2025 IEEE DELCON - International Conference on Recent Smart Technologies in Engineering for Sustainable Development | 979-8-3315-5864-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/DELCON68055.2025.11400401

Applying Vision–Language Models for Intelligent Multimodal Surveillance and Crime Detection

Nishchay Gupta
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
nishchaygupta40@gmail.com

Kunal
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
kunalmr2003@gmail.com

Madhav Sachar
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
madhavsachar24@gmail.com

Rupali Pandey
Department of Applied Science
Bharati Vidyapeeth's College of Engineering
New Delhi, India
pandeyrupali2401@gmail.com

Vishal Sharma
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi, India
vishalmtr@gmail.com

***Abstract*—With growing security concerns in the modern world, a requirement of a modern surveillance solution is generated. The current state of technology offers solutions in the form of custom built, fine-tuned neural networks to automate the process of detecting abnormal behaviour. These systems often rely on pretrained neural networks like ResNet, YOLOv5, etc. The research paper aims to explore new technology in the domain of smart surveillance to further the cause of a secure future for humanity as a whole. The proposed paper proposes the utilisation of pretrained Multi-Modal Models which can take Audio and Visual cues to identify Abnormal behaviour and further increase the pinpoint accuracy of the model. Other than the obvious advantage, these models like Gemini Vision offer Multimodal analysis of the input data therefore helping to further pinpoint the occurrence of abnormal behaviour in the input footage. The implemented system achieved an overall accuracy of 75%, with a precision of 93.3%, recall of 66.7%, and F1-score of 77.8% for crime detection, indicating strong performance in identifying actual criminal activities. By utilising the recent advancements in the field of deep learning the smart surveillance system offers an opportunity for the nation to develop into a more secure, peaceful state.**

***Index Terms*—Smart Surveillance, Gemini API, Anomaly Detection, Multimodal AI, Real-Time Monitoring, Scene Understanding, Crime Detection, Vision-Language Models**

I. Introduction

The increasing population on a global scale has put a significant strain on surveillance systems that exist in developing nations. The sheer amount of data to be processed poses a huge challenge to effective analysis of threats and notifying for immediate action and makes surveillance activities close to impossible for mere human beings to perform [1]. The current state of systems highlights the need for a more effective system, preferably automated [2], [3]. This need is fulfilled by modern research into effective systems powered by Machine Learning Algorithms and deep learning techniques such as Convolution Neural Network (CNNs) and Recurrent Neural Networks (RNNs) to detect disturbances in terms of criminal activity and classify them in relevant classes [4]

However, the usage of a custom neural network for each setup brings along new problems. The existing systems rely heavily on data hungry CNN architectures [5], [6]These architectures often use algorithms like You Only Look Once [7], [8](YOLO), for generating results, which require vast amounts of labelled data for successful training along with tremendous computational resources [9]. Other than the obvious disadvantage, the existing CNN systems only take visual cues such as frames extracted from videos and often miss out on additional inputs such as audio signals (e.g. screams, gunshots, etc.) which can help further pinpoint abnormal situations [10]

In this paper, we propose an alternative approach using a real-time, resource-efficient crime detection pipeline [11]. The pipeline leverages pretrained multimodal models; among the various options available, we selected Google's Gemini API [12], [13]due to its free accessibility, pretrained nature, and strong support for multimodal tasks. Our system analyses surveillance footage, and classifies activity into categories such as Theft, Assault, Vandalism, Cybercrime, and more. In general, the pre-trained multimodal models incorporate the use of Large Language Models (LLMs) to gain a comprehensive situational understanding of the input [14]–[16]. The Video

Large Language Models play an integral role in generating a detailed description of the visual input provided to the model [17], [18]. By incorporating both visual and potential audio cues, the model addresses the limitations of unimodal systems and provides robust, context-aware crime detection - [19], [20]

In summary, the paper makes the following contributions:

- Propose a complete, real-time crime detection system that uses pretrained multimodal models to combine both visual and audio signals.
- Use Google Gemini's pretrained multimodal model to create detailed scene descriptions, which will improve understanding and awareness of the situation.
- Apply multi-model cross-validation with several LLMs [21] to lower false alarms and improve the reliability of anomaly detection.
- Show better understanding of context and threat detection in changing real-world surveillance situations.

The Remaining Sections are arranged in the following manner. Section 2 includes a brief description of a relevant research paper entailing discussion on Surveillance technologies. Next, Section 3 Describes the proposed methodology of our solution. Following that, Section 4 explains the evaluation metrics and declares the experimental results. Finally, section 5 concludes discussion and discusses the future scope.

II. Related Works

Nowadays, many surveillance systems leverage deep learning techniques to enhance security and monitoring capabilities. Recent literature has explored applications such as weapon detection, face recognition, anomaly detection, and human interaction analysis. These studies form the foundation upon which our proposed multimodal pipeline builds. This section reviews key contributions to smart surveillance systems using custom neural networks, serving as a foundation for our work while highlighting the gap our approach addresses.

Mukto et al. proposed a modular real-time crime monitoring system integrating YOLOv5 for weapon detection, MobileNetV2 for violence classification, and LBPH for face recognition [22].This system achieved strong performance metrics including 86% F1-score for weapon detection, 95% F1-score for violence detection, and 97% accuracy for face recognition. However, its reliance on distinct models for each task and dependence on a custom dataset limit scalability and adaptability across diverse environments.

Jebur et al. introduced a generalized deep learning framework that combined MobileNetV2, InceptionV3, Inception-ResNetV2, and Xception to build a scalable anomaly detection system [23]. It reported accuracies of 97.99% on RLVS (violence), 83.59% on UCF (shoplifting), and 88.37% on a combined dataset without retraining for each anomaly type. While the model addresses scalability, it remains constrained to vision-based inputs.

Qasim and Verdu developed a video anomaly detection system using ResNet (18, 34, 50) for spatial features and Simple Recurrent Units (SRUs) for temporal modelling [24]. Their model achieved 91.44% accuracy, 91.71% precision, and 91.64% AUC on the UCF-Crime dataset, outperforming CNN-LSTM and CNN-GRU baselines. Despite its effectiveness, the system lacks contextual modalities such as audio, which limits its applicability in real-world scenarios.

Pooja et al. present a real-time intelligent video surveillance system that employs a Recurrent Neural Network (RNN) model combined with 3D convolutional and spatiotemporal autoencoders to detect abnormal behavior in surveillance footage [25]. The model was trained and tested using a set of video clips containing both regular and anomalous activities, achieving a detection accuracy of 96%. The system further integrates real-time audio communication for alert notifications, enhancing the immediate response capability. This study is particularly relevant to the current research as it emphasizes the use of multimodal inputs (video and audio) for anomaly detection, aligning with the Gemini API-based pipeline proposed in this paper, which also integrates scene interpretation and alert generation through multimodal data processing

Wu et al. proposed GL-AD [26], an anomaly detection model using a global-local attention mechanism and multi-task learning to improve classification of challenging video segments. Unlike our multimodal Gemini-based approach, GL-AD focuses solely on visual inputs and enhances performance through range scaling and focal loss. Achieving an AUC of 83.9% on the UCF-Crime dataset, it demonstrates the effectiveness of fine-grained visual attention in single-modality setups, complementing our multimodal strategy.

Nayak et al. proposed a captioning-based surveillance system that utilized VGG16 for feature extraction and an LSTM-based encoder-decoder architecture to generate real-time textual descriptions from CCTV footage [27]. Crime-related keywords such as "knife" and "thief" were used to detect threats. These captions were timestamped, searchable, and encrypted, offering a secure and efficient alternative to video storage. The use of natural language understanding closely aligns with our system's descriptive intelligence, which is powered by pretrained multimodal models like Gemini Vision.

Sidhu and Sharad introduced a smart surveillance framework combining video and audio signal processing to detect interpersonal crimes such as harassment, bullying, or assault [28]. The system activates conditional recording only when a "critical situation" is detected, preserving privacy while reducing memory usage. Using Mel-Frequency Cepstral Coefficients (MFCCs) for speech analysis and motion/emotion detection for visual input, the system identifies threats in real-time and issues alerts. This dual-sensory approach reinforces the value of combining multiple sensory streams for improved context-aware surveillance—a principle extended in our proposed architecture.

In addition to these foundational studies, recent studies have explored the integration of contextual intelligence and scalable solutions within surveillance systems.

Yilmazer and Karakose introduced a robust approach for abandoned object detection using YOLOv8 and DeepSORT

for real-time tracking, augmented by LLM-based scene explanations [29]. Their keyframe-based optimization aligns with the efficiency goals of our Gemini API pipeline while also introducing early forms of language-based interpretability.

Jeshmol P.J. and Binsu C. Kovoor talk about Video Question Answering (VideoQA) systems that employ multimodal data processing to analyse video content and return context-aware responses [30]. The system is also analogous to the proposed Gemini API-based surveillance pipeline that uses the same video processing algorithms to detect and respond to security threats by interpreting scenes. The emphasis on multimodal fusion in VideoQA is analogous to the fusion of visual and audio inputs in the proposed system, which helps it interpret complicated situations and detect abnormal behavior better.

Similarly, Sarpate et al. presented a real-time anomaly detection system combining YOLO and Keras frameworks, which, while effective in detection, showed limitations under occlusion and complex environments—challenges our multimodal pipeline seeks to overcome through Gemini's pretrained vision-text capabilities [7]

Himeur et al. tackled the challenge of generalization in deep models via transfer learning and domain adaptation [20]. Their findings emphasize the high training and labelling costs associated with traditional systems, reinforcing the value of prompt-based, pretrained models like Gemini that can operate without task-specific data.

In summary, the literature shows significant progress in surveillance systems using deep learning. This includes weapon detection, identifying unusual activity, and processing different types of data. However, most methods still rely on single-type inputs, require multiple specialized models, or struggle to adjust to various environments. Some studies have tried combining audio and visual data or using scene-captioning methods, but these efforts are limited and do not scale like pretrained multimodal models. The gap identified highlights the need for unified, context-aware systems that combine audio, visual, and text data. Future research should focus on using pretrained multimodal APIs, integrating cross-model verification, and ensuring real-time responses to provide scalable and dependable surveillance intelligence in changing situations.

III. Proposed Methodology

To address the growing demand for scalable, intelligent, and cost-effective surveillance, we apply our pipeline for real-time detection of crime from multimodal media analysis by LLMs. Our methodology integrates visual, audio, and contextual data using pre-trained models such as those offered by the Gemini API to enable end-to-end crime classification and real-time alert generation. When surveillance is up and running, the system grabs the live feed, extracts frames at intervals, then delivers those frames, along with a prompt, to the Gemini Vision model for analysis. It returns descriptive insights, which are subsequently parsed to determine whether suspicious behavior, such as an aggressive demeanour or a weapon, is present. In case a crime is detected, an email alert is sent in real time notifying about an anomalous behaviour.

A. Dataset Overview

To train, validate, and demonstrate the performance of the proposed smart surveillance system, we curated a dataset by collecting publicly available surveillance videos, audio clips, and images from online platforms (e.g., social media sites and open repositories). The dataset was carefully selected to cover a wide variety of normal and anomalous behaviours, providing diverse visual and audio inputs that mimic real-life scenarios. All media were indexed and categorized by crime type and whether or not a crime was taking place and activity labels where possible, to enable effective testing of the Gemini API-based crime detection pipeline. No personally identifiable information (PII) was collected during the preparation of the datasets, and all media were collected from publicly available, non-restricted sources to ensure ethical compliance. The collected dataset enabled the system to test numerous detection cases by simulating real-time operating conditions in all the modality (video, audio, image).

In practice, the experiments were conducted on a subset of 33 samples curated from publicly available datasets. This limitation arose due to financial constraints associated with the Gemini API, which incurs charges on a per-request basis. Despite the small sample size, the subset was strategically selected to represent both crime and non-crime activities under diverse conditions, including variations in lighting, camera angle, and crowd density.

The data included around 60% crime and 40% non-crime samples, resulting in a small class imbalance in favor of crime scenarios. This distribution was maintained on purpose to reflect real-world situations where the analysis of surveillance is frequently performed on crime events.

B. Data Processing Pipeline

The proposed surveillance system employs a structured data processing pipeline, illustrated in Fig. 1, for transparent multimodal media input processing and meaningful threat-related information extraction. The pipeline begins with media input and type recognition, where the system supports common surveillance media types like image files (.jpg, .png), video files (.mp4, .avi), and audio files (.wav, .mp3). Automatic file extension detection is employed to recognize the right processing path for the input.

Following identification, the inputs are pre-processed to be compatible with the Gemini Vision model. Images are resized, normalized, and re-formatted as necessary, while video streams are segmented into keyframes at periodic intervals of time so that temporal coverage is adequate. Audio streams are isolated from one another for individual inspection and, as necessary, transcribed into textual form utilizing a speech-to-text engine. This renders the visual as well as audio signals amenable to examination by large language models (LLMs) alike.

The pre-processed media is then processed through multimodal feature analysis, where each input is aligned with a structured prompt and thereafter submitted to the Gemini Vision and Language API for context-aware processing. The API returns structured JSON outputs relative to possible criminal activity, such as the type of anticipated crime, confidence scores, key visual or audio evidence, and contextual indicators.

To prevent redundant computation, the system uses content hashing and caching, which assigns the input a content-based unique hash. If a duplicate input has already been computed before, cached output is retrieved immediately, reducing both the computational load and the number of API requests. The function works best with batch inputs or duplicate queries.

A self-imposed cross-provider evaluation phase further improves reliability by enabling integration with several LLM providers, including OpenAI and Gemini. Outputs from various providers can be cross-compared to guarantee consistency, thus enhancing decision-making confidence and minimizing false positives or false negatives.

To round out the picture, the system conducts comprehensive risk profiling by combining evidence from multiple frames and cross-checking it against audio data. By combining temporal and multimodal evidence, the pipeline constructs a balanced risk profile that reflects the severity and probability of suspicious behavior.

Finally, the pipeline ends with a real-time crime notification module. This module automatically activates and sends an alert email to designated recipients whenever an incident is classified as a crime with medium or high confidence. Each email includes the event time, the inferred crime category, and a brief description of the supporting evidence. This helps with timely intervention and informed decision-making in real-world applications. The system uses email instead of SMS or instant messaging because professional surveillance contexts, such as corporate security teams and law enforcement agencies, typically rely on shared institutional mailboxes rather than individual phone numbers. This ensures that alerts go to a common address that multiple authorized personnel can monitor, allowing for coordinated responses, scalability across organizations, and maintaining a record of notifications. By fitting into existing communication workflows, the notification design improves practicality and accountability in real deployments.

C. Model Configuration and Experimental Setup

The proposed pipeline employs a hybrid inference strategy that combines the Google Gemini Vision API with lightweight local large language models (LLMs) for verification. The primary engine is the Gemini Vision API, which performs multimodal scene analysis, while the local verification models include TinyLlama-1.1B, OPT-1.3B, and quantized LLaMA-2-7B.

The main parameter settings were as follows: the maximum token length was set between 512 and 1024; the temperature parameter was fixed at 0.7; top-p sampling was set to 0.9;

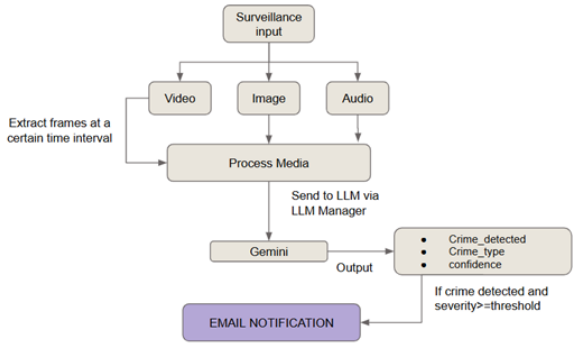

Fig. 1. Working pipeline for the proposed Smart Surveillance System using multimodal LLM.

video frames were sampled at a rate of 5 fps; and all input images were resized to 640×640 pixels.

All experiments were executed on an HP Laptop 15s-fq5xxx equipped with a 12th Gen Intel® Core™ i5-1235U CPU (12 cores, 1.3 GHz), 16 GB RAM, and integrated Intel Iris Xe Graphics. The software environment consisted of Windows 11 (64-bit), Python 3.10, PyTorch 2.2, Transformers 4.42, and SentenceTransformers 3.0. No GPU acceleration was used, and API calls were handled asynchronously to minimize latency.

IV. Results and Evaluation

To rigorously evaluate the performance of the proposed surveillance pipeline, we employ standard classification metrics widely used in machine learning research for measure of performance and detailed insights into class-specific behavior.

A. Evaluation Metrics

The Confusion Matrix classifies results by comparing predicted labels against actual labels. For a binary setup, it consists of:

- True Positive (TP): Crime correctly identified as crime.
- True Negative (TN): No-crime correctly identified as no-crime.
- False Positive (FP): No-crime incorrectly identified as crime.
- False Negative (FN): Crime incorrectly identified as no-crime.

$$\text{Confusion Matrix} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

The metrics are defined as follows:

Accuracy measures the proportion of correctly classified instances (both positive and negative) out of all predictions. It reflects the overall effectiveness of the model in making correct classifications.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision measures the proportion of correctly predicted positive instances among all predicted positives. It reflects the

reliability of positive classifications and the ability to minimize false alarms.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall, or sensitivity, measures the proportion of actual positive instances correctly identified. It reflects the ability of the model to capture relevant positive cases and minimize missed detections.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-Score is the harmonic mean of precision and recall, balancing false positives and false negatives. It provides a single performance measure for imbalanced datasets where both metrics are critical.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

B. Quantitative Results

The implemented smart surveillance system demonstrates integration of real-time video processing and Google's Gemini API to identify and analyse potentially suspicious or criminal behaviour. The pipeline was tested with varied input scenarios with varying lighting and activity conditions to evaluate its responsiveness, context awareness, and real-time applicability. In order to provide an extended view of model predictions in the Crime and No-Crime classes we have also incorporated the classification matrix (Confusion Matrix). The matrix provides a summary of true positives, true negatives, false positives, and false negatives, thus providing us with a view of unique misclassification patterns. This view supplements the quantitative measures by providing us with a view of the distribution of errors and indicating the model's stronger certainty when predicting crime events while maintaining high sensitivity to non-crime events. The classification matrix is shown in Fig. 2

The proposed system achieved an overall accuracy of 75%, correctly classifying the multimodal input in approximately three out of every four instances. A more specific evaluation shows that regarding the crime class, the model achieved precision of 75.8%, recall of 90.9%, and an F1-score of 82.6%, exhibiting a high likelihood that predicted crime events are actual crimes while missing some actual crime instances. However, the accuracy for the no-crime class was high in precision but low in recall, with precision of 93.3% and recall of 66.7%, F1-score 77.8%, indicating that most non-crime events are correctly detected, though some predicted non-crime events are actually crimes. The high recall against the crime class indicates that the model performs well in detecting actual crimes, minimizing the number of false negatives. The F1-score of 0.826 indicates a good trade-off between precision and recall against the crime class. But the performance for the no-crime class was slightly lower than crime class. The precision of 93.3% and recall metric of 66.7 % indicates strong sensitivity in detecting actual crime incidents, with occasional false negatives where normal activities are misclassified or overlooked due to contextual ambiguity. The high recall of the

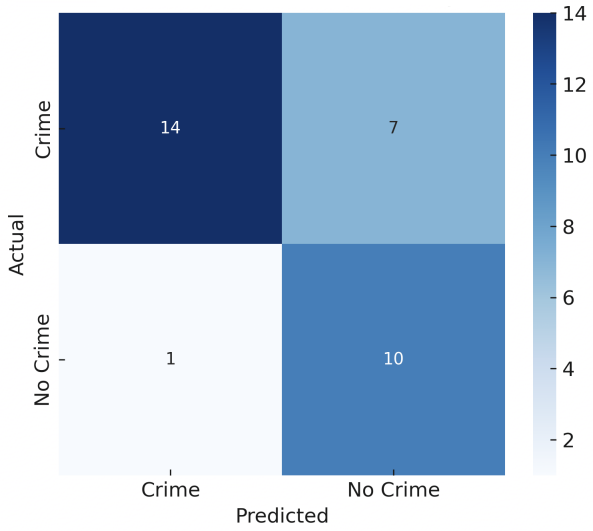

Fig. 2. Predictions from the Smart Surviellance system vs. Actual Labels

crime class by the model is a positive aspect in applications where criminal activity detection is crucial even at the cost of few false alarm rates. This type of feature is desirable in high-risk security scenarios where missing a crime would have catastrophic consequences. But the low recall (66.7%) for the No-Crime class indicates a deficiency in the model's ability to correctly identify all normal situations, leading to occasional misclassifications of benign activities as potential crimes. This behaviour suggests a slight bias toward predicting the existence of crime, which may arise from imbalanced training data, a conservative decision threshold favouring crime detection, or model design choices emphasizing sensitivity [31]. It is important to note that this relatively lower performance can also be attributed to the limited dataset size (33 samples), which restricted the model's ability to generalize. Because Gemini API usage incurs monetary costs, the experiments were constrained to a smaller dataset. Future work will focus on expanding the dataset to improve recall for non-crime classification and further enhance the overall reliability of the system.

V. Conclusion and Future Prospects

In this work, the feasibility of applying pretrained multimodal large language models, specifically Google Gemini, to real-time surveillance and crime detection was illustrated. The multi-stage pipeline integrates visual, audio, and contextual streams to achieve improved scene understanding and auto-generated threat alerts. The pipeline thus emphasizes scalability, cost-effectiveness, and ease of deployment. Therefore, enabling more intelligent surveillance across heterogeneous and resource-constrained environments. However, there is need for improvement for wider deployment, particularly in situations where the frequency of false alarms needs to be low. To reduce the computational burden and improve efficiency,

compressed formats and modular architectures are increasingly being adopted [32].

While the system achieves high recall in crime detection, there remain challenges to be resolved in reducing false positives for non-criminal activities. Future work includes expanding the training set to better reflect benign behavior, applying threshold calibration to maximize the sensitivity to specificity trade-off, and incorporating a secondary filtering model for verification as rich areas for future exploration. These advancements would enhance robustness, reduce spurious alarms, and bring the system closer to reliable, real-world deployment.

REFERENCES

[1] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," *Sensors*, vol. 23, no. 11, Art. no. 5024, 2023.

[2] S. Maliphol and C. Hamilton, "Smart Policing: Ethical Issues & Technology Management of Robocops," in *Proc. PICMET*, pp. 1–15, 2022.

[3] M. Afzal and P. Panagiotopoulos, "Smart Policing: A Critical Review of the Literature," in *Lecture Notes in Computer Science*, pp. 59–70, 2020.

[4] R. S. Mehsen, "Deep Learning Algorithm for Detecting and Analyzing Criminal Activity," *International Journal of Computing*, vol. 22, no. 2, pp. 248–253, 2023.

[5] J. Salido, V. Lomas, J. Ruiz-Santaquiteria, and O. Deniz, "Automatic Handgun Detection with Deep Learning in Video Surveillance Images," *Applied Sciences*, vol. 11, no. 13, p. 6085, 2021.

[6] A. Antoniou and P. Angelov, "A General Purpose Intelligent Surveillance System for Mobile Devices Using Deep Learning," in *Proc. IJCNN*, pp. 2879–2886, 2016.

[7] D. Sarpate, I. Tadas, R. Khaire, M. Antapurkar, and A. Sonone, "Unveiling Anomaly: Empowering Video Surveillance through Intelligent Anomaly Detection," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 11, no. 2, pp. 312–320, 2024.

[8] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] B. Kiran, D. Thomas, and R. Parakkal, "An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[10] Y. Ozkan and B. D. Barkana, "Forensic Audio Analysis and Event Recognition for Smart Surveillance Systems," in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, Woburn, MA, USA, pp. 1–6, 2019.

[11] M. Cieslik and C. Mura, "PaPy: Parallel and Distributed Data-processing Pipelines in Python," *arXiv preprint* arXiv:1407.4378, 2014.

[12] D. T. P. Phan, V. H. M. Doan, J. Choi, B. Lee, and J. Oh, "AADC-Net: A Multimodal Deep Learning Framework for Automatic Anomaly Detection in Real-Time Surveillance," *IEEE Trans. Instrumentation and Measurement*, vol. 74, Art. no. 5025713, 2025.

[13] Google AI for Developers, "API reference — Google AI for Developers," [Online]. Available: https://ai.google.dev/api (accessed: Sep. 1, 2024).

[14] S. Chatterjee, H. Shin, J.-M. Gil, and Y.-C. Byun, "RoadSitu: Leveraging Road Video Frame Extraction and Three-Stage Transformers for Situation Recognition," *Results in Engineering*, vol. 24, p. 103197, 2024.

[15] I. de Zarzà, J. de Curtò, and C. T. Calafate, "Socratic Video Understanding on Unmanned Aerial Vehicles," *Procedia Computer Science*, vol. 225, pp. 144–154, 2023.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[17] Y. Tang *et al.*, "Video Understanding with Large Language Models: A Survey," *arXiv preprint* arXiv:2312.17432, 2025.

[18] Z. Chen, Y. Qiu, L. Yang, B. Liao, and D. Cao, "Automatic generation of monitoring report based on large language model and knowledge graph inference," *Results Eng.*, vol. 25, p. 104795, 2025.

[19] P. Kumar, A. Mittal, and P. Kumar, "Study of robust and intelligent surveillance in visible and multi-modal framework," *Informatica (Slovenia)*, vol. 32, pp. 63–77, 2008.

[20] Y. Himeur, S. Al-Maadeed, H. Kheddar, N. Al-Maadeed, K. Abualsaud, A. Mohamed, and T. Khattab, "Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization," *Eng. Appl. Artif. Intell.*, vol. 119, p. 105698, 2023.

[21] W. Yin, Y. Xue, Z. Liu, H. Li, and M. Werner, "LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: Hurricane Harvey," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 137, p. 104423, 2025.

[22] M. M. Mukto, M. Hasan, M. M. A. Mahmud, I. Haque, M. A. Ahmed, T. Jabid, M. S. Ali, M. R. A. Rashid, M. M. Islam, and M. Islam, "Design of a real-time crime monitoring system using deep learning techniques," *Intell. Syst. Appl.*, vol. 21, p. 200311, 2024.

[23] S. A. Jebur, K. A. Hussein, H. K. Hoomod, L. Alzubaidi, A. A. Saihood, and Y. Gu, "A scalable and generalized deep learning framework for anomaly detection in surveillance videos," *arXiv preprint* arXiv:2408.00792, 2024.

[24] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," *Results Eng.*, p. 101026, 2023.

[25] P. Br and R. N., "Real-time intelligent video surveillance system using recurrent neural network," in *Proc. Int. Conf. on Procedia Computer Science*, vol. 235, pp. 1522–1531, 2024.

[26] Y. Wu, H. Sang, and F. Li, "Anomaly detection method of surveillance video based on global-local information," *Knowl.-Based Syst.*, vol. 317, p. 113530, 2025.

[27] S. Patwa, N. Nayak, S. Odhekar, and S. Roychowdhury, "Real time crime detection by captioning video surveillance using deep learning," *Int. J. Creative Res. Thoughts*, vol. 10, no. 7, pp. d367–d376, 2022.

[28] R. S. Sidhu and M. Sharad, "Smart surveillance system for detecting interpersonal crime," in *Proc. 2016 Int. Conf. Commun. Signal Process. (ICCSP)*, Melmaruvathur, India, pp. 2003–2007, 2016.

[29] M. Yilmazer and M. Karakose, "A new, robust, adaptive, versatile, and scalable abandoned object detection approach based on DeepSORT, dynamic prompts, and customized LLM for smart video surveillance," *Appl. Sci.*, vol. 15, no. 5, p. 2774, 2025.

[30] P. J. Jeshmol and B. C. Kovoor, "Video question answering: A survey of the state-of-the-art," *J. Vis. Commun. Image Represent.*, vol. 105, p. 104320, 2024.

[31] P. Vadlapati, "Autowatcher: A real-time context-aware security alert system using LLMs," *Int. J. Sci. Res. Eng. Manag.*, vol. 8, no. 11, pp. 1–6, 2024.

[32] Z. Zhang, S. Zheng, M. Qiu, G. Situ, D. J. Brady, Q. Dai, J. Suo, and X. Yuan, "A decade review of video compressive sensing: A roadmap to practical applications," *Engineering*, 2024.