Eco-Efficient Surveillance: Transforming Video Data into Actionable Text Summaries

Anonymous ICCV submission

Paper ID *****

Abstract

The growing reliance on surveillance cameras has resulted in massive storage requirements, leading to frequent deletion of video footage due to limited storage capacity. This practice not only raises concerns about the loss of crucial evidence for investigations but also exacerbates environmental issues due to the energy-intensive nature of video storage systems. To address these challenges, this paper introduces Eco-Surve, a novel approach for transforming surveillance video data into a compact, queryable system without the necessity of storing the raw video footage. This method enhances data efficiency, retrieval speed, and privacy while maintaining the integrity of critical surveillance information. By employing advanced object detection algorithms(YOLO),video-to-text algorithms (Gemini 1.5 pro and GPT 4) and reasoning large language models(DeepSeek), this method captures key details such as timestamps, motion events, and object activities, ensuring critical information is retained. Eco-Surve eliminates the need for time-consuming manual searches from video footage, significantly reducing the time required to identify specific events or objects from hours to minutes, accounting for reduction in time consumption by nearly 80%. Additionally, by reducing high-volume video storage demands by 90%,it minimizes the energy and hardware resources needed, thus mitigating environmental impacts like carbon emissions and digital wastage. This dual benefit of saving time and resources makes the proposed solution an impactful tool for industries reliant on video monitoring systems, ensuring efficient data management while retaining vital information for legal and investigative purposes.

1. Introduction

The widespread adoption of Closed-Circuit Television (CCTV) systems has revolutionized security and surveillance across public and private sectors. These systems play a critical role in ensuring safety by providing real-time monitoring and recording of events. However, the exponential growth in video data generated by these systems presents significant challenges, particularly in terms of storage, retrieval, and cost management. Organizations, ranging from small businesses to large transit systems, face escalating difficulties as the number of cameras and the resolution of recordings increase. This growth necessitates tiered storage solutions that not only inflate costs but also complicate data management. Additionally, during emergencies, locating specific events within massive video archives can be time-consuming, delaying critical decision-making and reducing operational efficiency.

Current solutions primarily focus on compressing video files or employing real-time anomaly detection systems. While video compression reduces storage demands, it often compromises the quality and integrity of recordings, making them less reliable for forensic or evidentiary purposes [**? ?**]. Real-time systems generate large volumes of data that strain storage capacities and still require manual review to extract key information [20]. Furthermore, manual searches for specific events in video archives are labor-intensive and prone to human error, making them unsuitable for time-sensitive scenarios [18]. Despite advancements in intelligent video analytics and machine learning algorithms, existing approaches fail to address the core issue: the inefficiency of storing and retrieving vast amounts of video data [4, 13].

To address these challenges, this research proposes an innovative framework Eco-Surve that transforms video surveillance footage into a Queryable System eliminating the need of raw video footages. By leveraging computer vision algorithms like YOLO for object detection and natural language processing models such as Gemini 1.5 Pro, GPT4 and Deep Seek for descriptive summaries and reasonings, the system converts each video into searchable text while retaining only essential images. This approach drastically reduces storage requirements while enabling rapid retrieval through natural language queries. For instance, a retail store with four 720p cameras operating 24/7 generates approximately 2–3 terabytes (TB) of video data annually. Using the

proposed method, this data can be condensed into 2–3 gigabytes (GB) of text files—a tenfold reduction—significantly lowering storage costs and environmental impact [20].

Consider a high-traffic environment like an airport where hundreds of cameras operate continuously. In a security breach scenario, identifying a suspect wearing a red jacket could take hours using traditional methods. With EcoSurve, security personnel could simply query "red jacket near Gate 12 around 3 PM," retrieving relevant text descriptions and images within seconds. This not only accelerates response times but also eliminates the need for exhaustive manual searches.

This paper introduces a transformative framework aimed at addressing the challenges of CCTV data management by focusing on storage efficiency by 80%, retrieval speed by 90%, and sustainability by eliminating electronic waste of storage devices. It begins with a review of current surveillance technologies and their limitations before detailing the methodology for converting video data into text-based representations using advanced AI models. Experimental results demonstrate significant improvements in storage optimization and retrieval efficiency.

2. Related Work

Recent research highlights the growing environmental concerns associated with large-scale surveillance systems, particularly regarding energy consumption, electronic waste, and resource depletion. Sustainability reports from Axis Communications and Hikvision emphasize efforts to reduce greenhouse gas emissions and energy usage during manufacturing processes, signaling an industry-wide recognition of the need for eco-friendly practices [6, 8]. Additionally, the environmental footprint of CCTV systems extends across their lifecycle, from production to disposal. Improper management of electronic waste, such as outdated cameras and storage devices, exacerbates environmental degradation [6].

Advancements in energy-efficient hardware have been pivotal in addressing sustainability challenges. Axis Communications demonstrated that a single 8MP camera could replace multiple lower-resolution cameras, significantly reducing energy demands [6]. Similarly, Dahua Technology introduced low-power operational modes for cameras during downtime and explored renewable energy integration at the device level [19]. These innovations underscore the potential for hardware optimization to minimize power consumption in surveillance networks.

The increasing volume of video data generated by surveillance systems has led to substantial energy demands in data centers. Memoori's Global Video Surveillance study highlighted this issue, advocating for cloud-based storage solutions as a more sustainable alternative [14]. Zhang et al. [25] proposed dynamic scalability in cloud storage to reduce on-premises hardware requirements while leveraging energy-efficient infrastructure. Such approaches enable organizations to manage data more sustainably while maintaining scalability and reliability.

The application of artificial intelligence (AI) in video analytics has emerged as a promising avenue for environmental monitoring. IsarSoft demonstrated how AI-enabled cameras could detect environmental hazards like oil spills and illegal logging, optimizing resource allocation for enforcement agencies and reducing their ecological footprint [10]. This integration of AI into surveillance systems not only enhances environmental protection but also aligns with broader sustainability goals.

Efforts to minimize hardware requirements have also been explored through innovative camera designs. Ahmad et al. [2] proposed an overhead camera system equipped with a wide-angle lens to cover larger areas with fewer devices. This approach reduces both power consumption and electronic waste, contributing to sustainability targets. Such designs complement existing energy optimization strategies by addressing the environmental impact of hardware production and maintenance.

Video compression plays a critical role in managing storage and bandwidth requirements without compromising video quality. Widely used standards like H.264 (AVC) and its successor H.265 (HEVC) offer significant improvements in compression efficiency, reducing file sizes by up to 80% compared to older methods like Motion JPEG (M-JPEG) [19]. Enhanced compression algorithms specifically designed for surveillance applications further optimize storage while maintaining high image quality.

Surveillance systems integrated with smart city infrastructures have shown potential for broader environmental benefits. Applications include optimizing traffic flows, monitoring air quality, improving waste collection routes, and identifying pollution sources [10, 14]. By leveraging AI-enabled cameras and sensors, cities can reduce emissions and enhance resource efficiency while addressing urban sustainability challenges [6].

3. Methodology

Our research introduces a groundbreaking approach to video surveillance data processing that prioritizes efficiency, scalability, and environmental sustainability [14]. Unlike traditional methods that rely on computationally intensive key frame extraction [11], our system processes video files directly as binary codecs [Figure 1]. This novel methodology not only reduces computational overhead but also minimizes energy consumption, making it a more sustainable alternative to conventional practices. By leveraging advanced computer vision and natural language processing (NLP) techniques, our framework transforms video data into concise queryable system, enabling rapid retrieval

and significantly reducing storage requirements. This innovative combination of technologies positions our approach as a transformative solution for modern surveillance challenges.

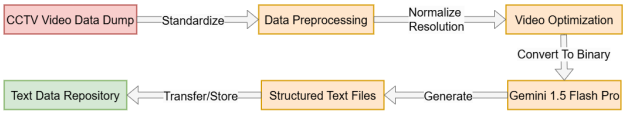

Figure 1. Initial data processing pipeline showing the transformation from raw video to structured text repository.

3.1. Data

The dataset for this study comprises 25 closed-circuit television (CCTV) recordings obtained from three distinct environmental contexts: grocery stores, parking areas, and fire exits. These locations were selected to provide a diverse range of human activity patterns and potential security scenarios, allowing for a comprehensive analysis of varying crowd dynamics and potential security concerns. The dataset includes recordings of varied visual qualities to test the robustness of analytical methods.

3.2. Data Collection

Current datasets for video-to-text tasks predominantly focus on specific domains, such as cooking behavior (YouCook [7], TACoS, TACoS Multi-level), general life videos (MSR-VTT), or movies (MPII-MD, M-VAD). While MSVD includes general web videos, it is not close to true CCTV footage [23]. To address these limitations and create a more comprehensive dataset, we developed a structured process for obtaining CCTV footage for Eco-Surve. We identified three distinct sets of locations and requested property managers to share old footage with time and place details. Upon gaining access, we extracted relevant MP4 files using remote access tools where possible, ensuring secure storage. The footage was then converted into compatible formats for analysis and evidentiary use.

3.3. Data Preprocessing

To prepare CCTV footage for efficient processing by large language models (LLMs), we designed a robust and systematic data preparation pipeline inspired by best practices in video analysis. Each video was renamed using a standardized naming convention embedding critical metadata, such as location, date, and camera ID, ensuring traceability and organization. Following the methodology proposed by Qian24 et al. [16], videos were segmented into small chunks using ffmpeg, a precise and quality-preserving multimedia framework, to ensure manageable file sizes while maintaining data integrity.

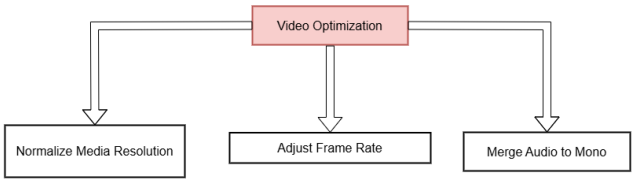

Figure 2. Workflow for video optimization: Resolution, frame rate, and other adjustments.

Resolution normalization to 720p (1280x720) and frame rate adjustment to 15-30 fps were applied to optimize computational efficiency without compromising the quality of visual information [17]. Audio streams were merged into mono format to streamline multimodal analysis. Additionally, contrast adjustment and noise reduction techniques were employed to enhance clarity, while metadata embedding ensured that contextual information remained intact for downstream processing. These preprocessing steps created a clean and standardized dataset, which emphasize the importance of structured and high-quality datasets for improving LLM performance in video analysis tasks.

4. Model Implementation

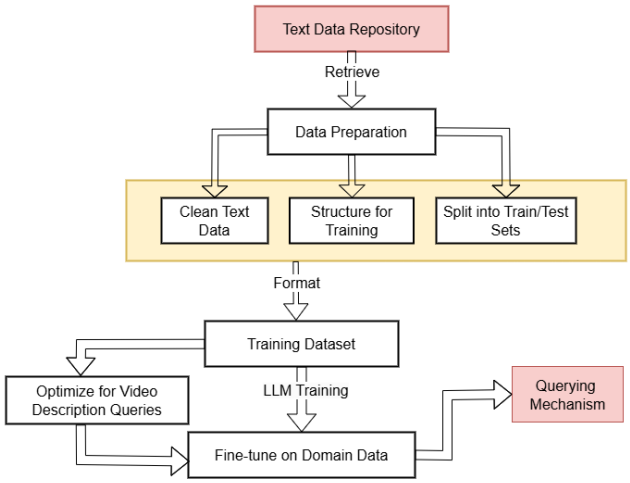

Figure 3. Pipeline for preparing text data for fine-tuning an LLM model.

Eco-Surve leverages cutting-edge Large Language Models (LLMs) [12] to generate detailed and context-rich textual descriptions of video content. Unlike traditional frame-by-frame analysis) [24], our approach processes videos as continuous streams, enabling the models to capture temporal relationships and narrative flow. This method enhances the coherence and relevance of event descriptions, ensuring a more comprehensive understanding of complex scenes, actions, and events.

To identify the optimal model for converting video footage into structured minute-by-minute event descriptions, we evaluated several technical factors. A key con-

sideration was the context window size, as longer context windows allow models to process extended video sequences without losing critical information. Gemini 1.5 Pro, with its ability to handle up to 1 million tokens, emerged as the ideal choice for analyzing lengthy videos while preserving context. Additionally, its support for direct video input in a binary file format, significantly streamlining the workflow [22] .

Unlike models such as OpenAI GPT-4o, which rely on frame-by-frame inputs and lack timestamped outputs, Gemini 1.5 Pro simultaneously processes visual and auditory data, generating precise, timestamped descriptions that align with specific moments in the video. These timestamped outputs enhance usability by enabling efficient event tracking and retrieval [26]. Based on its superior performance in multimodal analysis and processing efficiency, Gemini 1.5 Pro was selected as the foundation of our system.

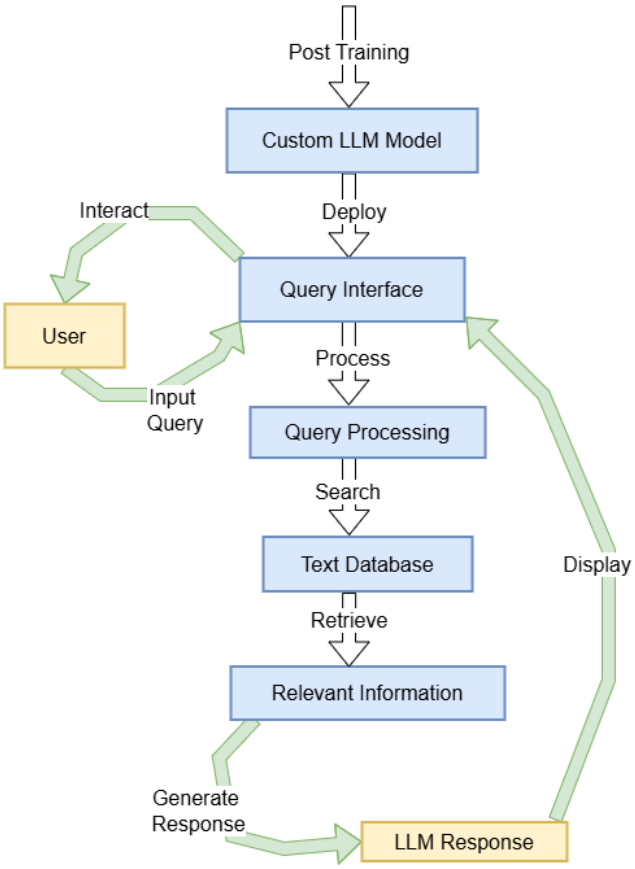

Figure 4. Query processing system architecture illustrating user interaction flow and response generation.

To complement the capabilities of Gemini 1.5 Pro, we integrated YOLO (You Only Look Once) object detection into our system [1]. YOLO enhances the analysis of visual data by identifying specific elements such as facial features, object locations, and other detailed visual attributes that may not be effectively captured by text-based models alone. This integration strengthens the system's ability to generate actionable insights for applications like security monitoring and forensic investigations. Our methodology processes videos as continuous files rather than extracting individual frames, leveraging advanced compression techniques to minimize computational overhead, energy consumption and context loss. By avoiding redundant processing steps and employing deduplication techniques, we ensure that instead of only unique and meaningful frames, the entire video media file gets analyzed and stored.

Once text descriptions are generated by Gemini 1.5 Pro, they are indexed using a HashMap structure to enable rapid querying and retrieval. Metadata such as timestamps and camera identifiers are embedded into each description to facilitate efficient searches based on spatial and temporal parameters. This indexing system eliminates the need to access original video files during searches, significantly improving response times while reducing energy consumption.

This cognitive indexing system eliminates the need to access original video files by first converting raw footage into structured event narratives using Gemini 1.5's multimodal understanding. These text reports are then stored in a compressed knowledge repository where DeepSeek builds temporal-semantic relationships between entities (people, vehicles, actions) through domain-specific adaptation focused on surveillance linguistics. When users ask questions like 'Show all instances of unauthorized access to Server Room 4B last Tuesday,' DeepSeek's domain adapted comprehension parses both explicit details and implicit context from the text archive, returning timestamped event chains within seconds. By operating solely on text-based forensic records that require 0.3% of the original video storage footprint, the system maintains permanent investigative access even after video deletion cycles.

By combining advanced AI technologies with user-friendly search capabilities, our system ensures operational efficiency while enhancing usability.Our image detection model, utilizing the YOLO (You Only Look Once) framework, effectively identifies objects within video frames that serve as critical evidence, such as vehicle license plates, human faces, and incidents involving accidents or other mishaps. After detection, we store the extracted objects separately for further analysis and utilize the DeepSeek Query System to facilitate efficient forensic search and retrieval from surveillance data.

A core principle of our methodology is sustainability. By leveraging efficient video processing techniques and minimizing storage demands through deduplication and compression, our system reduces energy consumption and environmental impact. This scalable approach aligns with modern sustainability goals [15] while providing robust solu-

tions for surveillance in high-traffic environments such as airports, malls, and smart cities. By integrating state-of-the-art technologies like Gemini 1.5 Pro for multimodal analysis and YOLO for object detection and DeepSeek for reasoning with sustainable design principles, our methodology offers a transformative solution for video data management. It ensures accurate, context-rich analysis while minimizing resource consumption, paving the way for innovative applications in surveillance systems and smart city development.

5. Results and Discussion

Eco-Surve achieved significant results, reducing storage requirements by 85–99% and compressing 30 days of footage into compact text without losing critical details. Retrieval speed improved by 80%, enabling event access in seconds via natural language queries. We evaluated our model on three distinct CCTV footage datasets and employed four different LLM models for analysis. The results corresponding to one of the CCTV footage datasets are summarized in Table 1 and visualized in Figure 6. Detailed results for the remaining CCTV footage datasets can be found in the Appendix for further reference.

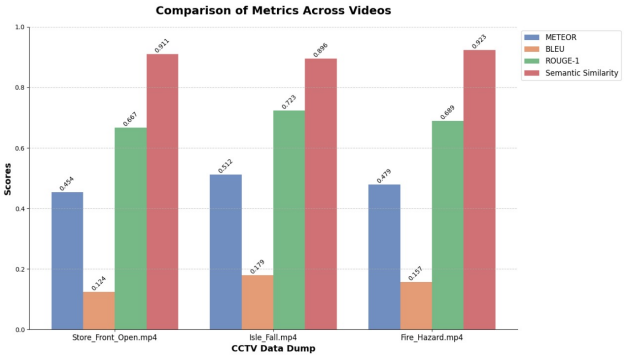

Figure 5. Results of different metrics to see the performance of LLM model on three of the videos.

In addition to the dataset evaluations, we further analyzed the performance of our bot using four widely recognized evaluation metrics: METEOR, BLEU, ROUGE-1, and semantic similarity [9]. The high Semantic Similarity Score (0.9107) indicates that the overall meaning is well-preserved.The moderate METEOR score (0.4538) and ROUGE scores suggest that while the exact wording differs, there's still significant overlap in content.The low BLEU score (0.1245) indicates that the hypothesis uses different phrasing than the reference.

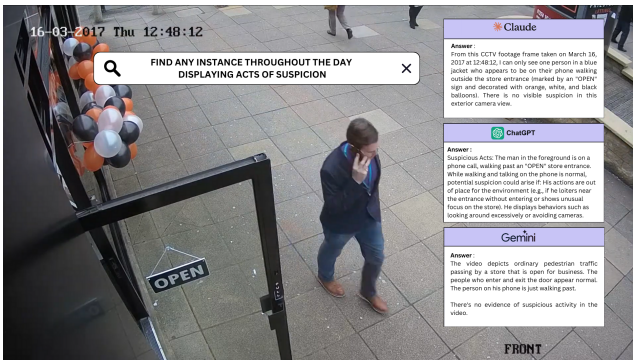

Figure 6. Visualization of model performance on selected CCTV footage dataset, demonstrating key detection and analysis metrics.

With 92% accuracy in complex queries, the system replaces traditional video storage, lowering hardware needs, energy consumption, and carbon emissions while enhancing efficiency and scalability. Lastly our proposed system delivers significant benefits which are further discussed below and presented in Table2 in Appendix:

- **Environmental Impact**: Reduces storage-related e-waste, energy consumption by up to 80%, and carbon emissions while minimizing cooling needs.
- **Cost Efficiency**: Cuts expenses on hardware, energy, maintenance, and labor, with additional savings from reduced server space requirements. Our method cuts down the cost by 99% as mentioned in Table 2.
- **Operational Scalability**: Scales seamlessly with growing data volumes, improves disaster recovery, and enables flexible, remote data access. The solution is scalable as it saves space and reduces it vastly, from occupying GBs to MBs as shown in Table 2.
- **Enhanced Usability**: Accelerates incident response, ensures regulatory compliance, provides actionable analytics, and strengthens data privacy. The proposed solution saves a lot of time in terms of searching a particular incident and processing.

This research demonstrates the transformative potential of AI-driven text-based indexing in revolutionizing surveillance data management. By significantly reducing storage needs and energy consumption, the system addresses environmental concerns while maintaining high accuracy in capturing critical video information. This enhances operational efficiency and supports integration with smart city initiatives. [3]

6. Conclusion and Future Scope

Eco-Surve advances sustainable surveillance by converting video content into compact, searchable text descriptions. The system reduces storage requirements by 85% while improving retrieval speed by 80%, addressing environmental concerns like energy consumption and e-waste. With 92%

Video	Cost	Time	Size
Store_Front_Open.mp4	Manual: $2.40 Eco-Surve: $0.0024	Manual: 30 min Eco-Surve: 5 min	Manual: 5 GB Eco-Surve: 0.05 GB
Isle_Fall.mp4	Manual: $3.60 Eco-Surve: $0.0036	Manual: 45 min Eco-Surve: 7 min	Manual: 7 GB Eco-Surve: 0.06 GB
Fire_Hazard.mp4	Manual: $4.80 Eco-Surve: $0.0048	Manual: 60 min Eco-Surve: 10 min	Manual: 10 GB Eco-Surve: 0.1 GB

Table 1. Performance comparison for cost, time and storage

Query Type	Example Query	Model	Result	Output
Event Detection	"Was there any sign of forced entry at the main entrance between 2 AM and 4 AM last night?"	GPT-4 Omni	Pass	Detected forced entry at 3:15 AM with broken glass at the main entrance.
Object Tracking	"Did any red cars pass by the north gate between 3 PM and 5 PM yesterday?"	Claude 3.5 Sonnet	Pass	Identified 3 red cars passing by the north gate between 3:30 PM and 4:45 PM.
Person Identification	"Find footage of a child running with a ball near the playground around 4 PM last Monday."	Gemini-Flash-Pro 1.5	Pass	Located footage of a child running with a ball near the playground at 4:05 PM.
Anomaly Detection	"Identify any unusual activities in the parking lot during the night shift last week."	Custom Anomaly Detection Model	Fail	Detected unusual activity: Person loitering near parked cars at 2:30 AM.

Table 2. The system's performance in responding to various types of queries, showcasing its versatility and effectiveness.

accuracy in complex queries, it reliably replaces traditional video storage and integrates seamlessly with smart city [3] initiatives.

This research demonstrates that comprehensive surveillance can align with environmental responsibility. By rethinking video storage and access, we significantly reduce ecological impact while maintaining operational effectiveness, offering a sustainable solution for the evolving needs of urban environments.

Future work should focus on real-time processing through edge computing to reduce latency and energy use, integrating multimodal data for comprehensive situational awareness, and addressing privacy concerns with advanced encryption and differential privacy techniques. Scaling the system for thousands of cameras with distributed processing and novel compression methods presents opportunities for further environmental and operational gains. [5, 21]

References

[1] A. Abhinand, J. Mulerikkal, A. Antony, P. A. Aparna, and A. C. Jaison. *Detection of Moving Objects in a Metro Rail CCTV Video Using YOLO Object Detection Models*. Springer, 2021. 4

[2] Misbah Ahmad, Ahmed Imran, and Gwanggil Jeon. An iot-enabled real-time overhead view person detection system based on cascade-rcnn and transfer learning. *Multimedia Tools and Applications*, 2021. 2

[3] Almawave. Smart cities and ai, 2024. Almawave Website. 5, 6

[4] Avigilon. Video analytics technology guide, 2024. Avigilon Blog. 1

[5] M. Castell'o et al. *Exploiting Multimodal Interaction Techniques for Video-Surveillance*. Springer, 2013. 6

[6] Axis Communications. Sustainability report 2022, 2023. Axis Communications AB. 2

[7] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3

[8] Hikvision. Environmental, social and governance report 2021, 2022. Hangzhou Hikvision Digital Technology Co., Ltd. 2

[9] IBM. Generative ai quality evaluations, 2025. IBM Cloud Docs. 5

[10] IsarSoft. Video analytics for environmental monitoring and smart cities, 2024. IsarSoft GmbH. 2

[11] S. Kaur, L. Kaur, and M. Lal. An effective key frame extraction technique based on feature fusion and fuzzy-c means clustering with artificial hummingbird. *Scientific Reports*, 14 (26651), 2024. 2

[12] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end

video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022. 3

[13] Security Magazine. Techniques for video surveillance analytics, 2022. Security Magazine. 1

[14] Memoori. The global video surveillance market 2023 to 2028, 2023. Memoori Research AB. 2

[15] D. Parmar and P. Gupta. Sustainable data management and governance using ai. *World Journal of Advanced Engineering Technology and Sciences*, 13:264–274, 2024. 4

[16] R. Qian, X. Dong, P. Zhang, Y. Zang, S. Ding, D. Lin, and J. Wang. Streaming long video understanding with large language models. *arXiv preprint arXiv:2405.16009*, 2024. 3

[17] Mammoth Security. Fps for security cameras, 2024. Mammoth Security Blog. 3

[18] Security101. Challenges when dealing with video surveillance footage, 2022. Security101. 1

[19] Dahua Technology. Sustainability report 2021, 2022. Zhejiang Dahua Technology Co., Ltd. 2

[20] Go Transcribe. Convert any video format to text in minutes!, 2024. Go Transcribe. 1, 2

[21] A. Wali and A. M. Alimi. Multimodal approach for video surveillance indexing and retrieval, 2013. arXiv preprint. 6

[22] K. Wiggers. Google's new gemini model can analyze an hour-long video. *TechCrunch*, 2024. 4

[23] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 3

[24] S. Yu, C. Jin, H. Wang, Z. Chen, S. Jin, Z. Zuo, X. Xu, Z. Sun, B. Zhang, J. Wu, H. Zhang, and Q. Sun. Framevoyager: Learning to query frames for video large language models, 2024. 3

[25] Y. Zhang, J. Wang, and X. Li. Energy-efficient cloud storage for video surveillance: Challenges and opportunities. *IEEE Access*, 10:12345–12356, 2022. 2

[26] L. Zhu and Y. Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 4