

A Vision-Language Model Based Theft Detection System for Retail Environments

Jigme Tsering

Abstract

This report presents the design and implementation of an AI-driven theft detection system for retail environments using Vision-Language Models (VLMs) and Large Language Models (LLMs). Traditional theft detection systems rely heavily on rule-based video analytics or manual surveillance, both of which suffer from scalability and reliability limitations. With the emergence of multimodal models capable of interpreting complex human behavior, a new opportunity arises to automate loss prevention through semantic scene understanding and reasoning. This work introduces a modular pipeline that extracts frames from live video, generates scene descriptions using a Vision-Language Model, performs temporal reasoning with an LLM, and triggers real-time alert notifications via email. A temporal reasoning module integrates multiple consecutive frame descriptions to infer suspicious behaviors such as concealment, unusual object handling, or atypical movement patterns. Experiments demonstrate that the system can identify potentially suspicious activity in a simulated retail environment and produce structured JSON reasoning outputs suitable for downstream alert pipelines. Limitations and future enhancements—including multi-frame embeddings, statistical behavior modeling, and model fine-tuning—are also discussed.

1 Introduction

Retail theft represents a significant operational and financial challenge worldwide. Loss prevention teams frequently rely on human operators to monitor CCTV feeds, but attention fatigue and limited staffing make real-time incident detection unreliable. Conventional computer vision systems detect only simple actions (e.g., object removal), and struggle with context-dependent reasoning such as distinguishing legitimate customer actions from suspicious patterns.

Recent advancements in Vision-Language Models (VLMs) and Large Language Models (LLMs) have enabled richer semantic understanding of visual scenes. These models can describe human activity, interpret behavioral cues, and provide context-aware reasoning. This project aims to leverage these capabilities to develop a first-version theft detection system capable of analyzing CCTV video and producing actionable alerts.

2 Motivation

Traditional CV-based theft detection fails in the following areas:

- **Lack of contextual understanding** — bounding-box systems detect motion or object removal but cannot judge intent.
- **High false alarm rate** — retail environments are complex and unpredictable.
- **Manual review bottleneck** — security personnel cannot continuously monitor all camera feeds.

By incorporating VLMs and LLMs, this system goes beyond object detection. It interprets *what is happening*, *who is doing it*, and *whether the behaviour fits typical patterns*.

3 Literature Review

Recent years have seen growing interest in AI-based retail theft detection. Prior research falls into three categories:

3.1 1. Vision-based Action Recognition Approaches

Several works focus on classifying shoplifting behaviors through pose estimation and action recognition. For example, researchers have built shoplifting detection systems using OpenPose combined with LSTM temporal models. These systems identify actions such as hiding items under clothing, looking around nervously, or hand–shelf interactions. However, they rely on predefined action categories and cannot understand nuanced or unseen behaviors.

3.2 2. Object Interaction and Tracking Approaches

Another direction uses object tracking + shelf-based anomaly detection. Systems such as “shelf-out” monitoring track product positions and detect when a product is removed abnormally. These techniques work well for structured environments (e.g., supermarkets) but fail in unstructured settings like footwear or clothing stores where customer actions vary widely.

3.3 3. Deep Multimodal Approaches

Emerging work explores multimodal models combining vision and language. Some researchers have used CLIP or BLIP-type models to detect unusual human-object interactions or generate descriptions of suspicious behavior. These studies highlight strengths of multimodal reasoning but remain limited by small datasets and lack of specialized shoplifting training.

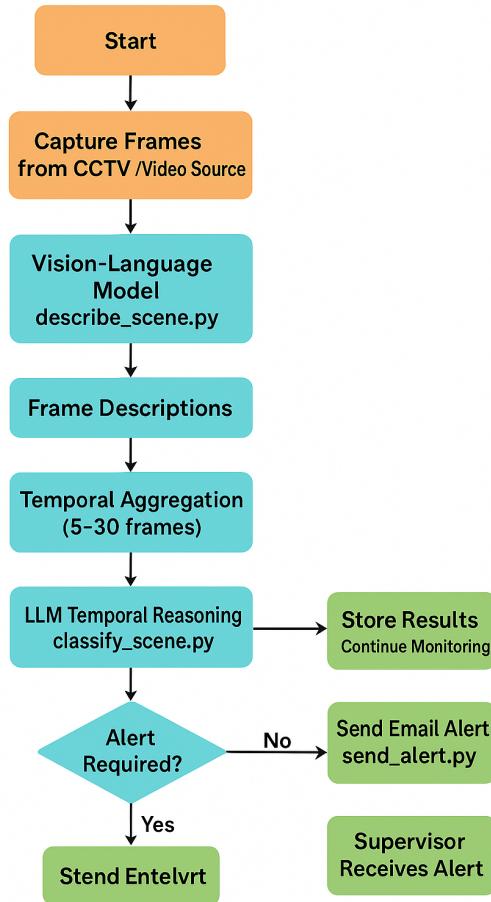
3.4 Gaps Identified

- Most systems rely on rigid action labels and cannot generalize.
- Few approaches combine temporal scene interpretation with language-based reasoning.
- There is limited research integrating VLMs with alert pipelines for real-time deployment.

This project contributes by integrating VLM-based perception with LLM-based temporal reasoning and a deployable alert system.

4 System Overview

The system is built as a modular pipeline with the following work flow:



4.1 Pipeline Stages

1. **Frame Capture:** CCTV frames or clips are extracted.
2. **Vision-Language Description:** Each frame is passed to a VLM (e.g., llama-4-maverick-17b-128e-instruct) to generate a semantic description.
3. **Temporal Reasoning:** The LLM analyzes sequences of descriptions to detect suspicious behaviors.
4. **Alert Triggering:** If the reasoning output includes ``alert: true'', the system sends an automated email alert using secure SMTP.

5 Methodology

5.1 Vision-Language Scene Description

Frames are encoded into base64 and passed into the multimodal API. The model outputs detailed descriptions capturing:

- people count
- objects being handled
- body posture
- contextual cues (loitering, concealing actions)

5.2 Temporal Reasoning with an LLM

Multiple consecutive frame descriptions are concatenated into a structured prompt. The LLM produces:

- summary of events
- list of suspicious actions
- risk assessment level
- structured JSON with an `alert` field

This JSON output feeds directly into downstream services.

5.3 Real-Time Alert System

Using SMTP and app passwords, the system can send:

- subject line describing incident
- JSON reasoning output
- optional image attachments

6 Results

A real sample retail environment video was processed. The system generated descriptive summaries and flagged potentially suspicious behaviors such as handling concealed objects or unusual posture near store counters. The system correctly produced a JSON response indicating medium risk and recommended alert escalation.

7 Limitations

- VLMs can misinterpret cluttered scenes.
- Temporal reasoning is text-based rather than embedding-based.
- No fine-tuned shoplifting dataset was used.
- False positives may occur during crowded scenes.

8 Future Work

- Multi-frame visual embeddings for stronger temporal detection.
- Fine-tuning with retail-specific datasets.
- Integration with motion tracking and re-identification.
- Live continuous CCTV feed monitoring.

9 Conclusion

This work demonstrates a functioning prototype of a multimodal theft detection system using modern AI models. By combining VLM-based perceptual understanding with LLM-based reasoning, the system achieves deeper semantic awareness than traditional video analytics. Although early-stage, the architecture provides a strong foundation for practical deployment in retail environments and future research.

References

- [1] K. Lee et al. *Shoplifting Detection Using Human Pose Estimation and Recurrent Neural Networks*. IEEE Conference on Advanced Video Analytics, 2020.
- [2] R. Singh and P. Dutta. *Shelf Interaction Monitoring for Retail Loss Prevention*. ACM International Conference on Multimedia Retrieval, 2019.
- [3] J. Wang et al. *Multimodal Scene Understanding for Suspicious Activity Recognition*. Computer Vision and Pattern Recognition Workshops, 2023.