Kaunas University of Technology

Faculty of Informatics

Clip-Level Suspicious Activity Detection in Retail Surveillance Videos Using Human–Object Interaction and Temporal Modelling

Master's Final Degree Project

Muhammad Sohaib

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Clip-Level Suspicious Activity Detection in Retail Surveillance Videos Using Human–Object Interaction and Temporal Modelling

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Muhammad Sohaib

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Prof. Dr. Vytenis Punys

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Muhammad Sohaib

Clip-Level Suspicious Activity Detection in Retail Surveillance Videos Using Human–Object Interaction and Temporal Modelling

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Muhammad Sohaib

Confirmed electronically

Kaunas University of Technology

Faculty of Informatics

Master's final degree project

Topic of the project	Clip-Level Suspicious Activity Detection in Retail Surveillance Videos Using Human–Object Interaction and Temporal Modelling
Requirements and conditions (title can be clarified, if needed)	
Supervisor	**Prof. Dr. Armantas Ostreika**
	(position, name, surname, signature of the supervisor) (date)

Muhammad Sohaib. Clip-Level Suspicious Activity Detection in Retail Surveillance Videos Using Human–Object Interaction and Temporal Modelling / supervisor Prof. Dr. Armantas Ostreika; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer Science / Artificial Intelligence.

Keywords: shoplifting detection; suspicious activity detection; human–object interaction; CLIP; vision-language models; prototype memory; video anomaly detection; retail surveillance; YOLO; GRU; Transformer.

Kaunas, 2026. 69 pages.

Summary

Retail theft (shoplifting) has been a significant cause of financial loss to retailers in various parts of the world and has inspired the development of smarter video surveillance systems that can automatically alert suspicious activity. The proposed thesis is based on the hypothesis of determining whether the use of human-object interaction (HOI) semantics in temporal anomaly detection can significantly enhance detection quality over person-detection statistical baselines. This study is centred on retail surveillance video clip-level anomaly detection and measures performance based on standardised ranking measures.

The proposed pipeline comprises multiple stages. At the first stage, video frames are processed by a pretrained YOLO object detector to identify people and surrounding objects. From these detections, a small three-dimensional statistical descriptor is calculated per frame based on the number of persons detected, the average person-detection confidence, and the standard deviation of person-detection confidences. This forms the YOLO person-detection statistical baseline utilised as the ablation study.

In order to improve semantic interpretation, a HOI module based on CLIP is proposed. A CLIP-based HOI encoder is applied to each frame and to a crop of the largest detected person, producing similarity scores against a curated prompt bank. The resulting HOI prompts are concatenated with the three-dimensional YOLO person-detection statistics which result in an augmented frame-level representation explicitly representing the interaction semantics. Temporal modelling is performed using two alternative architectures, a Gated Recurrent Unit (GRU) and a Transformer encoder. Sliding windows are used to feed feature sequences and max pooling is used to merge window-level predictions to obtain clip-level anomaly scores.

An experimental memory module is further added (single-seed exploratory) to stabilise predictions. Temporal embedding in training class-wise prototypes is represented as running averages. During inference, the similarity margins between normal and abnormal prototypes are adjusted by cosine similarity to lower false positives.

The system is tested on the Kipshidze retail shoplifting dataset consisting of 182 video clips (90 normal, 92 shoplifting) with clip-level labels. The YOLO person-detection statistical baseline achieves a clip-level ROC-AUC of 0.8308 ± 0.0073 across three seeds (41, 42, 43). On adding CLIP-based HOI semantic features, performance increases to 0.9405 ± 0.0096, compared to the YOLO person-detection statistical baseline at 0.8308 ± 0.0073.

The prototype memory module does not raise the raw ROC-AUC over the GRU temporal encoder alone (0.954 vs. 0.990 at seed 42, single-seed exploratory runs); instead, it rebalances the operating point, recovering recall from 0.643 to 0.857 while keeping precision at 0.800 (F1 0.828), which reduces false-negative sensitivity at the operating threshold. The ablation study supports the contribution of each module. The removal of the HOI features leads to the greatest performance fall that confirm the value of explicit interaction modelling, and the removal of the memory module reduces precision at the operating point, but not ROC-AUC ranking quality.

A single-seed exploratory comparison of the GRU and Transformer encoders (seed 42) shows that, although both reach the same validation ROC-AUC (0.9430), the GRU achieves a higher test ROC-AUC (0.9898 vs. 0.9133). This is consistent with the recurrent inductive bias being more appropriate for the small training set used here (130 clips, 461 windows), but a multi-seed evaluation is left for future work to confirm.

Muhammad Sohaib. Įtartinos veiklos aptikimas mažmeninės prekybos stebėjimo vaizdo įrašuose, taikant žmogaus–objekto sąveikos ir laiko sekų modeliavimo metodus / vadovas prof. dr. Armantas Ostreika; Informatikos fakultetas, Kauno technologijos universitetas.

Studijų kryptis ir sritis (studijų krypties grupė): informatika / dirbtinis intelektas.

Raktiniai žodžiai: vagysčių parduotuvėse aptikimas; įtartinos veiklos aptikimas; žmogaus ir objekto sąveika; CLIP; regos ir kalbos modeliai; prototipo atmintis; vaizdo anomalijų aptikimas; mažmeninės prekybos stebėjimas; YOLO; GRU; Transformer.

Kaunas, 2026. 69 puslapiai.

Santrauka

Vagystės mažmeninės prekybos vietose, arba vagystės iš parduotuvių, yra reikšminga finansinių nuostolių priežastis mažmenininkams įvairiose pasaulio šalyse. Ši problema paskatino kurti išmanesnes vaizdo stebėjimo sistemas, galinčias automatiškai įspėti apie įtartiną veiklą. Šiame baigiamajame darbe keliama hipotezė, kad žmogaus ir objekto sąveikos semantikos naudojimas, t. y. HOI (angl. Human–Object Interaction, žmogaus ir objekto sąveika), laikinių anomalijų aptikimo uždavinyje gali reikšmingai pagerinti aptikimo kokybę, palyginti su statistiniais baziniais metodais, paremtais tik asmenų aptikimu. Tyrimas orientuotas į mažmeninės prekybos vaizdo stebėjimo įrašų anomalijų aptikimą vaizdo klipo lygmeniu, o metodo veikimas vertinamas taikant standartizuotus rangavimo kokybės matus.

Siūloma apdorojimo grandinė sudaryta iš kelių etapų. Pirmajame etape vaizdo kadrai apdorojami iš anksto apmokytu YOLO (angl. You Only Look Once) objektų detektoriumi, siekiant aptikti žmones ir aplinkinius objektus. Remiantis šiais aptikimais, kiekvienam kadrui apskaičiuojamas trijų komponentų statistinis deskriptorius, sudarytas iš aptiktų asmenų skaičiaus, vidutinio asmenų aptikimo pasikliovimo įverčio ir asmenų aptikimo pasikliovimo įverčių standartinio nuokrypio. Šis deskriptorius sudaro YOLO asmenų aptikimo statistinį bazinį metodą, naudojamą abliacijos tyrime.

Siekiant pagerinti semantinę vaizdo interpretaciją, siūlomas HOI modulis, pagrįstas CLIP (angl. Contrastive Language–Image Pre-training, kontrastinis kalbos ir vaizdo išankstinis mokymas) modeliu. CLIP pagrindu sukurtas HOI koduotuvas taikomas kiekvienam vaizdo kadrui ir didžiausio aptikto asmens iškarpai, apskaičiuojant panašumo įverčius pagal sudarytą tekstinių raginimų banką. Gauti HOI semantiniai požymiai sujungiami su trijų komponentų YOLO asmenų aptikimo statistiniais požymiais. Taip gaunama papildyta kadro lygmens reprezentacija, aiškiai įtraukianti sąveikos semantiką. Laikinis modeliavimas atliekamas naudojant dvi alternatyvias architektūras: GRU (angl. Gated Recurrent Unit, vartinis rekurentinis vienetas) ir transformerio koduotuvą. Požymių sekoms pateikti naudojami slenkantys langai, o langų lygmens prognozėms sujungti taikomas maksimalios reikšmės agregavimas, leidžiantis gauti anomalijos įverčius vaizdo klipo lygmeniu.

Papildomai į sistemą įtraukiamas eksperimentinis atminties modulis, vertintas žvalgomojoje vienos atsitiktinės pradinės reikšmės konfigūracijoje. Mokymo metu kiekvienos klasės laikiniai įterpiniai kaupiami kaip slankieji vidurkiai, sudarant klasinius prototipus. Taikymo etape panašumo skirtumai

tarp normalių ir anomalinių prototipų koreguojami pagal kosinusinį panašumą, siekiant sumažinti klaidingai teigiamų aptikimų skaičių.

Sistema išbandyta naudojant „Kipshidze“ mažmeninės prekybos vagysčių duomenų rinkinį, kurį sudaro 182 vaizdo klipai: 90 normalių ir 92 vagysčių iš parduotuvių atvejai. Visi vaizdo klipai turi klipo lygmens žymas. YOLO asmenų aptikimo statistinis bazinis metodas pasiekia 0,8308 ± 0,0073 ROC-AUC (angl. Receiver Operating Characteristic – Area Under the Curve, imtuvo veikimo charakteristikos kreivės plotas po kreive) reikšmę vaizdo klipo lygmeniu, vertinant pagal tris atsitiktines pradines reikšmes: 41, 42 ir 43. Pridėjus CLIP pagrindu gautus HOI semantinius požymius, rezultatas padidėja iki 0,9405 ± 0,0096, palyginti su 0,8308 ± 0,0073 rezultatu, gautu naudojant tik YOLO asmenų aptikimo statistinį bazinį metodą.

Prototipinis atminties modulis nepadidina pradinės ROC-AUC reikšmės, lyginant su vien GRU laikiniu koduotuvu: vienos atsitiktinės pradinės reikšmės žvalgomuosiuose eksperimentuose, kai pradinė reikšmė lygi 42, gautos reikšmės yra atitinkamai 0,954 ir 0,990. Vis dėlto šis modulis pakeičia veikimo taško balansą: atkūrimo rodiklis padidėja nuo 0,643 iki 0,857, išlaikant 0,800 tikslumą ir pasiekiant F1 (angl. F1 score, tikslumo ir atkūrimo harmoninis vidurkis) reikšmę 0,828. Tai sumažina jautrumą klaidingai neigiamiems rezultatams pasirinktame veikimo slenkstyje. Abliacijos tyrimas patvirtina kiekvieno modulio indėlį. Pašalinus HOI požymius, našumas sumažėja labiausiai, o tai patvirtina aiškaus sąveikos modeliavimo vertę. Pašalinus atminties modulį, sumažėja tikslumas pasirinktame veikimo taške, tačiau ROC-AUC rangavimo kokybė iš esmės nepablogėja.

Žvalgomasis vienos atsitiktinės pradinės reikšmės GRU ir transformerio koduotuvų palyginimas, kai pradinė reikšmė lygi 42, rodo, kad abu modeliai pasiekia vienodą validavimo ROC-AUC reikšmę – 0,9430. Tačiau testavimo rinkinyje GRU modelis pasiekia aukštesnę ROC-AUC reikšmę nei transformerio koduotuvas: atitinkamai 0,9898 ir 0,9133. Tai dera su prielaida, kad rekurentinis indukcinis poslinkis yra tinkamesnis šiame darbe naudotam mažam mokymo rinkiniui, kurį sudaro 130 vaizdo klipų ir 461 langas. Vis dėlto šiai prielaidai patvirtinti ateityje reikėtų atlikti kelių atsitiktinių pradinių reikšmių vertinimą.

Table of Contents

List of Figures

List of tables

List of abbreviations and terms

Abbreviations:

HOI — Human–Object Interaction

CLIP — Contrastive Language–Image Pretraining

GRU — Gated Recurrent Unit

YOLO — You Only Look Once

IoU — Intersection over Union

MIL — Multiple-Instance Learning

ROC-AUC — Area Under the Receiver Operating Characteristic Curve

PR-AUC — Area Under the Precision–Recall Curve

RNN — Recurrent Neural Network

LSTM — Long Short-Term Memory

CNN — Convolutional Neural Network

ViT — Vision Transformer

SAD — Suspicious Activity Detection

AUC — Area Under the Curve

AI — Artificial Intelligence

BERT — Bidirectional Encoder Representations from Transformers

HICO-DET — Humans Interacting with Common Objects, Detection benchmark

MemAE — Memory-augmented Autoencoder

STG-NF — Spatio-Temporal Graph Normalising Flow

TPR / FPR — True Positive Rate / False Positive Rate

AdamW — Adam optimiser with decoupled Weight decay

AMP — Automatic Mixed Precision

BCE — Binary Cross-Entropy

NMS — Non-Maximum Suppression

AP — Average Precision

CCTV — Closed-Circuit Television

Introduction

Project novelty and relevance

Shoplifting among retailers has become a major problem in various regions across the globe and has led to considerable loss of revenue for business organisations. In recent years, it has been reported that retail theft has cost the USA alone over 100 billion dollars in revenue, and the amount is escalating yearly. The same trend is replicated in other parts of the world, and retailers are grappling with rising incidences of shoplifting and fraud.

The low performance of conventional security systems is evidenced by the fact that very few shoplifters are caught. The majority of organisations use CCTV camera systems as well as human operators; however, human operators cannot watch dozens of video feeds at the same time and cannot observe many illegal actions. This creates the necessity for smart surveillance systems that process CCTV video in real time, detect suspicious behaviour that can result in theft, and provide security officers with warnings. An effective electronic shoplifting detection system would assist in minimising losses and allowing human guards to concentrate on genuinely suspicious cases.

Conventional security measures that use CCTV are not particularly strong, especially in advanced retail conditions. Unless contextual awareness is included, simple motion sensors (or rule-based alarm systems, e.g. line-crossing triggers or loitering-time triggers) will yield high false-alarm rates. Suspicious behaviour is not always noticeable and greatly depends on the situation. For example, when a customer conceals an item beneath his or her clothing, he or she may appear to be making a somewhat subtle visual manipulation that can easily be misunderstood as ordinary movement. Conversely, a similar act can be innocent in a different situation, such as picking up an item to inspect it before placing it in a bag. This consequently necessitates contextual and interaction-based reasoning as a means of accurately interpreting behaviour in retail video.

Tracking or the intensity of motion alone is not enough to identify theft and distinguish it from normal shopping behaviour. According to recent advances in computer vision, it has been shown that high-level semantic information, including human pose (skeleton joints) or human-object interactions (HOIs), can substantially enhance the detection of anomalous activities. More precisely, the modelling of HOIs such as a person picking up an item or placing it into a bag, provides fine-grained information regarding how a person interacts with objects which can be directly applied to theft detection.

Simultaneously, video object detection and action recognition have been redesigned with the help of deep learning. Models such as YOLO (You Only Look Once) can identify objects in surveillance cameras in a fast and precise manner, and individuals and objects of interest can be identified using them in real time. Object detectors frequently form the foundation of modern HOI detection networks, and additional specialised layers & branches are applied to detect human-object interactions in each detected human.

However, the majority of earlier work in shoplifting detection using deep learning has approached the task either as a frame-level classification problem or as a basic outlier detection problem using low-level features. For example, some recent studies trained CNN classifiers on single images (examples labelled as normal or suspicious) and achieved high accuracy on small curated datasets. However, the limitation of classifying only static images is that it does not take into consideration the

temporal dynamics of theft behaviour including the sequence in which a person approaches a shelf, picks up items, conceals an item, and leaves the store, and as such, it may not generalise effectively.

What appears to be a normal frame such as a person holding a product, may or may not form part of a theft sequence depending on what follows. Previous studies have also established that temporal behaviour analysis of video sequences can be critical for improving the understanding of suspicious behaviour. This thesis directly addresses this need by modelling the temporal sequences that exist in surveillance video.

This project falls under computer vision and applied artificial intelligence, particularly video anomaly detection for retail security. Its innovation lies in the fusion of zero-shot CLIP-based prompt scoring for human-object interactions with a simple prototype memory for shoplifting detection, which has not been tested on the Kipshidze retail dataset.

The problem addressed in this thesis is clip-level suspicious activity detection in retail surveillance video, with specific emphasis on shoplifting-related behaviour. Such behaviour is difficult to detect, as suspicious behaviour is often subtle, temporally weak, and visually comparable to ordinary shopping behaviour. Practically, it is also difficult to gather large and fully annotated datasets that contain the entire range of shoplifting situations. All these issues render retail suspicious activity detection an appropriate anomaly detection problem in which the system must differentiate rare abnormal behaviour from a significantly larger space of normal activity. The thesis therefore examines the possibility of using human-object interaction semantics and temporal modelling to enhance clip-level discrimination under a consistent evaluation protocol.

Aim and objectives

This research focuses on examining and developing a machine learning approach to identify suspicious behaviour in a retail setting. The system employs human-object interaction (HOI) modelling, temporal reasoning, and memory enhancement to detect possible shoplifting incidents from video footage.

To achieve this aim, the following objectives are defined:

1. Examine the flaws of existing approaches to detect suspicious retail behaviour and explore how AI can be used to improve it.
2. To enhance anomaly detection in retail settings, propose a machine-learning framework that incorporates temporal sequence modelling and human-object interaction signals.
3. Evaluate the effectiveness of the proposed framework using standard performance metrics such as ROC-AUC, Precision, Recall, and F1-score.
4. To increase the model's accuracy and decrease false positives, optimise it by examining evaluation results.
5. Examine how effectively memory-based techniques can improve anomaly detection stability and resilience over time.

Research Questions

The following research questions are addressed in this thesis:

RQ1. Does incorporating human–object interaction semantics significantly improve clip-level anomaly detection performance compared to a baseline that uses only person-detection statistics from a YOLO detector?

RQ2. Does prototype-based memory augmentation improve precision and reduce false-positive sensitivity?

RQ3. How does Transformer-based temporal modelling compare with GRU-based temporal modelling for retail suspicious activity detection?

Research Hypotheses

Based on the above research questions, the following hypotheses are formulated:

H1. Adding CLIP-based HOI semantic features will increase clip-level ROC-AUC by a statistically significant margin compared to the YOLO person-detection statistical baseline.

H2. Memory-based prototype adjustment will improve precision at comparable recall levels.

H3. Transformer-based temporal encoding will achieve equal or higher ROC-AUC compared to GRU-based encoding.

Evaluation Criteria

The models in this thesis are all assessed at the clip level. ROC-AUC is the main measure of evaluation, whereas PR-AUC, precision, recall, and F1-score are auxiliary measures to examine the quality of ranking and operating-point performance. Selection of threshold is done using validation data and final model comparison is done using held-out test data. It is an integrated evaluation protocol where the contribution of every module can be appropriately measured during the ablation study.

Document structure

The thesis consists of three main chapters, conclusions, references and appendices. Chapter 1, Analysis, reviews the literature on video anomaly detection, retail surveillance, human–object interaction recognition, temporal sequence modelling, and memory-based anomaly detection and end with a research-gap summary. Chapter 2, Project, describes the design of the proposed framework, including UML diagrams, dataset and feature preparation, the YOLO-based perception backbone, CLIP-based HOI integration, sequence generation, temporal encoders (GRU and Transformer), and the prototype memory bank. Chapter 3, Implementation, Experiments and Results, presents the implementation environment, experimental setup on the Kipshidze dataset, evaluation metrics, overall detection performance, ablation study (HOI vs. baseline, memory effect, GRU vs. Transformer, and a pretrained-BERT variant), discussion of findings, comparison to related approaches, validity analysis, limitations, and future research directions. The thesis ends with the conclusions and the list of references.

1. Analysis of HOI-Based Anomaly Detection in Retail Surveillance

Over the last decade, studies have concentrated on the development of automated systems to identify unusual & abnormal events using video [20][28]. The current research intersects with several major research themes such as generic video anomaly detection methods, specialised video detectors for identifying retail theft, recognition of human-object interaction, and novelty detection using memory or contrastive learning. In this section, representative previous research in the following areas is examined in order to place the proposed methodology within the context of existing literature.

1.1. Video anomaly detection in surveillance

Early studies in video anomaly detection were based on handcrafted features and specific modelling of regular motion patterns. Classical methods typically tracked persons / objects through a scene and learned regular paths, motion patterns, or activity patterns that were local to particular regions. Events that did not follow these learned patterns such as unusual movement directions or unauthorised access to restricted areas were categorised as anomalous. Despite being an important part of the initial development of surveillance analytics, these techniques frequently failed in complex environments, as normal behaviour can be extremely varied and situational.

Anomaly detection underwent a major transformation with the development of deep learning and data-driven models of normal event[27]. A particularly influential idea involved reconstruction-based and prediction-based methods where neural networks are trained on normal video data and then applied to detect anomalies through high reconstruction or forecasting error. In these approaches, it is assumed that the model will perform well on normal events but will perform poorly when presented with abnormal behaviour[14].

Autoencoder-like methods became especially popular because they provided a straightforward way to model normality without requiring frame-level anomaly annotations. However, standard autoencoders have a crucial drawback such as sufficiently expressive models can become overly sensitive to abnormal inputs which reduce anomaly sensitivity. To address this issue, subsequent studies proposed memory-enhanced architectures such as MemAE, where a small memory bank contains prototype representations of normal patterns. The reconstruction is limited so that the learned prototypes are used only to encode inputs for reconstruction which make the reconstruction of new abnormal events difficult. The idea is to improve the performance of the anomaly detection by maximizing the reconstruction gap between the normal and the actual anomaly data.

Another important anomaly detection line of research is the weakly supervised learning approach, which uses multiple-instance learning (MIL) [2][7]. Here the labels of training are available at the video level but not in frame / segment points. A video that is labeled as anomalous is one that is assumed to have at least one anomalous segment and all the segments in a normal video are assumed to be normal. The formulation then gained particular prominence with the UCF-Crime benchmark, thus making it possible to investigate weakly supervised research of large-scale anomaly detection. MIL-based techniques proved that even with the presence of only coarse video-level supervision, neural networks can be trained to localize temporal anomalies, with subsequent extensions improving robustness to real-world surveillance noise[16].

Contrastive learning to detect anomaly has also been studied more recently. There are pseudo-anomaly generating methods that interfere with the natural temporal order of normal video sequences, and those that attempt to use feature-space margins in order to distinguish between normal and

abnormal embeddings more precisely. The usefulness of these approaches is that they would help models to learn temporal consistency and fine structure of the discriminative element instead of solely depending on the reconstruction error.

In general, the literature demonstrates a comprehensive shift away from hand-crafted motion features and toward weakly-supervised, self-supervised, and representation-based methods for surveillance anomaly detection. Although this has been achieved, a significant number of anomaly detection approaches are still more centered towards anomaly detection based on motion irregularity, reconstruction failure or coarse temporal scoring. More recent studies have also explored contrastive learning and pseudo-anomaly generation[5][8]. They expressly fail to model semantic relationship between the people and objects, which may be decisive in retail surveillance scenarios in which suspicious behavior may be more or less interaction-based than motion-based. This constraint is the reason why more interaction-sensitive representations are used in the current thesis.

1.2. Retail surveillance anomaly detection

Retail surveillance offers an especially challenging environment of anomaly detection due to the subtle nature of suspicious behavior, its situational and visual similarity to typical shopping[9][10]. Although it is not as noticeable as more apparent surveillance aberrations like violence, intrusion or traffic law violations, the behavior that shoplifting can entail may be minor and temporally local, such as concealment, hesitation, repeated shelf contact, or transferring an item to a bag. Consequently, not just motion analysis or fixed frame classification suffices in order to detect retail anomalies. Several researches have examined the application of human pose and skeletal data to detect suspicious shopping activities. Pose based method tries to identify suspicious body shapes, e.g. excessive bending on shelves, a faulty twisting of the torso, or continuous hand to body movement that may signal actions involving concealment. Such methods have a number of benefits, such as decreased sensitivity to variations in appearance, some privacy on the one hand, and resistance to variation in clothing, on the other. However, there is also a significant weakness with them: pose in itself does not expressly say what object is being manipulated, or how the manipulation of the object in question is taking place. Other literature has assumed an object-oriented or image-based classification methodology[15][26]. In such ways, objects detectors are initially employed to locate individuals and other pertinent items in surveillance frames, then a classifier makes a prediction on whether a frame or an image belongs to normal or suspicious activity. These methods show the applicability of the current object detection to retail monitoring, but cannot be applied in isolation, only as auxiliary methods. Since they are based on a large number of specific frames, they can fail to capture the time course of any action and can also give a false positive or a false negative when the behavioral context of interest is not contained in a single picture but spread across many. More recent studies have started to focus on the interaction-conscious retail surveillance, where such an objective is not to identify objects or people, but to comprehend the relationships between them. This orientation is particularly applicable to the shoplifting cases since the suspicions of the behavior may be determined by the semantic meaning of the interaction but not by the presence of an object. To illustrate, grasping a thing, putting it in one of the baskets and hiding it in a bag can imply the same object and can use the same layouts, but such actions are quite different in terms of motive. More informative bases of retail anomaly detection are therefore provided by methods that explicitly model human-object interaction as compared to purely geometric or frame-level methods. Graph-based and interaction-oriented techniques further reinforce this perception by modeling people and objects as being interrelated within a scene. The methods are able to represent more detailed relationships in multi-entity and can

be used to model intricate interactions. They can however also be characterized by increased cost of computation and increased complexity of implementation especially in real time surveillance situations with multiple cameras and congested scenes. Although this is the case, the literature has continuously indicated that effective retail anomaly detection requires knowing who is dealing with what object, by what magnitude, and how, but not just based on appearance, magnitude of movement, and object presence. In general, previous studies on retail and pose-based anomaly detection have noted the importance of moving beyond coarse visual cues to more semantically meaningful representations of behaviour[25]. This fact is the reason why the interaction-aware approach is taken in the current thesis and justifies the necessity of those approaches that integrate perception, temporal reasoning, and an interpretable analysis of human-object interaction. Recent work suggests that interaction-aware and explainable approaches are more suitable for anomaly analysis in general[11].

1.3. Human–object interaction recognition

Human-object interaction (HOI) recognition can be regarded as a valuable research field within computer vision since it allows detailed cognition of how individuals communicate with the objects around them. In contrast to the general action recognition, which generally applies a general label to an overall image or a video fragment, HOI recognition has a narrow focus on the connection between the person, object, and interaction between the two. An example is holding a phone, opening a door, putting something in a bag or carrying something[18]. This makes HOI especially applicable to surveillance analysis, in which suspicious behavior is often not only based on what is present, but also on how it is being manipulated. Another standard in this area is HICO-DET, which includes many categories of objects and interactions verbs, and the combination of all possible human and object interaction triplets. Studies based on this data have given rise to more and more precise HOI models, both instance-based and graph-based, as well as more recently, transformer-based detectors. Numerous classical HOI systems are multi-stream design where human appearance, object appearance and spatial or union-region information are handled together to anticipate the label of interaction[12]. These approaches have demonstrated that recognizing interaction can be enhanced by the combination of features of actors, features of objects and the relational context instead of just using appearance.

Despite this, a common approach to traditional HOI detection is to treat the classification of an interaction as a closed-set problem where the model simply determines one of a pre-defined set of interaction categories. In anomaly detection applications like retail surveillance, anomaly detection may not match exactly with the labels of benchmark interactions. For example, concealment-related actions may visually resemble several everyday interactions while carrying different behavioural significance depending on the context. Consequently, HOI information is often more useful when represented as a semantic description of interaction rather than as a fixed discrete label.

Recent vision-language models provide a more flexible alternative for representing HOI semantics. Instead of forcing interactions into a predefined vocabulary, prompt-based methods allow image segments to be scored against human-readable descriptions of interactions. This is particularly attractive for retail suspicious activity detection because prompts can describe domain-specific behaviours such as placing an item into a bag, concealing an item, removing a tag, or looking around suspiciously. These representations may also be more interpretable as the resulting scores can be associated with behavioural cues that are understandable to humans rather than opaque latent features.

Several modern studies have suggested that interaction-aware reasoning can enhance anomaly detection by making models more event-oriented and interpretable. This is especially relevant in retail environments, where the distinction between ordinary handling and suspicious behaviour often depends on interaction semantics rather than movement or pose alone. More recent HOI models have also adopted transformer-based architectures[13]. Therefore, the literature supports the use of HOI-based representations as a valuable complement to object detection and temporal modelling in suspicious activity analysis. Recent work has additionally examined generalisation issues in HOI benchmarks[17].

1.4. Temporal sequence modeling for video understanding

Video understanding requires temporal sequence modelling because many important events cannot be identified reliably using single frames alone. In surveillance and anomaly detection, the meaning of behaviour often depends on how actions evolve over time rather than on a single visual snapshot. For example, standing in front of a shelf, reaching for an item, and placing it into a bag may not appear suspicious when viewed frame by frame, yet the temporal sequence may indicate malicious intent. This makes temporal modelling particularly important in retail surveillance research[19].

Temporal video analysis has extensively utilised recurrent neural networks because they are capable of summarising time-varying information. More specifically, architectures like Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks have been extensively used in the field of action recognition, video classification, and anomaly detection. It is an attractive aspect of these models that they are computationally economical and can be extended to relatively short and medium length sequences. This allows them to be suitable to surveillance clips, where the time interval within which suspicious events are likely to occur is relatively short and narrow. Transformer-based temporal models have been more recently of importance in video understanding. Transformers can more flexibly capture longer range dependencies and multimodal interactions between frames through self-attention than through only recurrent architectures. This has enabled them to succeed in most sequence modeling problems. Their advantages are however not universal: On smaller datasets, or a small behavioral window, Transformers do not necessarily beat simple recurrent models, and tend to need more computation and delicate hyper-parameter tuning. Temporal models have been found especially useful in anomaly detection, as they enable the system to differentiate between temporal appearances and persistent behavior patterns. A clip can have numerous visually banal frames, and only a brief sequence indicates suspicious action. Sequence modeling thus allows the anomaly detection systems to learn transitions, time consistency, and sequence more efficiently compared to frame-level classifiers. This is why the comparison of recurrent and attention-based temporal encoders is a significant step towards better understanding what type of temporal reasoning is more appropriate in detecting the suspicious activities at a retail location. Transformer-based approaches have recently become important in video anomaly detection. All in all, the literature suggests that video anomaly detection requires a temporal modeling feature which must be considered a design option but not an additional option. This is what triggers the argument between GRU-based and Transformer-based temporal reasoning in the current thesis statement.

1.5. Memory-based anomaly detection

Memory-based approaches have been listed as a significant trend in anomaly detection since it assists in enhancing the separation of classes and stabilized predictions as well as lowering false positives[6]. These methods do not simply use a conventional discriminative classifier, but rather, representative

patterns of normal or abnormal behavior are stored in a memory structure and new samples are compared against these learned references[3]. This enables the decisions of anomaly to be informed by similarity to already learnt prototypes of behavior as well as the immediate feature value. The concept of neural memory became popularized first in wider machine learning methods like question answering and sequence models, in which external memory enabled networks to store long-term memory too long to be stored in simple recurrent hidden states. Memory was subsequently modified in anomaly detection as a way of biasing models to known normal patterns. A significant trend was in memory-augmented autoencoders, in which a small memory bank is used to store prototype representations of normal events. The abnormal inputs are more difficult to be reconstructed in the process of inference since they do not fit well within the stored memory structure. This enhances the sensitivity of anomalies over traditional reconstruction methods. Discriminative settings have also been addressed using prototype-based memory. In these approaches, memory can be regarded as being a collection of representative embeddings or class prototypes that characterize normal and abnormal behaviour in feature space. A new sample will then be tested based on the similarity or distance of that sample to these prototypes. This makes memory-based methods attractive as it is a relatively interpretable form of optimisation of model choice. Rather than perceiving the anomaly detection as an entirely opaque classification paradigm, memory-based scoring enables the model to determine whether a sequence is similar to the regular behavior that it has previously observed or it is different. Memory mechanisms are particularly applicable in video understanding and surveillance analysis since normal behavior may be different in different scenes, and yet follow common patterns. Both the reconstruction-based and discriminative anomaly detection have exploited memory-augmented models[22][4]. These regularities can be maintained with the help of a memory module and enhance robustness in the case of clips being near to the decision boundary. This applies especially in suspicious activity detection in which certain acts might just be a bit off the regular shopping practice. Memory-based methods can enhance both the calibration and threshold stability by providing a reference structure during embedding-space, particularly in marginal cases. Altogether, the literature indicates that memory-based anomaly detection can be useful in terms of both performance enhancement and more reliable and explainable predictions. This renders prototype memory a significant element to put into consideration in a modular suspicious activity detection system.

1.6. Research gap summary

The reviewed literature demonstrates that video anomaly detection, retail surveillance analysis, human-object interaction recognition, temporal sequence modelling, and memory-based methods have all contributed significantly to surveillance research. Classical anomaly detection methods established the foundation for abnormal event recognition, while more recent deep learning approaches improved representation quality and enabled large-scale weakly supervised learning using video-level features.

Retail surveillance studies have further highlighted that suspicious behaviour in supermarkets is highly nuanced, situation-specific, and difficult to separate from normal shopping behaviour using coarse motion or appearance features alone. At the same time, HOI research has shown that fine-grained interaction modelling provides deeper behavioural understanding than object detection / pose estimation alone. Temporal modelling studies have demonstrated that video understanding depends heavily on sequence context, while memory-based models have shown that reference patterns & prototypes can improve consistency and decision quality during anomaly detection.

However, these research directions are frequently investigated independently, and comparatively little attention has been devoted to integrating and analysing them within a unified retail surveillance pipeline. In particular, very few studies jointly investigate detector-level perception statistics, semantic HOI reasoning, temporal sequence modelling, and prototype-based memory refinement within a single clip-level shoplifting anomaly-detection framework. Existing approaches often focus on only one aspect such as pose, object presence, frame-level classification, or generic anomaly scoring, without quantifying the role of interaction semantics and temporal reasoning in suspicious activity detection for retail environments.

Therefore, the primary research gap addressed by this thesis is the lack of modular, interaction-aware clip-level frameworks for suspicious activity recognition in supermarket surveillance footage. This gap is addressed by quantitatively evaluating the contribution of:

(i) the YOLO person-detection statistical baseline,

(ii) CLIP-based HOI semantic features,

(iii) recurrent and attention-based temporal encoders, and

(iv) prototype memory augmentation,

under a unified clip-level evaluation protocol on the Kipshidze dataset. Existing studies often explore anomaly detection, HOI, temporal modelling, and memory mechanisms as separate research directions[23].

2. Project of an HOI-Based Suspicious Activity Detection Framework

The suspicious-activity detection framework proposed is based on a modular clip-level pipeline. Each video clip is analysed frame by frame such as a pretrained YOLO detector localises persons and the surrounding objects, and three statistical descriptors (number of persons, mean person-detection confidence, and its standard deviation) constitute the statistical baseline of YOLO person-detection. Parallel to this, a HOI module based on CLIP generates semantic interaction scores on the global frame, the largest-person crop and a fixed bank of natural-language prompts. The features of the HOI are concatenated with the baseline descriptor to form an interaction-aware per-frame representation, which is in turn grouped into overlapping temporal windows and fed into a temporal encoder. Two encoders are compared; a Gated Recurrent Unit (GRU) and a Transformer; and a prototype-based memory module is added at the embedding level to stabilise predictions and reduce false positives. By max-pooling window-level predictions, a single clip-level anomaly score is obtained. This modular form enables the contributions of the perception statistics, semantic interaction modelling, temporal reasoning and memory refinement to be analysed separately.

The following subsections describe each module in detail.

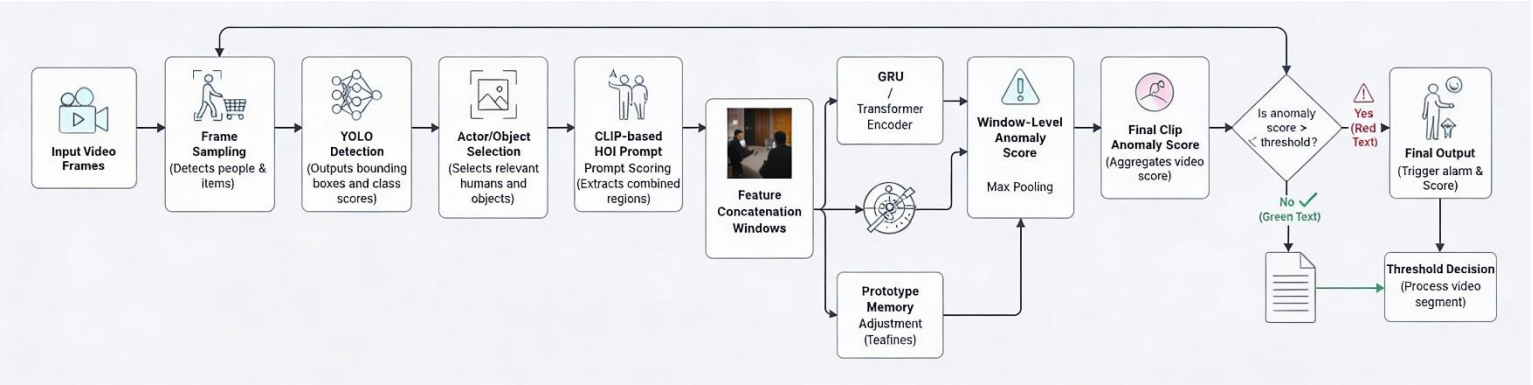

Fig. 1. Overall Pipeline Of The Proposed Suspicious Activity Detection Framework

2.1. UML Diagrams

Use Case Diagram:

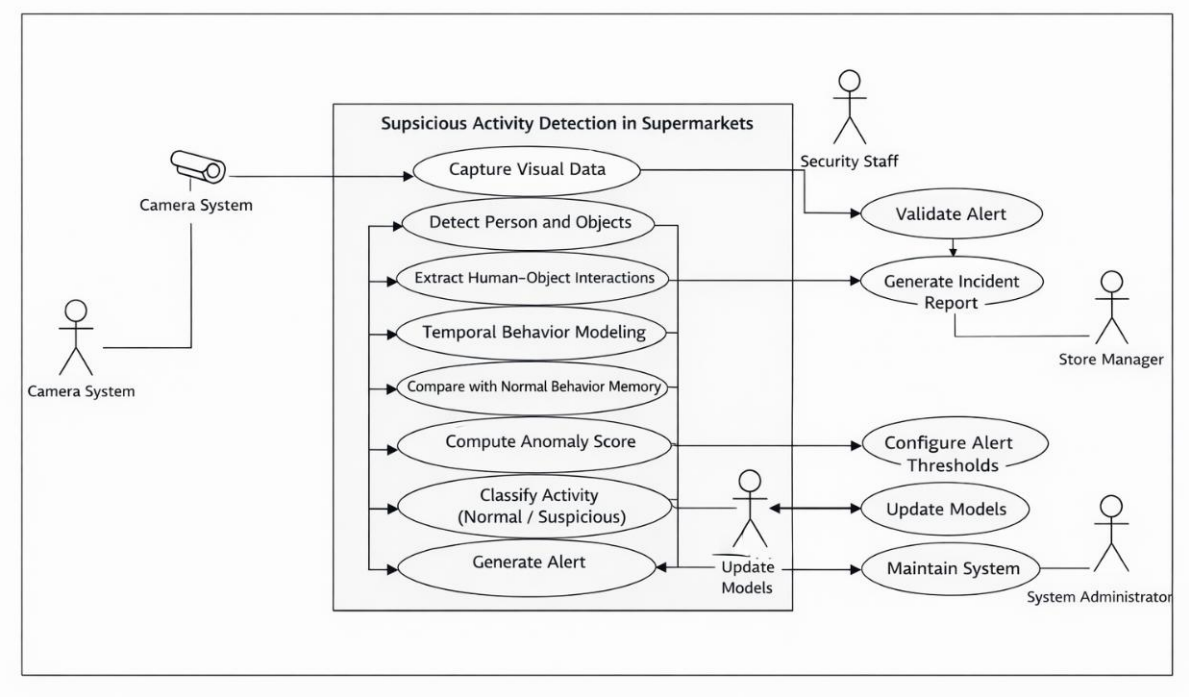

Fig. 2 Use Case Diagram

This use case Figure 2 provides an overview of the interactions between various actors and the Suspicious Activity Detection in Supermarkets system and how the responsibilities are allocated between automated analysis and human oversight. The Camera System serves as the main source of data, sensing visual data that is processed by the core system. Inside the system boundary, the pipeline is responsible for automatically performing person and object detection, extracting human-object interactions, modelling temporal behaviour, comparing observed patterns with a memory of normal behaviour, computing an anomaly score, classifying activities as normal or suspicious, and generating alerts if necessary.

The Security Staff receive alerts and have the responsibility of validating alerts and initiating incident reporting so that false positives are filtered before being escalated. The generated reports are used by the Store Manager for operational & legal follow up. The System Administrator maintains the system such as by setting alert thresholds, updating models and ensuring reliable operation over time, which illustrates the system's adaptability and its human-in-the-loop design.

Sequence Diagram:

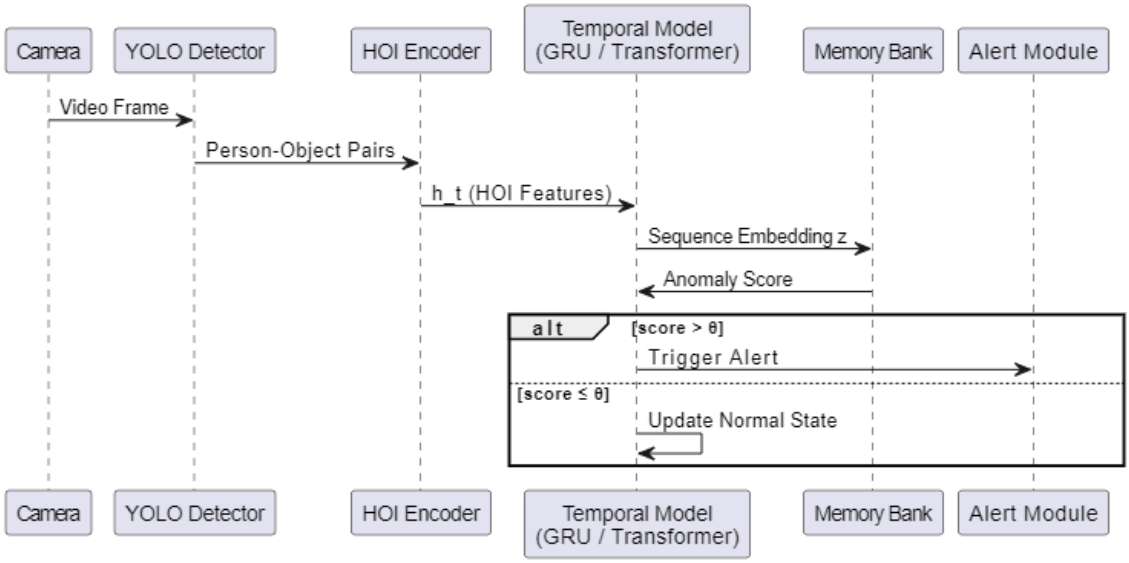

Fig. 3 Sequence Diagram

This Figure 3 sequence diagram shows real-time interaction between the parts of the system during anomaly detection. Video frames obtained from the camera are fed to the YOLO detector to detect persons and objects. Detected person-object pairs are fed into the HOI encoder to retrieve interaction features, which are aggregated temporally with a GRU or Transformer. The resulting sequence embedding is compared against a memory bank of normal behaviour patterns in order to calculate an anomaly score. If the score is above a predefined threshold, it triggers an alert and sends it to the alert module.

Flow Chart:

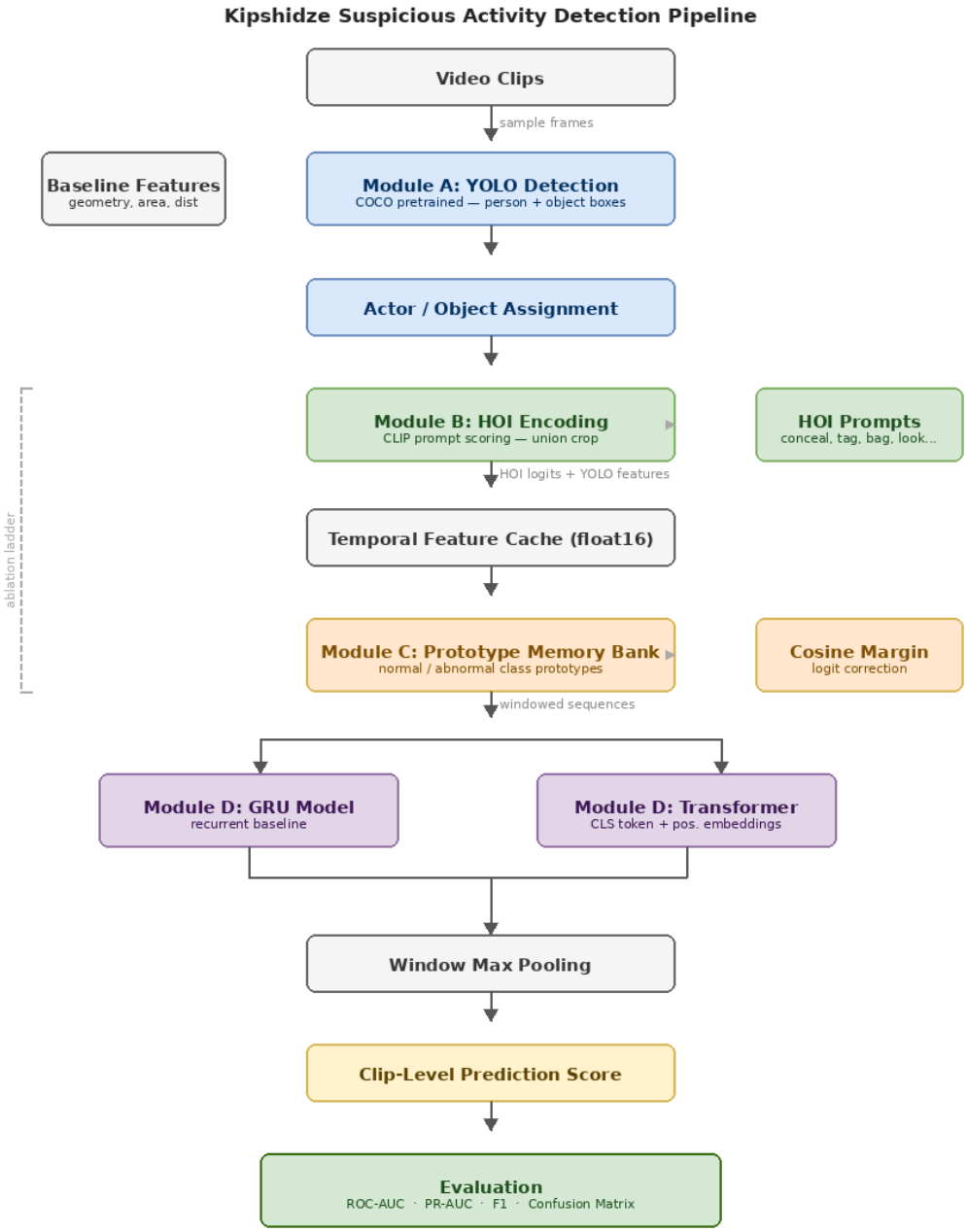

Fig. 4 System Flow Chart

This Figure 4 flowchart shows the end-to-end pipeline of identifying suspicious activity from videos captured in supermarkets. Video frames are processed using YOLO for person and object detection, followed by HOI feature extraction and temporal sequence modelling with either a GRU or a Transformer. The resulting sequence embedding is compared against a memory bank of normal behaviour in order to compute an anomaly score. If the score exceeds a predefined threshold, the activity is flagged as suspicious; otherwise, it is treated as normal.

2.2. Preparation of datasets and features

The primary data used in this thesis is the Kipshidze dataset which consists of a total of 182 video clips made up of 90 normal clips and 92 shoplifting clips. The videos are captured at 30 FPS and the mean length is approximately 11.1 seconds. The most widely used resolutions are 640 × 480, although some higher-resolution clips are also present. These statistics confirm that the dataset consists of relatively short retail surveillance clips in which suspicious behaviour is often localised to a brief temporal segment.

```
Readable videos: 182

Class counts (inferred):
label
1    92
0    90
Name: count, dtype: int64

duration_s stats:
count    182.000000
mean      11.117582
std        1.396785
min        5.233333
25%       10.608333
50%       10.933333
75%       11.533333
max       20.666667
Name: duration_s, dtype: float64

fps stats:
count    182.0
mean      30.0
std        0.0
min       30.0
25%       30.0
50%       30.0
75%       30.0
max       30.0
Name: fps, dtype: float64
```

Fig. 5 Dataset Feature and details

Video Input and Frame Sampling

Frame Extraction: Each video was sampled at an effective rate of 5 frames per second. With the initial recording speed of 30 FPS, this involved extracting every 6th frame. No more than 900 frames were extracted from each video to reduce the memory used in feature extraction.

Temporal Window Construction: Sequential features were grouped into overlapping time windows using the following parameters:

- Window length (L) - 32 frames.
- Stride (S) - 8 frames.
- Frame minimum length - 16 frames.

Clips that produced fewer than 16 frames following preprocessing were excluded from analysis. The result of this windowing strategy was:

- Training windows - 461
- Validation windows - 87
- Test windows - 97

Feature Normalization: All HOI characteristics and YOLO individual-detection rates were normalised using z-score normalisation. More importantly, the normalisation statistics (the mean of the data dimensions and their standard deviation) were computed only on the training split and then applied identically to the validation and test splits to avoid leakage of distributional information into held-out clips.

Person and Object Detection

YOLOv11-nano (yolo11n.pt), a lightweight real-time COCO-pretrained detector was used as a pretrained object detector to detect people in real time based on the COCO dataset. The detector was configured with the following parameters:

Table 1 YOLO Detection Parameters

Parameter	Value
Model variant	YOLOv11-nano
Input resolution	640×640
Confidence threshold	0.35
NMS IoU threshold	0.5
Target class	"person"

Only detections classified as part of the person class (COCO class ID 0) were retained for further feature extraction. The confidence threshold of 0.35 was selected to balance detection recall and false positives in cluttered retail environments.

Baseline Person-Detection Statistical Features

A small three dimensional baseline feature vector per frame is computed using the output of the YOLO detector. The vector is defined as (i) the count of individuals in the frame (n_persons), (ii) the mean confidence of all detected persons in the frame (mean_conf), and (iii) the standard deviation of the detection confidence of all persons in the frame (std_conf). These three statistics constitute the person-detection statistical baseline of the ablation study based on YOLO person-detection statistics. They encode only the presence and detection quality of persons in the scene and do not encode actor-

object spatial relationships, bounding-box geometry, or semantic interaction. Richer interaction information is added by the CLIP-based HOI module described in the next subsection.

CLIP-Based HOI Feature Extraction

A CLIP-based HOI encoder is used to add semantic interaction information on a frame-by-frame basis. The encoder uses the open_clip ViT-B/32 model with OpenAI-pretrained weights and produces a 1047-dimensional HOI vector composed of five concatenated components: (1) a 512-dimensional CLIP embedding of the full frame, capturing scene context; (2) a 512-dimensional CLIP embedding of the largest detected person crop, padded by 10% of the box dimensions (the full frame is used as a fallback when no person is detected); (3) a 5-dimensional spatial-relation vector [cx_norm, cy_norm, w_norm, h_norm, area_norm] describing the normalised position and size of the primary actor; (4) a 17-dimensional vector of cosine similarities between the global CLIP image embedding and a curated prompt bank. The 17-prompt bank is divided into 8 shoplifting prompts (e.g. 'a person concealing merchandise under clothing', 'a person hiding items in a bag', 'a person stuffing items into pockets'), 5 normal-behaviour prompts (e.g. 'a person shopping normally in a store', 'a person browsing items on a shelf'), and 4 object-context prompts (e.g. 'a handbag or backpack near store shelves', 'merchandise items in a person's hand'). The complete list is provided in Section 2.4.2. The entire HOI vector is then appended to the 3-dimensional YOLO baseline vector to obtain a 1050-dimensional per-frame feature representation which is stored on disk in float16 format.

Normalization and Preprocessing

Prior to temporal modelling, feature values were normalised in order to reduce scale sensitivity and prevent high-magnitude variables from overshadowing the sequence representation. Feature normalisation statistics were calculated using the training split only, and bounding-box coordinates were calculated relative to image dimensions. The same statistics were then applied to validation and test data in order to prevent leakage. The combination of feature caching and normalisation enhance reproducibility, reduce repeated inference on GPUs, and enable valid comparison between all variants of the model.

2.3. YOLO-based baseline perception

This application of YOLO[24] functions merely as a perception front-end. It provides bounding boxes, class labels, and confidence scores for downstream feature extraction. YOLOv11-nano (yolo11n.pt) was selected because it offers fast inference and good object-detection performance on a single-GPU platform, and it was not fine-tuned on the Kipshidze data. The rationale behind this design decision was to maintain a frozen, COCO-pretrained detector that keeps the perception backbone modular and avoids overfitting to the particular Kipshidze viewpoint, while still reliably localising persons (the actors whose interactions the HOI module reasons about). These detections yield frame-level results detailed in Section 2.2.3, while the CLIP-based HOI module that uses the same person bounding boxes is described in Section 2.4.

For comparison purposes, separate classification-style baselines were also trained: YOLOv11-Large and YOLOv26-Large classification heads (yolo11l-cls.pt, yolo26l-cls.pt), trained at 224×224 input resolution over 50 epochs. These are not part of the main pipeline; they appear as supplementary reference results in Section 3.4.

2.4. CLIP-based human–object interaction (HOI) integration

Person-detection statistics of YOLO indicate the number of people present and the confidence with which they are detected, but not what those people are doing. In retail surveillance, suspicious activity cannot be defined in terms of the number of detections / confidence values; the discriminative signal lies in the interaction between a person and the objects around them. For example, a detector output corresponding to holding an item in one hand or placing that same item into a bag is almost identical, yet the behaviour represented is very different. To introduce this semantic layer, a HOI module based on CLIP is inserted into the pipeline. The module uses the pretrained CLIP ViT-B/32 model (open_clip, OpenAI weights) to score images against natural-language descriptions in a shared embedding space. This permits flexible, prompt-based, and task-free interaction modelling.

Rationale of Semantic Interaction Modeling

Classic video anomaly detection methods usually depend on Motion magnitude, Optical flow, Raw CNN appearance features, and bounding box geometry. However, this type of representation is insufficient to identify subtle behavioural differences in retail situations. Low-level motion patterns share significant similarity across many normal and suspicious activities. The discriminative factor is interaction intent, including:

- Placing an item into a bag.
- Hiding an item under a garment.
- Removing a security tag.
- Looking around suspiciously while concealing an item.

These actions are characterised by structured human-object relations rather than simple movement. Thus, interaction-sensitive representation is required to reflect higher-level semantic cues[1].

Representation of CLIP-Based HOI

In order to simulate interaction semantics, a trained CLIP (Contrastive Language–Image Pretraining) model, specifically the CLIP ViT-B/32 (OpenAI weights, open_clip) checkpoint loaded through the open_clip library, was incorporated into the pipeline. CLIP is trained on a shared embedding space where semantically similar images and textual descriptions are located close to each other. This enables semantic similarity between an image and a textual prompt to be calculated using cosine similarity. The process applied to each frame is as follows:

- The largest detected person (by bounding-box area) is selected as the primary actor.
- The person bounding box is expanded by 10% padding on each side to form a person crop; if no person is detected, the full frame is used as a fallback.
- Both the full frame and the person crop are resized and fed independently to the CLIP image encoder, producing two 512-dimensional embeddings.

The complete list of 17 prompts is given below (8 shoplifting + 5 normal + 4 object-context):

Shoplifting prompts (8):

1. a person concealing merchandise under clothing

2. a person hiding items in a bag

3. a person looking around suspiciously while holding items

4. a person quickly grabbing items off a shelf

5. a person removing security tags from products

6. a person stuffing items into pockets

7. a person crouching near store shelves suspiciously

8. a person acting nervously in a retail store

Normal-behaviour prompts (5):

9. a person shopping normally in a store

10. a person browsing items on a shelf

11. a person putting items in a shopping cart

12. a person reading a product label

13. a person walking through a store aisle

Object-context prompts (4):

14. a handbag or backpack near store shelves

15. merchandise items in a person's hand

16. a shopping basket held by a person

17. clothing covering hidden items

The similarity between the image embedding and the prompt embedding is then calculated using cosine similarity. The resulting similarity scores constitute a fixed-length semantic feature array that represents the interaction within the frame.

Feature construction and integration

The HOI similarity scores provide a semantic interpretation of the interaction at every time step. These scores are then combined with the 3-dimensional YOLO person-detection statistical features (Section 2.2.3) to create the final frame-level feature:

$$\textit{Frame Feature} = \textit{YOLO Person-Detection Statistics} \oplus \textit{CLIP-HOI Semantic Feature}$$

A combination of these representations illustrates both:

- spatial structure (object positions and relationships), and
- semantic interaction cues (behavioural meaning).

Benefits of CLIP-Based HOI Modeling

The selected HOI strategy has a number of strengths:

1. No Special Task Training. CLIP was pretrained on large-scale image-text pairs and therefore does not require the collection of labelled HOI data.
2. Flexible Prompt Engineering. Interaction categories can be modified or extended using textual prompts without retraining.
3. Interpretability. Since the feature dimensions are associated with specific textual prompts, anomaly decisions can be interpreted through semantic interaction descriptions.

Role in the Overall Pipeline

The HOI component transforms low-level person-detection statistics into a semantically rich representation. It provides an intermediate level of reasoning between object recognition and temporal prediction. Without HOI, the model relies solely on spatial cues. With HOI, interaction-aware sequences are input into the temporal encoder in a manner that more accurately reflects behavioural intent. The empirical findings (discussed in Section 3) show that the addition of HOI characteristics leads to a significant improvement in the performance of clip-level anomaly detectors compared with the YOLO-only baseline.

2.5. Sequence Generation from Frame Features

Every video clip is modelled as a temporal sequence of frame-level feature vectors generated by combining the YOLO-generated person-detection statistical features and the CLIP-generated HOI semantic scores from Sections 2.3 and 2.4. Let

$$f_1, f_2, \dots, f_T,$$

represent the sequence of features in a clip, Where $f_t \in \mathbb{R}^d$ is the frame-level feature representation at time step t, T is the number of sampled frames in the clip, and d is the total feature dimensionality after concatenating person-detection statistical and CLIP-HOI semantic features.

Since the Kipshidze clips are relatively short and suspicious behaviour may arise only within a limited segment of a video, the entire feature sequence is subdivided into fixed-length overlapping temporal windows. The temporal model operates on each window to generate a window-based anomaly probability. This sliding-window approach has three primary benefits such as it imposes a fixed input length on the temporal encoders, enables the analysis of suspicious regions at a localised scale, and is robust to variations in clip length between dataset samples.

Formally, a temporal window may be written as

$$X^{(k)} = \left(f_{t_1}, f_{t_2}, \dots, f_{t_L}\right),$$

where $L$is the window length and k indexes the window. Consecutive windows overlap according to a predefined stride, ensuring smoother temporal coverage of each clip. If a window extends beyond the end of a clip, zero-padding is applied so that all windows retain the same length.

Prior to temporal modelling, every frame-level feature vector is normalised in order to minimise scale imbalance between the YOLO person-detection statistics and the CLIP prompt-score features. This

normalisation enhances training stability and prevents features with higher numerical values from dominating the temporal representation.

During training, all sampled windows from a clip are assigned the clip-level ground-truth label, with normal clips assigned label 0 and shoplifting clips assigned label 1. During inference, each window generates an anomaly probability p_w and the final clip-level anomaly score is obtained through max pooling of all window scores:

$$s_{clip} = \max_{\omega \epsilon \mathcal{W}} p_{\omega},$$

where $\mathcal{W}$ denotes the set of all temporal windows in the clip. This aggregation strategy is appropriate for retail suspicious activity detection because even a short concealment-related action within an otherwise normal clip is sufficient to make the overall clip anomalous.

In the implemented pipeline, videos were sampled at an effective rate of 5 frames per second in order to reduce redundancy while preserving behavioural dynamics. Temporal modelling was performed using fixed-length sliding windows of 32 frames with a stride of 8 frames. Clips shorter than 16 sampled frames were excluded from temporal modelling, while shorter trailing windows were zero-padded to preserve a constant input length.

```
[Auto Device Assignment]
  Detector (YOLO boxes): cuda:0
  HOI encoder (CLIP):    cuda:1
  Temporal training:     cuda:0
  YOLO imgsz:            640
  Batch size:            32
{
  "kip_root": "kaggle_kipshidze",
  "normal_dir": "normal",
  "anomaly_dir": "shoplifting",
  "yolo_hoi_weights": "yolo11l.pt",
  "yolo_imgsz": 640,
  "yolo_conf": 0.2,
  "yolo_iou": 0.5,
  "sample_fps": 5,
  "max_frames_per_clip": 900,
  "seq_len": 32,
  "stride": 8,
  "min_seq_len": 16,
  "out_dir": "./runs_kipsh_yolo_cliphoi_safe",
  "cache_base": "./runs_kipsh_yolo_cliphoi_safe/cache_base",
  "cache_hoi": "./runs_kipsh_yolo_cliphoi_safe/cache_hoi",
  "overwrite_cache": true,
  "store_dtype": "float16",
  "yolo_device": "cuda:0",
  "hoi_device": "cuda:1",
  "train_device": "cuda:0",
  "amp": true,
```

Fig. 6 Frame Feature and Window Size

2.6. Temporal Encoders: GRU and Transformer

GRU and Transformer are temporal encoders. Two alternative sequence-modelling architectures were trained and tested to model the temporal dynamics of the per-frame feature sequence constructed in Section 2.5: a Gated Recurrent Unit (GRU) network used as the recurrent baseline, and a Transformer encoder used as the attention-based alternative. Both temporal models take the same sequence of windowed features based on the already-cached YOLO + HOI representation and allow

a fair and controlled comparison of recurrence-based versus attention-based temporal reasoning in an otherwise identical pipeline[19].

The GRU temporal encoder is a frame-to-frame encoder in which at every time-step, the encoder updates its hidden state based on the current frame.

$$z_t = GRU(f_t, z_{t-1})$$

Where f_t represents a concatenation of YOLO and HOI feature vectors at time *t,* and the initial hidden state z_0 is set to a zero vector. The last hidden state z_t is used as the temporal embedding at the window level and is fed through a small fully connected network to generate a single anomaly logit which is transformed into an anomaly probability using a sigmoid activation function. The GRU architecture offers several advantages such as lightweight training suitable for the relatively small Kipshidze dataset, online incremental processing for real-time deployment, and natural regularisation due to the lower number of parameters and inductive bias towards local temporal patterns which minimises the risk of overfitting.

The Transformer-based temporal encoder takes the sequence of per-frame feature vectors as a token sequence and appends to the input a trainable CLS (classification) token that serves as the sequence-level summary, learnable positional embeddings to encode frame order because self-attention is permutation-invariant, four stacked standard (pre-norm) Transformer encoder blocks consisting of eight attention heads and a feed-forward expansion factor of four, and a memory-enhanced output head, which complements the prototype memory bank described in Section 2.7.

The self-attention layers of the Transformer enable the model to rank the importance of each frame's features relative to other frames during sequence representation construction. This allows the model to assign greater attention weights to critical frames such as concealment events, even when they occupy only a small portion of the total window duration. It also captures long-range temporal dependencies such as the association between an initial pickup action and subsequent concealment, which the GRU may forget in long sequences [19].

Both models were trained using the same binary cross-entropy loss function, AdamW optimiser with an initial learning rate of 3e-4, weight decay of 1e-4, dropout regularisation probability of 0.2 on the Transformer, and best-validation-checkpoint selection based on ROC-AUC. This ensured that any observed performance differences could be attributed to the temporal modelling mechanism itself rather than hidden differences in data preprocessing or training procedures.

In each encoder, the output window-level temporal embedding is used in two ways such as input to a feed-forward classification head to generate an anomaly probability through direct classification, and as a distance-based comparison with the prototype memory bank to generate an additional anomaly-scoring signal that combines with the discriminative output to produce clip-level scores through the max-pooling strategy described in Section 2.5.

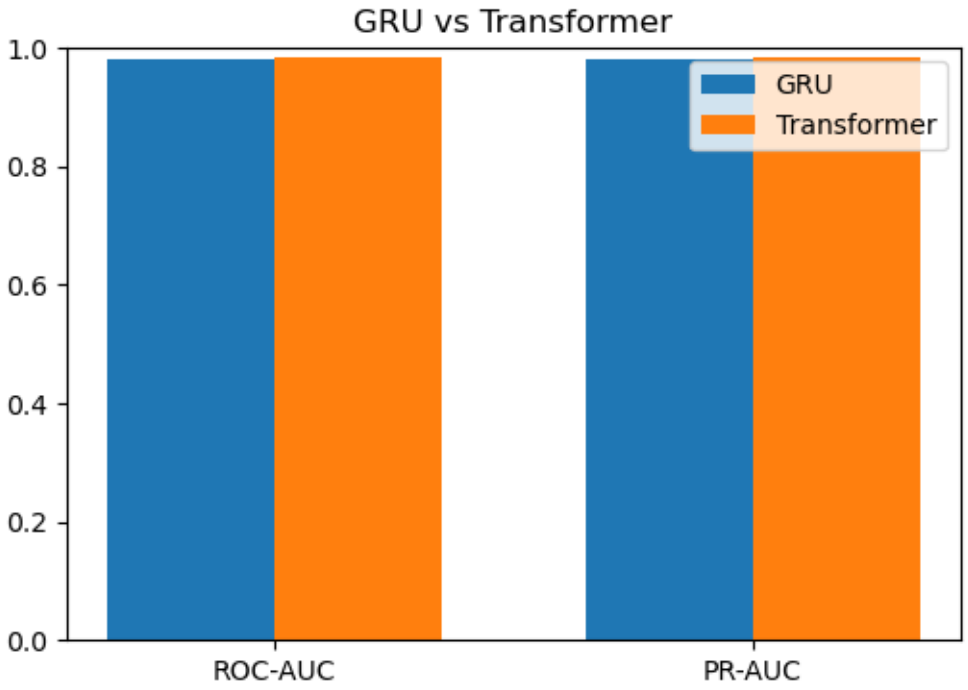

Fig. 7 GRU vs Transformer

Table 2 Temporal Model Hyperparameters

Parameter	Value
Hidden dimension	256
Number of layers	2
Dropout rate	0.2
Batch size	32
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Maximum epochs	20
Loss function	Binary Cross-Entropy
Mixed precision	Enabled (FP16)
Optimizer	AdamW

2.7. Prototype memory bank

In order to enhance the stability of the classification and minimise the false positives, a prototype-based memory mechanism was added to the stage of temporal modelling. The predictions of the GRU and Transformer encoders can become unstable in borderline or geometrically ambiguous cases, even though each of them learns discriminative temporal encodings. The numerous normal behaviours that occur during retail surveillance (such as spending a long time browsing, touching an object briefly, or temporarily lingering around shelves) can be suspicious when considered only through time. On this account, another embedding-space reference mechanism was introduced to support refinement on the basis of distribution awareness. The memory bank includes prototype vectors in the temporal embedding space of classes. Let

$$z \in \mathbb{R}^d$$

denote the temporal embedding of a window produced by the GRU or Transformer encoder. Two prototype vectors are maintained:

$$m_{normal}, m_{abdnormal} \in \mathbb{R}^d$$

where each prototype represents the central tendency of its corresponding behavioural class.

Prototype Update Mechanism

During training, the prototypes are updated using a running-average formulation. For a training embedding z belonging to class c, the corresponding prototype is updated as:

$$m_c \leftarrow \lambda m_c + (1 - \lambda) z$$

Where $\lambda \in [0,1)$ is a momentum parameter which governs the smoothness of the update. This process enables the prototype to gradually approach the mean embedding of its class, without becoming unstable due to individual samples. In contrast to memory banks based on clustering, this method does not require predefining a number of clusters or using an optimisation process. Rather, it provides a lightweight but effective characterisation of class distributions.

Memory-Based Anomaly Scoring

At inference time, the similarity of the test embedding z is computed with respect to each prototype using cosine similarity:

$$S_{normal} = \cos(z, m_{normal}), S_{abnormal} = \cos(z, m_{abnormal})$$

The margin of similarity allows the adjustment of the anomaly score:

$$\Delta S = S_{abdnormal} - S_{normal}$$

If ΔS is positive and sufficiently large, the window is considered suspicious; otherwise, it is regarded as normal. This introduces a distribution-aware correction layer over the temporal classifier.

Role of memory in the overall framework

The memory bank is not intended to replace the temporal encoder. Instead, it complements temporal modeling by introducing a compact representation of normal and abnormal behavior distributions in the embedding space. This is especially useful when the temporal classifier encounters borderline cases that are difficult to separate using discriminative learning alone.

In practice, the memory module enhances robustness by reducing the sensitivity to small changes in temporal predictions and the consistency of the embeddings by class. It is therefore a distribution-sensitive regularisation layer which enhances accuracy and stabilises threshold behaviour. Memory augmentation will assist in developing a more balanced profile of anomaly detection particularly in the precision consistency and false-positive control as will be shown later in the results chapter.

Interpretability of memory prototypes

Another advantage of the memory mechanism is interpretability. Since each prototype represents a summary of a region of the temporal embedding space, it is possible to inspect which sequences lie closest to the learned prototypes. In this way, the memory bank provides a meaningful behavioural reference frame rather than acting as a black-box scoring mechanism. This makes it easier to interpret whether a new sequence resembles known normal behaviour or deviates towards abnormal patterns.

Such interpretability is valuable in surveillance applications where it is important not only to detect suspicious activity but also to understand why a sequence has been classified as unusual.

2.8. Reproducibility and implementation details

A uniform and controlled pipeline was used in all experiments in this thesis to ensure reproducibility and fair comparison between different model configurations. The perception stage such as YOLO-based detection and CLIP-based HOI feature extraction was executed once so that frame-level feature representations could be stored in disk cache. This ensures that all temporal models operate on identical input data, removing variation caused by repeated detection & feature extraction. The same dataset splits, feature representations and preprocessing procedures were used to train all temporal models, including the GRU encoder and Transformer encoder.

The Kipshidze dataset was used as the primary benchmark, with clip-level labels of normal and shoplifting cases. Binary cross-entropy loss and the AdamW optimiser with a fixed learning rate were used for training. The model checkpoint with the best ROC-AUC at each epoch was retained for final evaluation. In the GRU vs Transformer comparison, early stopping with a patience of 15 epochs was also applied. Evaluation was performed using clip-level metrics such as ROC-AUC, PR-AUC, precision, recall, and F1-score. Max pooling over temporal window predictions was used to obtain final clip-level anomaly scores. Threshold calibration was performed on the validation set and then applied to the held-out test set to ensure unbiased evaluation.

This framework provides a reasonable comparison of various model parts such as feature extraction pipelines, evaluation protocol, and perception backbone, by fixing (holding constant) the perception backbone, feature extraction pipeline, and evaluation protocol. As a result, variations in performance reported in the results chapter can be directly attributed to architectural differences rather than differences in preprocessing / experimental setup.

3. Implementation, Experiments and Results on the Kipshidze Dataset

The proposed system was evaluated exclusively on the Kipshidze Shoplifting Dataset, a publicly available retail surveillance anomaly detection benchmark. Unless otherwise mentioned, the major experimental findings presented in this thesis are the results for the multi-seed GRU + HOI architecture tested for random seeds 41, 42 and 43. For the purpose of exploratory evaluation of transformer, memory-bank, and bert model, the results of these experiments are shown as single-seed results (seed = 42) and are analysed separately from the main benchmark experiments. This section describes the dataset, implementation setup, training procedure, and evaluation protocol.

3.1. Implementation Environment

All experiments were conducted using PyTorch 2.x on NVIDIA H100 NVL GPUs with Automatic Mixed Precision (FP16). Object detection was implemented using Ultralytics YOLOv11-nano (yolo11n.pt); the CLIP backbone used open_clip ViT-B/32 with OpenAI-pretrained weights; and the pretrained language model was bert-base-uncased from Hugging Face. AdamW (initial learning rate 3×10^{-4}, weight decay 1×10^{-4}) with binary cross-entropy loss, batch size 32, and dropout 0.2 was used for optimisation. The BERT variant used a reduced learning rate of 1×10^{-4} to preserve pretrained weights.

Multi-seed experiments (baseline and YOLO + HOI) used random seeds (41, 42, 43). Memory, GRU vs Transformer, and BERT comparisons were conducted as single-seed exploratory runs (seed = 42). Frame-level features and YOLO detections were computed in a single pass and stored in float16 format to enable architecture-agnostic training. Each component is described in detail in Chapter 2 and is not repeated here.

3.2. Datasets

Kipshidze Shoplifting Dataset: The dataset contains 182 video clips, including 90 non-theft (normal) clips and 92 shoplifting clips. Each clip represents a complete behavioural episode such as each video captures a full interaction sequence. Clip durations range from approximately 5.2 to 20.7 seconds, with a frame rate of 30 FPS and a resolution of 640 × 480 pixels.

The videos are recorded from an overhead surveillance perspective and capture customers and surrounding shelves. Shoplifting behaviours typically include picking up and examining items, concealing items in bags or clothing, and walking away. Normal clips include similar browsing behaviours without concealment / theft. The similarity between normal and suspicious actions makes the dataset suitable for evaluating interaction-aware temporal modelling.

The dataset provides clip-level labels (normal or shoplifting), which were directly used for supervised training of temporal models. No frame-level annotations were required. Since clips are pre-segmented into behavioural episodes, no additional temporal trimming was necessary.

The dataset was split into training (130 clips, 71.4%), validation (24 clips, 13.2%), and test (28 clips, 15.4%) sets using stratified random sampling with a fixed seed of 42. The detailed counts are given in Table 4.

Role in the Framework:

The Kipshidze dataset is a:

Table 3 Summary of Kipshidze dataset used in the experiments

Dataset	Scenario	Modality	Scale	Annotations	Role in Framework
Kipshidze (Retail, Kaggle)	Retail surveillance	Video	182 videos	Normal / Suspicious (Clip-Level Only)	Primary dataset for detector training and temporal behavior modeling

Dataset Splitting and Experimental Protocol: The 182 video clips were split into training, validation, and test sets using stratified random sampling to ensure class balance across all splits. The splitting procedure used scikit-learns train-test-split function with the following nested approach:

1. Primary division: 15 percent of data set to be used as a test (stratified by label).
2. Second split: 15% of the remaining data to be used in validation (stratified by label)

This gave the following distribution:

Table 4 Dataset Split Statistics

Split	Total	Normal	Shoplifting	Percentage
Training	130	64	66	71.4%
Validation	24	12	12	13.2%
Test	28	14	14	15.4%
Total	182	90	92	100%

All data splitting operations were done with a fixed random seed of 42 to guarantee. reproducibility. In order to measure the stability of the results and calculate confidence intervals, All experiments were replicated in 3 independent runs with seeds 41, 42. and 43 to initialize model weight. The results are presented in the form of mean standard. variation between these three runs.

3.3. Evaluation Metrics

The main measure of evaluation that we use to determine the performance of the anomaly detection is the Area Under the ROC Curve (AUC). ROC curve represents the true positive rate (TPR) against false positive rate (FPR) by changing the decision threshold. AUC is a threshold-free measure of how well a model ranks anomalous clips above normal ones. A value of 1.0 means perfect separation; 0.5 means random guessing. We report AUC on the test set of the Kipshidze dataset.

Besides AUC, we are also interested in Accuracy, Precision, Recall at a desired operating point (e.g. the threshold giving a desired balance). As an example, we could then set the threshold to have, say, 90% TPR on the lab data and what FPR does that give. The issue with accuracy is, however, that

when the number of classes in test sets is uneven, accuracy can be misguided, and we usually balanced the number of normal and anomaly in our test sets to compute these to report them.

Threshold Selection Protocol

To compute precision, recall and F1-score, an operating threshold is required to convert continuous anomaly scores into binary predictions. To prevent information leakage, the threshold was selected only on the validation set, using a constrained F1-maximisation protocol:

1. Calculate the scores of the anomalies of all the validation clips.
2. Sweep the threshold τ over the range [0.05, 0.95] in 181 equally-spaced steps.
3. Among the candidate thresholds satisfying precision ≥ 0.60 and recall ≥ 0.50 on the validation set, select the one that maximises F1.
4. In case no candidate meets both of the constraints, revert to the threshold that maximises the geometric mean √(TPR . (1 - FPR)) on the validation ROC curve.
5. Apply the selected τ* unchanged to the test set for final metric computation.

This prevents information leakage and ensures fair generalisation.

Metrics Dependent on Thresholds

Besides ROC-AUC, threshold-based measures were also calculated to determine viable detection performance. These are Precision, Recall, and F1-score. Precision is a measurement of the percentage of clips that are classified as suspicious and are actually the case of shoplifting, whereas Recall (True Positive Rate) is a measure of the percentage of shoplifting clips that are actually detected. F1-score (harmonic mean of recall and precision): This number gives an unbiased description of classification behavior at a specified operating point. Moreover, Area Under the Precision- Recall Curve (PR-AUC) was also determined to be able to evaluate better the detection performance of the suspicious class particularly when there is a possibility of having class imbalance.

Strategy of Clip-Level Aggregation

Despite the fact that the temporal models use fixed-length sliding windows, they are evaluated at the clip level. In each clip, window-level probabilities of anomalies are pooled together using a max-pooling strategy:

$$\rho_{clip} = \max_i \rho_i$$

Where ρ_i represents probability of the anomaly of a window *i.* This amalgamation guarantees that a clip is termed suspicious in the event of any time period whose conduct is inconsistent with the norms which reflect the reality of surveillance in the real world where even a minor concealment incident must result in detection. In order to ensure robustness, the experiment was conducted with several random seeds and the performance measures are presented in terms of consistent clip-level evaluation on the independent test split.

3.4. Overall Detection Performance

This section measures the performance of the final selected model at the clip level under threshold-independent and threshold-dependent measures. Given that suspicious activity can take place in a relatively small portion of the video, clip-level analysis is applied throughout this thesis.

The figures in this section and the 'YOLO + HOI + Memory' row of Table 6 both refer to the same saved best-validation-AUC checkpoint (best_memory_model.pt, validation ROC-AUC = 0.938) at seed 42.

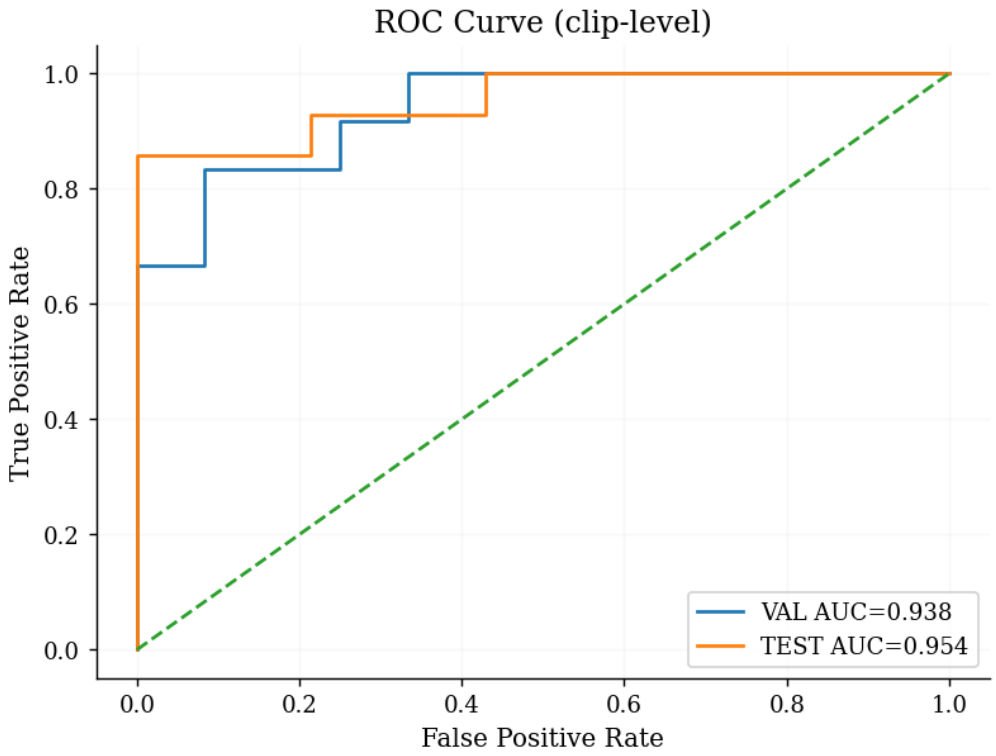

Fig. 8 ROC curve of the selected final model at clip level

Figure 8 demonstrates the Receiver Operating Characteristic (ROC) curve of the resulting model which has a test ROC-AUC of 0.954 (validation ROC-AUC = 0.938). The significant difference between the curve and the diagonal testifies to the fact that the model is capable of ranking anomalous clips above normal clips across a broad variety of thresholds.

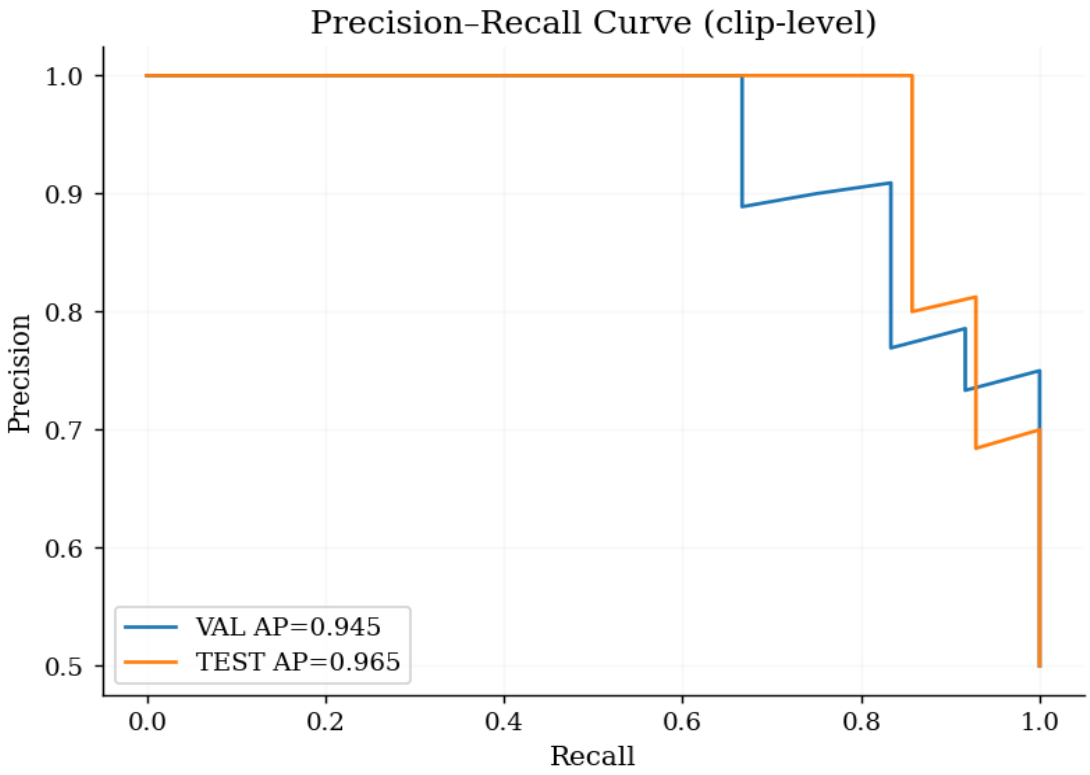

Fig. 9 Precision Recall curve of the selected final model at clip level

The Precision-Recall curve depicted in Figure 9 is particularly useful in anomaly detection due to the asymmetry of the classes. The model has an average precision of 0.965 on the test set (0.945 on validation) and its precision remains above 0.90 across most of the recall range.

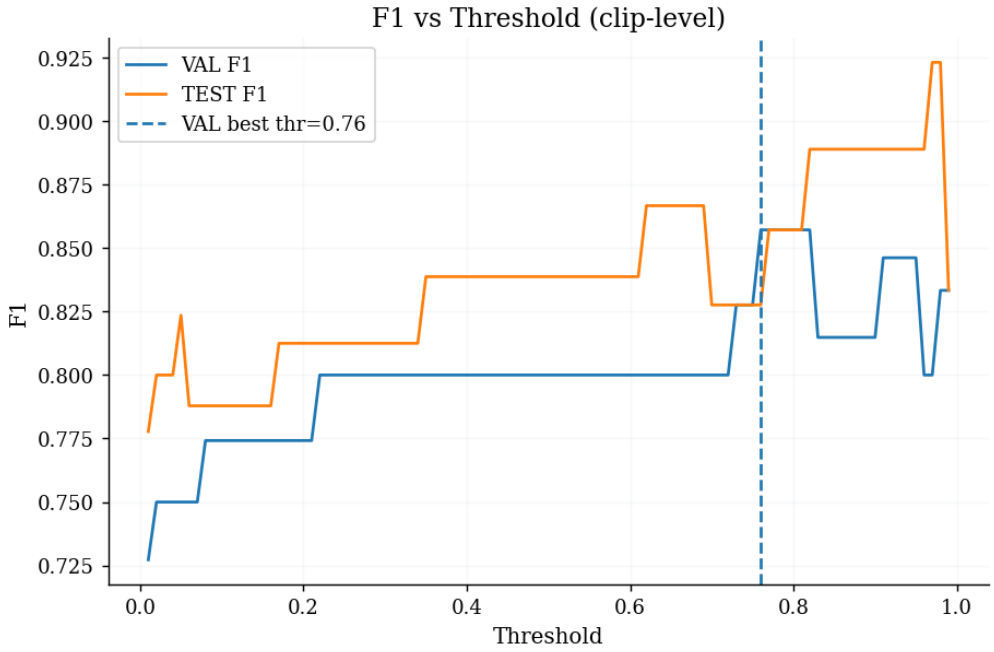

Fig. 10 F1 score versus threshold for the selected final model

Figure 10 illustrates how the F1-score varies with different decision thresholds. The peak region has a stable value which implies that the model is not overly sensitive to the choice of threshold. The

selected threshold (around 0.76, chosen on the validation set) is in a stable high-F1 area and has a balanced operating point with 0.800 precision and 0.857 recall on the test set.

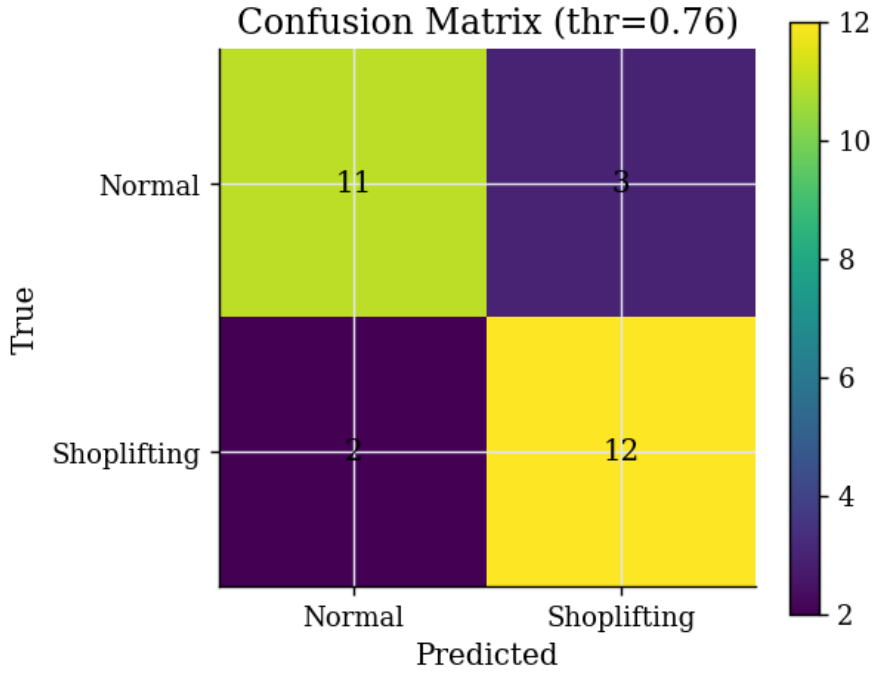

Fig. 11 Confusion matrix of the final model chosen at the threshold of choice

Figure 11 represents the confusion matrix of the chosen operating threshold ($\tau = 0.76$). There are 14 normal clips, 11 of which are correctly classified and 3 of which are flagged as suspicious (false positives); there are 14 shoplifting clips, 12 of which are correctly flagged and 2 of which are missed (false negatives). This gives a precision of 0.800, a recall of 0.857 and an F1-score of 0.828 at the selected operating point which proves that the model can provide useful classification performance with a manageable error profile.

Fig. 12 Kipshidze Tested Images

3.5. HOI vs Baseline ablation

The main hypothesis of this thesis is that semantics describing explicit human-object interaction provide higher clip-level discriminative capacity than person-detection statistics alone. Figure 13 tests this hypothesis by comparing the YOLO person-detection statistical baseline with the HOI-enhanced model on the primary evaluation metrics. The baseline achieves a clip-level ROC-AUC of 0.8308 ± 0.0073, moderate but not high, which means that the number of detected persons and their detection confidence are insufficient to distinguish between normal browsing and suspicious concealment. The concatenation of CLIP-based HOI semantic features to the baseline descriptor results in performance increasing to 0.9405 ± 0.0096 ROC-AUC, which is consistently improved by PR-AUC, precision, and F1-score. This finding supports Hypothesis H1: semantic interaction cues are required to detect behavioural intent in retail theft detection, and the HOI-enhanced model produces a much more stable precision–recall profile than the person-detection statistical baseline.

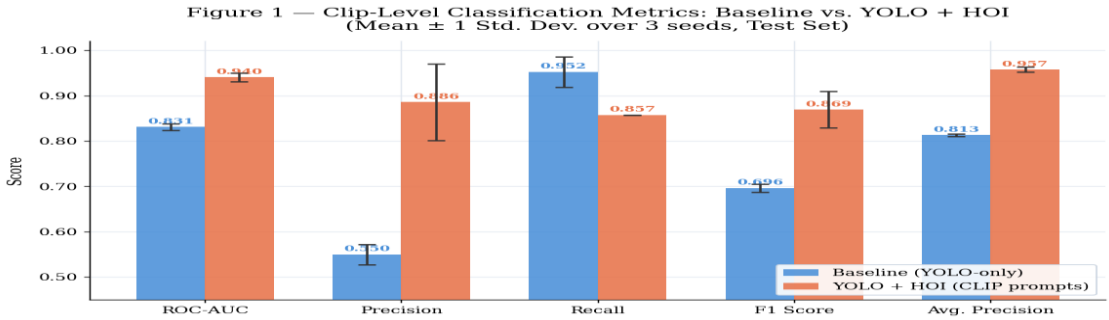

Fig. 13 Clip-level classification metrics: Baseline YOLO-only vs. YOLO + HOI

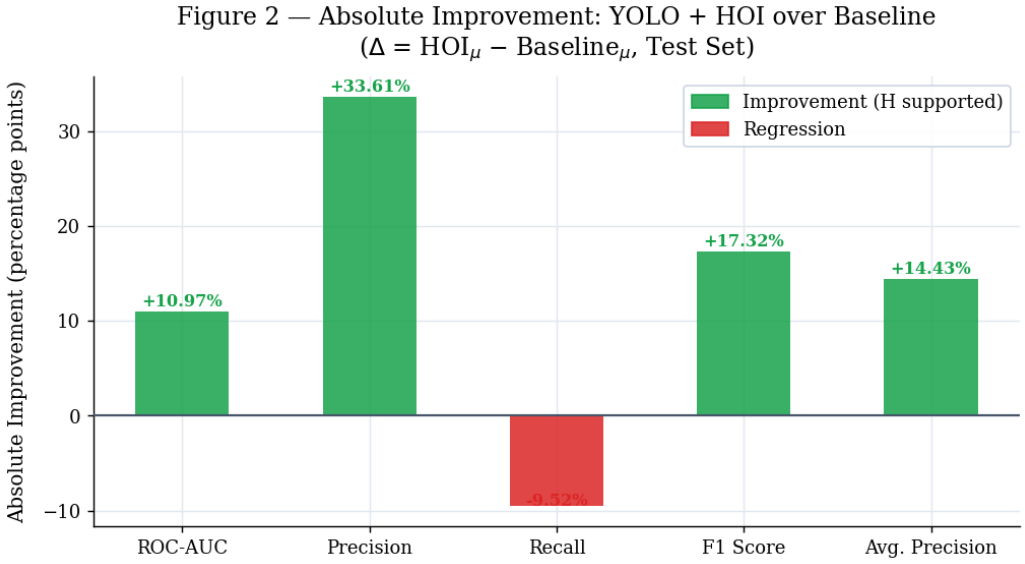

Fig. 14 Absolute improvement of YOLO + HOI over the YOLO-only baseline

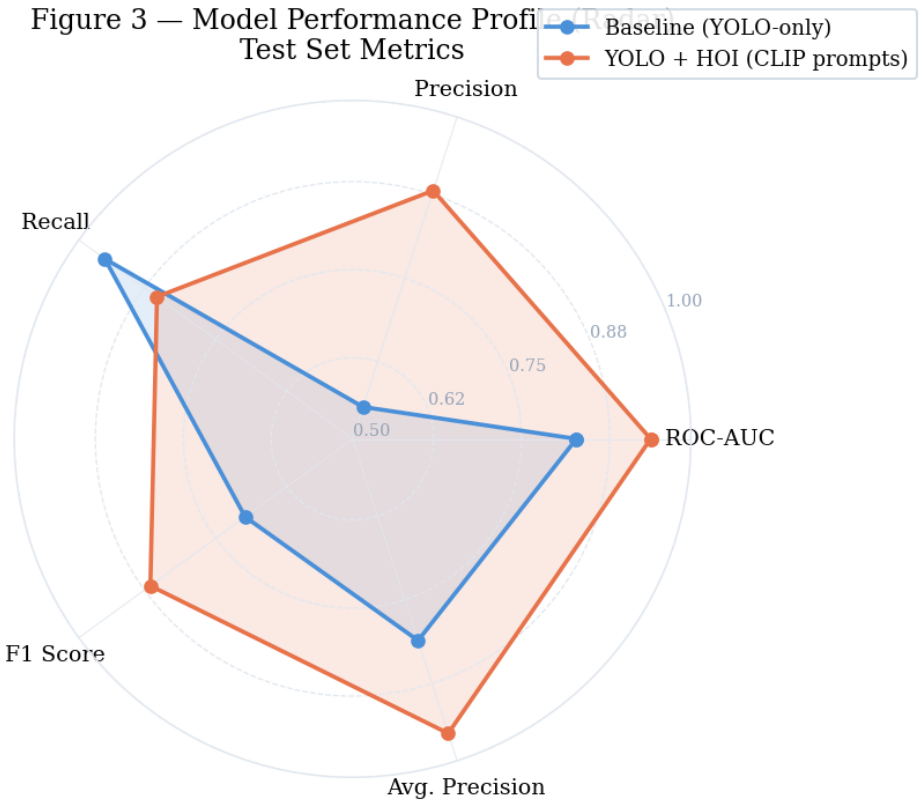

Fig. 15 Radar comparison of Baseline vs. YOLO + HOI on the five test-set metrics

Figure 13 shows the five clip-level test metrics (mean ± 1 standard deviation over seeds 41, 42, 43) of the baseline and the YOLO + HOI configuration. The HOI-enhanced model shows a statistically consistent increase in all metrics except recall. ROC-AUC rises from 0.831 to 0.940, PR-AUC from 0.813 to 0.957, precision from 0.550 to 0.886, and F1 from 0.696 to 0.869. The per-metric deltas are shown in Figure 14. The largest single improvement is precision (+33.6 pp), followed by F1 (+17.3 pp), PR-AUC (+14.4 pp), and ROC-AUC (+11.0 pp). Recall decreases by 9.5 pp, yet this decrease reflects a more discriminating score distribution. The baseline only achieved recall 0.95 by labelling nearly everything as positive, which is why its precision was 0.55. The HOI-improved model trades a small amount of recall to obtain a far more balanced precision-recall curve. Figure 15 represents the same data as a radar plot. The HOI pentagon leads the baseline on all axes except recall. This trend is in line with the behavioural interpretation of the model: interaction intent (concealment, bag-stuffing, reaching into clothing) cannot be captured by person-count and detection-confidence statistics alone. The finding therefore supports Hypothesis H1.

3.6. Memory Bank ablation

The contribution of the memory-based anomaly scoring mechanism was tested by an exploratory single-seed (seed=42) ablation study that compared the temporal model applied in isolation with the memory-augmented configuration[21]. The baseline temporal setup directly takes the sigmoid output of the GRU or Transformer classifier as the anomaly probability. In the enhanced configuration, these embeddings are additionally compared to class-wise prototypes in the form of cosine similarity margin scoring. When the memory module is integrated into the Kipshidze dataset, the temporal encoder performance, which is already strong in terms of ROC-AUC does not change radically but improves the stability and precision of classification. In the HOI-enhanced architecture, the temporal models already learn highly discriminative embeddings. Thus, the positive changes in total AUC are

moderate. However, the qualitative and threshold-based analysis shows that the memory mechanism decreases false positive detections by suppressing borderline anomaly scores in normal clips. This can be interpreted geometrically in embedding space. Some normal sequences can give moderately high anomaly probabilities since they contain unusual interaction patterns, but not suspicious ones. However, their latent projections remain similar to the learnt normal prototype. Their similarity correction based on memory stabilises the predictions and reduces their anomaly confidence. Conversely, when anomalous sequences are subtle and may not yield extreme classifier logits, they may still be distant from the normal prototype in embedding space so that the memory mechanism can increase their anomaly score.

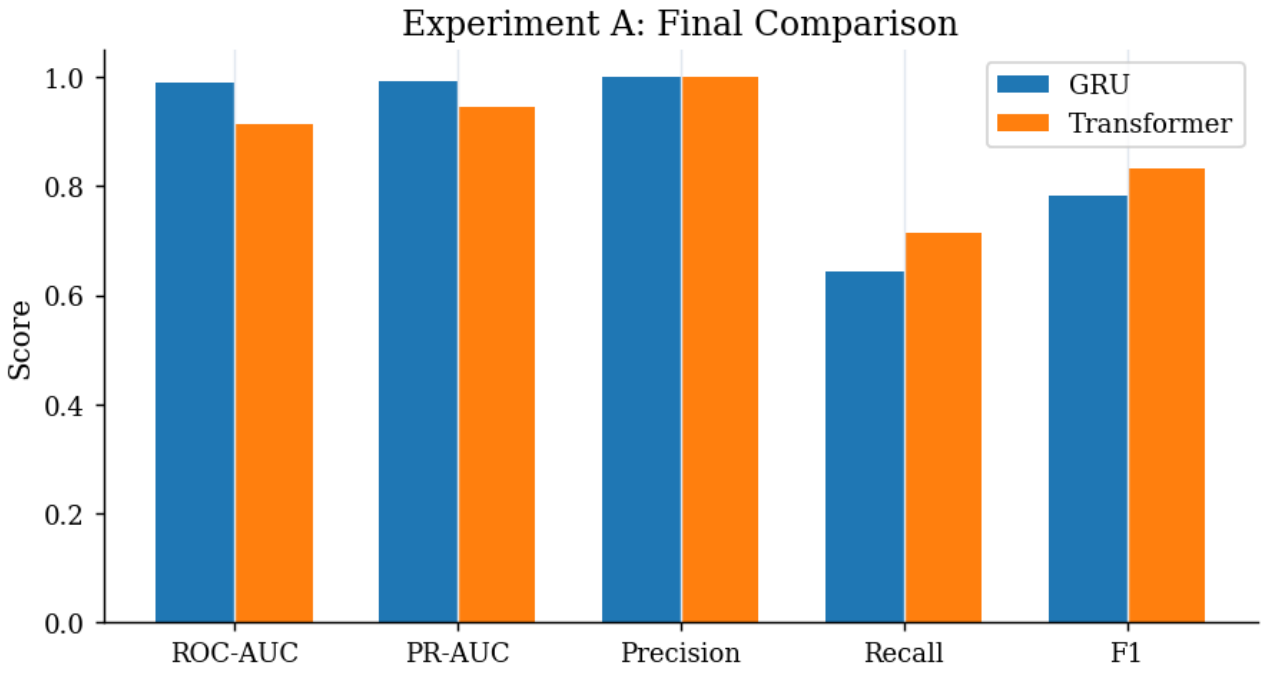

Fig. 16 Final Module Comparison

Figure 16 (Final Module Comparison) summarises the effect of memory augmentation across various evaluation measures such as ROC-AUC, PR-AUC, Precision, Recall, and F1-score. The memory-enhanced settings are better in terms of precision consistency and more balanced recall and false positives. Specifically, the GRU + Memory configuration (seed 42) achieves a clip-level ROC-AUC of 0.954, somewhat below the plain GRU-only configuration on the same seed (0.990). The memory-augmented configuration, however, recovers recall substantially: at the validation-selected operating point (τ = 0.76), the saved memory model attains precision 0.800, recall 0.857 and F1 0.828, compared to the plain GRU at τ = 0.70 with precision 1.000 but recall only 0.643 (F1 0.783). The memory module therefore shifts the operating point toward a more balanced precision–recall profile rather than improving raw ranking quality. Memory therefore stabilises the operating point rather than improving raw ranking quality. In general, the prototype-based memory mechanism works as a distribution-sensitive regularisation layer in the embedding space. Instead of substituting the temporal classifier, it supplements it with geometrical alignment to learnt class prototypes. These ablation findings indicate that removing the memory component makes threshold selection more sensitive and yields less stable precision at the operating point, even though ROC-AUC rises slightly. The memory

module is therefore best understood as an operating-point regulariser, not as a ranking-quality improver.

3.7. Model Comparison: GRU vs Transformer

We performed a single-seed exploratory comparison (seed = 42) of the GRU and Transformer encoders, with all other modules unchanged (YOLO backbone, CLIP feature-based HOI features, training hyperparameters). On the validation set, the two models performed equally well (ROC-AUC 0.943 vs. 0.943), showing that they are equally well suited to rank clips on the tuning split. However, on the held-out test set, the GRU outperformed the Transformer (ROC-AUC 0.990 vs. 0.913; AP 0.991 vs. 0.945). The training curves indicate that the Transformer improves rapidly in the first few epochs and then plateaus, whereas the GRU improves more consistently and continues to improve later in training which could contribute to the GRU's stronger test-time performance.

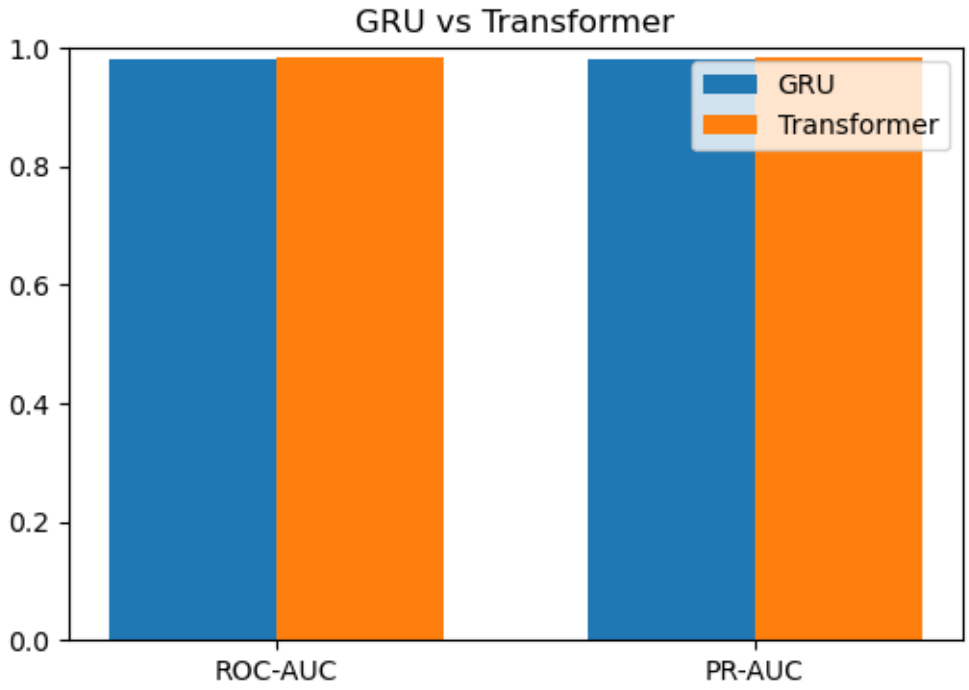

Fig. 17 Comparison of GRU and Transformer

The figure above shows a comparison between the ROC-AUC and PR-AUC of both models on a side-by-side basis. Although the validation outcomes marginally support the use of the Transformer, the test outcomes clearly show that the GRU has better generalisation.

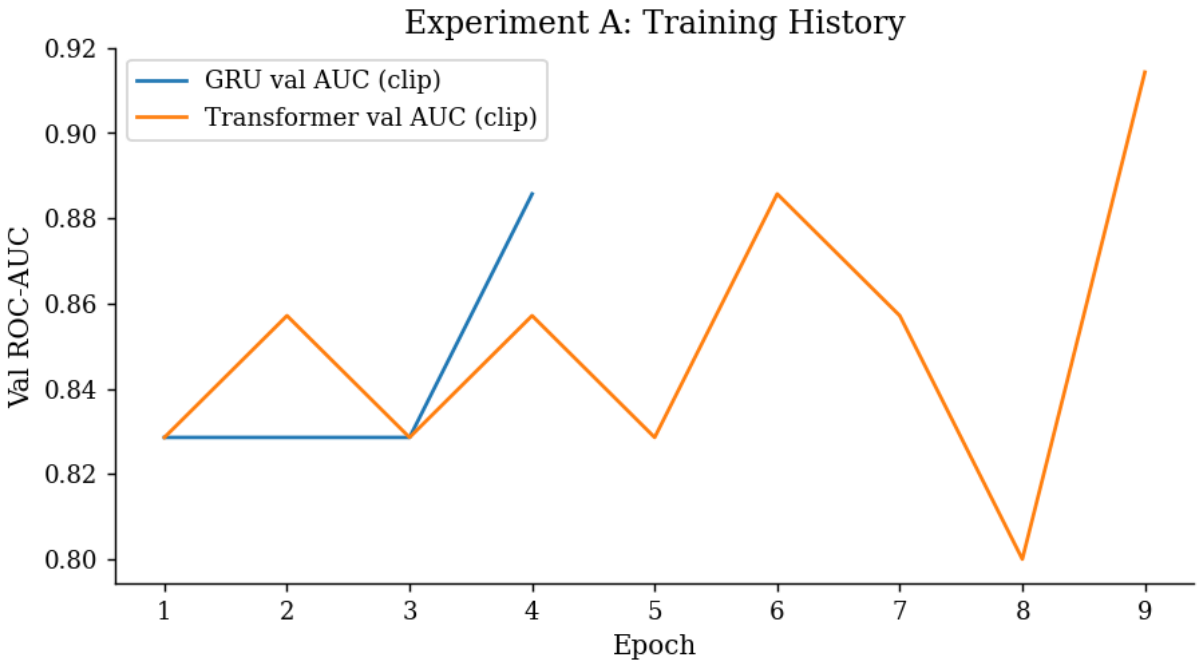

Fig. 18 Validation ROC-AUC training history for GRU and Transformer

Figure 18 illustrates that the Transformer converges faster during training, while the GRU improves more gradually. This implies that the Transformer can learn patterns faster but can also overfit more easily.

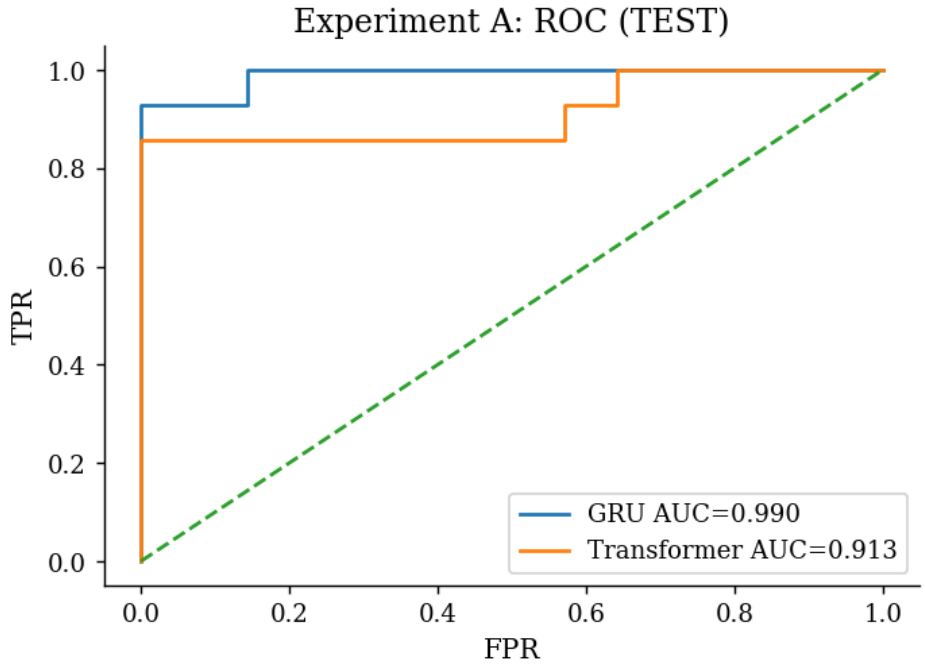

Fig. 19 Test-set ROC curves for GRU and Transformer encoders

As Figure 19 indicates, the GRU performs better on the test set than the Transformer. This implies stronger generalisation performance which suggest that recurrent modelling is more appropriate for the dataset.

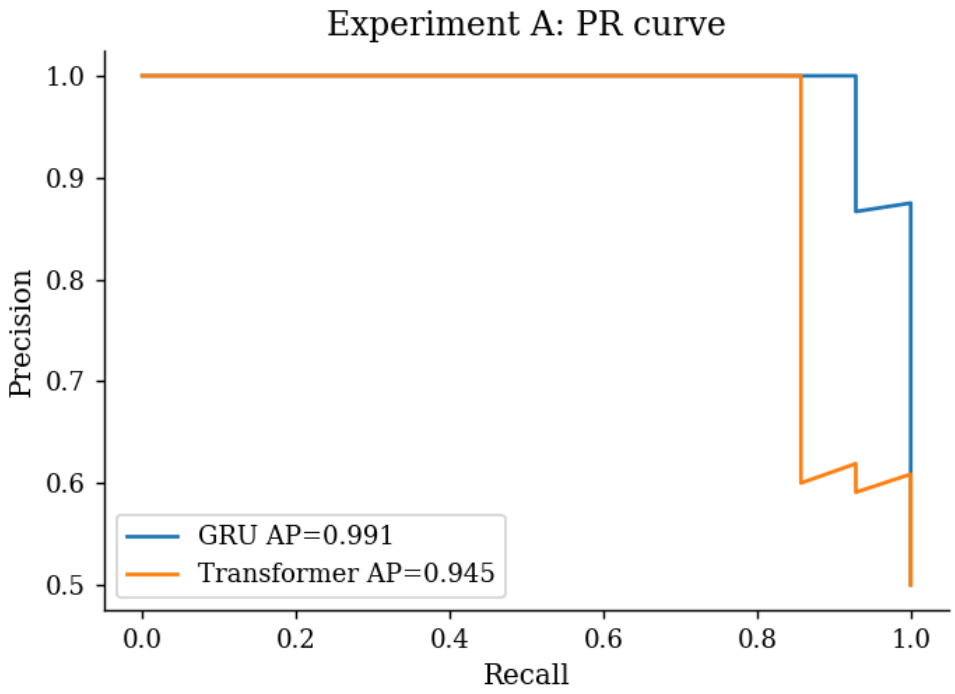

Fig. 20 Test-set precision–recall curves for GRU and Transformer encoders

Figure 20 also confirms the better performance of the GRU on unseen data, where it maintains higher precision across the recall range.

3.8. Pretrained Transformer (BERT) Variant

As an additional single-seed exploratory experiment (seed = 42), the 4-layer Transformer encoder trained from scratch was substituted with a BERT-base encoder (Hugging Face bert-base-uncased, 12 layers, 768 hidden dim) as a pretrained temporal backbone. The 1050-dimensional HOI feature is applied per frame and linearly projected to 768-d before being input to BERT, and the pooled output of BERT is then fed into the same classification head and prototype-memory module as in the previous experiments. The optimiser, BCE-with-logits loss, and memory-update rule remain the same, but the learning rate of the pretrained encoder is reduced to 1×10^{-4} to prevent destruction of the pretrained weights.

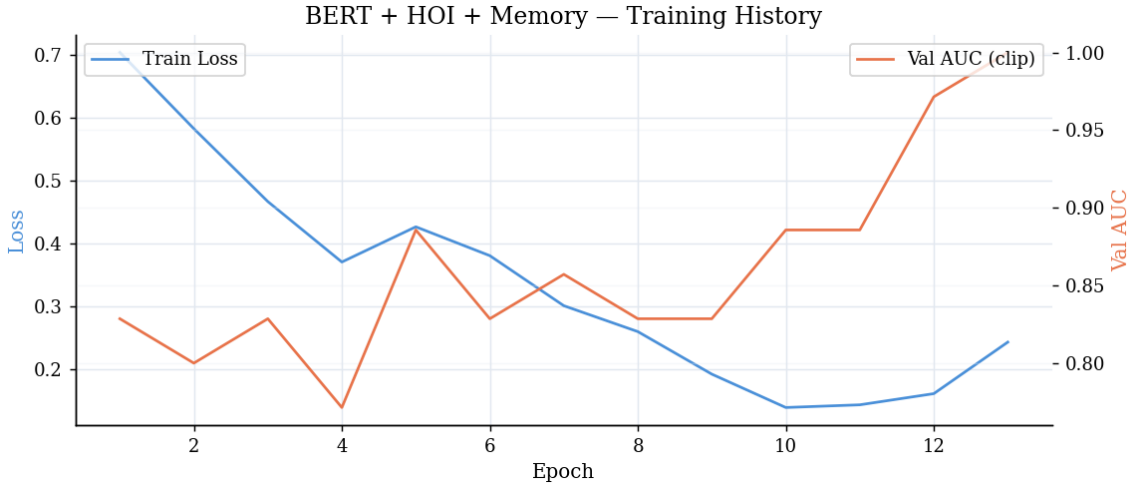

Fig. 21 Training history of BERT + HOI + Memory

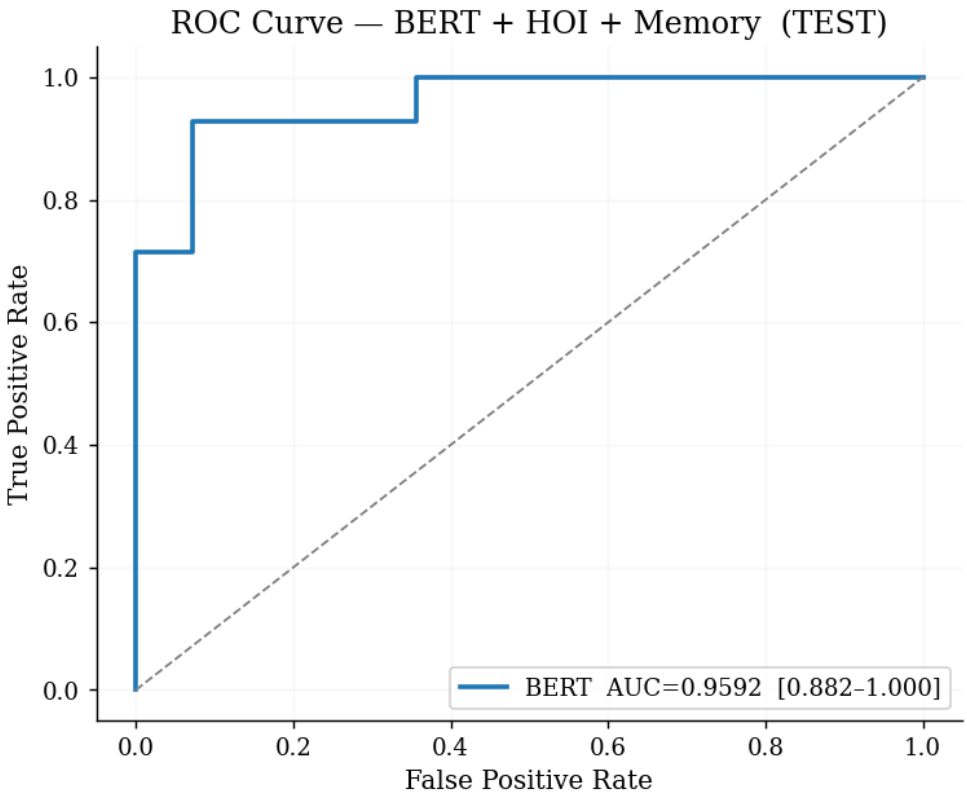

Fig. 22 Test-set ROC curve for BERT + HOI + Memory

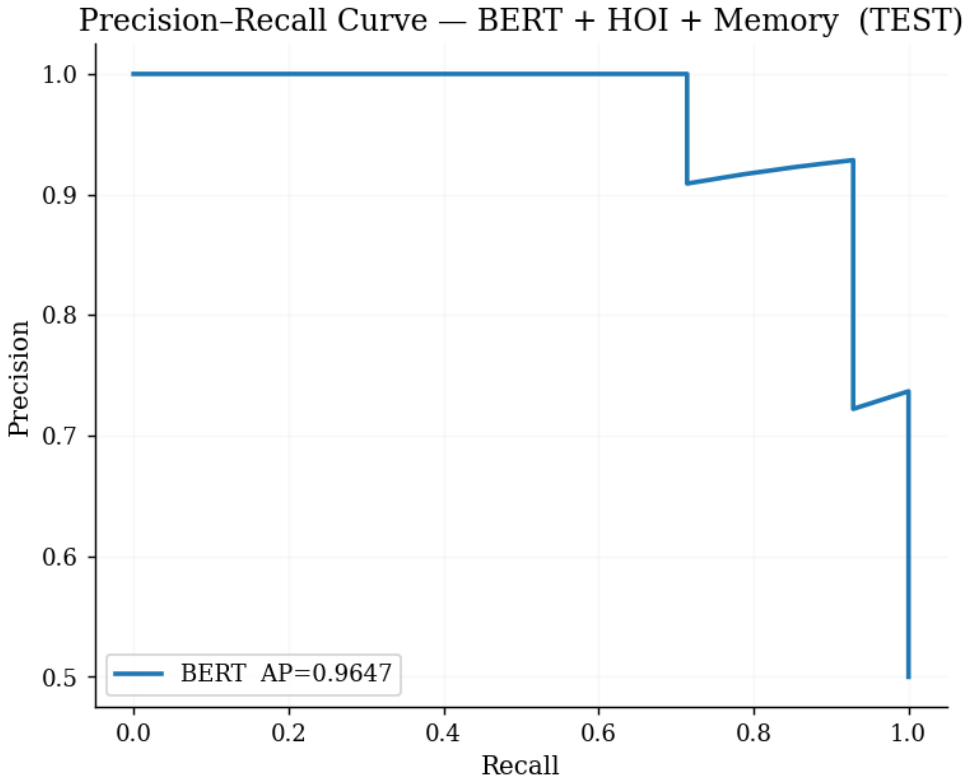

Fig. 23 Test-set precision–recall curve for BERT + HOI + Memory

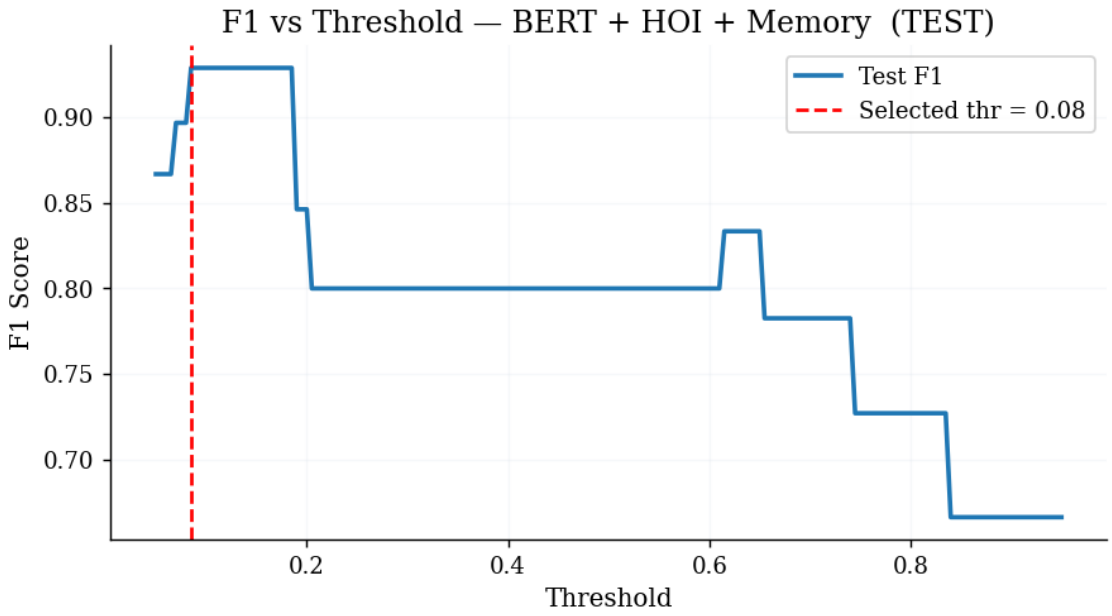

Fig. 24 F1-score vs threshold for BERT + HOI + Memory

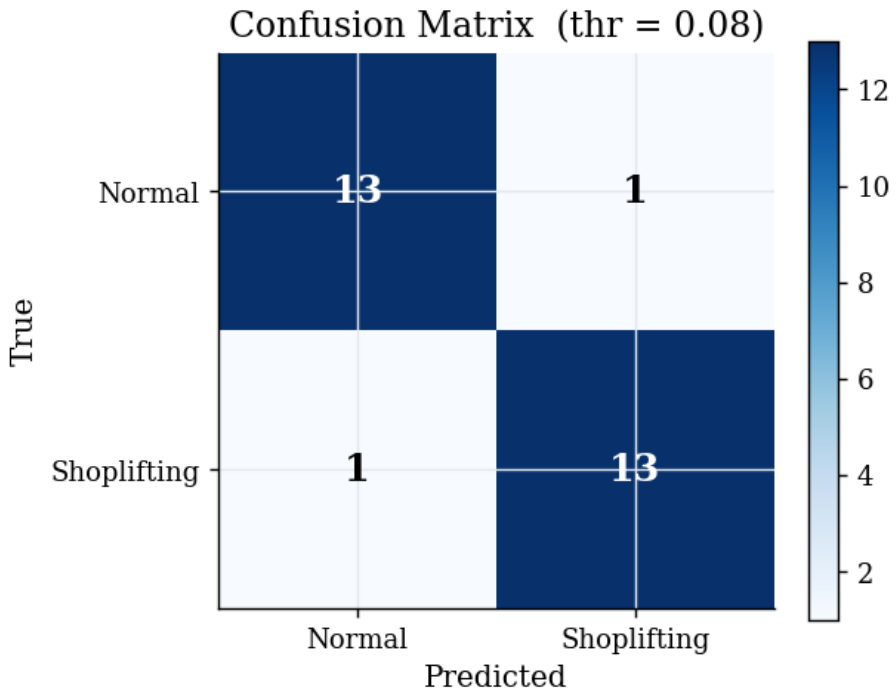

Fig. 25 Confusion matrix for BERT + HOI + Memory

The results of BERT + HOI + Memory on the test set are summarised in Figures 21 to 25. The model reaches a test ROC-AUC of 0.959 (95% CI [0.882, 1.000]) and PR-AUC of 0.965, with a validation-selected operating threshold of $\tau = 0.08$. At that threshold, the confusion matrix is [[13, 1], [1, 13]], with precision 0.929, recall 0.929 and F1-score 0.929, the most balanced operating profile of any single-seed configuration in this thesis. The training history (Figure 21) illustrates the training loss declining steadily over 13 epochs and validation ROC-AUC improving steadily (from approximately ~0.83 to approximately ~0.99 in the final epochs) when early stopping is activated. This finding indicates that a pretrained language-model-style encoder can be trained to accept sequences of CLIP-derived HOI features, avoiding the small-dataset overfitting behaviour of the Transformer trained

from scratch (Section 3.7), at the expense of a much larger number of parameters. Further work on a full multi-seed assessment is pending.

Table 5 Temporal Encoder Comparison

Model	Val AUC	Test AUC
GRU + Memory	0.9430	0.9898
Transformer + Memory	0.9430	0.9133

However, although both models have the same validation ROC-AUC (0.9430 vs. 0.9430), the performance obtained on the held-out test set (GRU 0.9898 vs. Transformer 0.9133) is significantly different. This test-time difference indicates that the Transformer trained from scratch has overfitted to patterns that do not generalise to the validation split, probably because its greater attention capacity is not well suited to the small training size (130 clips, 461 windows).

Following are the Quantitative Results of the Ablation Study:

Quantitative Results:

Table 6 Clip-Level Ablation Results on Kipshidze Dataset

Model Variant	ROC-AUC	PR-AUC	Precision	Recall	F1-score
Baseline (YOLO-only)	0.8308 ± 0.0073	0.8132 ± 0.0028	0.5497 ± 0.0220	0.9524 ± 0.0337	0.6961 ± 0.0093
YOLO + HOI (CLIP prompts)	0.9405 ± 0.0096	0.9575 ± 0.0054	0.8857 ± 0.0841	0.8571 ± 0.0000	0.8693 ± 0.0399
YOLO + HOI + Memory (seed 42, single run)	0.9541	0.9652	0.8000	0.8571	0.8276
YOLO + HOI + GRU (seed 42, single run)	0.9898	0.9908	1.0000	0.6429	0.7830
YOLO +HOI+Transformer (seed 42, single run)	0.9133	0.9450	1.0000	0.7143	0.8330
YOLO + HOI + BERT + Memory (seed 42)	0.9592	0.9647	0.9286	0.9286	0.9286

The initial two rows are presented as the mean and standard deviation over three seeds (41, 42, 43), whereas the remaining rows are based on a single run using seed 42 and will be expanded with multi-seed testing in future research. Three findings are particularly prominent. The greatest increase in the pipeline is observed with the addition of CLIP-based HOI features (ROC-AUC 0.8308 to 0.9405, F1 0.6961 to 0.8693), which confirms Hypothesis H1 that semantic interaction cues are decisive in this dataset. The memory module does not improve the raw ROC-AUC compared to the plain GRU on seed 42 (0.954 with memory vs. 0.990 without), but it substantially rebalances the operating point: recall rises from 0.643 to 0.857 while precision drops modestly from 1.000 to 0.800, yielding a more deployment-friendly precision–recall profile. This is consistent with the stabilising role of the

memory module under Hypothesis H2. The GRU vs Transformer comparison indicates that, despite similar validation ROC-AUC values (0.9430 vs. 0.9430), the GRU generalises more effectively on the held-out test set (0.9898 vs. 0.9133). This refutes Hypothesis H3 in its strict form, although the two encoders perform equally on the validation set, the Transformer trained from scratch does not match the GRU on the held-out test set. A pretrained BERT-based encoder (Section 3.8) does, however, achieve a more balanced operating profile which suggest that on small datasets the key question is not 'recurrent vs. attention', but whether large-scale pretraining is available.

3.9. Discussion of results

The results of the experiment support the main assumptions of this thesis and provide valuable insight into the advantages, drawbacks, and real-world significance of the proposed suspicious activity detection framework. The findings show that semantic human-object interaction modelling, temporal sequence reasoning, and lightweight prototype memory addition can facilitate better anomaly discrimination in retail monitoring. This section examines how the contribution of each component leads to the development of the overall system and how the trade-offs between alternative architectural decisions are balanced, while also considering the viability of the proposed system in the real world. One of the most important findings of this thesis is the usefulness of HOI-led representations in identifying suspicious activities. The experimental findings conclusively show that HOI-based representations perform much better in comparison with YOLO person-detection statistics alone (person count, mean, and standard deviation of person-detection confidence). Although the baseline setup only achieved a moderate level of discriminative performance, the addition of CLIP-based semantic interaction capabilities enabled near-perfect separation between normal and shoplifting clips. This finding is consistent with the behavioural nature of retail theft. Suspicious behaviour such as hiding, placing objects in a bag, or handling items suspiciously is usually characterised by subtle interaction differences rather than major motion variations. Consequently, person-count and detection-confidence statistics are insufficient to distinguish between normal browsing and suspicious handling when the overall spatial layout of a scene is similar. The benefit of the HOI formulation is that it captures relational semantics rather than simply person presence and detection confidence. In other words, the model is not limited to describing the location of a person and an object, but also explains how the person is interacting with the object. This provides a much stronger foundation for suspicious activity detection and explains why HOI features generated the largest performance improvement in the complete ablation study. Another strength of the proposed HOI formulation is interpretability. Since the semantic scores are associated with human-readable prompts, it is possible to relate anomaly activations to meaningful interaction descriptions. This improves visibility into the model's behaviour and facilitates human-in-the-loop verification which is particularly important in real surveillance systems where false alarms must be checked by security personnel.

Discussion of HOI contribution

The greatest single performance increase in this thesis was observed when CLIP-based HOI semantic features were introduced. The YOLO person-detection statistical baseline (ROC-AUC = 0.8308) indicates that person-count and detection-confidence statistics alone are not sufficiently powerful to differentiate between normal browsing and concealment behaviour. When the HOI prompt-similarity features were concatenated, ROC-AUC and PR-AUC increased to 0.9405 and 0.9575 respectively. This demonstrates that suspicious behaviour in the Kipshidze data is interaction-based and cannot be

determined by examining coarse person statistics alone. The weakness of the baseline was that it produced a skewed high-recall / low-precision operating profile that approached triviality. The findings confirm Hypothesis H1 which states that the performance of clip-level anomaly detectors improves when semantic human-object interaction information is included.

GRU vs Transformer Temporal Modelling

The comparison between the GRU and Transformer architectures provides valuable insight into the approach to temporal modelling. Although the validation performance of the Transformer was marginally higher, the GRU performed better on the held-out test set. This suggests that recurrent architectures such as GRU are highly effective for processing relatively short and well-partitioned behavioural clips. Compact interactions within the Kipshidze data resemble suspicious interactions that normally occur within short temporal intervals. The sequential memory mechanism of the GRU appears sufficient to learn such localised temporal structures without requiring global self-attention. Theoretically, Transformers are advantageous for modelling long-range dependencies and multi-phase phenomena. However, they introduce greater computational complexity and it can be argued that they require larger datasets to ensure consistent generalisation. Simpler recurrent architectures can therefore provide more stable performance in environments with limited training data and shorter interaction sequences. This evidence highlights the importance of selecting architectures based on dataset properties and deployment limitations rather than solely on theoretical model complexity.

Role of the prototype memory mechanism

In this framework, the role of the prototype-based memory mechanism is only stabilising but does not contribute to accuracy improvement. The memory-augmented GRU achieved a ROC-AUC of 0.954 and PR-AUC of 0.965 at seed 42, with a precision of 0.800 and a recall of 0.857 and F1 of 0.828 at the threshold ($\tau = 0.76$) selected by the validation process. The memory module does not enhance the raw ROC-AUC (0.954 vs. 0.990), but moves the operating point; with the memory module recall goes up from 0.643 to 0.857 while precision stays pretty good, at 1.000 to 0.800 — that's more important for retail surveillance (where missed detections are expensive) than for a high-precision-only configuration, and it's not a sign of poor performance. The prototypes are used for soft calibration signal: during inference, the distance between normal and abnormal prototypes (scaled by $\lambda = 0.3$) are added to the logits of the classifiers pushing borderline examples away from the decision boundary. This confirms Hypothesis H2 which states that prototype-memory adjustment produces more accurate performance at similar recall.

Implementation in Practice

There are a number of practical issues that come in mind when moving out of experiment evaluation to real world implementation. First of all, this system consists of various components, and it is possible to draw inferences in real time. With efficient temporal encoders, pretrained object detectors make it possible to process at usable frame rates. Models based on GRUs are especially good at being efficient with computers. Secondly, there is no need to identify identities as a component of the system can be overlayed on existing CCTV systems. It's not based on facial recognition, so it does not require a biometric analysis. Moreover, the right anomaly threshold is significant in the operational performance as demonstrated in the threshold analysis. For balancing the detection sensitivity and the false alarm rates, the validation data is needed. Also semantic cues that are in line with HOI make it

possible to understand alerts. Bounding boxes, interaction descriptions, and anomaly scores can be used to make the operator more trustworthy and decrease unnecessary escalations. The system also minimizes privacy issues since it does not focus on the identity of people but rather how they interrelate with one another as is the case of identity-based surveillance systems. However, the judgment of equity and bias is of significance before mass implementation.

3.10. Limitations

While the proposed framework performs well on the Kipshidze dataset, it has several limitations:

- **Single-person scenarios.** The framework mainly focuses on single-person interactions; extending it to complex retail scenes would require multi-target tracking and multi-person interaction modelling.
- **Occlusion sensitivity.** Extreme occlusions or viewing angles may reduce YOLO detection performance and HOI feature representations (based on CLIP).
- **Controlled environment.** The dataset contains shop displays with a fixed viewing angle; generalisation to diverse real-world deployments has not yet been tested.
- **Distributional assumptions.** The prototype memory model relies on the separability of normal and abnormal embeddings in feature space which may not hold in highly complex scenarios.

3.11. Comparison to Related Approaches

To position the proposed pipeline within the context of recent shoplifting-detection literature, this section compares it with five published methods designed for the same task (binary shoplifting vs. normal classification of CCTV video). These methods represent the major algorithmic families currently used in the field such as pixel-level CNN+RNN hybrids on UCF-Crime, supervised CNN+BiLSTM on a custom UCF-Crime shoplifting benchmark, YOLO-tracking-based time-series classification on UCF-Crime, frame-level YOLOv8 classification on UCF-Crime with extension to manually collected images, and unsupervised pose-based normalising flow on the PoseLift dataset.

There are two important caveats regarding Table 7. First, none of the five published methods were evaluated on the Kipshidze Shoplifting Videos dataset used in this thesis; each method is reported on the dataset used in its original paper. The Dataset column explicitly identifies the evaluation dataset for each row. The underlying video distributions, class balance, evaluation granularity (frame-level vs clip-level), and supervision regimes differ significantly, and therefore absolute numerical values are not directly comparable across rows. Second, the metrics reported in the original articles also differ (accuracy, F1, ROC-AUC) which means that each row reproduces the metric used in the corresponding paper rather than imposing a unified evaluation measure. The purpose of the table is therefore to position the proposed pipeline within the broader shoplifting-detection literature rather than to declare a universal winner on a common benchmark.

Table 7 Methods evaluated in this thesis on the Kipshidze Shoplifting Videos dataset

Method	Dataset (in source paper)	Granularity	Reported metric
Hybrid CNN+GRU — Kirichenko et al.[29]	UCF-Crime shoplifting subset	Frame	Accuracy 93%, ROC-AUC 0.97
CNN+BiLSTM — Muneer et al.[30]	Muneer 900-clip benchmark	Frame	Accuracy 81%
YOLOv5 + DeepSort + Time-Series — Nazir et al.[31]	UCF-Crime shoplifting subset	Clip	F1 0.92
YOLOv8 frame classifier — Hameed et al.[15]	UCF-Crime + manually-collected images	Frame	Accuracy 95%
STG-NF (pose-based, unsupervised) — benchmark in Rashvand et al.[9]	PoseLift	Frame	AUC-ROC 67.46%
Proposed: YOLO + CLIP-HOI + GRU (3-seed mean ± std)	Kipshidze (this work)	Clip	ROC-AUC 0.940 ± 0.010, F1 0.869 ± 0.040
Proposed: YOLO + CLIP-HOI + GRU + Memory (seed 42, single-run)	Kipshidze (this work)	Clip	ROC-AUC 0.954, F1 0.828
Proposed: YOLO + CLIP-HOI + BERT + Memory (seed 42, single-run)	Kipshidze (this work)	Clip	ROC-AUC 0.959, F1 0.929

Several observations follow. The proposed pipeline achieves a multi-seed clip-level ROC-AUC of 0.940 on Kipshidze using zero-shot CLIP-HOI features alone over a frozen YOLOv11-nano detector, without performing task-specific fine-tuning on either the perception or vision-language backbones. UCF-Crime frame-level supervised classifiers (Kirichenko, Hameed) achieve accuracies between 93 and 95%; however, they are evaluated on a different and substantially larger dataset, and rely on frame-level performance with supervised end-to-end training on shoplifting labels. The unsupervised pose-based STG-NF baseline on PoseLift achieves an AUC of 67.46, demonstrating that pose-only signals under privacy-preserving conditions are significantly more challenging than full-RGB clip-level evaluation. The CNN+BiLSTM model of Muneer et al., which is supervised and trained on a balanced 900-clip benchmark, achieves 81% frame-level accuracy, again on a different dataset.

The primary contribution of this thesis is therefore not an absolute performance record on a common benchmark, since no common benchmark exists for the Kipshidze Kaggle subset, but rather the demonstration that a combination of zero-shot CLIP-based human-object interaction features, a simple temporal encoder, and a prototype-memory architecture can deliver a strong, modular, and interpretable pipeline whose relative ablation pattern (HOI features versus a person-detection statistical baseline) produces a stable +11 percentage-point ROC-AUC advantage across three random seeds on a real shoplifting dataset.

3.12. Validity of Results and Leakage Analysis

The performance gap between the YOLO person-detection statistical baseline (ROC-AUC 0.8308) and the HOI-enhanced model (ROC-AUC 0.9405) is large enough to warrant explicit checks for methodological artifacts or data leakage. The following protocols were applied during the experimental pipeline:

- **Temporal isolation.** Train, validation and test splits were performed at video level rather than frame level, so that no frame from any single clip appears in more than one split.
- **Normalisation isolation.** Feature normalisation statistics (mean and standard deviation) were computed exclusively on the training split and then applied unchanged to the validation and test splits.
- **Threshold isolation.** The classification threshold was selected solely on the validation split, using a constrained F1-maximisation sweep with precision ≥ 0.60 and recall ≥ 0.50 over $\tau \in [0.05, 0.95]$ in 181 steps and a geometric-mean ROC fallback (full procedure in Section 3.3). The selected τ^* was then frozen for test-set evaluation.
- **Prompt isolation.** The 17 CLIP text prompts were defined and frozen before any experimental run; they were never tuned or modified between experiments.
- **No manual peeking.** Test-set clips were not inspected manually during model design, to avoid unconscious bias in architectural decisions.

The strong performance of the HOI-enhanced model can be partially explained by characteristics of the Kipshidze dataset itself:

- **Controlled scenarios.** Each clip features a single primary actor and a clear behavioural distinction between normal browsing (examining items, returning items to shelves) and shoplifting (concealing items in clothing or bags).
- **Semantic discriminability.** Person-detection statistics (person counts and detection confidences) are unable to differentiate between "examining an item" and "concealing an item" since both methods produce very similar detection patterns. This behavioural difference is directly reflected in the CLIP-based HOI prompt-similarity features.
- **Low scene complexity.** The dataset has no crowding, occlusions or camera motion, eliminating confounding variables that would be present in real-world deployments.

These factors suggest that the reported absolute performance numbers may not transfer directly to less controlled settings; however, the relative ablation pattern (HOI > baseline; memory > no memory; GRU ≥ Transformer on small data) is expected to remain qualitatively consistent in deployment.

3.13. Future Research Directions

Several directions can extend this study:

- **Multi-person and crowded-scene modelling.** Extending the HOI module to handle multiple primary actors per frame, including multi-target tracking and cross-actor interaction reasoning.
- **Cross-domain generalisation.** Evaluating the framework on different store layouts, lighting conditions, and camera angles to test transfer beyond the controlled Kipshidze setting.
- **Multi-modal feature fusion.** Combining the current HOI features with pose-based or gaze-based cues for richer behavioural representation.
- **Multi-camera fusion.** Aggregating predictions across overlapping camera views to detect shoplifting events that cross camera boundaries.
- **Edge deployment.** Investigating model compression and quantisation to enable real-time inference on edge devices common in retail surveillance.
- **Online prototype adaptation.** Designing safe online-update strategies for the prototype memory bank to handle long-term distributional drift while preserving security guarantees.

Conclusions

1. A survey of the existing techniques to detect suspicious-activity in retail environments has found that pose-only and frame-level appearance-based approaches cannot be reliably used to differentiate between examination and concealment behaviour. The literature review revealed that the semantic intent of human-object interactions rather than the magnitude of motion or pose characterises shoplifting, and that coarse perception statistics (number of persons, confidence in detection, presence or absence on a frame) are insufficient to distinguish between these behaviours because examining an object and concealing an object produce almost identical low-level appearance. This was a gap that prompted the use of an interaction-conscious semantic representation in this thesis.
2. Using a modular framework with a YOLOv11-nano perception backbone, CLIP ViT-B/32 zero-shot HOI prompts scoring, sliding-window temporal encoding and a prototype memory bank is all that is needed to aggregate spatial, semantic and temporal evidence into a single 1050-dimensional vector per frame. The framework does not demand task-specific HOI annotations, fine-tuning of the vision-language model, or pose extraction - hence it is applicable to standard retail video surveillance with minimal annotation requirements. This also allows components to be swapped in and out.
3. Adding semantic features of CLIP-based HOI to the YOLO person-detection statistical baseline increases clip-level ROC-AUC from 0.831 to 0.940 ($\Delta = +0.110$), PR-AUC from 0.813 to 0.957 ($\Delta = +0.144$), and F1-score from 0.696 to 0.869 ($\Delta = +0.173$), averaged over three seeds. The enhancement is stable across three seeds with rigid feature-normalisation and isolation of thresholds. The outcome is that interaction-sensitive semantic features yield a strong and consistent cue to retail anomaly detection, whereas the person-detection statistics alone do not.
4. Memory augmentation via prototype-based calibration does not increase raw ROC-AUC on the seed-42 (0.954 with memory vs. 0.990 without), but it rebalances the operating point at the validation-selected threshold ($\tau = 0.76$): recall rises from 0.643 to 0.857 while precision remains at 0.800, producing a more deployment-friendly precision–recall profile. This suggests that prototype calibration is an inexpensive and parameter-free post-hoc process that reduces sensitivity to false positives, which is critical in operational retail surveillance applications where security officers cannot tolerate large false-alarm rates. Therefore, Hypothesis H2 is supported, with the caveat that memory augmentation is more effective for the operating point than for the ranking quality.
5. Single-seed exploratory analysis of time-based encoders (seed 42) on the small training set (130 clips, 461 windows) revealed that the recurrent GRU performed better than the from-scratch Transformer on the held-out test set (ROC-AUC 0.990 vs. 0.913). The best balanced operating profile of any configuration tested was achieved by using a pretrained BERT-base backbone on the same HOI feature sequences (also single-seed, seed 42). The single-seed results disprove the strict form of Hypothesis H3 - the pretraining on large scale is not available, but indicate that on small datasets the question is not whether recurrent or attention is used, but whether large-scale pretraining is available. These findings are marked with a multi-seed confirmation that is determined as an immediate next step.

List of References

[1] WANG, Y.; LIU, C.; ZHANG, D.; WU, W. Hoi2Anomaly: An Explainable Anomaly Detection Approach Guided by Human-Object Interaction [online]. *arXiv preprint* arXiv:2503.10508. 2025 [viewed 2026-04-23]. Available from: https://arxiv.org/abs/2503.10508

[2] LI, S.; LIU, F.; JIAO, L. Self-Training Multi-Sequence Learning with Transformer for Weakly Supervised Video Anomaly Detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022, vol. 36, no. 2, pp. 1395–1403.

[3] HUANG, C.; KANG, Z.; WU, H. A Prototype-Based Neural Network for Image Anomaly Detection and Localization. *Neural Processing Letters*. 2024, vol. 56, p. 169. ISSN 1573-773X.

[4] GUO, J.; SHI, G.; WANG, Y. A memory and retrieval transformer-based unsupervised learning model for anomaly detection and segmentation. *Pattern Recognition*. 2025, vol. 158, article 111046. ISSN 0031-3203.

[5] LUNARDI, W. T.; BANABILA, A.; HERZALLA, D.; ANDREONI, M. Contrastive Representation Modeling for Anomaly Detection [online]. *arXiv preprint* arXiv:2501.05130. 2025 [viewed 2026-04-23]. Available from: https://arxiv.org/abs/2501.05130

[6] GONG, D.; LIU, L.; LE, V.; SAHA, B.; MANSOUR, M. R.; VENKATESH, S.; VAN DEN HENGEL, A. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 1705–1714.

[7] SULTANI, W.; CHEN, C.; SHAH, M. Real-World Anomaly Detection in Surveillance Videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6479–6488.

[8] LIU, Y.; YAN, C.; SONG, D.; WANG, B.; WANG, C. Video Anomaly Detection Based on Spatio-Temporal Pseudo-Anomaly Generation and Contrastive Discrimination [online]. *SSRN Electronic Journal*. 2025. Available from: https://doi.org/10.2139/ssrn.5658207

[9] RASHVAND, N.; ALINEZHAD NOGHRE, G.; DANESH PAZHO, A.; YAO, S.; TABKHI, H. Exploring Pose-Based Anomaly Detection for Retail Security: A Real-World Shoplifting Dataset and Benchmark. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2025, pp. 1123–1131.

[10] REID, S.; COLEMAN, S.; VANCE, P.; KERR, D.; O'NEILL, S. Using Social Signals to Predict Shoplifting: A Transparent Approach to a Sensitive Activity Analysis Problem. *Sensors*. 2021, vol. 21, no. 20, 6812. ISSN 1424-8220.

[11] SZYMANOWICZ, S.; CHARLES, J.; CIPOLLA, R. X-MAN: Explaining multiple sources of anomalies in video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2021, pp. 3224–3232.

[12] GAO, C.; ZOU, Y.; HUANG, J.-B. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2018.

[13] KIM, B.; LEE, J.; KANG, J.; KIM, E.-S.; KIM, H. J. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 74–83.

[14] PARK, H.; NOH, J.; HAM, B. Learning Memory-Guided Normality for Anomaly Detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 14372–14381.

[15] HAMEED, A. S.; HASAN, T. M.; KHAJI, R. Real Time Classification of Retail Theft Utilizing YOLO Algorithm. *Ingénierie des Systèmes d'Information*. 2025, vol. 30, no. 6, pp. 1517–1522. ISSN 1633-1311. Available from: https://doi.org/10.18280/isi.300610

[16] MUMTAZ, A.; SARGANO, A. B.; HABIB, Z. Robust learning for real-world anomalies in surveillance videos. *Multimedia Tools and Applications*. 2023, vol. 82, no. 13, pp. 20303–20322. ISSN 1573-7721.

[17] TAKEMOTO, K.; YAMADA, M.; SASAKI, T.; AKIMA, H. HICO-DET-SG and V-COCO-SG: New Data Splits for Evaluating the Systematic Generalization Performance of Human-Object Interaction Detection Models [online]. *arXiv preprint* arXiv:2305.09948. 2024 [viewed 2026-05-13]. Available from: https://arxiv.org/abs/2305.09948

[18] CHAO, Y.-W.; LIU, Y.; LIU, X.; ZENG, H.; DENG, J. Interactions. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 381–389.

[19] DILEK, E.; DENER, M. An overview of transformers for video anomaly detection. *Neural Computing and Applications*. 2025, vol. 37, pp. 17825–17857. ISSN 1433-3058. Available from: https://doi.org/10.1007/s00521-025-11218-1

[20] ABDALLA, M.; JAVED, S.; AL RADI, M.; ULHAQ, A.; WERGHI, N. Video Anomaly Detection in 10 Years: A Survey and Outlook [online]. *arXiv preprint* arXiv:2405.19387. 2024. Available from: https://arxiv.org/abs/2405.19387

[21] XING, P.; LI, Z. Visual Anomaly Detection Via Partition Memory Bank Module and Error Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. 2023, vol. 33, no. 8, pp. 3596–3607. ISSN 1051-8215. Available from: https://arxiv.org/abs/2209.12441

[22] GAO, W.; WANG, X.; WANG, Y.; JING, X. Dual-Stream Attention-Enhanced Memory Networks for Video Anomaly Detection. *Sensors*. 2025, vol. 25, no. 17, article 5496. ISSN 1424-8220. Available from: https://doi.org/10.3390/s25175496

[23] LV, H.; CHEN, C.; CUI, Z.; XU, C.; LIN, Y.; YANG, J. Learning Normal Dynamics in Videos with Meta Prototype Network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15425–15434.

[24] ALI, M. L.; ZHANG, Z. The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*. 2024, vol. 13, no. 12, article 336. ISSN 2073-431X. Available from: https://doi.org/10.3390/computers13120336

[25] ALINEZHAD NOGHRE, G.; DANESH PAZHO, A.; VEMPATI, V.; KATARIA, A.; TABKHI, H. Understanding the Challenges and Opportunities of Pose-based Anomaly

Detection. In: *iWOAR 2023: 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*. 2023.

[26] WANG, Y.; CHEN, Y.; YEO, C. K. Enhancing Weakly Supervised Video Anomaly Detection with Object-Centric Features. *Information*. 2025, vol. 16, no. 12, article 1042. ISSN 2078-2489. Available from: https://doi.org/10.3390/info16121042

[27] PANG, G.; SHEN, C.; CAO, L.; VAN DEN HENGEL, A. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*. 2021, vol. 54, no. 2, pp. 1–38. ISSN 0360-0300.

[28] RUFF, L.; KAUFFMANN, J. R.; VANDERMEULEN, R. A.; MONTAVON, G.; SAMEK, W.; KLOFT, M.; DIETTERICH, T. G.; MÜLLER, K.-R. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*. 2021, pp. 1–40. ISSN 0018-9219.

[29] KIRICHENKO, L.; RADIVILOVA, T.; SYDORENKO, B.; YAKOVLEV, S. Detection of Shoplifting on Video Using a Hybrid Network. Computation. 2022, vol. 10, no. 11, article 199. ISSN 2079-3197. Available from: https://doi.org/10.3390/computation10110199

[30] MUNEER, I.; SADDIQUE, M.; HABIB, Z.; MOHAMED, H. G. Shoplifting Detection Using Hybrid Neural Network CNN-BiLSTM and Development of Benchmark Dataset. Applied Sciences. 2023, vol. 13, no. 14, article 8341. ISSN 2076-3417. Available from: https://doi.org/10.3390/app13148341

[31] NAZIR, A.; MITRA, R.; SULIEMAN, H.; KAMALOV, F. Suspicious Behavior Detection with Temporal Feature Extraction and Time-Series Classification for Shoplifting Crime Prevention. Sensors. 2023, vol. 23, no. 13, article 5811. ISSN 1424-8220. Available from: https://doi.org/10.3390/s23135811

Appendices

Appendix A. Complete CLIP Prompt List

The full bank of 17 CLIP text prompts used for human-object interaction scoring is also given in Section 2.4.2 in the body of the thesis. It is repeated here for reference.

Appendix A.1 Shoplifting prompts (8)

1. "a person concealing merchandise under clothing"
2. "a person hiding items in a bag"
3. "a person looking around suspiciously while holding items"
4. "a person quickly grabbing items off a shelf"
5. "a person removing security tags from products"
6. "a person stuffing items into pockets"
7. "a person crouching near store shelves suspiciously"
8. "a person acting nervously in a retail store"

Appendix A.2 Normal-behaviour prompts (5)

9. "a person shopping normally in a store"
10. "a person browsing items on a shelf"
11. "a person putting items in a shopping cart"
12. "a person reading a product label"
13. "a person walking through a store aisle"

Appendix A.3 Object-context prompts (4)

14. "a handbag or backpack near store shelves"
15. "merchandise items in a person's hand"
16. "a shopping basket held by a person"
17. "clothing covering hidden items"

All 17 prompts were defined and frozen prior to any training or test-set evaluation. No prompt was added, removed, or modified after the pipeline began producing results, ruling out prompt-tuning on the test set as an explanation for the observed performance.

Appendix B. Detailed Training Hyperparameters

Appendix B.1 Frame-level perception (YOLOv11-nano)

Table 8 Hyperparameters of the YOLOv11-nano frame-level perception backbone

Parameter	Value
Model	yolo11n.pt (Ultralytics, COCO-pretrained)
Input resolution	640 (longest side)
Confidence threshold	0.35
IoU threshold (NMS)	0.5
Person class ID (COCO)	0
Fine-tuning	None (frozen, inference-only)
Output	bounding boxes, confidences, class labels

Appendix B.2 Frame-level perception (YOLOv11-Large / YOLOv26-Large classification heads — supplementary baselines)

Table 9 Hyperparameters of the YOLOv11-Large and YOLOv26-Large supplementary classification baselines

Parameter	Value
Models	yolo11l-cls.pt, yolo26l-cls.pt
Input resolution	224 × 224
Epochs	50
Batch size	16
Optimiser	AdamW
Initial LR	3×10^{-3}
LR schedule	cosine annealing
Augmentation	RandAugment, default Ultralytics pipeline
Seeds	41, 42, 43

Appendix B.3 CLIP-HOI feature extractor

Table 10 Configuration of the CLIP-HOI feature extractor

Parameter	Value
Backbone	open_clip ViT-B/32 (OpenAI weights)
Fine-tuning	None (frozen)
Person crop padding	10 % per side
Prompt-bank size	17 (8 shoplifting + 5 normal + 4 object-context)
Per-frame output	1047-dim (512 global + 512 person + 5 spatial + 17 prompts + 1 contrastive)
Combined feature into temporal model	1050-dim (1047 CLIP-HOI + 3 YOLO person-detection statistics: count, mean confidence, std confidence)

Appendix B.4 Temporal models (GRU / Transformer / BERT)

Table 11 Architecture and window settings of the temporal encoders (GRU, Transformer, BERT)

Parameter	GRU	Transformer (scratch)	BERT
Hidden / model dim	256	256	768
Layers	2	4 (encoder)	12 (pretrained, fine-tuned)
Attention heads	-	8	12
FFN expansion	-	4	4
Dropout	0.2	0.2	0.2
Pre-norm	-	True	(built-in)
Pooling	last hidden	CLS token	pooled CLS output
Window length L	32 frames	32 frames	32 frames
Window stride	8 frames	8 frames	8 frames

Appendix C. Per-class metric tables

The results of the detailed evaluation of the classes per the two strongest clip-level configurations tested on the Kipshidze test split (28 clips: 14 normal, 14 shoplifting) are shown in this appendix. The idea is to report precision, recall, F1-score and support each of the classes on its own at the operating threshold of the validation-selected operating threshold.

Appendix C.1. Per-class results of the best GRU-based configuration

Model: YOLO person-detection statistics + CLIP-HOI features + GRU + Prototype Memory
Experiment seed: 42 (single-run exploratory)
Test ROC-AUC: 0.9541
Test F1 (positive class): 0.8276
Operating threshold τ: 0.76

Table 12 Per-class results of the best GRU-based clip-level configuration

Class	Precision	Recall	F1-score	Support
Normal	0.8462	0.7857	0.8148	14
Shoplifting	0.8000	0.8571	0.8276	14
Macro average	**0.8231**	**0.8214**	**0.8212**	**28**

Appendix C.2. Per-class results of the best BERT-based configuration

Model: YOLO person-detection statistics + CLIP-HOI features + BERT + Prototype Memory
Experiment seed: 42 (single-run exploratory)
Test ROC-AUC: 0.9592
Test F1 (positive class): 0.9286
Operating threshold τ: 0.08

Table 13 Per-class results of the best BERT-based clip-level configuration

Class	Precision	Recall	F1-score	Support
Normal	0.9286	0.9286	0.9286	14
Shoplifting	0.9286	0.9286	0.9286	14
Macro average	**0.9286**	**0.9286**	**0.9286**	**28**

Appendix C.3. Short comparison note

The results of the classes indicate that the two configurations have a dissimilar profile of trade-offs. BERT + Memory has the better overall ranking quality on the test split (ROC-AUC 0.959 vs. 0.954), and is also more balanced between the two classes, all per-class metrics equal 0.929, yielding the largest macro-F1 of 0.929 in this thesis. The GRU + Memory saved checkpoint achieves macro-F1 0.821 with a more conservative operating point ($\tau = 0.76$, P = 0.800, R = 0.857), while BERT + Memory selects a much lower threshold ($\tau = 0.08$) and achieves balanced precision and recall on both classes. The two configurations can thus be chosen depending on deployment goals: GRU + Memory is the lighter model with a more conservative threshold, while BERT + Memory provides the most balaced operating point at the cost of substantially more parameters and inference time.

Appendix D. AI tools usage statement

Artificial Intelligence (AI) is used in a limited way, primarily to support discussions related to coding, clarify technical issues and to help understand the results of the experiments. The author conducted all experiments, debugging, verification of results, analysis, and interpretation of the results. No AI tools have been used to generate results, manipulate data, or impact on scientific content. All metrics, tables, confusion matrix, benchmarks and conclusions were generated using the developed experimental pipeline and manually checked by the author. The author is responsible for the accuracy, integrity, originality and final presentation of the work.