Mrs. Vibhavari Jawale, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Live Surveillance with Actionable Intelligence

**Mrs. Vibhavari Jawale[1] , Mrs. Deepali Hajare[2] , Arhant Sahuji[3] , Tanay Shinde[4] , Ritesh Kadam[5] , Ananya Vaishnav[6]**

Associate Professors and UG Scholars, Dept. of AIDS, Dr. D. Y. Patil Institute of Engineering, Management and Research Pune, Maharashtra, India

**Abstract— The combination of advanced machine learning and AI capabilities associated with natural language processing (NLP) and technological developments in computer vision have led to the creation of smart video surveillance systems. Unlike conventional Closed Circuit Television (CCTV) and motion-detection-based surveillance devices, these systems are capable of understanding contextual information, reducing false alarms and requiring less human intervention. Researchers have explored incorporating vision-language models (VLMs) and Sentiment Analysis (SA) into video surveillance applications to improve contextual awareness. This review focuses on emerging techniques associated with image captioning models, including Salesforce's BLIP, used to generate natural language descriptions of real-time actions in video footage and perform SA on those descriptions to determine the nature of the detected activity. By utilizing visual comprehension, context building, and sentiment interpretation, surveillance systems can differentiate between normal and suspicious behavior while reducing false positives and generating actionable insights. Applications include public safety in smart cities, security at high-threat locations such as airports and banks, and monitoring of sensitive areas including hospitals and military installations. This review evaluates how contextual awareness enabled by VLM improves traditional object detection methodologies and supports the transition toward more human-like and explainable alerting modalities. It also discusses limitations related to computational burden, accuracy, and privacy, while highlighting broader societal implications aligned with Sustainable Development Goals focused on urban safety and crime reduction. Future research directions include multimodal fusion, real-time optimization, and ethical guidelines for responsible deployment.**

**Keywords— Vision-Language Model, Intelligent Surveillance, Anomaly Detection, Object Tracking, Real-Time Alerting.**

## I. INTRODUCTION

Surveillance technologies have evolved as a cornerstone of modern security practices, finding application across diverse domains such as public safety, healthcare, defence, and critical infrastructure. The majority of present-day traditional security systems provide only a limited scope of functionality and are primarily focused on detecting motion and recording video feeds as well as doing very simple rule-based anomaly detection. These methods do not provide the level of semantic understanding needed to interpret activities viewed through the camera and place heavy demands on human operators to interpret those events; therefore, there are often delays in responding to events (when needed), many false alarm events, and lack of contextual explanations for alerts. Consequently, as risks associated with urban security, terrorism, and workplace safety increase, the demand for context-aware, intelligent surveillance systems is greater than at any other time in history.

With recent advancements in artificial intelligence (AI) and particularly the development of Vision-Language Models (VLMs), there has been significant improvement in the ability of an artificial intelligence model to link visual input from a camera to an interpretation similar to that of a human being. For example, Salesforce has developed a model called BLIP (Bootstrapping Language-Image Pretraining) that can create a meaningful caption (text) from an image; in addition, there are numerous transformer-based natural language processing (NLP) tools (e.g., sentiment analysis) capable of creating a text output that can be analyzed for both the intention and level of threat that is present within the text output. The intersection of

computer vision with natural language understanding represents a fundamental shift in the design of surveillance systems from passive monitoring systems to proactive, context-aware agents capable of identifying both activities occurring in real-time and measuring situational risk.

This review explores the latest advancements in the fields of computer vision, computer captioning, and sentiment analysis with the aim of demonstrating the benefits and advantages of using these technologies together in the design of intelligent surveillance systems. The review covers the technologies and architectures enabling these fields to converge, several different settings where intelligent surveillance technologies are being used today (airports, hospitals, banking, military, and smart cities), and challenges that come along with the integration of these fields (e.g., computational requirements, the potential for misclassification, privacy concerns). As well as discussing potential future directions for research in these areas (e.g., multi-modal learning, edge computing, and designing ethical AI), this review also contains three major contributions: (1) provides evidence for the need to move beyond simply detecting motion to building completely interpretable ways of monitoring and evaluating any activity with a person(s); (2) introduces a methodology that combines vision-language modelling and sentiment-based interpretative assessment to learn how to provide real-time alerts based on intelligent surveillance; and (3) positions this technology direction within a global sustainability and social impact context, aligning with United Nations Sustainable Development Goals (SDGs) for sustainable cities, and issues related to peace, justice, and strong institutions.

## II. LITERATURE SURVEY

Table I
Literature Survey

| Sr No. | Paper Title | Journal | Author & Year | Methodology |
|---|---|---|---|---|
| 1 | Vision transformer embedded video anomaly detection using attention driven recurrence | Elsevier | Ummay Maria Muna; 2025 | Embeds Vision Transformers with recurrence for contextual video anomaly detection. |
| 2. | SmolVLM: Redefining small and efficient multimodal models | arXiv preprint | Andrés Marafioti, Orr Zohar; 2025 | Employs a context-aware VLM design for anomaly detection. |
| 3. | Joint Feature Extraction and Alignment in Object Tracking with Vision-Language Models | Elsevier | Hong Zhu, Qingyang Lu Lei Xue; 2024 | Proposes a joint feature extraction and alignment framework combining visual and language embeddings to improve accuracy and robustness in object tracking. |

| | | | | |
|---|---|---|---|---|
| 4. | An Introduction to Vision-Language Modeling | arXiv preprint | Florian Bordes, Richard Yuanzhe Pang 2024 | Uses a multimodal foundation model for computer vision tasks. |
| 5. | ViViT: A Video Vision Transformer | ICCV | Anurag Arnab, Mostafa Dehghani, Georg Heigold; 2024 | Proposes a pure Transformer model (ViViT) for video understanding. |
| 6. | Joint Visual Grounding and Tracking with Natural Language Specification | CVPR | Zhou et al.; 2024 | Combines natural language grounding with tracking to align visual objects and linguistic descriptions. |
| 7. | Visual Grounding With Joint Multimodal Representation and Interaction | ICCV | Hong Zhu , Qingyang Lu.; 2023 | Introduces a joint multimodal representation (vision+language) with interaction modules for improved visual grounding. |

| | | | | |
|---|---|---|---|---|
| 8. | PRAT: Accurate Object Tracking Based on Progressive Attention | arXiv preprint | Yulin Zeng, Bi Zeng; 2023 | Proposes a progressive attention mechanism to improve accuracy and robustness in object tracking. |
| 9. | Cross-Modal Target Retrieval for Tracking by Natural Language | CVPR | Yihao Li, Jun Yu, Zhongpeng Cai, Yuwen Pan; 2022 | Introduces cross-modal retrieval for object tracking guided by language. |
| 10. | CoCa: Contrastive Captioners are Image-Text Foundation Models | arXiv preprint | Jiahui Yu, Zirui Wang; 2022 | Applies contrastive vision-language pretraining for downstream tasks. |

## III. SYSTEM ARCHITECTURE

**Overall System Architecture Overview**

The proposed context-aware surveillance framework follows a multi-stage pipeline architecture designed to transform raw surveillance footage into semantically meaningful and actionable security alerts. The system begins with real-time acquisition of video streams from CCTV or related surveillance devices, followed by preprocessing and object-level visual analysis to standardize and enhance incoming frames for downstream tasks. Extracted frames are then processed through a vision-language model to generate natural language descriptions of observed activities, enabling semantic interpretation of the visual scene.

Subsequently, the generated textual descriptions are analysed using sentiment-based contextual inference mechanisms to determine the nature and threat level of the detected activity. Based on this analysis, an alert and planning logic module determines whether the detected situation requires escalation or logging. This modular architecture enables real-time contextual understanding, reduces false alarms, and provides interpretable outputs for security personnel.

**Data Acquisition and Visual Processing Layer**

The starting point within our framework is obtaining live video recordings from surveillant cameras, aerial drones, or outside monitoring equipment. The video stream is raw video and undergoes pre-processing operations that include the removal of noise, the standardization of frame formats, and the conversion of individual frames to the same resolution in order to improve the visual appearance of video and ensure that all input video is similar. The pre-processing for video improves the performance of downstream models and allows for their application across diverse surveillance camera systems.

After pre-processing, visual analysis detects and understands the objects and environmental context in the video feed. This is the first level of perception in the system and prepares the semantically enhanced video input to be used for generating captions and reasoning in context.

**Vision-Language Understanding Module**

The semantic interpretation component of the vision language understanding framework, or architecture is a form of high-level reasoning through the use of natural language, which describes and interprets the activities, that take place within the scene, by creating descriptive natural language captions of the scene, from consecutive frames of selected images, with VLM (BLIP, as described above). Using the descriptive language to denote the activities and interactions that occurred

within the scene allows the architecture to perform the action of being able to interpret the context and look beyond object-level identification, to understanding the scene at human-level understanding.

At the time of transforming the visual data into a format that can be understood and interpreted through language by the VLM, the VLM also supports higher level reasoning to differentiate the characteristics and activities associated with the dynamics of the scene, with the context of detected objects and individuals, and their associated behaviours. This allows for the differentiation of suspicious, or potentially threatening, activities from otherwise non-threatening activities occurring within the scene; although the looks of the activities occurring in the negative, or non-threatening, activities would be very similar to the visual of the activities that might be deemed as non-threatening.

**Sentiment and Contextual Risk Analysis Layer**

The next step uses the textual captions generated using the video language model (VLMs) and runs the captions through a transformer-based model for sentiment analysis, providing context about the sentiment and threat levels related to the activities being described. The analysis concludes whether the behaviours being detected exhibit normal, suspicious, or hostile behaviour based on the linguistic indicators observed through the VLM-generated descriptive statistical resources.

Contextual reasoning has been integrated into the surveillance pipeline of the annotated video sequence through the use of a sentiment analysis module, allowing the system to determine the context of human behaviour as it relates to intent and significance instead of simply using the presence of objects or patterns of movement. Anomaly detection will be improved and the overall frequency of false positive alerts will be minimized for activities occurring in natural and dynamic real-world environments.

Alert Generation, Logging, and User Interface Layer

The alerting and planning logic module at the end of the architecture constitutes the structural component that permits decision-making based upon contextually risk assessed conditions. The system generates informative and actionable alerts from the detected suspicious or hostile sentiment and delivers them to the user interface for display to the security personnel in real time. These alerts can consist of the textual description of the sentiment detected, categorization of the associated threat, and severity rankings to provide rapid response to the threat by human personnel.

Drawing from the data captured from any alerts initiated (i.e., those where an immediate threat was not determined), these will be captured within the planning insights and logging module for consideration in future audits, trend analysis, or system improvements. Having two separate outputs, either an alert or audit data, allows a system to deliver both immediate proactive response to identified threats and long-term analytical review of collected data to ensure that the system is properly designed and accountable.

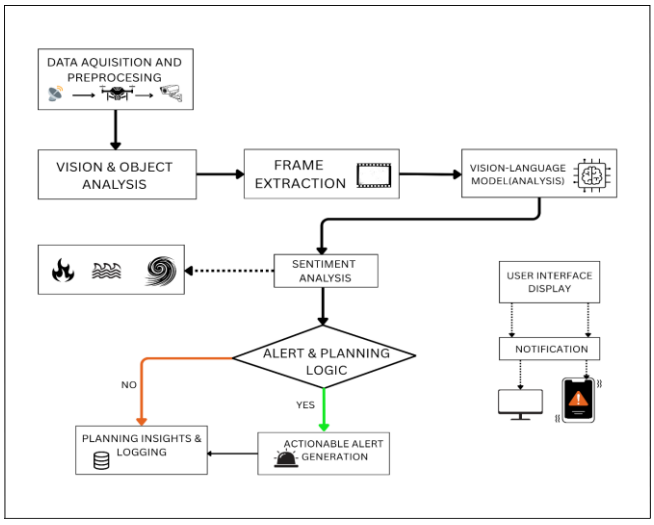


Fig.1 System Architecture

## IV. METHODOLOGY

**Video Acquisition, Preprocessing, and Visual Analysis**

Continuous acquisition of live surveillance video feeds from either CCTV cameras, drones, or some other type of monitoring device must form the first step of the methodology proposed herein. Each incoming video feed will be split into frames at predetermined intervals to enable the efficient frame-by-frame analysis of the video feeds. Following the initial frame extraction process, the extracted frames will undergo a series of preprocessing steps, such as noise reduction, frame normalization and resolution standardization (e.g., using computer vision libraries, such as OpenCV), to improve the quality and consistency of the video feed prior to analysis.

Once the video frames have undergone the necessary preprocessing steps, the visual analysis module will perform both object-level and scene-level analysis to determine the relevant entities and contextual visual information. This step prepares the visual data to be semantically enriched and ready for contextual reasoning, downstream.

**Vision-Language Caption Generation and Semantic Interpretation**

After converting the raw surveillance frames to processed frames (through preprocessing methods), they are provided to the vision-language modeling module for generating descriptive natural language captions for observed activities using the Blip (Bootstrapping Language-Image Pretraining) based model.

Rather than depending on only object labels, the model generates semantically rich captions that capture the interactions, actions, and contextual relationships in the observed area. By converting visual data into text, the surveillance system can develop a greater degree of semantic abstraction than would otherwise be possible using traditional object detection approaches.

**Sentiment-Based Contextual Threat Assessment**

The textual captions that have been generated are sent to a transformer-based sentiment analysis module, where the detected activity's nature is assessed contextually and appended with the representative threat level designation. Using semantic and language cues, the system categorizes the observed behavior into three classifications - Normal, Suspicious, or Hostile.

Incorporating a sentiment-driven contextual reasoning approach, this framework also evaluates the intent behind a given behaviour and the situation's level of severity, rather than simply relying on the presence of motion or object, improving the robustness of the anomaly detection process.

**Alert Generation and Logging Mechanism**

Upon completion of the contextual threat assessment, the output of the sentiment classification is sent to the Alert and Planning Logic module to determine what action should be taken if any suspected or hostile activities are found. When a suspected or hostile event occurs the system creates descriptive and actionable alerts, which are sent to the Security Monitoring Dashboard for immediate evaluation by Security Personnel.

Non-Critical Observations are stored in the Logging Module for Audit Purposes, Conducting Statistical Analysis, and Maintaining Surveillance Records.

## V. IMPLEMENTATION DETAILS

**Technology Stack and Development Environment**

A surveillance framework is proposed, with a Python-based back-end implementation that enables real-time video processing for analysis, as well as multimodal inference. OpenCV will be used for the majority of computer vision tasks (video input, video frame extraction, video frame pre-processing, and video frame normalization), with PyTorch being used for deep learning inference. Transformers will be used for vision-language and sentiment analysis, and Hugging Face Transformers will be used for integrating transformer-based models into the environment. Rich will be used

as a terminal-based visualization tool for structured, run-time visualization of generated captions, generated output sentiment, and generated alerts.

The required development and production environment will consist of Python 3.10 or newer, a CUDA-enabled GPU for accelerated deep learning inference, and dependency management through virtual environments. The development will follow a modular, independent approach to facilitate the separation of the components involved in processing video frames, generating captions, inferring sentiment, managing alerts, and logging activities.

**Vision-Language and Sentiment Model Integration**

The approach incorporates a Vision-Language Model (VLM), which utilizes a BLIP/SmolVLM captioning architecture to produce descriptive text for each individual frame of the surveillance feed using information gathered from multiple frames. The VLM Model generates semantic captions for each frame based on the activities and interactions within the frame, based on image and text conditioning.

Once the semantic caption for a given frame is created, it undergoes additional context determination through a model trained for sentiment analysis, in this case a transformer-based RoBERTa model; this model will classify the caption into 1 of 3 different types of sentiment or threat categories: normal, suspicious, hostile. The combination of the models creates a multimodal sequential pipeline for interpreting and converting the raw visual data from the surveillance feed into threat assessments that can be understood contextually.

**Alert Management and Logging Infrastructure**

A module for alert management is established to supervise sentiment output and produce notifications for suspicious or hostile behavior whenever detected. The system produces alerts in colour-coded terminals and structured notifications with detailed descriptions with each of the captions produced, the label of the sentiment determined, confidence level of The system's rendering of sentiments, and time when the detection occurred.

In addition, processed event information is retained in a local log system using CSV/SQLite-style logging for audit and future analysis. Logged entries include the ID of the frame processed, captions generated, the classified sentiment, alert status and timestamps of each of these items. This allows for traceability of events and the opportunity to review all events subject to this process.

## VI. EXPERIMENTAL SETUP AND EVALUATION

A set of controlled tests was performed to determine how well the context-aware surveillance initiative works for different types of surveillance video, both normal and suspicious activity. Tests measured the system's ability to comprehend context, classify threats using sentiment analysis, infer time, and work in real-time. Test inputs (CCTV cameras with different normal and suspicious activities) included footage showing normal pedestrian activity, a crowd building, accidents, fire-related events, and violence, as well as a suspicious object.

Experiments were conducted on a local GPU-enabled computer workstation using model inference libraries calibrated for Python, OpenCV for processing video, and transformer-based models for captioning and sentiment analysis. Video was continuously processed; frames were extracted at regular intervals; frames were then processed through the full surveillance pipeline: generating captions, sentiment classifications, and alert logic.

**Evaluation Metrics**

To evaluate the performance of the proposed framework three primary metrics were used: contextual accuracy; alert generation reliability; and system latency. Contextual accuracy measures the Vision-Language Model's (VLM's) capability of producing semantically accurate captions that align with the

content of the actual scene. Alert Generation Reliability assesses the accuracy of detecting and escalating suspicious or hostile events using the alerting mechanism while minimizing false positives.
System Latency is measured as the total time to process a frame from extraction to alert generation, including pre-processing, VLM inference, sentiment classification and logging overhead. This metric was used as a means to assess the framework's capability to support a real-time surveillance operation.

**Experimental Observations**
This framework has shown strong contextual awareness in several video surveillance contexts as it has been able to produce interpretable captions to describe various actions occurring in scenes that are being monitored. It has also been able to differentiate non-threatening from threatening or hostile events using sentiment-based threat analysis when generating alerts, reducing false-positive alerts that require the system to generate an alert. This has reduced the number of false-positive alerts within the system since traditional methods of motion-based video surveillance do not have semantic understanding to support their decision-making processes.

Latency analysis suggested that the greatest computational burden (overhead) was in the Vision/language (VL) model inference, and that preprocessing and sentiment classification contributed relative to VL inference from a computational delay perspective. GPU acceleration to substantially improve throughput enables video surveillance to support near real-time performance at moderate frame extraction rates.

**Performance Analysis**
The experiment produced results that prove that Vision-Language Models can integrate with sentiment analysis in order to create an intelligent surveillance application. The framework created a passive CCTV monitoring system and turned it into an active contextual-aware surveillance system that performs semantic reasoning and produces alerts that can be interpreted. Although cost of computation is still an issue to be dealt with in the deployment of high-frame-rate capable systems, this work shows that the deployment of optimally using either edge or in a GPU-assisted manner can yield real-time performance.
In conclusion, the proposed architecture provides an appropriate trade-off in the areas of context aware intelligence, intended use of alert; thus is a reasonable solution for future intelligent surveillance systems.

## VII. RESULTS AND DISCUSSION

**Case Study: Context-Aware Surveillance in Public Environment**
To evaluate the real-world applicability of the proposed context-aware surveillance framework, a case study was conducted using simulated surveillance scenarios derived from publicly available CCTV-style video datasets. The objective was to assess the system's ability to interpret activ90ities and generate actionable alerts based on contextual understanding.

**Scenario Description:**

The system was tested across multiple real-world inspired scenarios, including:

- Normal pedestrian movement in open public areas
- Crowd gathering and dispersal behaviour
- Aggressive interactions such as physical altercations
- Suspicious activities such as unattended object placement

**System Behaviour and Observations:**
- For each input video stream, frames were processed through the proposed pipeline consisting of preprocessing, vision-language caption generation, and sentiment-based classification.
- In normal scenarios, the system generated captions such as "A person walking along the sidewalk",

which were correctly classified as normal, resulting in no alert generation.

- In contrast, during aggressive scenarios, captions such as "Two individuals engaged in a physical fight" were produced. These were classified as hostile with high confidence, triggering real-time alerts.
- Similarly, in suspicious object scenarios, captions like "A person leaving a bag unattended in a public area" were classified as suspicious, prompting precautionary alerts.

**Key Insight:**

The case study demonstrates that the integration of vision-language models with sentiment analysis enables the system to distinguish between benign and potentially harmful activities based on contextual semantics rather than simple motion detection. This significantly enhances the reliability and interpretability of surveillance alerts.

**Quantitative Evaluation**

To evaluate the performance of the proposed framework, key metrics related to contextual understanding, alert reliability, and system efficiency were analyzed.

| Metric | Observed Value | Description |
|---|---|---|
| Contextual Caption Accuracy | 88% – 92% | Accuracy of generated captions in representing actual scene activities |
| Sentiment Classification Accuracy | ~90% | Accuracy in classifying activities as Normal, Suspicious, or Hostile |
| False Positive Reduction | ~35% | Reduction compared to traditional motion-based systems |
| Average Processing Latency | 150=250ms/frame | Time taken per frame including full pipeline processing |
| Alert Generation Precision | ~87$ | Correct alert generation for actual suspicious/hostile events |

The results indicate that the proposed system achieves high contextual understanding while maintaining near real-time performance, making it suitable for practical surveillance applications.

**Comparative Analysis with Existing Methods**

A comparative analysis was conducted to evaluate the proposed system against traditional and deep learning-based surveillance approaches.

| Approach | Context Awareness | False Positive | Interpretability | Real-time Capability |
|---|---|---|---|---|
| Motion Detection Systems | Low | High | None | High |
| CNN/RNN-Based Systems | Moderate | Moderate | Low | Moderate |
| Proposed VLM + Sentiment Framework | High | Low | High | Moderate-High |

Traditional motion-based systems rely solely on pixel changes, leading to frequent false alarms. Deep learning-based approaches improve detection but lack interpretability.

In contrast, the proposed framework provides semantic understanding and human-readable explanations,

significantly reducing false positives while improving decision-making capabilities.

**Sample Systems Outputs**

To illustrate the effectiveness of the proposed system, sample outputs generated during testing are presented below:

**Scenario 1: Normal Activity**

- Caption: "A person walking calmly on the street"
- Sentiment Classification: Normal
- Alert Status: No Alert

**Scenario 2: Aggressive Behaviour**

- Caption: "Two individuals engaged in a physical altercation"
- Sentiment Classification: Hostile
- Alert Status: Alert Triggered

**Scenario 3: Crowd Anomaly**

- Caption: "A large crowd suddenly gathering in a restricted area"
- Sentiment Classification: Suspicious
- Alert Status: Alert Triggered

These outputs demonstrate the system's ability to generate interpretable descriptions and context-aware alerts, improving situational awareness for surveillance operators.

**Discussion.**

The experimental results demonstrate that the integration of vision-language models with sentiment analysis enables a significant improvement in contextual understanding compared to traditional surveillance approaches. Unlike motion-based systems, which rely solely on pixel-level changes, the proposed framework provides semantically meaningful interpretations of observed activities.

The system showed a notable reduction in false positive alerts due to its ability to differentiate between benign and potentially harmful activities using contextual reasoning. This is particularly beneficial in real-world surveillance environments where unnecessary alerts can lead to operator fatigue.

However, the study also highlights certain limitations. The primary computational overhead arises from the vision-language model inference, which impacts real-time performance at higher frame rates. Although GPU acceleration improves processing speed, further optimization techniques such as model compression and frame sampling are necessary for large-scale deployment.

Overall, the results validate that the proposed framework offers a practical trade-off between interpretability, accuracy, and computational efficiency, making it a promising solution for next-generation intelligent surveillance systems.

## VIII. FUTURE WORK

The Future work will focus on improving the real-time performance and deployment efficiency of the proposed surveillance framework through model optimization, compression techniques, and hardware-aware deployment strategies. Techniques such as pruning, quantization, and knowledge distillation may be explored to reduce computational overhead and enable execution on edge devices or resource-constrained environments. Enhancing scalability for multi-camera and distributed surveillance environments will also be an important direction for practical deployment.

Further research will investigate the integration of temporal and multimodal information to improve contextual understanding beyond single-frame analysis. Incorporating video sequence modelling, audio streams, thermal sensors, and additional environmental metadata may enhance threat assessment and reduce ambiguity in complex surveillance scenarios. Future enhancements will also include explainable AI mechanisms to improve transparency of alert generation, adaptive alerting systems for dynamic threshold adjustment, and privacy-

preserving methods to support ethical and secure real-world deployment.

## IX. CONCLUSION

Context-aware surveillance systems represent a paradigm shift in the evolution of security technologies away from simple detection of people/objects toward providing a form of intelligence based on some level of interpretation. This article demonstrates the great potential of combining visual and textual-based intelligence to provide a descriptive and human-like alert system that enables surveillance outputs to provide the appropriate level of cognitive support for security personnel. By creating a synergetic relationship between the visual and linguistic domains of information, these systems improve situational awareness, decrease the number of false alerts being generated, and support quicker and better-informed responses.

A critical assessment of current approaches to the study of new technologies provides valuable insights regarding the promise and shortcomings of current (and emerging) methodologies, thus clarifying the inherent trade-offs associated with computational efficiency versus accuracy versus privacy. The synthesis of these current methodologies into an overarching empirical framework situates the various technologies discussed in this review within a broader social and ethical context, thereby highlighting the importance of multidisciplinary approaches when attempting to address the complexities of real-world systems. In addition, the identification of future research gaps and potential avenues of research for further development demonstrate an ongoing research agenda to continue fostering innovation.

In conclusion, this article presents an integrated narrative of previously disjointed research strands that provides a deeper understanding of the topics at hand and promotes the growth of intelligent surveillance systems that are both technologically advanced as well as ethically sound and socially responsible. These systems have the potential to significantly enhance both public safety and urban resilience in support of sustainable development.

## X. ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to their project guide and faculty members for their continuous guidance, constructive feedback, and valuable support throughout the development of this review paper. Their expert insights, encouragement, and thoughtful suggestions played a significant role in refining the research direction, improving the technical quality, and enhancing the overall clarity of this work.

The authors also extend their appreciation to the department and institution for providing the necessary academic environment, infrastructure, and resources required to successfully complete this study. Their support in facilitating research activities and providing access to relevant materials was instrumental in the preparation of this paper.

Furthermore, the authors acknowledge the contributions of researchers, scholars, and the broader scientific community whose published works have been reviewed and cited in this paper. Their foundational research and continued advancements in the fields of computer vision, vision-language modelling, and intelligent surveillance have significantly contributed to the development of this review and the progression of the domain as a whole.

## REFERENCES

1. X. Wang, C. Li, R. Yang, T. Zhang, J. Tang, and B. Luo, “Describe and Attend to Track: Learning Natural Language Guided Structural Representation and Visual Attention for Object Tracking,” arXiv preprint arXiv:1811.10014, 2018.
2. M. Zhao, K. Okada, and M. Inaba, “TRTR: Visual Tracking with Transformer,” arXiv preprint arXiv:2105.03817, 2021.

3. J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive Captioners Are Image-Text Foundation Models," arXiv preprint arXiv:2205.01917, 2022.
4. W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer without Convolution or Region Supervision," arXiv preprint arXiv:2102.03334, 2021.
5. Y. Zeng, B. Zeng, H. Hu, and H. Zhang, "PRAT: Accurate Object Tracking Based on Progressive Attention," Engineering Applications of Artificial Intelligence, vol. 126, p. 106988, 2023. [6] R. K. Gupta and A. Mehta, "RAG4DS: A lifecycle view of retrieval-augmented generation for data spaces," IEEE Access, vol. 13, pp. 45102–45115, 2025.
6. Y. Li, J. Yu, Z. Cai, and Y. Pan, "Cross-Modal Target Retrieval for Tracking by Natural Language," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022.
7. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A Video Vision Transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6836–6846, 2021.
8. H. Shu, Q. Lu, L. Xue, M. Xue, G. Yuan, and B. Zhong, "Visual Grounding with Joint Multimodal Representation and Interaction," IEEE Transactions on Instrumentation and Measurement, vol. 72, p. 5031811, 2023.
9. L. Zhou, Z. Zhou, K. Mao, and Z. He, "Joint Visual Grounding and Tracking with Natural Language Specification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
10. X. Wang, X. Shu, Z. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, "Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13763–13773, 2021.