INFOVISION
#AccelerateDigital

Revolutionizing Video Intelligence:

Harnessing AI-Driven Solutions for Video Analysis

Leveraging Generative AI to Enhance Surveillance, Retail Management, and Public Safety.

The advancement of Generative AI and its applications across industries has led to obtaining crucial insights into a growing number of use cases. This paper discusses a novel approach to harnessing the power of Generative AI in Video Analysis, where the user can easily query a specific video's data through a mobile and website interface.

The solution proposed here highlights incident detection capabilities in various settings, including archived footage, and object/anomaly detection from videos. This paper highlights the in-depth scene analysis provided by this solution by utilizing Generative AI to perform Scene Understanding in various settings and provide a communication channel through Natural Language Processing and easy-to-understand input/output modalities like a traditional website or a mobile app.

Executive Summary

In today's fast-paced digital world, businesses and public institutions face unprecedented challenges in managing and analyzing vast volumes of video data. The need for accurate, efficient, and scalable video analysis solutions has never been greater. Current models, such as YOLO-based object detection, demand extensive training, high-performance servers, and lengthy retraining periods —significantly strain resources, time, and costs.

InfoVision's innovative partnership with Intel and VMWare presents a breakthrough solution: an AI-driven, video-based Large Language Model (VideoLlama) that offers unparalleled offline video insights. The ability to summarize and analyze video footage — whether detecting theft, anomaly detection, or analyzing events — empowers businesses to make data-driven decisions faster and more cost-effectively.

This document aims to present a cutting-edge solution designed to transform industries that rely on video analysis, including retail, public safety, insurance, healthcare, and more. It targets business executives, IT professionals, and security leaders seeking innovative ways to streamline operations, improve decision-making, and stay ahead of technological advances in AI-driven video intelligence. Through real-world use cases and performance insights, this paper highlights how InfoVision's Generative AI solution can enhance operational efficiency, safeguard assets, and unlock new growth opportunities.

Introduction

Background

In the digital age, businesses and institutions are inundated with immense quantities of video data, which presents both a significant challenge and an opportunity. Retailers, public safety organizations, healthcare facilities, and insurance companies are all tasked with sifting through extensive video footage to identify critical events and incidents, whether it's detecting shoplifting or analyzing accidents post-incident. Existing video analysis models, such as YOLO-based object detection, have demonstrated potential but come with substantial drawbacks. These models require highly specialized data, long training times, and expensive hardware setups, making them inefficient and costly to implement and scale.

InfoVision, in collaboration with Intel and VMWare, has developed a Proof of Concept (PoC) to address these pain points by creating a next-generation, AI-powered solution for video analysis. By utilizing Generative AI models trained on video data, this solution introduces groundbreaking advancements in video intelligence, allowing for more accurate and efficient event detection and analysis, regardless of the volume of video data involved.

Problem Statement

Today, many businesses struggle with the time, resources, and infrastructure required to effectively analyze video footage. Traditional object detection models, such as those based on YOLO, must undergo lengthy and expensive training processes, are constrained by GPU-powered high server requirements, and often need to be retrained from scratch with new data. These models' limitations hinder their scalability and inflate operational costs.

The challenge lies in the inefficiencies of existing video analysis methods, which lead to limited object and event detection, increased operational expenses, and a general inability to scale to meet the demands of dynamic industries. The primary problem this POC seeks to solve is providing a more agile, accurate, and cost-effective solution for video analysis across various verticals that is relatively easy to train and runs with hardware with optimal configuration.

Objective

The primary goal of the POC is to create an advanced video analysis solution powered by a video-based Large Language Model (VideoLlama) that drastically reduces the effort, cost, and time associated with training and deploying video intelligence systems. This solution aims to outperform existing models in accuracy and efficiency, offering a flexible, scalable alternative for businesses requiring video analysis.

The solution can be applied to a wide range of industries, including but not limited to:

- **Retail:** Detecting theft, monitoring customer flow, and loss prevention.
- **Public Safety:** Surveillance and incident response.
- **Healthcare:** Safety compliance and staff performance analysis.
- **Insurance:** Accident verification, fraud detection, and claim validation.
- **Smart Cities:** Traffic management, infrastructure monitoring, and public safety enhancement.
- **Security:** Monitoring for crowd management and protection.

Technology Overview

Technology Description:

- Leverages cutting-edge Generative AI technology with a video-based Large Language Model (VideoLlama).
- Enables extraction of actionable insights from video data through natural language-based queries.
- Processes archived footage to find specific video snippets related to events, objects, or behaviors.
- Powered by Intel Xeon on-prem servers with custom low-level programming for optimized video processing and reduced latency.
- Capable of scene summarization and frame-by-frame analysis with minimal human intervention.
- Utilizes edge computing and on-premises solutions, ensuring seamless integration into existing infrastructures.
- Provides insights and post-processing capabilities without requiring extensive hardware upgrades.

Relevance and Advantages:

- Streamlines video analysis process significantly compared to traditional YOLO-based models.
- Reduces model preparation effort by up to 50% through task-oriented fine-tuning.
- Automates surveillance, incident detection, and post-event analysis with 20% higher accuracy than existing YOLO-based solutions.
- Intel Xeon processors enhance performance with hardware-level optimizations, ensuring minimal latency during complex video processing.
- Integrates VMWare's virtualization services, offering scalability and security without compromising performance.
- Provides a flexible, scalable, and cost-effective alternative to current video analysis models.
- Ideal for businesses across multiple industries requiring operational efficiency and data-driven decision-making.

POC Description

Architecture and Design

The Proof of Concept (POC) architecture revolves around a hybrid approach that leverages Intel Xeon processors for optimized video processing and VMWare's virtualization services for scalability and deployment flexibility. The core of the system is a video-based large language model (VideoLlama) that operates on a task-oriented fine-tuning mechanism.

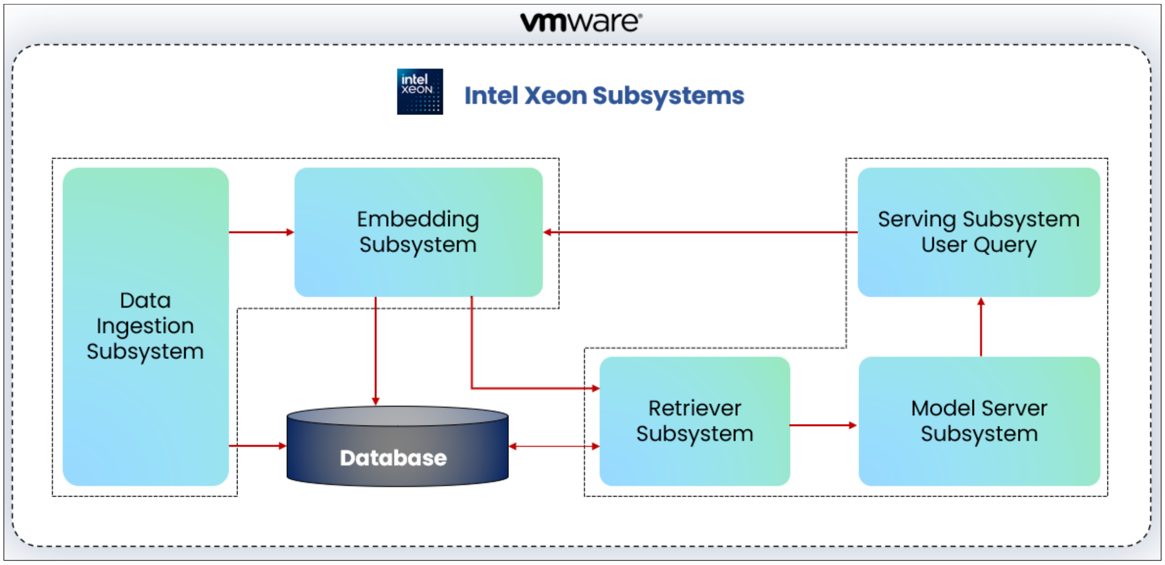

Figure 1: High-level Architecture of Visual RAG

This solution consists of a cohesive network of subsystems that work in real-time sync with minimal latency and high efficiency. These subsystems are grouped together and hosted on multiple Intel Xeon machines.

The Data Ingestion Subsystem and the Embedding Subsystem are hosted on Xeon machine 1 and the Retriever Subsystem, the Model Server Subsystem, and the Serving Subsystem are hosted on Xeon machine 2. These sets of subsystems send and receive data based on the user query and interact with a Database that stores all the relevant metadata.

The Serving Subsystem becomes the main front-facing point of contact for the user to be able to interact with this solution. Powered by the Generative AI capabilities of the Large Language

Model (LLM) hosted on the Model Server Subsystem, this solution proves to be very useful and powerful in extracting relevant video data and metadata and providing the most appropriate responses with scene descriptions.

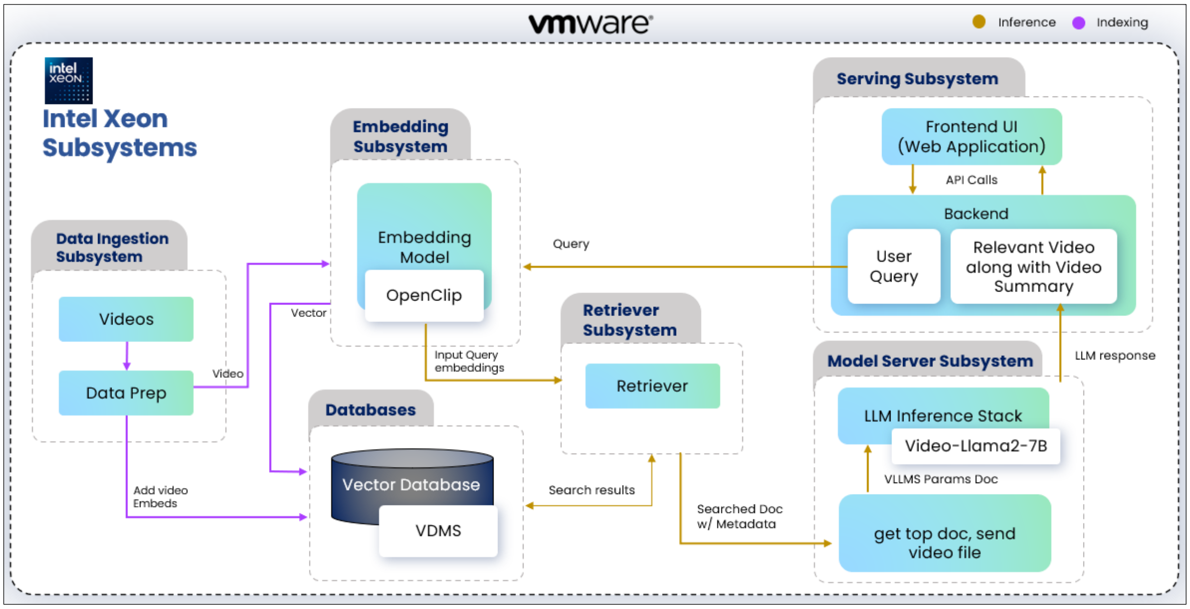

Figure 2: Detailed System diagram with Intel components

Figure 2 illustrates the process by which this solution handles an incoming user query:

Data Ingestion Subsystem: Pre-recorded video feeds are streamed into the system via edge devices or on-prem servers. The ingestion layer preprocesses the video and prepares it for the AI model.

Serving Subsystem: Users interact with the system through a user-friendly web and mobile interface. They can query the system using natural language to search for specific incidents or behaviors within the video data.

Embedding Subsystem: This layer processes user queries by converting them into text embeddings. Similarly, video feeds that were previously processed are converted into video embeddings. These video embeddings are then sent to the vector database.

Vector Database: The generated video embeddings will be stored along with the video feeds in the Vector Database.

Retriever Subsystem: The generated user query embeddings are sent to the retriever Subsystem and the retriever layer utilizes these user query embeddings to look for the requested/matching video snippets from the vector database based on the stored video embeddings and sends this metadata along with retrieved video to the Model Server Subsystem.

Model Server Subsystem: Here the video-based LLM (VideoLlama) processes these inputs and generates the appropriate scene description response to the user query by processing the video metadata and sending it back to the user interaction layer.

This generated response (Video summary), and the relevant video are then displayed to the user on the application website.

Through scene understanding, the LLM can generate insights from the video footage it parses. These insights set this solution apart because it goes beyond just object recognition and highlights key information and important things happening in the video. It can provide observations and solutions to any situation it understands from the video footage.

Deployment

The architecture supports edge computing as well as on-premises and cloud-based deployment, ensuring security and scalability across different business environments.

The design principles focused on modularity, scalability, and efficiency. By breaking down the system into independent modules, the architecture allows for flexible upgrades and quick adjustments to the system based on the specific needs of different verticals.

Functional Capabilities

The POC offers several powerful functional capabilities:

- **Video Analysis:** The system processes stored video feeds identifying events such as theft, accidents, and suspicious behaviors with negligible latency.
- **Post-Processing and Summarization:** Users can request summaries of pre-recorded video footage, allowing for quick identification of incidents without manually reviewing hours of footage.
- **Natural Language Querying:** Users can interact with the system via natural language, asking questions such as “Show me instances of shoplifting last week” or “Display any traffic accidents during the afternoon.”
- **Scalable Deployment:** The system can be deployed on the edge, on-premises, or in the cloud, making it adaptable to various business needs.
- **Actionable Alerts and Insights:** The POC provides customizable-automated alerts to security personnel or store managers when specific incidents are detected, enabling immediate response.

Business Impact

Business Case Analysis

The POC directly addresses the problem of inefficient video analysis by offering a solution that significantly reduces the time, cost, and resources required for monitoring and post-processing of video footage. The AI-powered model is capable of detecting incidents and anomalies 20% more accurately while requiring 50% less training effort than traditional YOLO-based models. This improvement translates to a reduced need for expensive infrastructure and a faster time-to-market for video intelligence systems, especially for industries that rely heavily on video surveillance and monitoring.

For retail, security, healthcare, and public safety businesses, processing and analyzing video footage without the need for costly server farms or extensive retraining is a game-changer. The solution provides an immediate and actionable response to incidents, improving both operational efficiency and customer satisfaction.

Value Proposition

The POC presents significant quantitative and qualitative benefits, offering a strong value proposition for businesses across multiple verticals:

- **Efficiency Gains:** The system reduces video analysis efforts by up to 50%, translating into cost savings on both personnel and hardware.
- **Enhanced Accuracy:** With a 20% improvement in accuracy over existing models, the POC increases the reliability of event detection, reducing false positives and improving response times.
- **Scalability:** The solution's modular architecture ensures scalability, making it adaptable to a wide range of business sizes and deployment environments.
- **Cost Savings:** The reduction in training time and infrastructure needs lowers overall costs for businesses, providing a measurable ROI within a short period of adoption.

Comparison between YOLO and Video-based LLM

Features comparison:

Feature	YOLO	Video LLM-based Scene Description
Object detection	The primary functionality is to seamlessly recognize and localize objects with good accuracy.	Able to recognize most of the general items in the scene and produce a good scene description.
Contextual understanding	Trained to only detect objects in each frame without any scene understanding. It would take multi-model architecture to understand the scene.	Ability to understand the situation and generate responses as required by the user.
Real-time performance	Fast and detects objects accurately in real-time.	Slower but can understand the intention of the user and provides accurate circumstantial data.

Action-based events	The purpose of the YOLO is just to detect the objects in the frame but not to understand the intention of the scene.	Produces a greater understanding of the scenario and provides an accurate description.
Architecture	Specifically designed for object detection	General-purpose language model
Knowledge base	Trained on large datasets of specific objects tailored fit to the user's requirements.	Trained on massive amounts of text data, allowing them to access and process information from the real world.
Data availability	Procurement and processing of data is a laborious task that must be performed as a primary step in training a customized YOLO model.	A Plethora of easily available data is found online which can play a vital role in fine-tuning the LLMs.
Training expenditure	Powerful GPUs will be required to train models efficiently which can cost us thousands of dollars.	LLMs are very powerful language models, which can produce desired results with just fine-tuning – leading to lowered hardware costs.
Adaptability	Ability to detect objects in each frame.	Can be fine-tuned for various tasks – translation, summarization and question answering.

YOLO vs Scene Understanding feature comparison

Object Detection vs. Scene Understanding: YOLO excels at detecting objects within frames, but video-based large language models (LLMs) go further by understanding the scene context and responding accurately to user intent.

Contextual Awareness: YOLO lacks scene-level awareness and focuses only on identifying objects, while video-based LLMs offer a deeper understanding of the scenario, generating more nuanced descriptions.

Real-Time Performance: YOLO performs faster in real-time object detection, but video-based LLMs prioritize accuracy and circumstantial insights over speed, making them ideal for detailed analysis.

Data and Training: YOLO requires labor-intensive data procurement and powerful GPUs for training, whereas LLMs leverage abundant text data, reducing training costs and enhancing adaptability.

Adaptability and Purpose: While YOLO is specifically designed for object detection, video-based LLMs can be fine-tuned for a wide range of tasks, including summarization, translation, and question-answering.

Training Costs: LLMs are more cost-effective, requiring only fine-tuning rather than full-scale training on expensive hardware, unlike YOLO's resource-intensive training requirements.

Cost Comparison

Steps involved in training customized YOLO model	Human hours required (2000 images for 10 objects)
Data Procurement	40 hours
Data Refinement	24 hours
Data Annotation	200 hours
Data Training	10 hours
	274 hours = $27,400

YOLO vs Scene Understanding Cost comparison

Multi-model YOLO Classifier Approach	Video LLM-based Scene Description Approach
Customized model: $27,400 x 4 = $109,600	Fine-tuned model: $13,700
Training Hardware: $2000 x 3 = $6,000	Hardware: $12 x 24 = $300
Model validation: $90,000	Model validation: $90,000
Deployment Hardware (On-prem/Cloud): $60,000	Deployment Hardware (On-prem/Cloud): $60,000
TOTAL: $265,600	**TOTAL: $164,000**

Use cases

- **Retail Security Analysis:**
    - **Theft Detection:** Store managers can quickly search for instances of theft by asking the system to show video snippets where shoplifting or suspicious activities are detected from the stored footage.
    - **Employee Performance:** Evaluate employee behavior and performance during customer interactions or in handling merchandise.
- **Insurance Claims:**
    - **Accident Verification:** Insurance companies can use your solution to verify claims by identifying and reviewing specific incidents such as car accidents or property damage.
    - **Fraud Detection:** Detect fraudulent claims by analyzing video footage for inconsistencies with reported incidents.
- **Event Summarization:**

  - **Conference Highlights:** Automatically generate highlights from conference recordings, focusing on key moments like speaker presentations, panel discussions, and audience interactions.
  - **Sports Recaps:** Create summaries of sports events by extracting key plays, goals, or significant moments from full game recordings.
- **Safety Compliance:**
  - **Workplace Safety Audits:** Companies can review video footage to ensure compliance with safety protocols and identify any hazards that need addressing.
  - **Construction Site Monitoring:** Monitor construction sites for safety violations or to document progress.
- **Education and Training:**
  - **Lecture Summarization:** Generate lectures or training session summaries focusing on important topics and discussions.
  - **Skill Assessment:** Evaluate training sessions by identifying key moments where trainees demonstrate specific skills.
- **Healthcare Monitoring:**
  - **Staff Efficiency:** Ensure staff adhere to protocols and efficiently respond to patient needs, review their behaviors, and provide feedback for improvement.
- **Event Security:**
  - **Crowd Management:** Monitor large events for crowd control, identifying potential safety hazards or disturbances from recorded clips and improving overtime.
- **Smart City Applications:**
  - **Public Safety:** Enhance public safety by monitoring public spaces and auto-reporting events with video snippets for prompt action.
  - **Infrastructure Monitoring:** Detect issues with city infrastructure, such as broken streetlights or damaged property.

Technical Details

Integration capabilities

The solution is designed with a high degree of flexibility to integrate seamlessly into existing systems without disrupting operations. Key integration features include:

- **APIs and SDKs:** The system offers a set of RESTful APIs and SDKs that allow easy integration with existing video management systems, security platforms, or retail management software. This enables businesses to incorporate AI-driven video analysis into their current workflows without extensive redevelopment.

- **Edge Computing Compatibility:** The architecture supports edge computing, allowing video feeds from edge devices (e.g., security cameras, IoT sensors) to be processed locally. This reduces the strain on central servers and ensures low-latency performance.
- **Cloud and On-Prem Integration:** The POC can be deployed on cloud infrastructures (e.g., AWS, Azure) or on-premises systems, depending on the enterprise's preference. It integrates with both virtualized environments (through VMWare) and physical servers, making it adaptable to different deployment strategies.
- **Existing Camera Compatibility:** The solution is compatible with a wide range of existing camera setups, eliminating the need for hardware replacement. It can be configured to work with various camera protocols and feeds, providing flexibility across different environments.

Scalability and performance metrics

The POC is designed to scale horizontally and vertically based on the enterprise's needs, providing excellent performance under various loads:

- **Horizontal Scaling:** The system can handle large volumes of video data concurrently by distributing workloads across multiple edge devices or server clusters. Architecture supports scaling across multiple regions, allowing global businesses to deploy the solution in different locations without degradation in performance.
- **Vertical Scaling:** The system can be optimized with more powerful hardware (e.g., additional processors and enhanced GPU support) to handle more complex tasks and larger data sets. This enables analysis even in high-traffic environments such as airports or large retail stores.
- **Performance Metrics:** Initial tests have shown that the system processes video 20% more accurately than YOLO-based models and with up to 50% less training time. In real-world scenarios, the POC has demonstrated the ability to efficiently process multiple videos and detect multiple events concurrently.

Conclusion

This document has outlined InfoVision's groundbreaking Proof of Concept (POC) development and capabilities for AI-powered video intelligence. Designed to address the growing need for video analysis across various industries, the solution provides a scalable, accurate, and cost-effective alternative to traditional models. By leveraging Intel's Xeon processors, VMWare's virtualization services, and the VideoLlama model, this POC delivers a 20% improvement in accuracy and a 50% reduction in training time compared to YOLO-based systems.

The architecture's flexibility ensures compatibility with existing systems, allowing businesses to adopt the technology without extensive infrastructure changes. Furthermore, its robust security features guarantee that sensitive video data is protected, making it suitable for healthcare, public safety, and retail environments. Through real-world use cases, the POC has demonstrated its ability to enhance operational efficiency, improve safety, and streamline decision-making.

Looking ahead, the future evolution of this POC will continue to be shaped by emerging technologies such as Edge computing, 5G, and Generative AI, ensuring it remains at the forefront of innovation in video intelligence. As businesses increasingly seek data-driven insights and proactive solutions, this POC stands poised to transform video analysis, delivering superior results and significant value across diverse verticals.

This solution redefines how businesses and public institutions harness the power of video data, offering a scalable and effective path to enhanced operational efficiency, safety, and growth.

Sources:

1. Article Title: "Unlocking New Frontiers: The Synergy of Audio Transcripts using Video Intelligence API and Generative API"

Page Title: "How video analytics and generative AI will reshape industries"

Link: https://cloud.google.com/blog/products/ai-machine-learning/how-video-analytics-and-generative-ai-will-reshape-industries

2. Article Title: "Real Time Video Analytics with Generative AI"

Link: https://www.xenonstack.com/blog/real-time-video-analytics

3. Paper Title: "A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming"

Link: https://arxiv.org/html/2404.16038v1

4. Article Title: "Enhacing CCTV Security with Large Language Models and Computer Vision"

Link: https://specialitweapons.com/enhancing-cctv-security-with-large-language-models-and-cv/

Credits:

Abhiram Kalidindi

Noumika Balaji

Ria Ghosh

Pratyoosh Patel