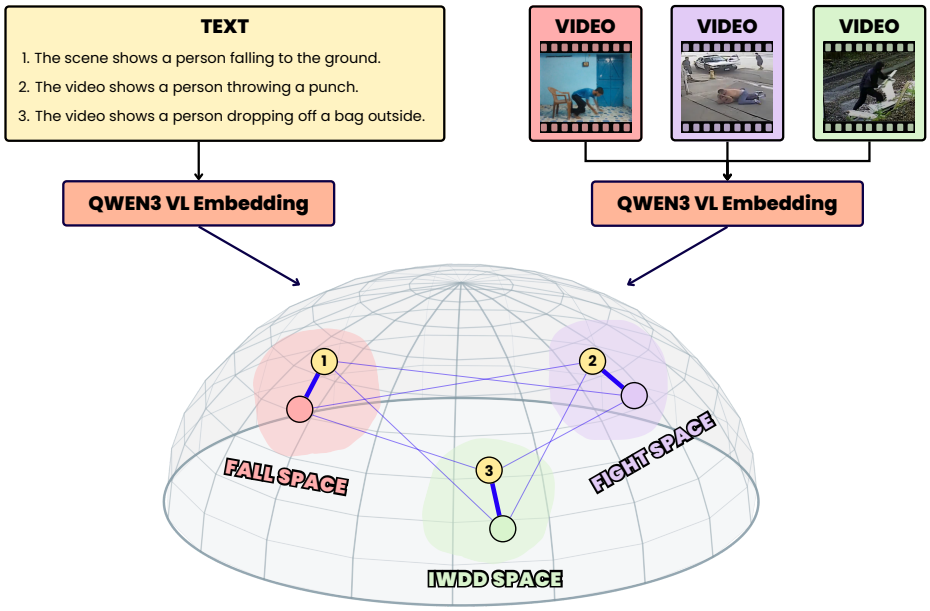

Fig. 2. A unified multimodal representation space is illustrated, where the Qwen3-VL-Embedding model aligns text and video data for tasks such as fall detection, fight recognition, and illegal waste dumping detection. By mapping both modalities into a shared representation space, the model enables meaningful connections between textual event descriptions and their corresponding visual patterns in video sequences.

encoding of temporal, vertical, and horizontal dimensions, ensuring that the sequential nature of an event is accurately captured within the visual tokens. The 2B model maps these inputs into a 2048-dimensional embedding space, adopting a sequence length capacity of up to 32K tokens to accommodate complex video contexts. The final representation is extracted from the last hidden state of the language model decoder, specifically targeting the final padding token (`<|endoftext|>`), which serves as a global semantic descriptor for the entire multimodal input.

Instruction-aware multi-event alignment. To balance computational overhead with the preservation of temporal context, incoming video streams V are partitioned into 20-second chunks, sampled at a consistent rate of 1 FPS. Each segment is processed by the foundation model to produce a representative visual vector $\mathbf{e}_v \in \mathbb{R}^{2048}$. This vector captures the sequential dynamics inherent in activities like the sudden descent of a fall or the chaotic motion of a fight.

The system aligns the model's internal world knowledge with specific surveillance tasks through instruction-aware prompting. In addition to the primary textual query T, the framework allows for the specification of a system prompt I, which acts as a high-level guiding instruction to steer the model's internal

reasoning and orient the video embedding projection toward the intended task context. The selection of the reference texts is the result of an extensive, but automatic, search process. Taking advantage of the decoupled nature of the proposed framework, we evaluated several hundred textual prompt candidates per task, comparing them against the same video embeddings. The optimal prompt for each task was identified using task-specific private validation sets. Once the best-performing instruction was locked, it was evaluated on public test sets to ensure generalizability; these final performance metrics are reported in Section 4. For the three primary application scenarios, the automatic procedure selected the following optimized textual queries T to generate the reference embeddings $\mathbf{e}_t \in \mathbb{R}^{2048}$:

- T_{FALL} = `"The scene shows a person falling to the ground."`
- T_{FIGHT} = `"The video shows a person throwing a punch."`
- T_{IWDD} = `"The video shows a person dropping off a bag outside."`

The detection of an event is formulated as the calculation of the cosine similarity (S_c) between the video embedding $\mathbf{e}_v$ and the reference embedding $\mathbf{e}_t$:

$$S_c(\mathbf{e}_v, \mathbf{e}_t) = \frac{\mathbf{e}_v \cdot \mathbf{e}_t}{|\mathbf{e}_v||\mathbf{e}_t|} \tag{1}$$

Classification is governed by a deterministic decision boundary based on a threshold τ, which can be calibrated specifically for each scenario. Consequently, a video is classified as a positive event if the similarity exceeds the threshold:

$$Y = \begin{cases} 1 & \text{if } S_c(\mathbf{e}_v, \mathbf{e}_t) > \tau \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

By utilizing this query-driven inference, the system functions as a modular event detector. New scenarios can be integrated with minimal effort by simply adding a new textual query T and a corresponding guiding instruction I to the pipeline, significantly reducing maintenance and deployment costs typical of traditional surveillance infrastructures.

4 Results

In this section, we provide a comprehensive evaluation of the proposed framework. First, we define the evaluation metrics used to assess the temporal accuracy of event detection. Subsequently, we describe the experimental setup, including the datasets employed for each surveillance task, and discuss the quantitative results.

4.1 Event-based detection metrics

Detection performance is evaluated by comparing the predicted timestamp p with the ground truth g. The timestamp p is defined as the midpoint of the

Table 1. Per-task comparison of methods across quality and model configuration metrics. Best results are in **bold**, second-best are underlined. Within each task, methods are grouped by name and ordered by ascending F_1-score. Thresholds were set to $\tau = 0.5$ for FALL and FIGHT, and $\tau = 0.55$ for IWDD as optimal values on the private validation set. Temporal windows (T_b, T_a) are reported in seconds; larger windows are used for FIGHT due to the longer video duration and higher temporal overlap between events.

	Method	Metrics			Configuration		
		P	R	F_1	τ	(T_b, T_a)	System Prompt
FALL	TCNTE [41]	0.846	0.916	0.904	–	–	–
	OVED	0.545	**1.000**	0.705	0.50	[5,10]	Given a video, determine whether a person falls during the scene.
		0.974	0.962	**0.968**	0.50	[5,10]	Represent the user's input.
FIGHT	**OVED**	0.559	0.660	0.606	0.50	[20,20]	Given a video, determine whether a fight occurs during the scene.
		0.698	**0.740**	**0.718**	0.50	[20,20]	Represent the user's input.
IWDD	GARBERUS [32]	0.540	0.540	0.540	0.55	[5,10]	–
	AGH-EVS [7]	0.470	0.740	0.570	0.55	[5,10]	–
	OVED	0.705	0.860	0.775	0.55	[5,10]	Represent the user's input.
		0.789	**0.900**	**0.841**	0.55	[5,10]	Given a video, determine whether a person dumps trash during the scene.

first 20-second chunk in which a positive event is detected (e.g., at 30 s if the detection occurs at the second chunk of the video). To account for minor temporal misalignments, early detections are tolerated within a margin of T_b seconds, while detections occurring more than T_a seconds after g are considered invalid.

A detection is counted as a true positive (TP) if it occurs within a positive video and falls within the valid temporal window $(g - T_b) \leq p \leq (g + T_a)$. Detections in positive videos that fall outside this interval are treated as both False Positives (FP) and False Negatives (FN), as they fail to correctly identify the event. Following these definitions, standard event-level metrics, namely Precision (P), Recall (R), and F_1 score, are calculated as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{F}_1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{3}$$

4.2 Performance analysis

The effectiveness of the proposed OVED service has been validated across three heterogeneous surveillance scenarios using the following datasets:

- **Fall Detection**: Evaluated on the GMDCSA24 dataset [1], comprising 81 fall events and 79 normal activity clips.
- **Fight Recognition**: Conducted using the UCF-CRIME dataset, specifically filtering for "Fight" as the positive class (50 clips) against 147 normal video sequences.

– **IWDD**: Evaluated on the Mivia-IWDD Private Test Set [7], consisting of 100 balanced videos (50 dumping events and 50 normal samples).

Table 1 presents a quantitative evaluation of the proposed OVED framework against state-of-the-art (SOTA) task-specific architectures, incorporating an ablation study on system prompt engineering across varying semantic complexities. For the FALL and FIGHT benchmarks, the transition from the default system prompt (`"Represent the user's input."`) to task-specific instructions paradoxically yields a performance degradation, characterized by a significant drop in precision; this suggests that for tasks with high inherent visual-textual alignment already well-represented during model training, overly descriptive prompts may induce a predictive bias. In contrast, the custom system prompt proves to be significantly more effective for downstream tasks that may not have been extensively seen during training, such as IWDD. In this scenario, providing tailored instructions facilitates granular disambiguation and yields a substantial performance gain, specifically improving the F_1-score by approximately 7% (from 0.775 to 0.841). Ultimately, OVED demonstrates superior generalization capabilities, significantly outperforming SOTA baselines by achieving F_1-scores of 0.968 in the FALL task (surpassing TCNTE's 0.904) and 0.841 in IWDD (exceeding AGH-EVS's 0.570).

4.3 Computational analysis

To assess the operational feasibility of the framework in real-world smart city deployments, we conducted a performance analysis focusing on inference latency and VRAM utilization. The experiments were performed on a workstation equipped with an NVIDIA RTX 2070 GPU (8 GB VRAM). Table 2 summarizes the results for different spatial resolutions and clip lengths, processed at a fixed rate of 1 FPS. The scalability of the system is determined by two primary constraints: the maximum number of service instances that can be hosted in the GPU memory ($N_{services}$) and the number of concurrent video streams each instance can process within the temporal window ($N_{streams}$).

Following the 20-second chunking strategy defined in our methodology, we calculate the capacity for a 1080p surveillance setup as follows:

1. **Service Capacity**: Given the 8 GB (8192 MiB) capacity of the RTX 2070 and an inference memory requirement of $\approx$ 4800 MiB per 1080p service instance, the GPU can host $N_{services} = \lfloor 8192/4800 \rfloor = 1$ active service.
2. **Throughput per Service**: For a 20-second clip, the inference time is $T_{inf} \approx 2.28$ seconds. The number of streams handled by a single service is $N_{req} = \lfloor 20/2.28 \rfloor = 8$ streams.
3. **Total Scalability**: The total number of concurrent streams is $N_{streams} = N_{services} \times N_{req} = 8$ concurrent streams.

In a lower-resolution scenario (640x360), while the memory constraint still limits the deployment to one or two services, the reduced inference time (1.58

Table 2. Hardware performance and scalability on NVIDIA RTX 2070 (8GB). All tests were conducted at a sampling rate of 1 FPS.

Video Info		**VRAM**	**Inference**	**Max Concurrent**
Resolution	Clip Length	(MiB)	Time (s)	Streams ($N_{streams}$)
640x360	10 s	3170	0.72	13
640x360	20 s	4100	1.58	12
640x360	30 s	4780	2.05	14
1920x1080	10 s	4800	2.26	4
1920x1080	20 s	4800	2.28	8
1920x1080	30 s	4800	2.34	12

s) allows each service to handle up to 12 concurrent streams. This modular architecture allows for flexible resource allocation, confirming that the foundation model approach is compatible with edge-centralized hybrid infrastructures.

5 Conclusions

In this work we introduced a unified framework for real-time and multi-event detection in video surveillance based on foundation models. Starting from the limitations of traditional task-specific approaches, which usually require dedicated architectures, annotated datasets, and retraining for each new scenario, we proposed a query-driven methodology based on video-text similarity, where both video chunks and textual event descriptions are represented in a shared embedding space, allowing heterogeneous events to be detected by simply changing the reference text.

Our proposed framework was implemented within a Video-Analysis-as-a-Service architecture, where the OVED service acts as a modular semantic engine for surveillance event detection. By processing video streams as temporal chunks and comparing their embeddings with optimized textual prompts, the system supports flexible and on-demand inference without task specific fine-tuning. This design enables the integration of new event categories with minimal configuration effort, while preserving compatibility with scalable edge and cloud environments.

The evaluation analysis demonstrated the effectiveness of the proposed approach across three different surveillance scenarios of fall detection, fight recognition, and illegal waste dumping detection. The results show that this framework can achieve competitive or superior performance compared with specialized methods, and highlighting the importance of semantic formulation in the reference text, confirming that textual queries play a key role in the alignment between visual events and language descriptions. These results suggest that foun-

dation models can effectively support the next generation of adaptive, extensible, and intelligent video surveillance systems for smart city environments.

References

1. Alam, E., Sufian, A., Dutta, P., Leo, M., Hameed, I.A.: Gmdcsa-24: A dataset for human fall detection in videos. Data in Brief **57**, 110892 (2024). https://doi.org/https://doi.org/10.1016/j.dib.2024.110892, https://www.sciencedirect.com/science/article/pii/S2352340924008552
2. Alanazi, T., Muhammad, G.: Human fall detection using 3d multi-stream convolutional neural networks with fusion. Diagnostics **12**(12), 3060 (2022)
3. Almorsy, S., Torki, M.: X3d-based illegal waste dumping detection with temporal localization. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
4. Anderson, D., Luke, R.H., Keller, J.M., Skubic, M., Rantz, M., Aud, M.: Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Computer vision and image understanding **113**(1), 80–89 (2009)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
6. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Icml. vol. 2, p. 4 (2021)
7. Bouwmans, T., Greco, A., Pierard, S., Ricciardi, A.V., Sansone, C., Van Droogenbroeck, M., Vento, B.: Illegal waste dumping detection. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 539–548 (2026)
8. Cerasuolo, C., Ricciardi, A.V., Rocco, D., Saldutti, S., Vento, B., Vitale, A.: AI-VaS: Empowering edge video analytics with cloud-based services. In: 5th National Conference on Artificial Intelligence (Ital-IA 2025). CEUR Workshop Proceedings, vol. 4121. Trieste, Italy (June 2025)
9. Datta, A., Shah, M., Lobo, N.D.V.: Person-on-person violence detection in video data. In: 2002 International conference on pattern recognition. vol. 1, pp. 433–438. IEEE (2002)
10. Delussu, R., Putzu, L.: A lightweight temporal detection framework for illegal waste dumping in real surveillance footage. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Gao, M., Zheng, F., Yu, J.J., Shan, C., Ding, G., Han, J.: Deep learning for video object segmentation: a review. Artificial Intelligence Review **56**(1), 457–531 (2023)
13. Garcia, E., Villar, M., Fáñez, M., Villar, J.R., De La Cal, E., Cho, S.B.: Towards effective detection of elderly falls with cnn-lstm neural networks. Neurocomputing **500**, 231–240 (2022)

14. Gehrig, M., Scaramuzza, D.: Recurrent vision transformers for object detection with event cameras. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 13884–13893 (2023)
15. Hassan, N., Hesham, S., Fares, S., Hesham, M., Shaheen, L., Halim, I.: Intelligent Surveillance Systems for Smart Cities: A Systematic Literature Review, pp. 135–147 (01 2022). https://doi.org/10.1007/978-981-16-2877-1_14
16. Ji, K., Liu, J., Hong, W., Zhong, L., Wang, J., Chen, J., Chu, W.: Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In: 45th international ACM SIGIR conference on research and development in information retrieval. pp. 949–959 (2022)
17. Kairanbay, M.: Vision-language temporal analysis for illegal waste dumping detection in surveillance videos. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
18. Kibet, D., So, M.S., Kang, H., Han, Y., Shin, J.H.: Sudden fall detection of human body using transformer model. Sensors **24**(24) (2024). https://doi.org/10.3390/s24248051, https://www.mdpi.com/1424-8220/24/24/8051
19. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., Lin, J.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking (2026), https://arxiv.org/abs/2601.04720
20. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence **36**(1), 18–32 (2013)
21. Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: Swinbert: End-to-end transformers with sparse attention for video captioning. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 17949–17958 (2022)
22. Liu, X., Song, Y., Li, X., Sun, Y., Lan, H., Liu, Z., Jiang, L., Li, J.: Efficient partitioning vision transformer on edge devices for distributed inference. In: 2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS). pp. 286–296 (2025). https://doi.org/10.1109/ICDCS63083.2025.00036
23. Lu, N., Wu, Y., Feng, L., Song, J.: Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data. IEEE journal of biomedical and health informatics **23**(1), 314–323 (2018)
24. Mabrouk, A.B., Zagrouba, E.: Spatio-temporal feature using optical flow based distribution for violence detection. Pattern Recognition Letters **92**, 62–67 (2017)
25. Maitre, J., Bouchard, K., Gaboury, S.: Fall detection with uwb radars and cnn-lstm architecture. IEEE journal of biomedical and health informatics **25**(4), 1273–1283 (2020)
26. Mumtaz, N., Ejaz, N., Habib, S., Mohsin, S.M., Tiwari, P., Band, S.S., Kumar, N.: An overview of violence detection techniques: current challenges and future directions. Artificial intelligence review **56**(5), 4641–4666 (2023)
27. Negre, P., Alonso, R.S., González-Briones, A., Prieto, J., Rodríguez-González, S.: Literature review of deep-learning-based detection of violence in video. Sensors **24**(12), 4016 (2024)
28. Nguyen, N.T., Phung, D.Q., Venkatesh, S., Bui, H.: Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 2, pp. 955–960. IEEE (2005)

29. Ramzan, M., Abid, A., Khan, H.U., Awan, S.M., Ismail, A., Ahmed, M., Ilyas, M., Mahmood, A.: A review on state-of-the-art violence detection techniques. IEEE Access **7**, 107560–107575 (2019)
30. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Robust video surveillance for fall detection based on human shape deformation. IEEE Transactions on circuits and systems for video Technology **21**(5), 611–622 (2011)
31. Sanjalawe, Y., Fraihat, S., Abualhaj, M., Al-E'Mari, S.R., Alzubi, E.: Hybrid deep learning for human fall detection: A synergistic approach using yolov8 and time-space transformers. IEEE Access **13**, 41336–41366 (2025). https://doi.org/10.1109/ACCESS.2025.3547914
32. Scarrica, V., Placitelli, A.P., Staiano, A.: Garberus: Three-headed video classifier to guard against illegal waste dumps. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
33. Selva, J., Johansen, A.S., Escalera, S., Nasrollahi, K., Moeslund, T.B., Clapés, A.: Video Transformers: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 12922–12943 (Nov 2023). https://doi.org/10.1109/TPAMI.2023.3243465, https://ieeexplore.ieee.org/document/10041724/
34. Su, C., Wei, J., Lin, D., Kong, L., Guan, Y.L.: A novel model for fall detection and action recognition combined lightweight 3d-cnn and convolutional lstm networks. Pattern analysis and applications **27**(1), 3 (2024)
35. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). pp. 1–6. IEEE (2017)
36. Sun, M., Ren, J., Zhang, Y.: Edgevpr: Transformer-based real-time video person re-identification at the edge. In: 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS). pp. 13–24 (2024). https://doi.org/10.1109/ICDCS60910.2024.00011
37. Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., Xu, C.: Video understanding with large language models: A survey. IEEE Transactions on Circuits and Systems for Video Technology **36**(2), 1355–1376 (2026). https://doi.org/10.1109/TCSVT.2025.3566695
38. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025), https://arxiv.org/abs/2502.14786
39. Ullah, F.U.M., Obaidat, M.S., Ullah, A., Muhammad, K., Hijji, M., Baik, S.W.: A comprehensive review on vision-based violence detection in surveillance videos. ACM Computing Surveys **55**(10), 1–44 (2023)
40. Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., Jiang, Y.G.: Svformer: Semi-supervised video transformer for action recognition. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 18816–18826 (2023)
41. Yu, X., Wang, C., Wu, W., Xiong, S.: A real-time skeleton-based fall detection algorithm based on temporal convolutional networks and transformer encoder. Pervasive and Mobile Computing **107**, 102016 (2025)
42. Zbuce, A.D., Laza, L., Ticarat, A., Moisi, E., Popescu, D.: Hybrid temporal-spatial novelty detection for illegal waste dumping in surveillance video. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2026)