SmartEyes: Plug-and-Play Event Detection for Retail Loss Prevention

Pi-Wei Chen[1], Jerry Chun-Wei Lin[4], Barış Fahri Kahrıman[1], Zih-Ching Chen[2], Rafał Cupek[1], Marek Drewniak[3]

[1]Silesian University of Technology
[2]BakuAI [3]Nvidia
[4]Aiut
{pi-wei.chen,baris-kahriman,rafal.cupek}@polsl.pl, jcwlin.tw@gmail.com, virginiac@nvidia.com, mdrewniak@aiut.com.pl

Abstract

Event detection is essential for surveillance, particularly in retail loss prevention, where accurate and timely monitoring is critical. Vision Language Models (VLMs) provide strong generalization but are inefficient at processing full video streams and are prone to hallucinations induced by redundant frames. We present **SmartEyes**, a plug-and-play system for real-time retail surveillance. SmartEyes introduces the **Perception Cognition Focusing (PCF)** framework, which combines lightweight perception with semantic triggering to isolate two keyframes (customer contact and departure) and constrains the VLMs to a focused differencing task. This design reduces hallucination by 44% compared to vanilla VLMs. From the demonstrated retail application, the proposed perception-to-reasoning pipeline is general and directly extends to industrial environments that require reliable event detection and real-time decision-making. Our demo includes a user-friendly Region of Interest (ROI) selection interface and live CCTV monitoring, producing accurate alerts within 1–2 seconds on a single RTX 4080 GPU. This lightweight framework design enables efficient deployment to broader industrial applications.

Introduction

Event detection plays a central role in modern surveillance systems, powering applications in public safety, traffic monitoring, and retail analytics. In high-end retail environments, it is impractical to assign staff to monitor every customer's movement. Automated systems that detect when items are taken from shelves can therefore reduce labor costs, improve operational efficiency, and support timely interventions for loss prevention. The dynamic nature of retail environments also suggests a promising extension to industrial settings, where reliable perception and event recognition are essential for real-time decision-making.

Recent advances in Vision Language Models (VLMs) provide strong generalization for video understanding without task-specific fine-tuning, reducing annotation costs (Wang et al. 2022, 2023, 2024). However, their direct use in retail is limited by two challenges (Figure 1, upper part). First, inefficiency: even mid-sized VLMs (Team et al. 2024; Liu et al. 2023; Yang et al. 2025) with 7B to 12B parameters must repeatedly process redundant clips from a continuous stream, causing wasted computation and latency on consumer GPUs such as RTX 4080. Second, hallucination: continuous analysis of raw video increases false positives (Li, Im, and Fazli 2025; Leng et al. 2024), which hinders deployment when accuracy and responsiveness are critical.

We introduce **SmartEyes**, a plug-and-play event detection system for retail loss prevention. SmartEyes centers on **Perception Cognition Focusing (PCF)**, which routes lightweight perception outputs into targeted VLM reasoning. PCF performs semantic compression by extracting only two keyframes, representing customer contact and departure from the shelf, so the VLM executes a focused differencing task rather than full-stream reasoning. This design greatly reduces hallucination compared to vanilla VLM on raw video, as shown in Table 1. Combined with an intuitive interface for defining ROIs, SmartEyes adapts to diverse store layouts and delivers real-time, accurate alerts on consumer-grade GPUs.

SmartEyes Overview

SmartEyes is composed of two main components:

- **Adaptive ROI Label Interface:** To adapt to any scene or store layout, SmartEyes provides an intuitive interface to define ROI directly from CCTV views. Powered by Segment Anything Model (SAM) (Kirillov et al. 2023), the interface supports pixel-level segmentation, allowing users to flexibly and precisely specify shelves or other sensitive areas.
- **Perception Cognition Focusing (PCF):** This framework structures the detection pipeline into three stages: (i) a lightweight perception module that detects and tracks people in real time, (ii) a semantic trigger module that isolates keyframes of customer–shelf interaction, and (iii) a cognition module where the VLM compare the contact frame ($k1$) against the departure frame ($k2$) to check whether the person is leaving with a new item. When such a change is detected, the system issues an alert to notify the store staff. The PCF framework is described in detail in the following subsection.

PCF Framework: Although VLMs show strong zero-shot video understanding, they often occur hallucination when it is applied to continuous streams, where redundant frames

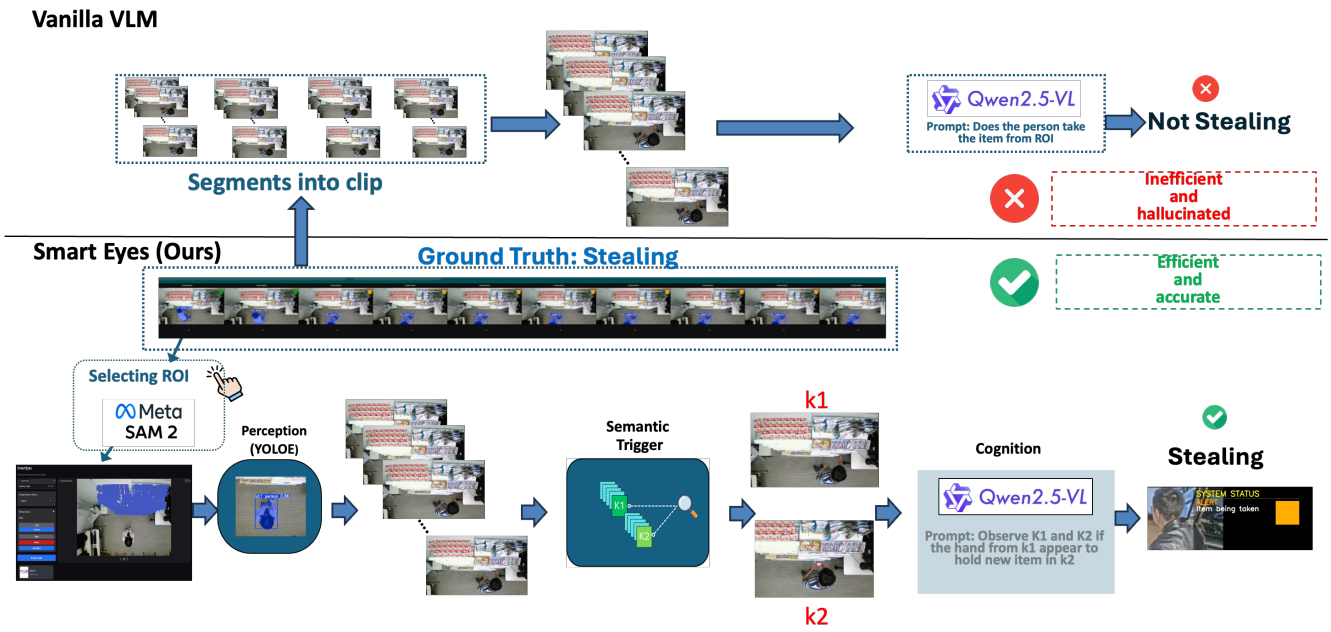

Figure 1: The upper part illustrates the limitations of directly feeding raw video into a VLM, which requires processing every frame of a video clip. The lower part shows the proposed SmartEyes framework, which improves efficiency by inferring only on two keyframes (*k1*, *k2*) while mitigating hallucinations and improving the efficiency .

encourage over-reliance on language priors rather than visual evidence (Leng et al. 2024; Li, Im, and Fazli 2025). Our **PCF** framework addresses this by filtering video into only the most informative keyframes, so the VLM performs a constrained differencing task grounded in observable change. This focused reasoning minimizes hallucination while preserving the efficiency needed for real-time retail monitoring. PCF consists of three modules:

- **Perception module:** A lightweighted text prompt YOLOE (Wang et al. 2025) detector performs real-time person detection and tracking. Each individual is assigned a consistent tracking ID and a pixel-level segmentation mask, providing spatial–temporal grounding for subsequent analysis.
- **Semantic Trigger Module:** Acting as the core alignment mechanism, which performs semantic compression by monitoring spatial interactions between tracked persons and ROI. A contact event (*k1*) is triggered when the segmentation mask of a tracked person first overlaps with the ROI boundary, and a departure event (*k2*) is triggered when the overlap ceases. These two keyframes capture the minimal sufficient visual evidence for theft detection, filtering out redundant frames, and reducing the VLMs hallucination.
- **Cognition module:** The VLMs receives only *k1* and *k2* and is prompted to compare the person's hand status across them. The task is framed as: *"Does the hand contain a new item at departure (k2) compared to contact (k1)?"* This constrained differencing forces the VLM to make decisions strictly from visual evidence, improving robustness and reducing hallucination, as shown in Table 1.

Method	Accuracy	FP	FN
SmartEyes (Ours)	**82%**	**16%**	**2%**
Vanilla Qwen 2.5 VL (7b)	38%	52%	10%

Table 1: Performance comparison between PCF (ours) and a vanilla VLMs baseline (All used Qwen 2.5 VL model). FP represents False Positive while FN represents False Negative

Demonstration and Evaluation

Demonstration. Our demo shows the plug-and-play workflow. A user first defines an ROI on a live CCTV feed using the SAM-powered interface. Once set, SmartEyes begins real-time monitoring and issues alerts within 1–2 seconds of a potential theft event. The complete process is shown in the demo video.

Evaluation. We evaluated SmartEyes on the MERL Shopping Dataset (Singh et al. 2016). Using Qwen-VL (7B) on a single RTX 4080 GPU, our PCF-based method reached 82% accuracy, while the vanilla model showed frequent false positives and negatives. These results demonstrate that PCF enables reliable real-time event detection on affordable hardware, making it promising for AGV-related extension for real-time event detection or monitoring.

Acknowledgments

This work was co-funded by the European Union HORIZON TMA MSCA Doctoral Networks / HORIZON-

MSCA-2023-DN-01 / project TUAI / grant agreement N° 101168344. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Co-funded by the Statutory Research funds of the Department of Distributed Systems and Informatic Devices, Silesian University of Technology, Gliwice, Poland (Grants BK-244/RAu8/2025 02/110/BK_25/1036). We also acknowledge the support of tools such as Gemini and ChatGPT for proofreading, and implementation.

References

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13872–13882.

Li, C.; Im, E. W.; and Fazli, P. 2025. Vidhalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13723–13733.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in neural information processing systems*, volume 36, 34892–34916.

Singh, B.; Marks, T. K.; Jones, M. J.; Tuzel, C. O.; and Shao, M. 2016. A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1961–1970.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

Wang, A.; Liu, L.; Chen, H.; Lin, Z.; Han, J.; and Ding, G. 2025. YOLOE: Real-Time Seeing Anything. *arXiv preprint arXiv:2503.07465*.

Wang, J.; Ge, Y.; Yan, R.; Ge, Y.; Lin, K. Q.; Tsutsui, S.; Lin, X.; Cai, G.; Wu, J.; Shan, Y.; et al. 2023. All in One: Exploring Unified Video-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6598–6608.

Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2024. Videoagent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 58–76. Springer.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191*.

Yang, A.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Huang, H.; Jiang, J.; Tu, J.; Zhang, J.; Zhou, J.; et al. 2025. Qwen2.5-1M Technical Report. *arXiv preprint arXiv:2501.15383*.