

Contrastive and Generative Vision Language Models for Detecting Anomalies and Violence in Human Monitoring

ホセ アレハンドロ アベジャネダ ゴンザレス

<https://hdl.handle.net/2324/7396250>

出版情報 : Kyushu University, 2025, 博士 (情報科学) , 課程博士
バージョン :
権利関係 :



Contrastive and Generative Vision Language Models for Detecting Anomalies and Violence in Human Monitoring



Jose Alejandro Avellaneda Gonzalez

Graduate School of Information Science and Electrical Engineering

Kyushu University

A thesis submitted for the degree of

Doctor of Information Science

September 2025

Abstract

Anomaly and violence detection in human monitoring is a critical area of research with wide-ranging applications in surveillance, healthcare, and human-robot interaction. These tasks involve identifying rare, context-dependent events, often in the absence of labeled data. Recent advances in Vision-Language Models (VLMs) have introduced promising tools for learning from multimodal content without supervision. However, existing contrastive VLMs such as CLIP rely on massive image-text datasets and are not designed for one-class learning, making them unsuitable for small-scale anomaly detection. On the other hand, generative VLMs, while capable of semantic interpretation, inherit social and contextual biases from large language models, limiting their reliability in sensitive applications such as violence detection.

To address these limitations, this thesis proposes an improved contrastive VLM approach that enables multimodal self-supervised learning for anomaly detection when only a small amount of normal data is available for training. To this end, we first address the limitations of existing methods in anomaly detection by introducing CMSSL-AD (**C**ross-**M**odal **S**elf-**S**upervised **L**earning for **A**nomaly **D**etection), a contrastive VLM method designed for one-class learning in human monitoring. It maximizes mutual information between image and text views, capturing context-dependent patterns without relying on anomaly labels. This framework improves the structural coherence of normal samples by aligning multimodal representations during training.

Then, we propose M3VAD (**M**ulti**M**odal **M**ulti**V**iew Self-**S**upervised Learning for **A**nomaly **D**etection), a method that extends this contrastive paradigm by incorporating both multimodal and multiview objectives. The multimodal objective enhances alignment between image and text to improve the clustering of normal behavior, while the multiview objective increases representation consistency within each modality. When trained under constrained conditions on real-world anomaly datasets, M3VAD outperforms CNN-based and ViT-based CLIP variants, achieving ROC-AUC improvements of up to 0.174.

Building on the same vision-language modeling perspective, the final contribution of this thesis focuses on violence detection. To address the limitations of generative VLMs in sensitive classification scenarios, this thesis introduces VIVID (**V**ision-**L**anguage **I**ntegration for **V**iolence **I**dentification and **D**etection), a retrieval-augmented

framework for zero-shot violence detection. Instead of relying solely on generative outputs, VIVID incorporates external definitions of violence through a context-aware retrieval mechanism. These definitions are then aligned with the visual content during inference to support accurate classification. The method also includes a keyframe reordering step to prevent semantic loss during representation. Without task-specific fine-tuning, VIVID achieves the highest F1-score among generative VLM-based models on four out of five benchmark datasets, with performance gains of up to 0.131.

In summary, this thesis presents a unified framework for anomaly and violence detection in human monitoring using contrastive and generative VLMs. CMSSL-AD and M3VAD enable multimodal contrastive learning under limited supervision, enhancing detection accuracy in one-class scenarios. VIVID extends generative models with retrieval-based reasoning to support bias-aware zero-shot violence classification. Together, these contributions advance the state of multimodal learning by improving robustness, interpretability, and generalization in both general and sensitive monitoring settings.

Acknowledgments

I would like to express my sincere gratitude to the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan for awarding me the scholarship that enabled me to pursue my doctoral studies in Japan. I am also grateful to Kyushu University, the Graduate School of Information Science and Electrical Engineering (ISEE), Suzuki Laboratory, and Professor Einoshin Suzuki for providing the academic environment and resources that made this research possible.

I would like to thank Professor Tetsu Matsukawa for his assistance with technical matters related to computer systems, software, and network configurations, as well as for his perspective on the revisions of various research materials. I am especially thankful to my fellow PhD students for sharing their materials, code, and experimental frameworks, which served as a valuable foundation for my own work.

I am also grateful to my friends in Japan, who provided not only companionship but also a sense of home in a new and unfamiliar country. Their support outside of academic life offered me the balance and perspective needed to overcome the many challenges I faced.

I would also like to express my deepest gratitude to Alejandra Rojas for her patience, encouragement, and constant support. Her presence brought strength and stability, especially during the most demanding and uncertain periods, and her understanding allowed me to remain focused and continue moving forward.

I wish to thank my family for their patience, love, and encouragement throughout these years. Despite the distance and time difference, they always made me feel close and supported, allowing me to maintain strong family ties while pursuing my academic goals. Their continued presence gave me the strength to persevere and bring this work to completion.

Finally, I would like to acknowledge my own dedication and resilience throughout this long and demanding process. Completing this thesis required sustained effort and determination, and I am grateful to have overcome each obstacle along the way.

Contents

Abstract	i
Acknowledgments	iii
Contents	v
List of Figures	ix
List of Tables	xii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Related Works	5
1.3.1 Anomaly Detection	5
1.3.1.1 Traditional Methods: Classification, Reconstruction, and Clustering	5
1.3.1.2 Contrastive Vision-Language Models	6
1.3.2 Violence Detection	7
1.3.2.1 Traditional and Deep Learning Methods	7

1.3.2.2	Generative Vision-Language Models	8
1.4	Our Contribution	10
1.5	Thesis Organization	12
2	Cross-Modal Self-Supervised Feature Extraction for Anomaly Detection in Human Monitoring	13
2.1	Overview	13
2.2	Problem Formulation	14
2.3	Proposed Method	15
2.3.1	Overview	15
2.3.2	Relating modalities	16
2.3.3	Learning an anomaly detector	20
2.4	Experiments	20
2.4.1	Datasets	20
2.4.2	Design of the experiments	21
2.4.3	Experimental results	23
2.5	Conclusion	30
3	Detecting Anomalies in Human Monitoring Based on Multimodal Multi- view Self-Supervised Learning	31
3.1	Overview	31
3.2	Problem Formulation	32
3.3	Proposed Method	35
3.3.1	Overview	36
3.3.2	Multimodal Objectives	37
3.3.2.1	Cross-modal Objective	38
3.3.2.2	Prototypical Objective	40

3.3.3	Multiview Objectives	41
3.3.4	Objective Loss	43
3.3.5	Anomaly Score	45
3.4	Experiments	46
3.4.1	Datasets	47
3.4.2	Experimental setup	48
3.4.2.1	Detailed Architecture	48
3.4.2.2	Parameter setting	50
3.4.2.3	Compared Methods	52
3.4.3	Results and analysis	58
3.4.4	Component Analysis and Evaluation	68
3.4.4.1	Impact of Support Set Size on Stability	68
3.4.4.2	Effect of Clustering Refinement on Anomaly Detection	71
3.4.4.3	Effectiveness of Multimodal and Unimodal Objectives	72
3.5	Conclusion	78
4	Leveraging Vision Language Models for Understanding and Detecting Violence in Videos	79
4.1	Overview	79
4.2	Problem formulation	80
4.3	Proposed method	81
4.3.1	Overview	81
4.3.2	Key-frame selection	83
4.3.3	Violence detection	86
4.3.3.1	Modality translator	87

CONTENTS

4.3.3.2	Definition Retriever	88
4.3.3.3	Output Generator	92
4.4	Experiments	93
4.4.1	Datasets	93
4.4.2	Implementation details	95
4.4.3	Experimental results	96
4.4.3.1	Comparison with LLM-based methods	96
4.4.3.2	Comparison with traditional hand-crafted features and deep learning-based methods	99
4.4.3.3	Ablation study	100
4.5	Conclusions	103
5	Conclusions and Future Work	104
5.1	Conclusions	104
5.2	Future Work	105
	Bibliography	107
	Published Papers	124

List of Figures

2.1	The architecture of the CMSSL-AD training.	16
2.2	Comparison of ROC curves, PR curves, ROC-AUC, and AUPRC scores on the two datasets.	24
2.3	2D Visualization with t-SNE [1] of the cross-modal features extracted by CMSSL-AD and the features extracted by VGG16+Word2Vec pre- trained models [2,3] for the HAM test split [4]	25
2.4	Examples of images and their captions generated by Denscap [5] ex- tracted from the HAM dataset [4]. The rectangles show the extracted regions. (a) and (b) are test images which contain an anomaly of a person riding a bicycle in a room. (a) is considered a single-region anomaly since the woman and the bicycle are contained in the ex- tracted region.	28
2.5	Training and test images captured on different days from the HAM dataset. (a) is a training sample of a woman with a teddy bear and its caption. (b) is a test sample of a woman with a teddy bear. (e) and (d) are the heatmap and the wrong localization of the highest activation point obtained by [6], respectively.	28

LIST OF FIGURES

2.6	Comparison of the ROC-AUC between our proposed method (a) and Dong et al. [6] (b) with different amounts of data (5 trials per amount of data).	29
3.1	Pipeline of M3VAD.	35
3.2	Multi-modal objectives.	38
3.3	Multi-View objectives.	44
3.4	Comparison of ROC curves on the four datasets.	56
3.5	Comparison of PR curves on the four datasets.	57
3.6	2D Visualization with t-SNE [1] of the features extracted by (a) M3VAD, (b) ScatSimCLR [7], and (c) CLIP [8] for the HAM test split [4]. . .	59
3.7	Comparison of the attention maps for normal images (a), (b), and (c) for HAM, BM, and AnoVG, respectively.	60
3.8	Comparison of the attention maps for abnormal images for (a) HAM, (b) BM, and (c, d) AnoVG.	61
3.9	Comparison in terms of ROC-AUC and number of trainable parameters	63
3.10	Examples of abnormal events from the UCSDped2 [9] dataset. A skater is seen on the pedestrian walkway, highlighted by a red arrow. .	65
3.11	Examples of normal and abnormal events from the UCSDped2 [9] dataset along with their captions. On the left: a normal event with pedestrians on the walkway. In the middle: an abnormal event with a skater on the walkway. On the right: an abnormal event with a bicycle being ridden on the walkway.	67
3.12	Distributions of the ROC-AUC score on the datasets for different support sets sizes.	70

LIST OF FIGURES

3.13 Comparison with OCL methods in terms of ROC curves and ROC-AUC scores.	73
3.14 Training and test images from the HAM dataset taken on different days. On the left: a training sample showing a woman with a teddy bear on the table. On the right: a test sample of the same woman with the teddy bear on the table. The test image’s caption is inaccurate but still relates to the woman and the teddy bear.	76
3.15 Comparison of anomaly scores in terms of ROC curves and ROC-AUC scores.	77
4.1 Architecture of the proposed method.	84
4.2 Example of handling bias.	91
4.3 Example in which VIVID successfully handles the biases of the LLMs.	98
4.4 Example in which the inclusion of a violent-related definition improves performance compared to the baseline methods.	98
4.5 Example where occlusion caused by the pool table lights partially blocking the violent scene, along with lack of illumination, resulted in the LLM failing to detect the violent scene correctly.	99

List of Tables

2.1	Comparison of the methods for both datasets	27
3.1	Datasets summary.	48
3.2	Comparison in terms of ROC-AUC and AUPRC scores.	55
3.3	Mean and Variance of ROC-AUC Scores for 90% Support Set Across Datasets.	70
3.4	Ablation study of the key components of M3VAD.	75
4.1	Dataset description.	94
4.2	Comparison with LLM-based methods in terms of Accuracy and F1- Score.	97
4.3	Comparison with hand-crafted features and deep learning-based meth- ods in terms of accuracy. A hyphen indicates that the results are not provided in the corresponding paper.	101
4.4	Comparison of our method's variants in terms of Accuracy.	102

Chapter 1

Introduction

1.1 Background

Anomaly detection is a critical area of research in human monitoring and visual content analysis, with applications in public safety, healthcare, and human–robot collaboration [10–13]. The goal is to identify behaviors or patterns that deviate from expected norms, under conditions where anomalies are rare, context-dependent, and difficult to label. In practical scenarios such as surveillance systems, early and accurate detection of abnormal events can enable timely intervention and reduce risk.

A particularly important case of anomalies involves violent behaviors, such as physical assaults or aggressive encounters [14]. Although these events are sparse and ambiguous, they can have serious consequences in safety-critical settings. From a technical perspective, detecting such instances presents similar challenges to general anomaly detection: limited data, high intra-class variation, and the need for fine-grained contextual understanding [15].

Recent advances in Vision Language Models (VLMs) offer a promising paradigm for addressing these challenges by jointly learning from visual and textual information.

VLMs integrate visual and language modalities into shared representation spaces, allowing models to capture semantic relationships and contextual dependencies that are often difficult to infer from visual data alone [16]. Rather than relying solely on supervised learning or single-modality self-supervised learning, VLMs incorporate diverse training strategies such as contrastive alignment, masking, generative modeling, or pretrained backbone mappings. This flexibility has enabled VLMs to achieve strong performance across various vision-language tasks, including classification, captioning, retrieval, and complex reasoning. In human monitoring contexts, VLMs are particularly attractive due to their potential to handle limited supervision, support semantic interpretation, and enhance generalization across diverse scenarios.

Among the different VLM paradigms, two families are particularly relevant for human monitoring tasks addressed in this thesis. Contrastive VLMs, such as CLIP [8], align paired image-text data using contrastive objectives, learning a joint embedding space in which semantically related visual and textual inputs are close together. This virtue makes them promising for tasks like anomaly detection. However, they typically require large-scale datasets and are not designed for one-class learning, which limits their applicability in small-scale scenarios. In contrast, generative VLMs, such as BLIP-2 [17] and InstructBLIP [18], process visual inputs and generate textual outputs using a large language model (LLM). These models are designed to interpret and describe visual content. They are well-suited for tasks such as violence detection, where semantic understanding and interpretability are essential. However, they often inherit social and contextual biases from their language components, making their outputs less reliable in sensitive environments [19, 20].

This thesis builds on these observations to investigate contrastive and generative VLMs as complementary tools for detecting anomalies and violence in human monitoring. Despite the growing use of VLMs in vision-language tasks, their application to

human monitoring remains limited by challenges in context reasoning, data scarcity, and ethical reliability. These limitations are examined in the following section.

1.2 Motivation

Human monitoring tasks such as anomaly detection and violence detection remain highly challenging due to the rarity, ambiguity, and context sensitivity of the events of interest [20–23]. In both domains, behavior that may seem abnormal or threatening in one setting may be entirely acceptable in another. For example, taking someone else’s bag may appear inappropriate or suspicious in a workplace, especially if the individuals are not closely acquainted. In contrast, the same action may be entirely normal within a household, such as when a family member picks up a spouse’s or sibling’s belongings. This fact illustrates how identical actions can be interpreted very differently depending on social roles and situational context [14]. These subtleties demand models that go beyond surface-level appearance and incorporate contextual understanding into behavior recognition.

An important challenge in anomaly and violence detection is the limited availability of annotated data. Anomalies are unpredictable and often cannot be comprehensively labeled in advance, including violent incidents, which are infrequent, difficult to document, and ethically sensitive, which further complicates dataset construction. This scarcity constrains the applicability of conventional supervised approaches and highlights the need for models that can generalize from unlabeled data. Furthermore, both domains require systems that can reason about context to differentiate between benign and unexpected behaviors that may appear visually similar.

In anomaly detection, reconstruction-based approaches, such as autoencoders and generative models, often fail to capture local irregularities or relational context within

a scene [6, 24–30]. Other methods that rely on unimodal pre-trained visual features struggle to generalize across diverse settings [4, 6, 31]. More recently, VLMs such as CLIP [8] have shown promise in learning robust multimodal representations by aligning image and text features. However, these models require large-scale paired datasets and are not designed for situations where only normal data is available in limited quantity. These constraints limit their applicability to real-world anomaly detection settings characterized by small-scale and one-class learning conditions.

In violence detection, conventional methods relying on handcrafted features or supervised visual classifiers have limited robustness across varying environments, camera angles, and video resolutions [32–37]. Deep learning models have improved performance by learning directly from visual data [15, 38–41], yet they remain constrained by the scarcity and censorship of violent content. The diversity of violent actions and the ethical concerns surrounding their depiction further restrict dataset availability, reducing generalizability in practical deployments.

Generative VLMs, such as BLIP-2 and InstructBLIP [17, 18], offer a promising alternative by allowing models to produce natural language descriptions of visual content. This capability is particularly useful for tasks like violence detection, where interpretability and semantic reasoning are crucial. However, generative VLMs inherit biases from their pretraining data and language models [19, 20]. As a result, they may misclassify complex or socially sensitive scenes. For instance, a VLM might label a video of a police officer restraining a suspect as non-violent due to encoded assumptions about authority, even though the scene involves physical force that may reasonably be interpreted as violent [14]. Such misinterpretations highlight the importance of improving contextual reliability and ethical sensitivity in vision-language modeling.

Taken together, these challenges point to the need for vision-language approaches that address shared problems in human monitoring while adapting to task-specific con-

straints. This thesis is motivated by the goal of developing VLM-based methods for anomaly and violence detection that operate effectively under limited supervision and in context-sensitive settings.

1.3 Related Works

1.3.1 Anomaly Detection

1.3.1.1 Traditional Methods: Classification, Reconstruction, and Clustering

Research in anomaly detection has relied on three main methodological paradigms: classification-based, reconstruction-based, and clustering-based approaches.

Classification-based methods aim to learn classifiers from labeled training data and assign test data to either normal or anomalous categories. Multimodal extensions of these methods have demonstrated improved performance over unimodal variants [21–23], but typically rely on pre-trained unimodal feature extractors. As a result, they fail to model contextual dependencies across modalities. A notable exception is CMAD [42], which learns to align representations across modalities to detect cross-modal inconsistencies. However, CMAD requires labeled training data, which limits its usefulness in real-world anomaly detection settings where anomalies are rare and diverse.

Reconstruction-based methods detect anomalies by measuring reconstruction errors. Techniques based on autoencoders [24–27] and GANs [28–30, 43, 44] can detect visual anomalies but struggle with local or contextual variations. More recent work has applied self-supervised inpainting frameworks [6, 45] that reconstruct missing image regions to learn contextual information. GIAD [6], for example, improves contextual sensitivity by masking and restoring semantically salient regions. However, it does not

incorporate language information and therefore cannot leverage cross-modal semantics.

Clustering-based methods group normal training data and detect outliers based on their deviation from learned clusters [46–49]. Some approaches use hierarchical clustering, e.g., BIRCH [50], and deep captioning features [5] to detect region-level anomalies [4, 31]. While effective, these systems rely on unimodal pre-trained models, e.g., VGG-16 [2], Word2Vec [3], that may lack coherence across modalities. This misalignment limits their ability to capture joint visual-linguistic context, which is often crucial in human behavior analysis.

1.3.1.2 Contrastive Vision-Language Models

SSL allows models to learn useful representations without relying on labeled training data. Among SSL techniques, contrastive learning methods have been particularly influential, including unimodal approaches such as SimCLR [51], NNCLR [52], and CURL [53]. These models optimize embeddings by bringing semantically similar examples (positive pairs) closer together and pushing dissimilar ones (negative pairs) apart. While effective at learning unimodal representations, whether visual or linguistic, these methods fail to account for multimodal contextual information critical in human monitoring applications [54].

VLMs, such as CLIP [8] and ALIGN [55], extend contrastive learning to multimodal settings. These models learn joint representations by aligning image and text embeddings using large-scale paired datasets, e.g., CLIP is trained on 400 million image-text pairs. Their semantic alignment capabilities offer potential advantages in anomaly detection, where interpreting context and context-dependent visual-linguistic relationships is essential. However, their effectiveness depends heavily on the scale

and diversity of training data.

These large-scale models are not well suited for domain-specific or small-scale scenarios, which are typical of anomaly detection in surveillance or healthcare contexts. Even in one-class learning settings, where only normal data is used for training, acquiring enough diverse samples remains a challenge. Moreover, contrastive VLMs are trained for broad visual-linguistic alignment using large-scale paired data, but they are not optimized to model the contextual dependencies or structured relationships within individual scenes typically needed for anomaly detection.

Some recent works have attempted to adapt contrastive SSL to smaller-scale domains. For example, ScatSimCLR [7] replaces the standard encoder with a handcrafted invariant network and trains a small adapter module, improving sample efficiency. Yuan et al. [56] introduce multimodal contrastive objectives and tag supervision, but focus only on visual representation learning. These approaches demonstrate promising adaptations of contrastive frameworks but do not fully exploit the joint semantic representations available through vision-language alignment.

1.3.2 Violence Detection

1.3.2.1 Traditional and Deep Learning Methods

Violence detection has traditionally relied on manual feature extraction methods, which isolate spatial, temporal, or spatiotemporal characteristics from video data to classify scenes as violent or non-violent [32–37]. Spatial features capture static properties such as body pose or the presence of weapons, while temporal features analyze motion patterns over time, including sudden gestures or physical contact. Spatiotemporal approaches integrate both types to recognize complex interactions across frames. Although these handcrafted features are interpretable, their performance often suffers in

diverse real-world settings due to variation in camera angles, environments, and video quality.

Deep learning methods have improved over traditional approaches by learning spatiotemporal patterns directly from raw video data [15,38–41]. Convolutional and recurrent neural networks are commonly used to capture local appearance and long-range temporal dependencies, allowing for more flexible representations of violent behavior. Despite these advances, deep models face persistent challenges. Violence is inherently rare and context-sensitive, and the limited availability of annotated video datasets constrains model generalization. Furthermore, collecting violent footage for training is complicated by legal, ethical, and censorship-related concerns, which further reduce data diversity and scale.

As a result, even high-capacity models trained in supervised settings may overfit to narrow definitions of violence and fail to adapt to novel scenes, context-dependent interactions, or varying cultural interpretations of aggression. These limitations motivate the exploration of alternative approaches that combine visual understanding with semantic reasoning and contextual flexibility, gaps that vision-language models may help address.

1.3.2.2 Generative Vision-Language Models

Generative VLMs represent a promising direction for violence detection due to their ability to describe visual content using natural language and to provide interpretable outputs. These models integrate visual encoders with LLMs and generate captions or textual reasoning based on visual inputs. Their semantic expressiveness offers potential advantages in domains where both accurate classification and interpretability are essential.

Recent generative VLMs, such as BLIP-2 [17], InstructBLIP [18], and X-InstructBLIP [57], bridge the modality gap between vision and language by using a Querying Transformer (Q Former [17]) to connect frozen image encoders with LLMs. BLIP-2 introduces this architecture to enable language-based reasoning from visual inputs without extensive retraining. InstructBLIP [18] extends the framework through instruction tuning, improving the model’s ability to follow natural language prompts. X-InstructBLIP [57] builds on this improvement by incorporating cross-modal alignment mechanisms that support more general visual tasks, including video input, with minimal modality-specific adjustment.

Other models have been proposed to explicitly handle video-based generative tasks. Video LLaMA [58] introduces a dual-branch architecture that converts both video frames and audio signals into structured queries interpretable by LLMs, enabling multimodal temporal understanding. Video LLaVA [59] extends image-based vision language alignment to video by creating a unified feature space from sequential frames, which is then used for the subsequent instruction. LLaMA VID [19] represents each video frame using two complementary tokens, one for scene-level context and the other for detailed visual content, allowing the LLM to reason across both types of information simultaneously.

While generative VLMs offer interpretability and flexibility, they also present important limitations. First, they reflect pretraining biases from large-scale Web corpora, which can encode social and cultural assumptions that distort the interpretation of sensitive actions. Second, safety constraints designed to prevent the generation of harmful or inappropriate content can suppress relevant information. For instance, a VLM might label a video of a police officer restraining a suspect as non-violent due to encoded assumptions about authority, even though the scene involves physical force that may reasonably be interpreted as violence [14, 20]. Third, fine-tuning bias occurs when LLMs

are aligned too closely with specific video datasets during training, limiting their ability to generalize to unfamiliar or divergent scenarios [19]. These limitations highlight the need for models that are semantically expressive, ethically aware, and capable of context-sensitive violence detection.

Taken together, these prior works highlight both the progress and the limitations of current approaches to human monitoring. Traditional and deep learning methods often struggle with the complexity of context-dependent behaviors, especially under conditions of limited supervision. While contrastive and generative VLMs offer promising tools for capturing semantic and multimodal information, they remain constrained by data requirements, representational biases, and generalization gaps. This thesis builds on these insights to develop contrastive and generative VLM-based methods that are tailored to the practical and ethical challenges of anomaly and violence detection.

1.4 Our Contribution

This thesis develops complementary contrastive and generative VLM approaches for detecting anomalies and violence in human monitoring scenarios, where behaviors are context-dependent, data is scarce, and interpretability is critical. To address these shared challenges, we propose three VLM-based methods that leverage the joint modeling of visual and language data to support context-aware, data-efficient, and interpretable detection in low-resource and ethically sensitive settings.

First, we propose Cross-Modal Self-Supervised Learning for Anomaly Detection (CMSSL-AD) [60], a contrastive VLM approach that improves representation learning in settings where only normal behavior is available for training. CMSSL-AD maximizes mutual information between visual and textual features to enhance cross-modal alignment and capture scene-level semantics. Designed specifically for one-

class anomaly detection, it enables context-aware feature learning from limited unlabeled data by leveraging caption-guided supervision within a contrastive framework.

Building on this foundation, we introduce Multimodal Multiview Self-Supervised Learning for Anomaly Detection (M3VAD) [61], an enhanced contrastive VLM method that incorporates multiview learning objectives in addition to multimodal alignment. M3VAD jointly optimizes visual-textual correspondence and intra-modal consistency, improving both the clustering of normal instances and the robustness of learned features. It also introduces a novel anomaly scoring function that integrates prototype-based distance with cross-modal similarity, enabling more reliable detection in real-world small-scale scenarios.

Finally, we present VIVID (Vision-Language Integration for Violence Identification and Detection) [62], a generative VLM framework for interpretable violence detection. VIVID integrates vision-language models and large language models with an external knowledge base through a retrieval-augmented context mechanism, enabling contextual reasoning without task-specific fine-tuning. By comparing visual inputs with semantically grounded definitions of violence, VIVID produces explanation-driven outputs while mitigating inherited, safety-induced, and modality-alignment biases that affect generalization in socially sensitive situations.

Together, these contributions demonstrate how contrastive and generative VLM paradigms can be tailored to address the shared demands of anomaly and violence detection. Through extensive evaluation across multiple datasets, this thesis advances a unified framework for semantic reasoning in human monitoring under practical and ethical constraints.

1.5 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we delve into cross-modal self-supervised feature extraction for anomaly detection in human monitoring, presenting an overview, the problem formulation, and our proposed method, followed by detailed experimental results and a summary. Chapter 3 addresses detecting anomalies in human monitoring based on multimodal multiview self-supervised learning. This chapter includes an overview, the problem formulation, the proposed method with multimodal objectives and anomaly score calculation, and comprehensive experiments including datasets, experimental setup, results, an ablation study, and a concluding summary. Chapter 4 explores leveraging vision language models for understanding and detection of violence in videos. This chapter covers the problem formulation, the proposed method, including key-frame selection and violence detection mechanisms, and detailed experiments, followed by a summary. Finally, Chapter 5 presents the conclusions of the thesis and discusses potential future work based on the findings.

Chapter 2

Cross-Modal Self-Supervised Feature Extraction for Anomaly Detection in Human Monitoring

2.1 Overview

This chapter addresses the problem of context-sensitive anomaly detection in human behavior monitoring, building on the challenges outlined in Section 1.2. While anomaly detection has broad applications in surveillance, healthcare, and human–robot interaction, existing methods often rely on unimodal representations and struggle to capture contextual relationships across modalities.

As reviewed in Section 1.3, traditional classification, reconstruction, and clustering approaches face notable limitations in real-world settings, especially when anomalies are rare, ambiguous, and dependent on scene context. Multimodal anomaly detection methods have emerged to address these limitations, but many of them still depend on

unimodal pre-trained features, which may fail to yield coherent visual and linguistic representations. For instance, recent methods that combine image features and caption texts [4, 31] do not model cross-modal interactions directly, limiting their ability to learn semantically grounded anomaly representations.

To overcome these challenges, this chapter introduces Cross-Modal Self-Supervised Learning for Anomaly Detection (CMSSL-AD), a contrastive VLM that jointly trains visual and textual encoders to maximize mutual information between modalities. By leveraging unlabeled data and aligning features across modalities, CMSSL-AD improves context-aware representation learning. This enriched feature space is then used to enhance clustering-based anomaly detection, enabling more accurate identification of unexpected behaviors in complex monitoring scenarios.

2.2 Problem Formulation

We adapt the target problem of Hatae et al. [31] to anomaly detection at the image level. Let the image be $\mathbf{H}_i$ with $i = 1, \dots, I$, where I is the total number of images. The j -th region caption $\mathbf{C}_{i,j}$ of $\mathbf{H}_i$ is extracted as in [4, 31, 54], using region captioning with Densecap [5]. We define an image caption as $\mathbf{C}_i = \text{CONCAT}(\mathbf{C}_{i,1}, \dots, \mathbf{C}_{i,J})$, with $j = 1, \dots, J$, where J is the number of region captions.

We tackle the problem as a one-class anomaly detection task, i.e., where there are only normal data during the training phase, as in [4, 31, 54]. We define the training data as $\mathcal{D}^{\text{tra}} = \{(\mathbf{H}_1, \mathbf{C}_1, y_1), \dots, (\mathbf{H}_n, \mathbf{C}_n, y_n)\}$, which only contains normal examples $y_i = 0, i = 1, 2, \dots, n$, where n is the number of training images. y_i and y_i^* are the class labels given to the corresponding image-caption pair. Label 0 represents the normal class while 1 the abnormal class. We use the test data defined as: $\mathcal{D}^{\text{tst}} = \{(\mathbf{H}_1^*, \mathbf{C}_1^*, y_1^*), \dots, (\mathbf{H}_m^*, \mathbf{C}_m^*, y_m^*)\}$, where m is the number of test examples.

$\mathcal{D}^{\text{tst}}$ contains normal and abnormal examples $y_i^* \in \{0, 1\}$. Intuitively, an anomaly is an image-caption pair $(\mathbf{H}_t^*, \mathbf{C}_t^*)$ never observed in the training data ($y_t^* = 1$).

We intend to detect the following kinds of anomalies.

- An **anomalous object** [31] represents an object highly dissimilar to the objects in the training data.
- An **anomalous action** [31] represents an action highly dissimilar to the actions in the training data.
- An **anomalous position** [31] represents objects located at a highly unlikely position in the training data.
- An **anomalous background** [63] is an action or an activity which is allowed in one area but forbidden in another.

These contextual anomalies refer to individual anomalous instances situated in a specific context given by the interactions between humans and objects. Therefore, addressing the challenge of capturing contextual information is crucial in anomaly detection tasks. We adopt, as evaluation measures, the area under the ROC curve (ROC-AUC), the area under the precision-recall curve (AUPRC), the F1 score, and the number of false positives/negatives (FP/FN).

2.3 Proposed Method

2.3.1 Overview

Our method consists of two steps: relating modalities and learning an anomaly detector. In the first step, we train a cross-modal feature extractor using contrastive SSL

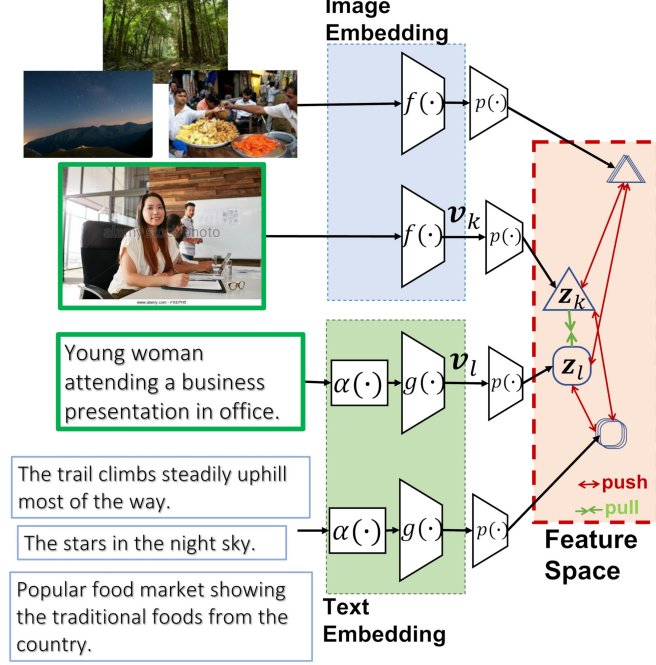


Figure 2.1: The architecture of the CMSSL-AD training.

(CMSSL-AD). In the second step, we follow Hatae et al. [31] by using BIRCH [50] to create a CF tree to model contextual information across modalities. Figure 2.1 shows the architecture of the CMSSL-AD training.

2.3.2 Relating modalities

As backbones of CMSSL-AD, we use two neural network-based encoders: $f(\cdot)$ and $g(\cdot)$ for the visual and language modality, respectively, both of which extract vector representations. In Eqs. (2.1) and (2.2), $\mathbf{v}_k$ and $\mathbf{v}_l$ are visual and language feature vectors, respectively, and Θ_f and Θ_g are their parameters. We transform caption $\mathbf{C}_i$ into a vector using a pre-trained function $\alpha(\cdot)$, which extracts contextual information from text sentences. This vector is processed by $g(\cdot)$, obtaining the final language embedding $\mathbf{v}_l$ which is directly comparable with visual embedding $\mathbf{v}_k$ in the feature

space.

$$\mathbf{v}_k = f(\mathbf{H}_i; \Theta_f) \quad (2.1)$$

$$\mathbf{v}_l = g(\alpha(\mathbf{C}_i); \Theta_g) \quad (2.2)$$

As the primary loss function, we use the normalized temperature-scaled cross-entropy loss (NT-Xent loss) function [51] because minimizing it also guarantees maximizing a lower bound on the mutual information between the input and the representation [64]. To apply the NT-Xent loss to a multi-modal scenario, we propose to adapt it by creating mini-batches which contain samples of both modalities. Since the objective is to maximize the mutual information between modalities to learn the contextual information across them, we relate both modalities in the loss function and propose to calculate it by defining an anchor as a sample of the visual or language modality (modality anchor) when the positive example is its modality counterpart. That is, if the image is the anchor, then the positive example is its caption, and vice versa. As negative examples, we use all the other images and captions in the mini-batch that do not correspond to the modality anchor.

$$\ell(\mathbf{z}_k, \mathbf{z}_l) = -\log \frac{e^{\frac{\text{sim}(\mathbf{z}_k, \mathbf{z}_l)}{\tau}}}{\sum_{q=1}^{2N} \mathbb{1}_{[q \neq k]} e^{\frac{\text{sim}(\mathbf{z}_k, \mathbf{z}_q)}{\tau}}} \quad (2.3)$$

In Eq. (2.3), **sim**, N , and $\mathbb{1}$ represent the cosine similarity function, the number of image-caption pairs in the mini-batch, and the indicator function, respectively. τ is a temperature parameter which allows scaling the cosine similarities between modalities. It is proved by Chen et al. [51] that a projection head is beneficial to define the contrastive loss. Thus, for a positive pair of examples $(\mathbf{v}_k; \mathbf{v}_l)$, the loss function is calculated over its projections into lower-dimensional space $\mathbf{z}_k$ and $\mathbf{z}_l$. The final loss

$\mathcal{L}_{NT_Xent}$ is computed across all positive pairs, both $(\mathbf{z}_k; \mathbf{z}_l)$ and $(\mathbf{z}_l; \mathbf{z}_k)$, in a mini-batch as is shown in Eq. (2.4).

$$\mathcal{L}_{NT_Xent} = \frac{1}{2N} \sum_{k,l=1}^N [\ell(\mathbf{z}_k, \mathbf{z}_l) + \ell(\mathbf{z}_l, \mathbf{z}_k)] \quad (2.4)$$

Existing contrastive learning methods critically rely on specific data augmentation to favor certain sets of features over others [64]. In traditional unimodal approaches [51–53], a positive sample is the one which belongs to the same distribution as the anchor, while the negative sample has a different distribution. These characteristics can be guaranteed by selecting the data augmentations. However, when using multi-modal data, it is not possible to control the distribution of positive or negative examples because the distribution is given by the modality itself. Thus, we need an additional loss which minimizes the distance between the two probability distributions in the feature space. We propose to use the Jensen-Shannon divergence [65] as the loss [66, 67], which is a symmetrical similarity measure between the two distributions. This means that the value of image embeddings distribution P from text embeddings distribution Q is highly similar as Q from P in the mini-batch. The JS-divergence is defined in Eq. (2.5), where KL represents the Kullback-Leibler divergence [68].

$$\mathcal{L}_{JS} = \frac{1}{2} \left(KL \left[P \parallel \frac{P+Q}{2} \right] + KL \left[Q \parallel \frac{P+Q}{2} \right] \right) \quad (2.5)$$

The primary and secondary loss functions are combined into a global loss $\mathcal{L}$ as is shown in Algorithm 1, in which the factor γ is the weight for the Jensen-Shannon loss.

Algorithm 1: SSL-FE’s main learning algorithm.**Require:** batch size N , constant τ

```
1: for sampled minibatch  $\{\mathbf{H}_k, \mathbf{C}_k\}_{k=1}^N$  do
2:   for all  $k \in \{1, \dots, N\}$  do
3:     // Visual modality
4:      $\mathbf{v}_k = f(\mathbf{H}_k)$ 
5:      $\mathbf{z}_k = p(\mathbf{v}_k)$ 
6:     // Language modality
7:      $\mathbf{v}_l = g(\alpha(\mathbf{C}_k))$ 
8:      $\mathbf{z}_l = p(\mathbf{v}_l)$ 
9:   end for
10:  Define  $\ell(\mathbf{z}_k, \mathbf{z}_l)$  as in Eq. (2.3)
11:  Compute  $\mathcal{L}_{NT-Xent}$  via Eq. (2.4)
12:  Compute  $\mathcal{L}_{JS}$  via Eq. (2.5)
13:   $\mathcal{L} = \mathcal{L}_{NT-Xent} + \gamma \mathcal{L}_{JS}$ 
14:  Update networks  $f$ ,  $g$ , and  $p$  to minimize  $\mathcal{L}$ 
15: end for
16: return encoder networks  $f(\cdot), g(\cdot)$ 
```

2.3.3 Learning an anomaly detector

Different from Hatae et al. [31], which uses a concatenation of unimodally pre-trained features, we combine the modality features extracted by the CMSSL-AD in step one. The cross-modal feature vector $\mathbf{F}$ is created by concatenating the modality features $\mathbf{v}_k$ and $\mathbf{v}_l$, which were trained simultaneously. In Eq. (2.6), the symbol $\parallel$ represents the concatenation operator.

$$\mathbf{F}_i = \mathbf{v}_k \parallel \mathbf{v}_l \quad (2.6)$$

A cross-modal feature vector $\mathbf{F}$ is judged anomalous if its distance to the closest cluster in the CF tree is greater than a predefined threshold ϕ [4, 31].

2.4 Experiments

2.4.1 Datasets

In the first step, we train our CMSSL-AD with the Conceptual Captions dataset [69]. In the second step, we evaluate the performance of our anomaly detection method on two real-world datasets: Human Activities Monitoring (HAM) [4] and Behavior Monitoring (BM) [63].

Conceptual Captions dataset [69] contains 3.3 million images with weakly-associated descriptive captions, which were automatically collected from the WWW by taking as caption the Alt-text HTML attribute. Each image contains only one caption. The dataset has 51 201 total unique words in the captions. The average number of words per caption is 10.3, the median and the standard deviation are 9.0 and 4.5, respectively.

The training and validation split consists of 3 318 333 and 15 840 image-caption pairs, respectively.

HAM [4] was constructed from video clips taken by a mobile robot which patrolled around several designated positions in a room. This dataset includes various anomalies, such as a girl riding a bike, and a man holding a teddy bear or a bag belonging to another person. The training data was divided into two parts. The first part contains 11 777 image-caption pairs. They were collected over three days, during which the robot patrolled around three different spots in the room. The second part contains 22 318 image-caption pairs. They were collected over six days, during which the robot patrolled around six different spots in the room. The training split contains 34 095 image-caption pairs corresponding to nine days of monitoring. The test split consists of 2631 image-caption pairs, of which 2103 and 528 belong to the normal and abnormal classes, respectively.

BM [63] was constructed from a large-scale human monitoring dataset of around 100 hours of video clips. A mobile robot navigated to locations designated by a program to monitor diverse human behaviors in the laboratory. It includes a wide range of anomalies of many human behaviors, such as eating and sleeping in the working area (defined as abnormal activities) and eating and sleeping in the resting area (defined as normal activities). It contains 5548 normal images for training, as well as 585 normal images and 106 abnormal images for testing.

2.4.2 Design of the experiments

In step one, to train our CMSSL-AD we used 15% of the data (around 500 000 image-caption pairs) from the Conceptual Captions dataset [69]. We trained it from scratch for

50 epochs by using an ADAM optimizer [70] with the following parameters: $lr=0.001$, $\beta_1=0.9$, $\beta_2=0.999$, and $\epsilon=1e-08$.

The proposed method was conceived with the intention that, during the step described in Section 2.3.3, it can be used in real-time human monitoring applications such as the one described in [4]. For that reason, we use a subset of the Conceptual Captions dataset [69], a small batch size, and an architecture with the fastest backbones as possible. We used Alexnet [71] because it is faster than similar CNN architectures such as ResNet [72], pre-trained DistilBERT model [73], which is a smaller and faster representation of BERT [74], and MLP with a single hidden layer instead of other more complicated architectures such as CNN or transformers [75].

For the Alexnet network [71] used as $f(\cdot)$, we removed the final layer to obtain 2048-dimensional $\mathbf{v}_k$. For $g(\cdot)$, the multi-layer perceptron (MLP) has one hidden layer and a 2048-dimensional output layer to obtain $\mathbf{v}_l$. For the DistilBERT model [73] used as $\alpha(\cdot)$, we concatenate the last four hidden layers, obtaining a 3072-dimensional sentence vector, which is used as the input of $g(\cdot)$. For projection head $p(\cdot)$, we selected an MLP with one hidden layer to obtain $\mathbf{z}_k$ and $\mathbf{z}_l$ as shown in Eq. (2.7), where σ is a ReLU function. This projection head is shared between text and image representations and has a 256-dimensional output layer.

$$\begin{aligned}\mathbf{z}_k &= p(\mathbf{v}_k) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{v}_k) \\ \mathbf{z}_l &= p(\mathbf{v}_l) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{v}_l)\end{aligned}\tag{2.7}$$

On the other hand, for each image, we use an online data augmentation as follows: we resize the image to 256x256 pixels. We perform a random resizing and cropping, obtaining a 224x224 pixel image which matches the input layer of the Alexnet network [71]. We perform a random horizontal flip with a probability of 50%, and finally, a color jitter with a probability of 80% or random grayscale with a probability of 20%.

Additionally, we use a batch of 256 image-caption pairs, i.e., $N=256$.

In step two, for the clustering algorithm BIRCH [50], we set the parameter $\theta = 0.05$, which represents the threshold of the diameter of a leaf node when the CF tree is built. We set the number J of region captions to 5. We implemented our approach using the Pytorch¹ framework and Python on Ubuntu 20.04 equipped with two NVIDIA GeForce RTX 3090 and an i9-7900X CPU.

We conducted two kinds of experiments to compare our method with the baselines, through the detection quality and performance. As baseline methods, we used the proposals made by Hatae et al [31], Fadjrimiratno et al. [4], Zang et al. [63], Dong et al. [6], and Radford et al [8]. The first three proposals are region-based approaches, but we only detect anomalies at the image level. For those three baseline methods, an anomalous image is defined as an image which contains at least one anomalous region as in [6] and [4]. We also keep the hyperparameter values as suggested in their original reports. For the last proposal, we apply CLIP [8] to anomaly detection. First, we extract the features by using pre-trained CLIP-RN² models, which have the most similar architecture to our proposal. Second, we train an anomaly detector as proposed in Section 2.3.2.

2.4.3 Experimental results

As shown in Fig. 2.2a and Fig. 2.2b, for the first experiments, our method achieves an ROC-AUC score of 0.969 for the HAM Dataset, which corresponds to an improvement of 0.235, 0.165, 0.135, 0.074, and 0.059 over the baseline methods Hatae et al. [31], Fadjrimiratno et al. [4], Zhang et al. [54], Dong et al. [6], and Radford et

¹<https://pytorch.org/>

²We used all the pre-trained CLIP-RN models available [8]: ResNet-50, ResNet-101, ResNet-50X4, ResNet-50X16, and ResNet-50X64. However, due to space limitation, we only report in Section 2.4.3 the best results achieved by using the ResNet-50X64 model.

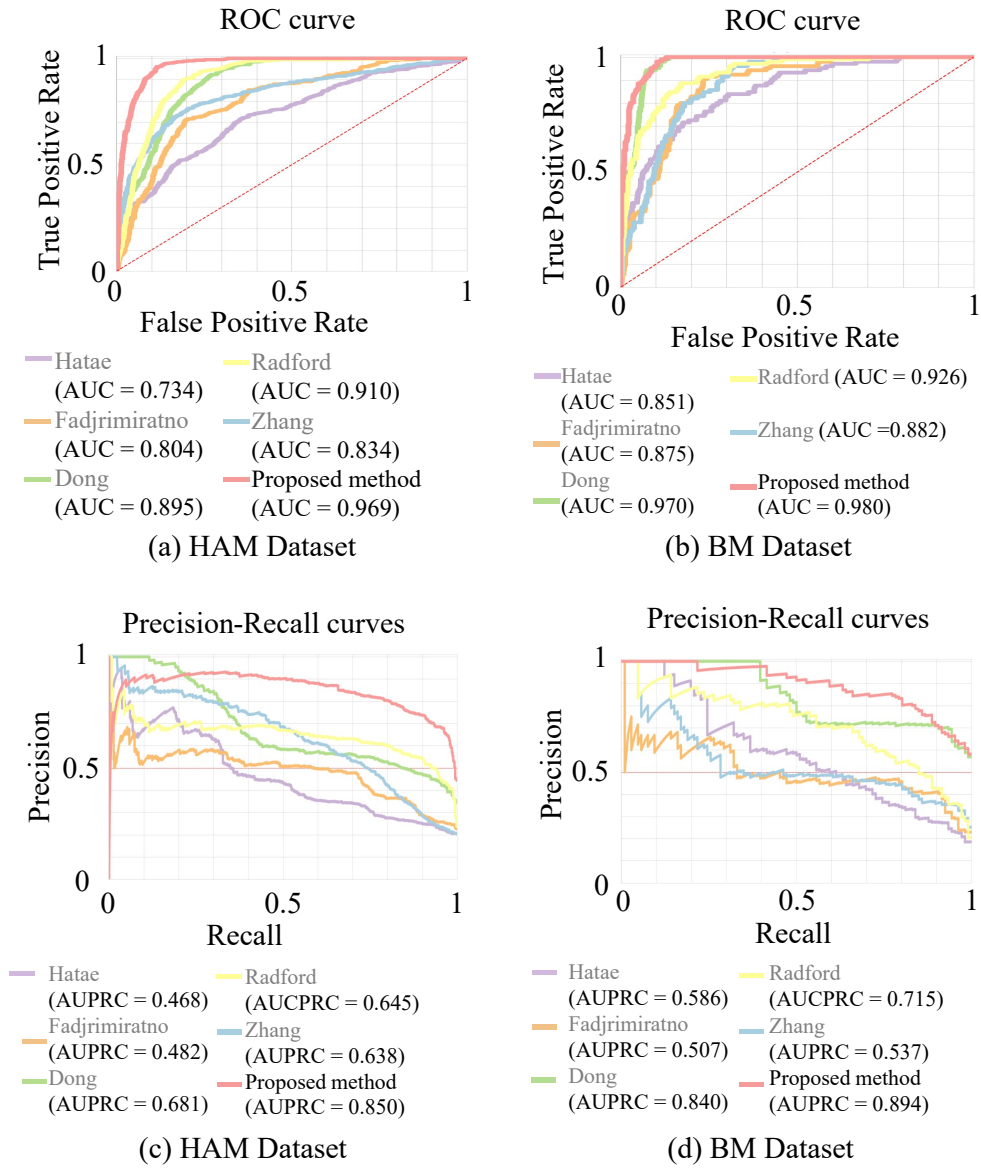


Figure 2.2: Comparison of ROC curves, PR curves, ROC-AUC, and AUPRC scores on the two datasets.

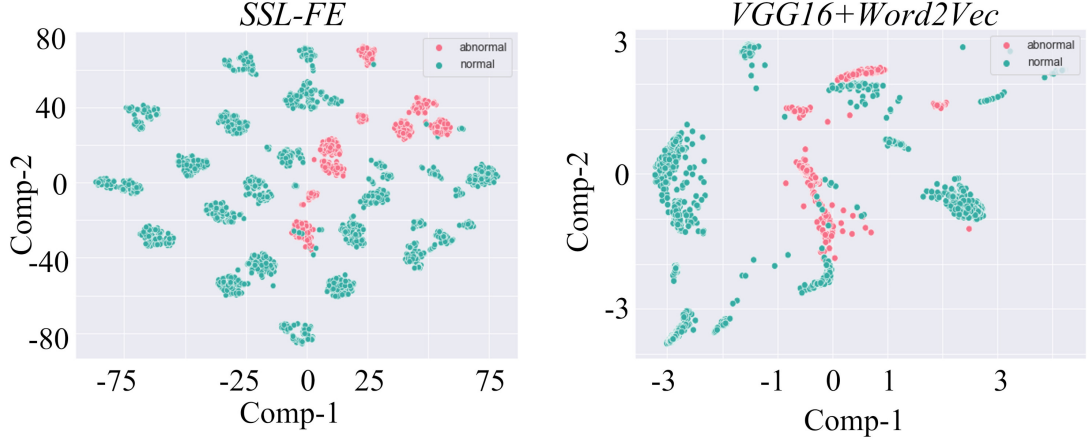


Figure 2.3: 2D Visualization with t-SNE [1] of the cross-modal features extracted by CMSSL-AD and the features extracted by VGG16+Word2Vec pre-trained models [2,3] for the HAM test split [4]

al. [8], respectively. For the BM Dataset, our method achieves an ROC-AUC score of 0.980, which corresponds to an improvement over the baseline methods of 0.129, 0.105, 0.098, 0.01, and 0.054, respectively. Our ROC curves dominate the others, indicating that our method in both datasets always outperforms the baselines.

Fig. 2.2c and Fig. 2.2d show the PR curves and the average precision score. In the PR curves for both datasets, the value of precision of our method does not decrease drastically, unlike other methods. Specifically, our method maintains a higher precision than the baselines when the recall value is closer to 1. Our method achieves an AUPRC score of 0.850 for the HAM Dataset, which corresponds to an improvement over the baseline methods Hatae et al. [31], Fadjrimiratno et al. [4], Zhang et al. [54], Dong et al. [6], and Radford et al. [8] of 0.382, 0.368, 0.212, 0.169, and 0.205, respectively. For the BM Dataset, our method achieves an AUPRC score of 0.894, which corresponds to an improvement over the baseline methods of 0.308, 0.387, 0.357, 0.054, and 0.179, respectively. These results indicate that in the HAM dataset, the abnormal class is more difficult to detect than in the BM. It should also be noted that our method better

handles abnormal examples than the baseline methods.

The reason for the good performance of our method resides in its capability to extract contextual information across modalities, which is reflected in its high-level cross-modal features. In Fig. 2.3, we show a comparison between the features extracted by our CMSSL-AD and the features extracted by combining unimodal pre-trained models used in [31] and [4]. Our cross-modal features create better-defined groups in the latent space compared to unimodal pre-trained features. Those better-defined groups help to improve the detection performance of abnormal samples.

We also compare our method with the baselines. Following Dong et al. [6], the statistics were obtained under the methods’ best performance settings. We chose the value of ϕ that has the largest geometric mean (G-Mean) for each method. Table 2.1 shows a comparison of the false positives, negatives, and statistics for both datasets. For the HAM dataset, our method outperforms the baseline methods and has the least number of false positives and false negatives. It also has the largest value in the F1-score, precision, and recall. For the BM dataset [63], the method of Fadjrimiratno et al. [4] achieved the smallest value of false negatives and therefore the largest value of recall. However, it also has the largest value of false positives and the lowest precision. These results mean that in this case, the method of Fadjrimiratno et al. tends to classify normal samples as abnormal. On the contrary, our method shows better generalization capability and a consistent behavior in the values of precision and recall. Therefore, it has the best F1-score value. Furthermore, it has the smallest value of false positives as well as the highest precision.

Fig. 2.4 shows an example in which our method outperforms the baselines which used unimodal pre-trained features. Fig. 2.4a shows an example in which ours and the baseline methods successfully detected as an anomaly. However, even when Fig. 2.4b is almost visually identical to Fig. 2.4a, Hatae et al. [31], Fadjrimiratno et al. [4], and

Table 2.1: Comparison of the methods for both datasets

	Human Activities Monitoring Dataset					Behavior Monitoring Dataset				
	FN	FP	F1-Score	Recall	Precision	FN	FP	F1-Score	Recall	Precision
Hatae	151	743	0.460	0.716	0.338	30	104	0.531	0.716	0.422
Fadjrimiratno	154	412	0.571	0.710	0.478	10	137	0.566	0.905	0.412
Zhang	124	431	0.592	0.765	0.483	21	106	0.572	0.801	0.445
Dong	42	558	0.618	0.920	0.465	21	34	0.755	0.801	0.714
Radford	54	395	0.680	0.898	0.547	12	101	0.624	0.886	0.482
Ours	20	242	0.796	0.962	0.679	19	21	0.813	0.820	0.805

Zhang et al. [54] failed to detect it as an anomaly. The reason is that those methods do not refine the feature space and have difficulties in capturing contextual information across modalities, as shown in Fig. 2.3. Ours and Dong et al. [6] do not have this limitation, but [6] is sensitive to viewpoint variations and is devised when images of the same spot are captured from different angles. An example is shown in Fig. 2.5, where Dong et al. [6] incorrectly localize the region of interest and give more relevance to regions which are less important. Our method is more robust to this problem because the language modality remains invariant to those variations, i.e., the caption of an image for the same situation should be similar regardless of the position where the image was taken, as shown in Fig. 2.5a and Fig. 2.5b.

The results of the second experiments are shown in Fig. 2.6. The values of ROC-AUC for both methods become higher as the size of the data in the training split increases. However, the results obtained by Dong et al. method [6] have lower means and bigger standard deviations for each number of examples. When in the training split there is a small number of examples, there are few variations in viewpoints, and thus the results are affected negatively. As a consequence, if a training split contains simi-

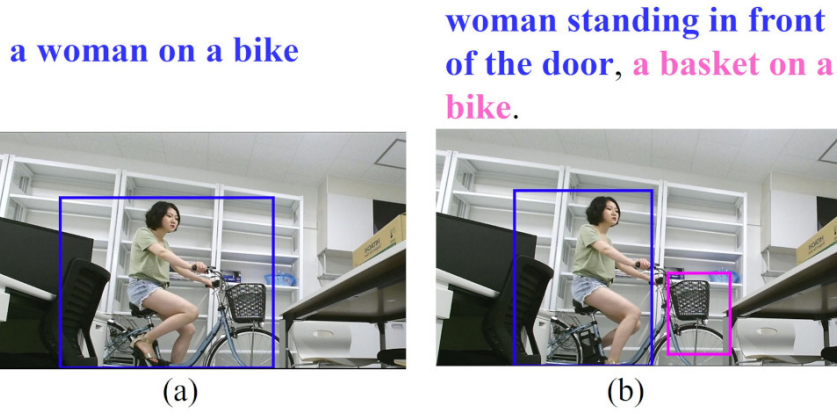


Figure 2.4: Examples of images and their captions generated by Densecap [5] extracted from the HAM dataset [4]. The rectangles show the extracted regions. (a) and (b) are test images which contain an anomaly of a person riding a bicycle in a room. (a) is considered a single-region anomaly since the woman and the bicycle are contained in the extracted region.

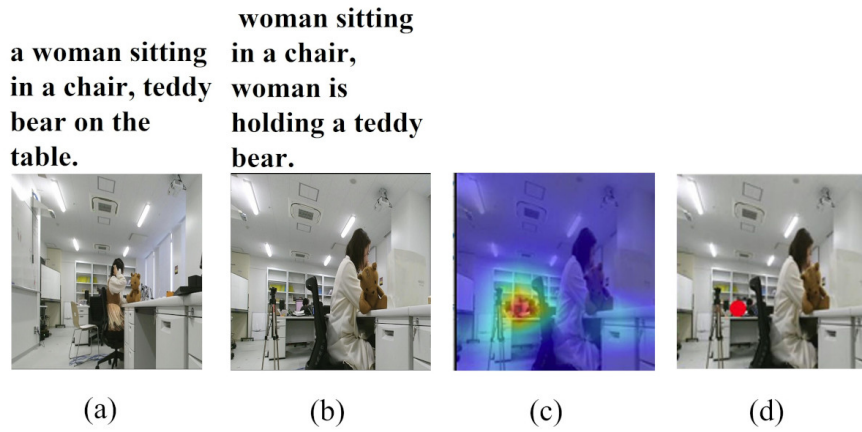


Figure 2.5: Training and test images captured on different days from the HAM dataset. (a) is a training sample of a woman with a teddy bear and its caption. (b) is a test sample of a woman with a teddy bear. (c) and (d) are the heatmap and the wrong localization of the highest activation point obtained by [6], respectively.

lar images to the test split, then the detection performance is high and the ROC-AUC value is equal to or higher than 0.8. However, if this training split contains images with different viewpoints, the ROC-AUC reaches values lower than 0.4. On the contrary, our method is more robust to this problem due to the inclusion of the language modality, as explained before. It has more consistency and obtains higher values of the mean and lower values of the standard deviation for each training split as shown in the box plots of Fig. 2.6.

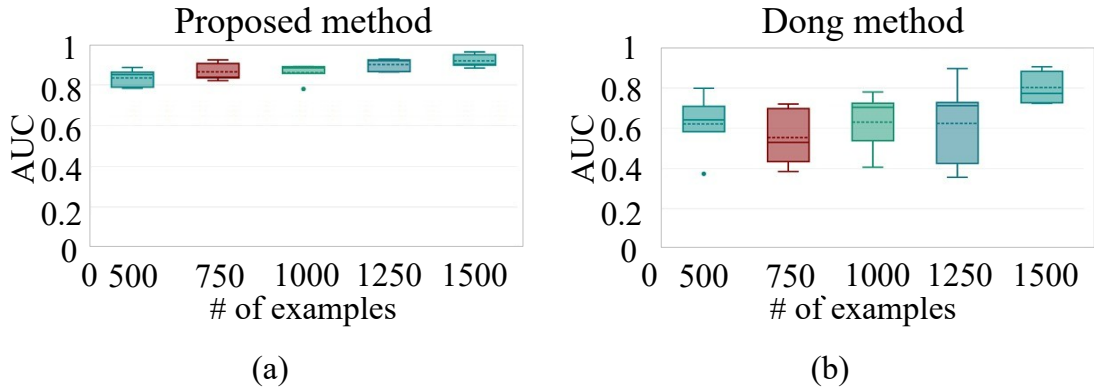


Figure 2.6: Comparison of the ROC-AUC between our proposed method (a) and Dong et al. [6] (b) with different amounts of data (5 trials per amount of data).

In terms of computational complexity, we compared our method with the CLIP-RN models [8]. Our architecture contains around 65 million parameters and is faster to train compared with the CLIP-RN simplest model (ResNet-50), which contains around 102 million parameters. We trained our model with 2 GPUs, 0.5 million examples for 50 epochs, and 256 batch sizes, which took around 7 days. On the other hand, the CLIP-RN model, which achieved the best results in our experiments (ResNet-50X64), according to Radford et al. [8], was trained with 592 GPUs, 400 million examples for 32 epochs, and 32 768 batch sizes, which took 18 days.

2.5 Conclusion

In this chapter, we propose a contrastive VLM method that extracts multimodal features that capture contextual information. Our features facilitate learning the distribution of normality for anomaly detection in human monitoring.

Experiments on two real-world robotic datasets demonstrate the superiority of our method over state-of-the-art baselines.

There are a number of directions which can be explored for future research, including (1) combining the CMSSL-AD training with the clustering process through multiple tasks; (2) proposing a cross-modal attention mechanism to discard parts of the feature vectors in each modality which do not contribute to the detection, and (3) determining the level of contribution of the image regions by further analyzing their captions and their regions.

Chapter 3

Detecting Anomalies in Human Monitoring Based on Multimodal Multiview Self-Supervised Learning

3.1 Overview

This chapter builds on the method introduced in Chapter 2 and addresses the challenges of context-sensitive anomaly detection in human behavior monitoring under small-scale training conditions, as outlined in Section 1.2. Two major challenges complicate this task: the scarcity and diversity of anomalous data, which makes it difficult to collect representative training examples, and the need to capture context-dependent interpretations of behavior. As discussed in Section 1.3, even subtle contextual cues can determine whether an action is interpreted as normal or abnormal [4].

As discussed in Section 1.3, conventional methods for anomaly detection include reconstruction-based and clustering-based approaches. While these methods have shown

potential in human monitoring, they often rely on unimodal features and fail to capture cross-modal context, limiting their applicability to settings that require semantic reasoning across images and language. These limitations motivate the development of improved multimodal approaches such as the one presented in this chapter.

To address these issues, we propose M3VAD, a contrastive VLM designed to improve representation learning in small-scale, context-sensitive settings. M3VAD combines multimodal contrastive objectives to align image and text representations, with unimodal multiview learning to enhance intra-modal robustness. This design improves the clustering of normal behaviors while maintaining semantic consistency across modalities. We further introduce a hybrid anomaly scoring function that integrates distance to normal prototypes with cross-modal similarity, enhancing detection performance. Finally, through evaluation on four human monitoring datasets, we demonstrate that M3VAD significantly outperforms baseline models, confirming its effectiveness under real-world constraints.

3.2 Problem Formulation

The problem we tackle is anomaly detection in human behavior monitoring, addressed as a one-class anomaly detection problem. Our proposal is a one-class learning method, where the model is trained on normal behavior data, and during testing, it must detect whether new, unseen instances represent normal or anomalous behavior [76]. This setup assumes that only normal data is available in the training phase, and anomalies must be identified by comparing test samples against the learned representations of normal behavior [4, 31, 54].

Let the training dataset $\mathcal{D}^{\text{tra}} = \{(\mathbf{H}_i, \mathbf{C}_i, y_i)\}_{i=1}^n$ consist of n normal image-caption pairs. Each image $\mathbf{H}_i$ represents visual information, and each caption $\mathbf{C}_i$ provides a

corresponding textual description. The label y_i indicates the class of the sample, with $y_i = 0$ for all samples, as they all represent normal behavior.

In the **training phase**, the model learns to capture the distribution of normal behaviors from positive samples only, in a shared feature space formed by both image and caption modalities [31]. The objective is to learn a representation that generalizes the normal behavior, enabling the model to detect deviations from this normality in the test phase.

In the **test phase**, the model is presented with a new image-caption pair $(\mathbf{H}_i^*, \mathbf{C}_i^*)$, where the label y_i^* is unknown. The goal is to classify the test sample as either normal ($y_i^* = 0$) or anomalous ($y_i^* = 1$) based on its deviation from the learned normal behavior as in Chapter 2. In this phase, anomalies are identified using an anomaly score which quantifies the deviation of a test sample from the learned normal behavior. If this score exceeds a predefined threshold ϕ , the test sample is classified as anomalous:

$$A(\mathbf{H}_i^*, \mathbf{C}_i^*) \geq \phi. \quad (3.1)$$

Here, $A(\cdot, \cdot)$ reflects the similarity between the test sample and normal data, while ϕ represents a predefined threshold value, which determines whether a test sample is considered anomalous. Here, ϕ varies between 0 and 1 to evaluate the model's performance across all its values.

To assess the effectiveness of our model, we evaluate it using standard metrics such as the Receiver Operating Characteristic (ROC) Curve and the area under the ROC curve (ROC-AUC) score. Additionally, the Precision-Recall (PR) curve is recommended for highly skewed domains where the ROC curve may provide an excessively optimistic view of the performance [77]. For this reason, given the rarity of anomalies in the test set, Precision-Recall (PR) curves and Area Under the PR Curve

(AUPRC) are particularly informative metrics for evaluating the detection performance of our method.

The ROC curve is computed by calculating the true positive rate and the false positive rate across different threshold values ϕ for the anomaly score in Equation (3.1). The ROC curve is obtained by varying ϕ from 0 to 1, with each threshold producing a corresponding pair of true positive and false positive rates. The ROC-AUC score is the area under the ROC curve, which summarizes the model’s ability to distinguish between normal and anomalous samples. It provides an aggregate measure of performance, indicating the probability that a randomly chosen anomalous instance is ranked higher than a randomly chosen normal instance by the anomaly detection model [78].

Similarly, the PR curve is computed by plotting precision against recall at different ϕ values in Equation (3.1). The AUPRC is the area under the PR curve and provides a comprehensive evaluation of the model’s ability to detect anomalies while maintaining high precision [77].

Thus, the problem is defined as follows: learn a representation of normal behavior from the training data and classify test samples as either normal or anomalous.

The anomalies to be detected include the following types:

- An **anomalous object** [31] represents an object highly dissimilar to the objects in the training data.
- An **anomalous action** [31] represents an action highly dissimilar to the actions in the training data.
- An **anomalous position** [31] represents objects located at a highly unlikely position in the training data.
- An **anomalous background** [63] represents an action or an activity that is al-

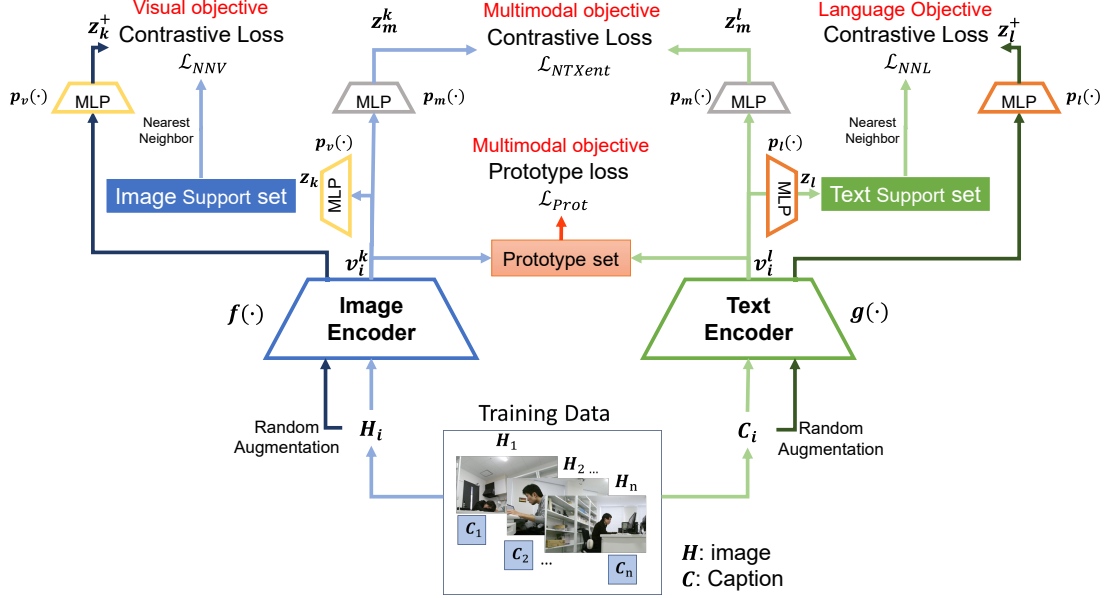


Figure 3.1: Pipeline of M3VAD.

lowed in one area but forbidden in another.

3.3 Proposed Method

This Section presents our proposed method: **M3VAD: MultiModal MultiView Self-Supervised Learning for Anomaly Detection**. Our goal is to learn a robust representation of normal human activities from image-caption pairs and detect anomalies at test time. The approach combines four main contrastive objectives: two *multimodal objectives* (cross-modal and prototypical) and two *unimodal objectives* (image- and text-based multiview objectives). These objectives yield high-quality multimodal feature representations, even with small-scale training data.

3.3.1 Overview

M3VAD operates in two main phases. In the **training phase**, the model learns multi-modal representations of input image-caption pairs by combining multiple contrastive learning objectives. These objectives simultaneously enhance the quality of unimodal and multimodal representations and group the data into clusters that represent normal behavior. Each cluster is associated with a *prototype*, defined as the centroid of the cluster’s embeddings. These prototypes serve as compact representations of normal behavior, simplifying the anomaly detection process.

The overall training process is guided by a unified objective function, which integrates multiple contrastive objectives:

$$\mathcal{L}_{M3VAD} = \lambda_m \mathcal{L}_{NT_Xent} + \lambda_p \mathcal{L}_{proto} + \lambda_v \mathcal{L}_{NNV} + \lambda_l \mathcal{L}_{NNL}, \quad (3.2)$$

where $\mathcal{L}_{NT_Xent}$ is the cross-modal objective, $\mathcal{L}_{proto}$ is the prototypical objective, and $\mathcal{L}_{NNV}$ and $\mathcal{L}_{NNL}$ are the multiview objectives for images and captions, respectively. The hyperparameters λ_m , λ_p , λ_v , and λ_l balance the contributions of each component. This formulation ensures that the learned representations capture both cross-modal consistency and unimodal robustness. Detailed explanations of each objective are provided in Sections 3.3.2 and 3.3.3, while the overall loss function is further elaborated in Section 3.3.4.

In the **test phase**, the model evaluates a new image-caption pair by computing an anomaly score based on two key measures: (1) the similarity of the sample to the learned prototypes and (2) the alignment between image and caption features in the shared latent space. This scoring process, which quantifies how much the sample deviates from normal behavior, is explained in Section 3.3.5.

At the core of M3VAD are two neural network encoders that transform the inputs

into high-dimensional feature embeddings for anomaly detection. These encoders and their specific implementations are detailed in Section 3.3.2. Figure 3.1 illustrates the overall pipeline.

3.3.2 Multimodal Objectives

To effectively learn from multimodal data, M3VAD integrates two core objectives: the cross-modal objective, which aligns embeddings across modalities, and the prototypical objective, which structures the embeddings into distinct clusters to represent normal behavior. For our purposes, a modality refers to a specific form of data [79], such as visual data (images) or language data (captions). Images encode spatial and visual information, while captions provide descriptive language that complements the image content, capturing context and semantics.

The visual encoder $f(\cdot)$ processes images $\mathbf{H}_i$ into visual embeddings $\mathbf{v}_i^k$, and the language encoder $g(\cdot)$ processes captions $\mathbf{C}_i$ into language embeddings $\mathbf{v}_i^l$. These embeddings are obtained as follows:

$$\mathbf{v}_i^k = f(\mathbf{H}_i; \Theta_f), \quad (3.3)$$

$$\mathbf{v}_i^l = g(\mathbf{C}_i; \Theta_g), \quad (3.4)$$

where Θ_f and Θ_g represent the parameters of the visual and language encoders, respectively. These embeddings serve as the foundation for the subsequent objectives.

The cross-modal objective ($\mathcal{L}_{NT-Xent}$) aligns embeddings across modalities by ensuring that embeddings of corresponding image-caption pairs are closer to each other than to unrelated pairs. This characteristic promotes consistency and contextual relevance between the two modalities.

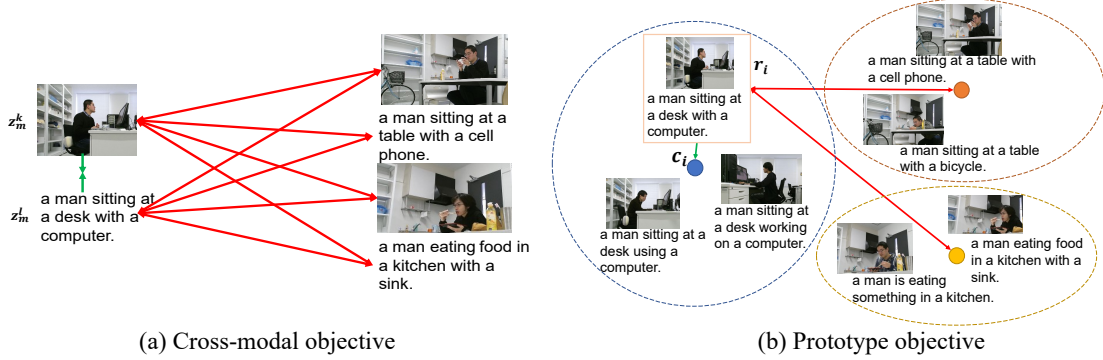


Figure 3.2: Multi-modal objectives.

The prototypical objective ($\mathcal{L}_{proto}$) groups multimodal embeddings into clusters, each represented by a prototype, or centroid, to capture diverse patterns of normal behavior. This structured clustering enhances the model’s ability to differentiate between normal and anomalous samples.

The mathematical formulations and implementation details of these objectives are elaborated in the following subsections.

3.3.2.1 Cross-modal Objective

To achieve contextual consistency between the two modalities, we adopt and adapt the normalized temperature-scaled cross-entropy loss (NT_Xent) [51] to the cross-modal domain. The cross-modal objective focuses on learning a shared latent space where visual and textual representations that belong to the same sample are closely aligned, while unrelated samples are well-separated. Unlike conventional NT_Xent setups that target within-modality consistency [51], our adaptation extends its applicability to cross-modal alignment, a crucial requirement for anomaly detection in multimodal contexts.

We use a shared projection head to map the raw embeddings ($\mathbf{v}^k$ and $\mathbf{v}^l$) into a

lower-dimensional latent space ($\mathbf{z}^k$ and $\mathbf{z}^l$), as this measure has been shown to stabilize contrastive learning and improve alignment quality across modalities [51].

We propose to define an anchor as a sample from either the visual or language modality (*modality anchor*), with its positive counterpart being the corresponding sample from the other modality. Specifically, when the image serves as the anchor, its corresponding caption is treated as the positive sample, and vice versa. For negative samples, we consider all other images and captions within the mini-batch that do not correspond to the anchor. These measures ensure that the model learns to pull together embeddings of related image-caption pairs while pushing apart embeddings of unrelated pairs. An example of the cross-modal objective is shown in Figure 3.2.a

Let $\mathbf{z}^k$ and $\mathbf{z}^l$ be lower-dimensional projections of $\mathbf{v}^k$ and $\mathbf{v}^l$, obtained through a shared projection head. For a mini-batch of N image-caption pairs, the cross-modal objective for a positive pair $(\mathbf{z}_m^k, \mathbf{z}_m^l)$ is defined as:

$$\ell(\mathbf{z}_m^k, \mathbf{z}_m^l) = -\log \frac{e^{\frac{\text{sim}(\mathbf{z}_m^k, \mathbf{z}_m^l)}{\tau}}}{\sum_{m=1}^{2N} \mathbb{1}_{[q \neq k]} e^{\frac{\text{sim}(\mathbf{z}_m^k, \mathbf{z}_m^q)}{\tau}}}, \quad (3.5)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature parameter, and $\mathbb{1}_{[q \neq m]}$ is an indicator function that excludes the positive pair from the denominator. Averaging over all positive pairs in the mini-batch, the cross-modal loss is:

$$\mathcal{L}_{NT_Xent} = \frac{1}{2N} \sum_{m=1}^N [\ell(\mathbf{z}_m^k, \mathbf{z}_m^l) + \ell(\mathbf{z}_m^l, \mathbf{z}_m^k)]. \quad (3.6)$$

By minimizing $\mathcal{L}_{NT_Xent}$, we maximize mutual information between the modalities, encouraging the learned representations to explicitly encode cross-modal consistency.

3.3.2.2 Prototypical Objective

To create well-separated clusters that represent normal behavior, we define a *prototypical objective*, which minimizes the distance between a sample and its assigned prototype while maximizing its distance from prototypes of other clusters. A prototype is defined as the centroid of a cluster in the multimodal embedding space, serving as a compact representation of semantically similar embeddings [64].

It is a common practice to use a clustering algorithm to create groups and use them as a model to represent normal behavior [4, 31, 48]. However, traditional methods do not guarantee the creation of well-separated clusters, i.e., clusters in which each object is significantly closer to other objects in the same cluster than to objects in any other cluster [80]. This limitation can reduce the accuracy of anomaly detection systems.

Li *et al.* [81] proposed ProtoNCE, which combines instance-wise and prototypical contrastive learning. ProtoNCE incorporates multiple granularities of prototypes by clustering the data M times per epoch with varying cluster sizes. While this approach has been effective in unimodal scenarios, adapting it to multimodal settings introduces computational challenges due to the higher dimensionality of the embedding space. To address these challenges, we propose two modifications. First, we eliminate the granularity by clustering the data only once per epoch instead of multiple times. Second, we use the BIRCH algorithm [50], which is computationally efficient and dynamically determines the number of clusters.

The multimodal representation of an image-caption pair is defined as the concatenation of its visual and textual embeddings:

$$\mathbf{r}_i = \mathbf{v}_i^k \parallel \mathbf{v}_i^l, \quad (3.7)$$

where $\mathbf{v}_i^k$ and $\mathbf{v}_i^l$ are the visual and textual embeddings, respectively, and $\parallel$ denotes

concatenation.

At the beginning of each training epoch, we randomly sample a subset of the training data to form a prototype set. Clustering this subset with BIRCH produces multi-modal prototypes $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$. The prototypical loss function is defined as:

$$\mathcal{L}_{prot} = \frac{1}{N} \sum_{i=1}^N -\log \frac{e^{\text{sim}(\mathbf{r}_i \cdot \mathbf{c}_i)/\tau}}{\sum_{j=1}^d e^{\text{sim}(\mathbf{r}_i \cdot \mathbf{c}_m^j)/\tau}}, \quad (3.8)$$

where N is the batch size, τ is a temperature parameter, and d is the number of negative prototypes ($m \neq i$). The numerator represents the similarity between a sample and its assigned prototype, while the denominator sums over similarities to all prototypes, including negatives. By minimizing $\mathcal{L}_{proto}$, we refine clusters to ensure that samples are close to their assigned prototypes and distant from others. An example of the prototypical objective is shown in Figure 3.2.b

3.3.3 Multiview Objectives

To enhance the robustness of unimodal representations, we introduce two separate contrastive objectives: one for the image encoder ($\mathcal{L}_{NNV}$) and one for the caption encoder ($\mathcal{L}_{NNL}$). These objectives leverage diverse perspectives of the same data, referred to as views, to improve representation quality.

We define a view as a distinct representation of the same data. For visual data, views are generated through augmentations such as random cropping and horizontal flipping. For language data, views include transformations like synonym replacement or random deletion. These augmentations preserve the semantic meaning of the data while introducing variations that help the model learn invariances. By incorporating multiple views, we simulate the diversity typically seen in larger datasets, enabling the

model to generalize better despite being trained on small-scale data. This approach enhances the model’s ability to detect subtle deviations in anomaly detection tasks while maintaining contextual integrity [82].

In addition to standard augmentations, we also consider the nearest neighbor of each sample as another view. The nearest neighbor provides a semantically meaningful representation of the data within its modality, which introduces further diversity in the contrastive pairs. For instance, an image’s nearest neighbor in the feature space captures structural and contextual similarities that complement the original data, while a caption’s nearest neighbor captures linguistic parallels. By using nearest neighbors alongside augmentations, we improve the model’s ability to learn robust, semantically rich representations.

The motivation for including nearest neighbors lies in their role in ensuring modality independence. If the original image were directly paired with its corresponding caption as the anchor-positive pair, this inclusion could create cross-modal dependencies. By using the nearest neighbor of the original image as the anchor and its augmented variant as the positive, we restrict the connections to within the same modality. Similarly, for language embeddings, the nearest neighbor of a caption serves as the anchor. This strategy assumes that the nearest neighbor of an image is not inherently tied to the nearest neighbor of its paired caption, effectively preventing interference between the modalities. This strategy ensures the model learns disentangled, modality-specific representations, which are crucial for capturing intra-modality variations robustly. Additionally, nearest neighbors diversify the contrastive pairs, mitigating overfitting and encouraging the learning of semantically meaningful and structurally consistent clusters within each modality [52]. Our approach builds upon the method proposed by Dwibedi et al. [52], which originally demonstrated the effectiveness of nearest-neighbor-based contrastive learning for visual data. Here, we extend it to include language data, lever-

aging the embedding space’s generality for both modalities.

Formally, for an image embedding $\mathbf{z}_k$, the nearest neighbor is defined as:

$$NN(\mathbf{z}_k, \mathbf{Q}_v) = \underset{\mathbf{a} \in \mathbf{Q}_v}{\operatorname{argmin}} \|\mathbf{z}_k - \mathbf{a}\|_2, \quad (3.9)$$

where $\mathbf{Q}_v$ is the support set for image embeddings. Similarly, for a caption embedding $\mathbf{z}_l$, we have:

$$NN(\mathbf{z}_l, \mathbf{Q}_l) = \underset{\mathbf{b} \in \mathbf{Q}_l}{\operatorname{argmin}} \|\mathbf{z}_l - \mathbf{b}\|_2, \quad (3.10)$$

where $\mathbf{Q}_l$ is the support set for caption embeddings. The support sets $\mathbf{Q}_v$ and $\mathbf{Q}_l$ are subsets of the training data sampled at the beginning of each training epoch to reduce computational costs.

The contrastive loss for the image modality and language modality are defined as:

$$\mathcal{L}_{NNV} = \frac{1}{N} \sum_{n=1}^N -\log \frac{e^{\operatorname{sim}(NN(\mathbf{z}_k, \mathbf{Q}_v), \mathbf{z}_k^+)/\tau}}{\sum_{q=1}^N \mathbb{1}_{[q \neq k]} e^{\operatorname{sim}(NN(\mathbf{z}_k, \mathbf{Q}_v), \mathbf{z}_q)/\tau}}, \quad (3.11)$$

$$\mathcal{L}_{NNL} = \frac{1}{N} \sum_{n=1}^N -\log \frac{e^{\operatorname{sim}(NN(\mathbf{z}_l, \mathbf{Q}_l), \mathbf{z}_l^+)/\tau}}{\sum_{q=1}^N \mathbb{1}_{[q \neq l]} e^{\operatorname{sim}(NN(\mathbf{z}_l, \mathbf{Q}_l), \mathbf{z}_q)/\tau}}. \quad (3.12)$$

where $\mathbf{z}_k^+$ is an augmented version of $\mathbf{z}_k$, τ is the temperature parameter, and the denominator sums over all negatives.

By minimizing these losses, we ensure the model learns robust, intra-modality representations while maintaining scalability and efficiency. An example of the multiview objectives is shown in Figure 3.3

3.3.4 Objective Loss

In our model, we integrate four complementary contrastive loss functions, as detailed in the Overview section (Equation (3.2)). These losses include the cross-modal loss

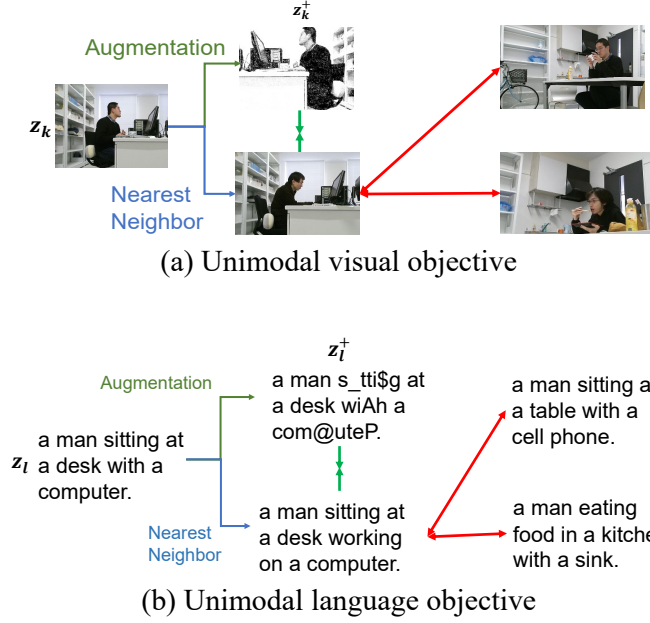


Figure 3.3: Multi-View objectives.

($\mathcal{L}_{NT_Xent}$), the prototypical loss ($\mathcal{L}_{proto}$), and the visual and language nearest-neighbor losses ($\mathcal{L}_{NNV}$ and $\mathcal{L}_{NNL}$). Together, they address distinct challenges and collectively form the backbone of the M3VAD training process.

The cross-modal loss, $\mathcal{L}_{NT_Xent}$, ensures alignment between image and caption representations in a shared latent space. By adapting the NT_Xent loss [51] to a multi-modal context, we introduced the concept of an anchor-modality pair, where the positive sample is drawn from the paired modality. This adaptation maximizes cross-modal consistency and mutual information while preserving contextual relevance.

The prototypical loss, $\mathcal{L}_{proto}$, promotes the formation of well-separated clusters in the multimodal latent space. By simplifying ProtoNCE [81] to calculate the loss only once per mini-batch, we intend to reduce computational overhead without sacrificing the quality of the clustering. This efficiency ensures scalability while maintaining robust cluster representations, even for small-scale data.

The multiview losses, $\mathcal{L}_{NNV}$ and $\mathcal{L}_{NNL}$, enhance the robustness of unimodal representations by leveraging both standard data augmentations and nearest-neighbor sampling. Following the framework of nearest-neighbor contrastive learning [52], we extend this approach beyond the visual domain to include the language modality, using semantic neighbors to diversify contrastive pairs. This extension ensures the model learns disentangled, modality-specific features while mitigating the risk of overfitting and improving generalization.

3.3.5 Anomaly Score

In the **test phase**, the anomaly score is computed for each incoming image-caption pair $(\mathbf{H}_i^*, \mathbf{C}_i^*)$ to quantify its deviation from the learned normal behavior. As outlined in the Section 4.2, Equation 3.1, an image-caption pair is detected as anomalous if its anomaly score exceeds a predefined threshold ϕ . This score integrates two complementary perspectives: the proximity to learned prototypes and the cross-modal consistency of the sample.

The **prototypical anomaly score** A_d evaluates the similarity of the test sample to the nearest prototype in the latent space. Prototypes are the centroids of clusters learned during training and represent normal behavior. Following [4, 31], A_d is calculated as the Euclidean distance between the test embedding $\mathbf{r}_i^*$ and the closest prototype $\mathbf{p}$ in the prototype set $\mathcal{P}$:

$$A_d(\mathbf{r}_i^*, \mathcal{P}) = \min_{\mathbf{p} \in \mathcal{P}} \|\mathbf{r}_i^* - \mathbf{p}\|_2. \quad (3.13)$$

The **cross-modal anomaly score** A_s captures the alignment between the image and caption embeddings in the shared feature space. This score reflects the degree of cross-modal consistency, as enforced during training. For an image-caption pair $(\mathbf{z}_m^{*k}, \mathbf{z}_m^{*l})$, where $\mathbf{z}_m^{*k}$ and $\mathbf{z}_m^{*l}$ are the respective embeddings of the image and caption, A_s is defined

as:

$$A_s(\mathbf{z}_m^{*k}, \mathbf{z}_m^{*l}) = 1 - \frac{\mathbf{z}_m^{*k} \cdot \mathbf{z}_m^{*l}}{\|\mathbf{z}_m^{*k}\|_2 \|\mathbf{z}_m^{*l}\|_2}. \quad (3.14)$$

The final anomaly score A combines these two components, weighted by hyperparameters λ_d and λ_s , which control their respective contributions:

$$A = \lambda_d A_d + \lambda_s A_s. \quad (3.15)$$

This formulation balances the two perspectives. While A_d evaluates the sample’s alignment with the overall distribution of normal behavior, A_s focuses on the internal consistency between the image and caption. Together, these components provide a robust framework for detecting anomalies in complex multimodal data.

As explained in Section 3.2, we evaluate our model by using the Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, and their respective area under the curve (AUC) scores. These metrics provide comprehensive insights into the model’s anomaly detection capabilities across different application scenarios.

3.4 Experiments

In this Section, we systematically evaluate the proposed method, **M3VAD**, for image-based anomaly detection in human monitoring tasks. The evaluation aims to assess its detection performance by comparing it with baseline methods, particularly under small-scale data conditions. Another important aspect of the analysis is the sensitivity of M3VAD to variations in the amount of data used for nearest-neighbor and prototype sets, providing insights into its robustness. Additionally, the effectiveness of the clustering refinement strategy is examined by investigating whether the prototypical contrastive objective enhances anomaly detection compared to standard one-class

learning methods. Finally, the contributions of both multimodal and unimodal objectives are analyzed to determine their role in improving detection accuracy.

To provide a wide evaluation, we first describe the datasets and the experimental setup, including the architectural details, parameter configurations, and the selection of baseline methods. It is followed by an analysis of the detection results, highlighting the advantages of M3VAD over existing approaches. Lastly, additional analyses are conducted to investigate the impact of training configurations, the refinement of clusters, and the effectiveness of individual components in shaping the final detection performance.

3.4.1 Datasets

We evaluate our method on four human-centric anomaly detection datasets: **HAM**, **BM**, **AnoVG**, and **UCSD Ped2**, each capturing different contexts of human monitoring. Following the one-class anomaly detection paradigm, only normal examples are used in the training set, and normal/abnormal examples appear in the test set.

The **HAM** and **BM** datasets were introduced in Section 2.4.1. In this section, we introduce two additional datasets—**AnoVG** and **UCSD Ped2**. All four datasets are used in the experimental evaluations conducted in this Chapter.

AnoVisualGenome (AnoVG) [54] was constructed after analyzing images that can be considered abnormal situations from the Visual Genome dataset [83], which provides dense annotations for regions on over 108K images. AnoVG includes several kinds of human activities in inappropriate environments as contextual anomalies, such as a person watching TV on the street or a person sitting on a couch on the beach. It contains 1427 normal images for training, as well as 218 normal images and 31 abnormal images for testing.

Table 3.1: Datasets summary.

	HAM		BM		AnoVG		UCSD Ped2	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Training Data	13453	0	5548	0	1427	0	2550	0
Test Data	2103	528	590	106	218	31	374	1636

UCSD Anomaly Detection Dataset (UCSD Ped2) [9] is a video dataset that was acquired with a stationary camera overlooking pedestrian walkways. The anomalies in this dataset are labeled by frame, meaning that each frame is annotated to indicate whether it contains an abnormal event. These anomalies are varied: common anomalies in the dataset include bikers, skaters, small carts, and people walking across a walkway or in the surrounding grass. The training split contains 16 video samples with a total of 2 550 frames and does not contain any abnormal events. The test split contains 12 video samples with a total of 2 010 frames, of which 374 frames are normal and 1 636 frames are abnormal.

Table 3.1 summarizes the size of the training and test splits of the three datasets.

3.4.2 Experimental setup

3.4.2.1 Detailed Architecture

The proposed method, M3VAD, processes both images and captions to extract multimodal features that improve anomaly detection. For caption generation, we use GIT [84], a modern image captioning model that generates one caption per image, aiming to capture the overall scene context. Unlike prior works relying on region-level captioning, e.g., DenseCap [5], GIT produces image-level captions more suitable for our direct image-caption approach.

Our proposed method is designed to detect anomalies using limited training data. Given the constraints of small-scale datasets, we require a lightweight and effective architecture for both image and text modalities. For the textual modality, we require a lightweight transformer-based model that captures semantic meaning efficiently. We select DistilBERT [73], which retains most of BERT’s [74] accuracy while being significantly faster and more efficient. Each generated caption $\mathbf{C}_i$ is first transformed into an embedding using DistilBERT. The last four hidden layers are concatenated to form a 3072-dimensional feature vector, which is then processed by an MLP projection head $g(\cdot)$ to obtain the final 2048-dimensional language representation $\mathbf{v}_i^l$. Therefore, Equation (3.4) is respectively replaced with

$$\mathbf{v}_i^l = g(\alpha(\mathbf{C}_i); \Theta_g), \quad (3.16)$$

where $\alpha(\cdot)$ represents the pre-trained DistilBERT function and Θ_g denotes the trainable parameters of $g(\cdot)$.

Instead of relying on large-scale vision transformers (ViTs) [85], which demand extensive data for meaningful feature learning as we explained in Section 1.3, we select a computationally efficient encoder optimized for small datasets. For the visual modality, we adopt ScatSimCLR [7] as the backbone. Unlike traditional CNNs, ScatSimCLR incorporates Scattering Networks (ScatNet) [86], which apply fixed wavelet transforms to extract stable and interpretable features before passing them to a ResNet-based encoder. This incorporation reduces the number of trainable parameters while maintaining strong representation-learning capabilities [86]. The image $\mathbf{H}_i$ is first transformed by ScatNet, and the resulting low-dimensional feature map is then passed through a small ResNet-based encoder $f(\cdot)$ with eight residual blocks, yielding a 2048-dimensional visual embedding $\mathbf{v}_i^k$. Therefore, Equation (3.3) is respectively replaced

with:

$$\mathbf{v}_i^k = f(\beta(\mathbf{H}_i); \Theta_f), \quad (3.17)$$

where $\beta(\cdot)$ represents the fixed ScatNet transformation and Θ_f refers to the trainable parameters of $f(\cdot)$.

The multimodal representation of each image-caption pair is then constructed by concatenating the visual and textual embeddings:

$$\mathbf{r}_i = \mathbf{v}_i^k \parallel \mathbf{v}_i^l. \quad (3.18)$$

This results in a 4096-dimensional joint representation in which anomalies are identified. Additionally, to facilitate contrastive learning, we introduce three separate projection heads, each mapping embeddings into a 256-dimensional latent space as shown in Figure 3.1:

- $p_m(\cdot)$ for multimodal contrastive objectives,
- $p_v(\cdot)$ for unimodal image contrastive objectives,
- $p_l(\cdot)$ for unimodal language contrastive objectives.

Each projection head consists of a single-layer MLP. The purpose of these projection heads is to refine feature separability in the contrastive learning process, as demonstrated in SimCLR [51].

3.4.2.2 Parameter setting

To ensure the robustness of unimodal representations, we apply data augmentations for both image and text modalities in the contrastive learning framework. Specifically,

we perform augmentation on the positive views $\mathbf{z}_k^+$ and $\mathbf{z}_l^+$. The image augmentation pipeline consists of random cropping, horizontal flipping, color jittering, and grayscale conversion. For captions, we introduce textual perturbations, including simulated typos, OCR errors, and word deletions [87]. These augmentations enhance invariance within each modality and help the model generalize better.

However, for multimodal contrastive learning, no augmentation is applied to ensure strict cross-modal consistency. Since multimodal objectives rely on exact image-caption pairs, altering one modality could distort the alignment, reducing the effectiveness of cross-modal representation learning. This design ensures that the model learns contextual consistency between image and caption features without unnecessary distortions.

We train our method from scratch with each dataset for 100 epochs using the Adam optimizer [70] with the following parameters: learning rate $lr = 1 \times 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The batch size is set to $N = 256$ image-caption pairs. For prototype-based clustering, we employ the BIRCH algorithm [50] with a clustering threshold of $\theta = 0.05$, where θ represents the diameter threshold of a leaf node in the clustering feature (CF) tree.

We set the trade-off parameters of the contrastive losses $\lambda_m, \lambda_p, \lambda_v = 1$, and $\lambda_l = 0.5$ to mitigate the effect of wrong captioning. For the final anomaly score, we combine the prototypical distance score A_d and the cross-modal similarity score A_s , assigning equal trade-off weights of $\lambda_d = 0.5$ and $\lambda_s = 0.5$. This ensures that both structural similarity to normal clusters and cross-modal coherence contribute equally to anomaly detection. Our approach was implemented using the PyTorch framework¹ and Python on Ubuntu 20.04, running on an NVIDIA GeForce RTX 3090 GPU with an Intel i9-7900X CPU.

¹<https://pytorch.org/>

3.4.2.3 Compared Methods

To evaluate the effectiveness of M3VAD, we compare it against state-of-the-art methods in contrastive self-supervised learning and image-level anomaly detection. In Chapter 2, we performed anomaly detection at the image level and relied on region captions, allowing a fair comparison with methods that analyze anomalies at the region level. These methods detect anomalous regions by analyzing spatial relations between objects rather than evaluating entire images. However, in this Chapter, we use image captions, meaning anomaly detection now operates at the image level. This fundamental change makes direct comparisons with region-based anomaly detection methods inappropriate.

The methods proposed by Hatae *et al.* [31], Fadjrimiratno *et al.* [4], and Zhang *et al.* [54] perform anomaly detection by modeling object-region relationships. Instead of considering all possible object-to-object relations, these methods introduce constraints to selectively focus on a subset of region relations. Fadjrimiratno *et al.* [4] analyzes dual-region relations, capturing anomalies based on pairwise inconsistencies, while Zhang *et al.* [54] employs a graph-based approach to model semantic region dependencies. Although these constraints reduce computational costs, they also limit the ability to capture more complex, unexpected contextual anomalies.

A further limitation of these methods is their dependence on regions generated by a deep captioning model. If the captioning model fails to detect certain regions, merges distinct objects into a single bounding box, or overlooks background elements contributing to contextual understanding, the anomaly detection process is inherently compromised. This reliance on automatic region generation constrains their ability to generalize beyond the predefined region representations.

In contrast, M3VAD detects anomalies at the image level, learning multimodal

representations from entire image-caption pairs rather than predefined regions. This broader perspective allows the model to capture anomalies in object placement, background context, and global semantic inconsistencies that region-based methods might overlook.

In Chapter 2, we also demonstrated that refining a shared multimodal feature space provides better performance on anomaly detection compared to multimodal methods that use separate unimodal pre-trained models. We showed that contrastive self-supervised learning effectively aligns image and text representations, outperforming methods that simply rely on independently trained vision and language encoders. Since this finding has already established the advantage of cross-modal refinement, we do not need to reintroduce comparisons with methods that exclusively use pre-trained unimodal representations.

Instead, we evaluate M3VAD against C-SSL methods and image-level anomaly detection baselines. As an unimodal baseline, we include ScatSimCLR [7], a self-supervised learning model that applies contrastive learning only to image data. It leverages a wavelet-based ScatNet to enhance feature learning under small-scale data constraints. It includes different architectures, such as ScatSimCLR_8, ScatSimCLR_12, ScatSimCLR_16, and ScatSimCLR_30, where the number denotes the number of ResBlocks in the network. These variations allow for different levels of model complexity and adaptation to varying dataset sizes.

For multimodal contrastive learning, we compare against CLIP [8], which employs contrastive learning to align image and text representations in a shared space. CLIP serves as a strong benchmark for learning joint image-text representations. We evaluate multiple CLIP architectures, including CLIP_ViT.L.14, CLIP_ViT.B.16, CLIP_ViT.B.32, and CLIP_RN50. In these models, ViT.L.14 and ViT.B.16 refer to Vision Transformers [85] of different sizes, with "L" representing a large model and "B" a base model,

while the numbers indicate the patch size (e.g., 14×14 or 16×16). The RN50 variant employs a ResNet-50 architecture for image encoding.

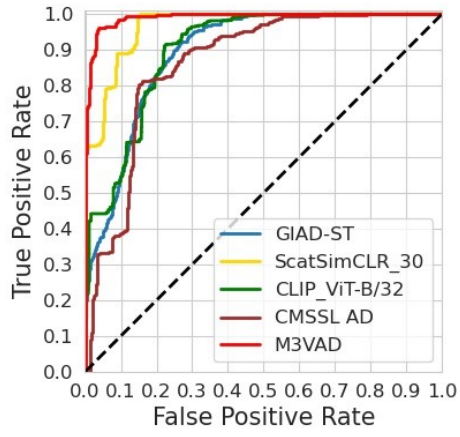
For image-level anomaly detection, we also evaluate GIAD-ST [6], an inpainting-based method that reconstructs missing image regions to assess consistency. However, we exclude GIAD-ST from UCSD Ped2 experiments due to a fundamental mismatch between the dataset structure and the method’s requirements. Unlike other datasets, UCSD Ped2 does not contain predefined regions of interest, which means that GIAD-ST would require an additional preprocessing step to generate these regions.

While automatic region proposal methods exist, their reliability on low-resolution surveillance footage remains a challenge, as they often struggle to detect small or occluded objects due to the lack of detailed appearance features [88, 89]. Studies have shown that traditional region proposal networks underperform when dealing with small objects, leading to poor localization accuracy in low-resolution images [89]. Since UCSD Ped2 primarily consists of low-resolution surveillance footage, applying such methods could introduce inconsistencies in the generated object regions, which would directly impact GIAD-ST’s anomaly detection capability. Additionally, UCSD Ped2 was not originally designed for localized region-based anomaly detection, making direct comparisons with GIAD-ST less meaningful.

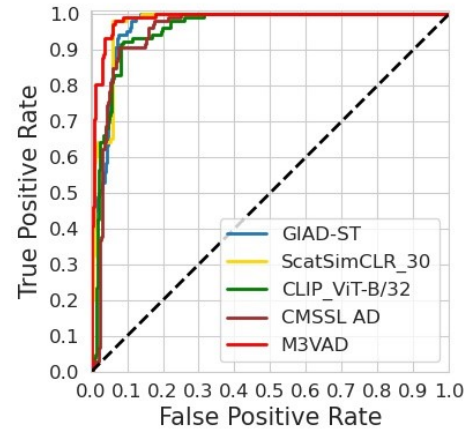
To ensure a consistent evaluation, we apply BIRCH clustering [50] as in [4, 31] to all learned feature representations. This choice ensures that anomaly scores are computed using the same method across different baselines, making comparisons fair and controlled. All models, including M3VAD and baselines, are trained from scratch using only the normal training data from each dataset to reflect the one-class learning paradigm, where anomalies are identified based on their deviation from the normal feature distribution.

Table 3.2: Comparison in terms of ROC-AUC and AUPRC scores.

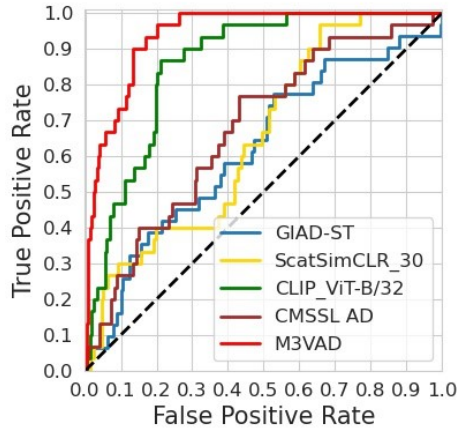
Dataset Model		ROC-AUC				AUPRC			
		HAM	BM	AnoVG	UCSD Ped2	HAM	BM	AnoVG	UCSD Ped2
CMSSL_AD [60]		0.865	0.953	0.677	0.529	0.527	0.646	0.228	0.824
GIAD-ST [6]		0.895	0.970	0.621	-	0.681	0.840	0.211	-
S.S.C. [7]	ScatSimCLR_8	0.881	0.902	<u>0.698</u>	0.514	0.711	0.716	0.209	0.824
	ScatSimCLR_12	0.886	0.965	0.635	0.546	0.793	<u>0.90</u>	0.181	0.818
	ScatSimCLR_16	0.848	0.966	0.663	0.452	0.741	0.847	0.181	0.771
	ScatSimCLR_30	<u>0.967</u>	<u>0.973</u>	0.654	0.427	<u>0.892</u>	0.848	<u>0.210</u>	0.755
CLIP [8]	CLIP_RN50	0.816	0.890	0.843	0.562	0.543	0.620	0.470	0.865
	CLIP_ViT.L.14	0.881	0.953	0.875	0.650	0.549	0.673	0.510	0.887
	CLIP_ViT.B.16	0.878	0.948	<u>0.885</u>	<u>0.671</u>	0.606	0.641	<u>0.512</u>	<u>0.89</u>
	CLIP_ViT.B.32	<u>0.900</u>	<u>0.956</u>	0.860	0.620	<u>0.70</u>	<u>0.697</u>	0.395	0.875
M3VAD		0.990	0.987	0.940	0.713	0.960	0.905	0.662	0.907



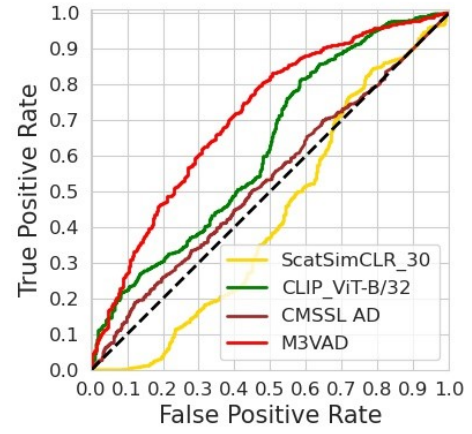
(a) HAM Dataset



(b) BM Dataset

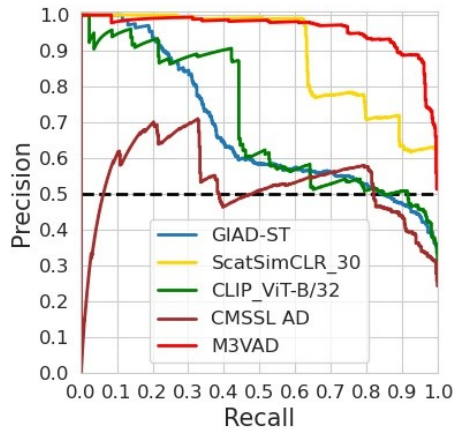


(c) AnoVG Dataset

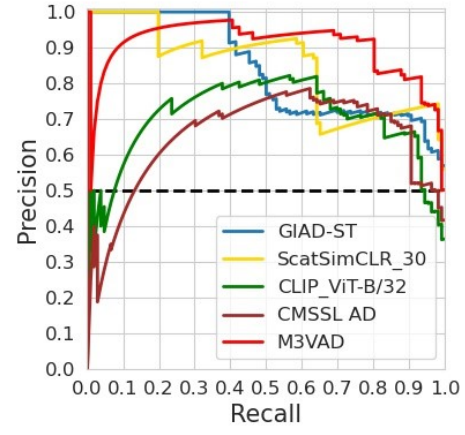


(d) UCSDped2 Dataset

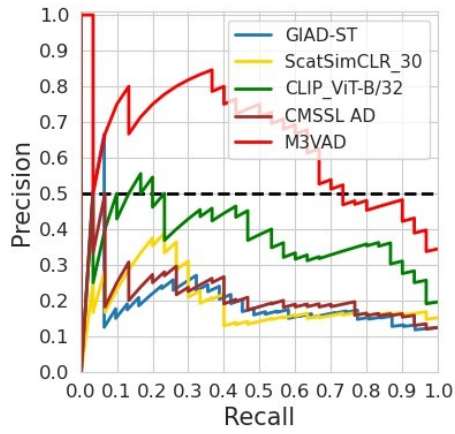
Figure 3.4: Comparison of ROC curves on the four datasets.



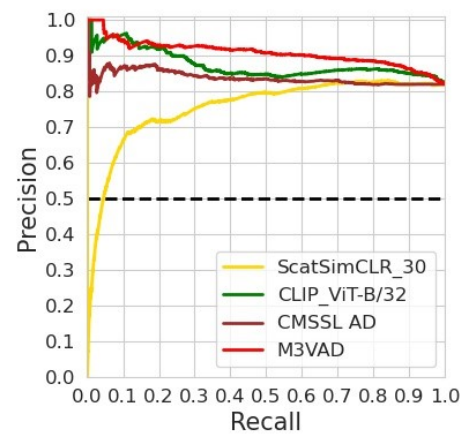
(a) HAM Dataset



(b) BM Dataset



(c) AnoVG Dataset



(d) UCSDped2 Dataset

Figure 3.5: Comparison of PR curves on the four datasets.

3.4.3 Results and analysis

Fig. 3.4, Fig. 3.5, and Table 3.2 show the ROC curves, PR curves, and ROC-AUC/AUPRC scores of M3VAD compared with the baselines on the four datasets, respectively. We have the following findings based on the results:

1. M3VAD outperforms all the baseline methods on the HAM, BM, and AnoVG datasets, achieving 0.023 – 0.174, 0.014 – 0.097, and 0.055 – 0.319 improvements in terms of ROC-AUC scores, respectively. These results validate the superiority of our method in identifying anomalies at the image level. The key factor behind this improvement would be the model’s ability to extract high-quality multimodal representations. This advantage can be attributed to the integration of both multimodal contrastive learning and unimodal multiview objectives.

To illustrate this advantage, Figure 3.6 presents a 2D visualization of the learned multimodal representations using the t-SNE method [1]. t-SNE is a non-linear dimensionality reduction technique that preserves local structures while providing a visualization of high-dimensional data in a lower-dimensional space. This method is particularly useful for understanding how well a model separates normal and anomalous samples in the learned feature space [90].

Figure 3.6 confirms this advantage, where M3VAD produces well-separated, compact clusters compared to ScatSimCLR_30 and CLIP_ViT.B.32. The effectiveness of M3VAD is evident in the distinct clusters, which are visible in the t-SNE projection, demonstrating that our approach better structures the normal feature space while isolating anomalies more effectively. In contrast, the feature embeddings obtained from ScatSimCLR_30 appear more scattered, likely due to its unimodal nature, which lacks cross-modal contextual supervision. This nature makes it harder to refine feature representations, leading to a less structured

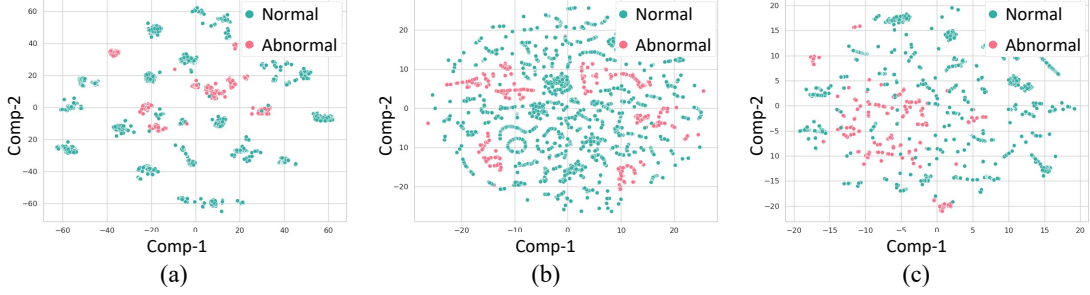


Figure 3.6: 2D Visualization with t-SNE [1] of the features extracted by (a) M3VAD , (b) ScatSimCLR [7], and (c) CLIP [8] for the HAM test split [4].

latent space. On the other hand, CLIP_ViT.B.32 produces entangled clusters, which can be attributed to its large model capacity relative to the small dataset size. ViTs typically require extensive data to develop well-separated feature spaces, as they rely on large-scale self-attention mechanisms that struggle with limited training data [85].

2. M3VAD outperforms all baseline methods in terms of AUPRC scores across the HAM, BM, and AnoVG datasets, achieving improvements of 0.068 – 0.433, 0.005 – 0.285, and 0.149 – 0.48, respectively. These results suggest that M3VAD is more reliable in distinguishing between normal and abnormal cases while maintaining higher precision, particularly at high recall values.

Figure 3.5 provides evidence of this advantage by illustrating the Precision-Recall (PR) curves. In the HAM and BM datasets (Figures 3.5.a and 3.5.b), M3VAD consistently achieves higher precision than the baselines, even as recall approaches 1.0. These findings suggest that M3VAD maintains its reliability in detecting abnormalities while minimizing false positives. Similarly, in UCS-Dped2 (Figure 3.5d), M3VAD achieves the highest PR curve across most recall values, outperforming the baseline methods.

The results from Figure 3.5 also show that anomaly detection in AnoVG is chal-

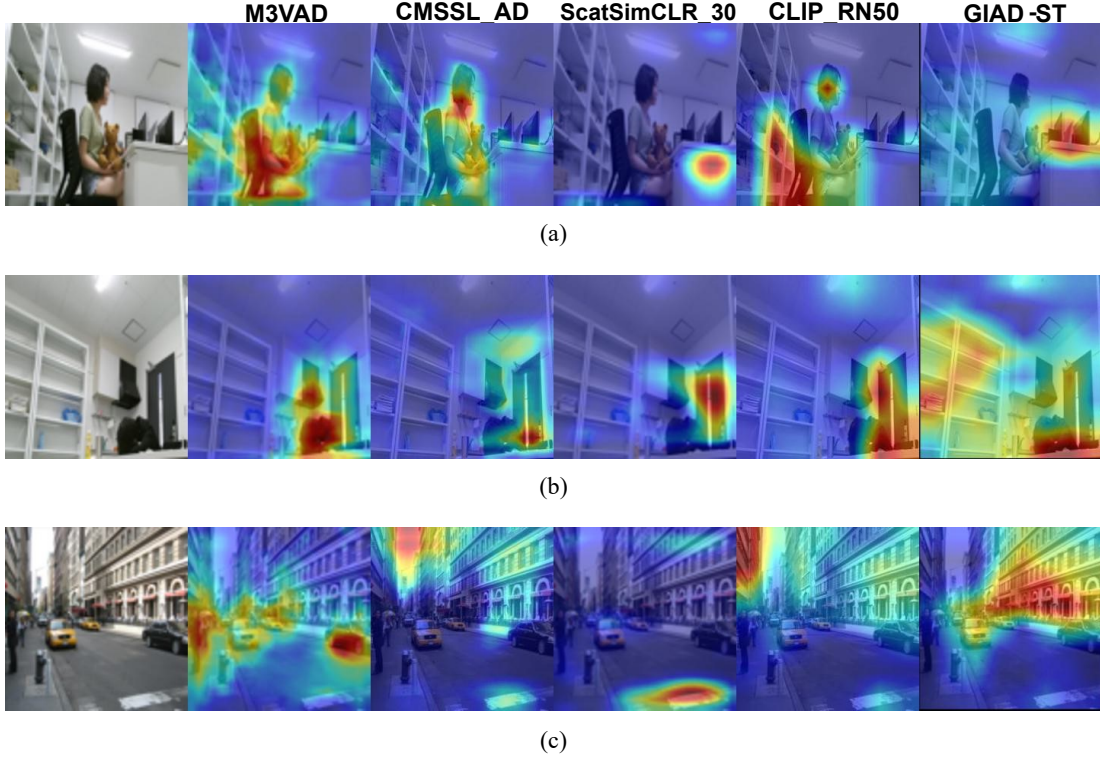


Figure 3.7: Comparison of the attention maps for normal images (a), (b), and (c) for HAM, BM, and AnoVG, respectively.

lenging compared to the other datasets. This observation is supported by the larger AUPRC gap between M3VAD and the baselines in Figure 3.5c, where the precision of baseline methods fluctuates significantly across recall values. The qualitative results in Figure 3.8 further support this claim by demonstrating that baseline models often focus on irrelevant background features, whereas M3VAD correctly attends to human-related elements, which are crucial for anomaly detection in human monitoring scenarios.

The Grad-CAM [91] visualizations in Figure 3.8 illustrate this advantage. While M3VAD accurately identifies anomalous human actions, baseline models frequently misplace their attention. For example, in Figure 3.8.c, where a person

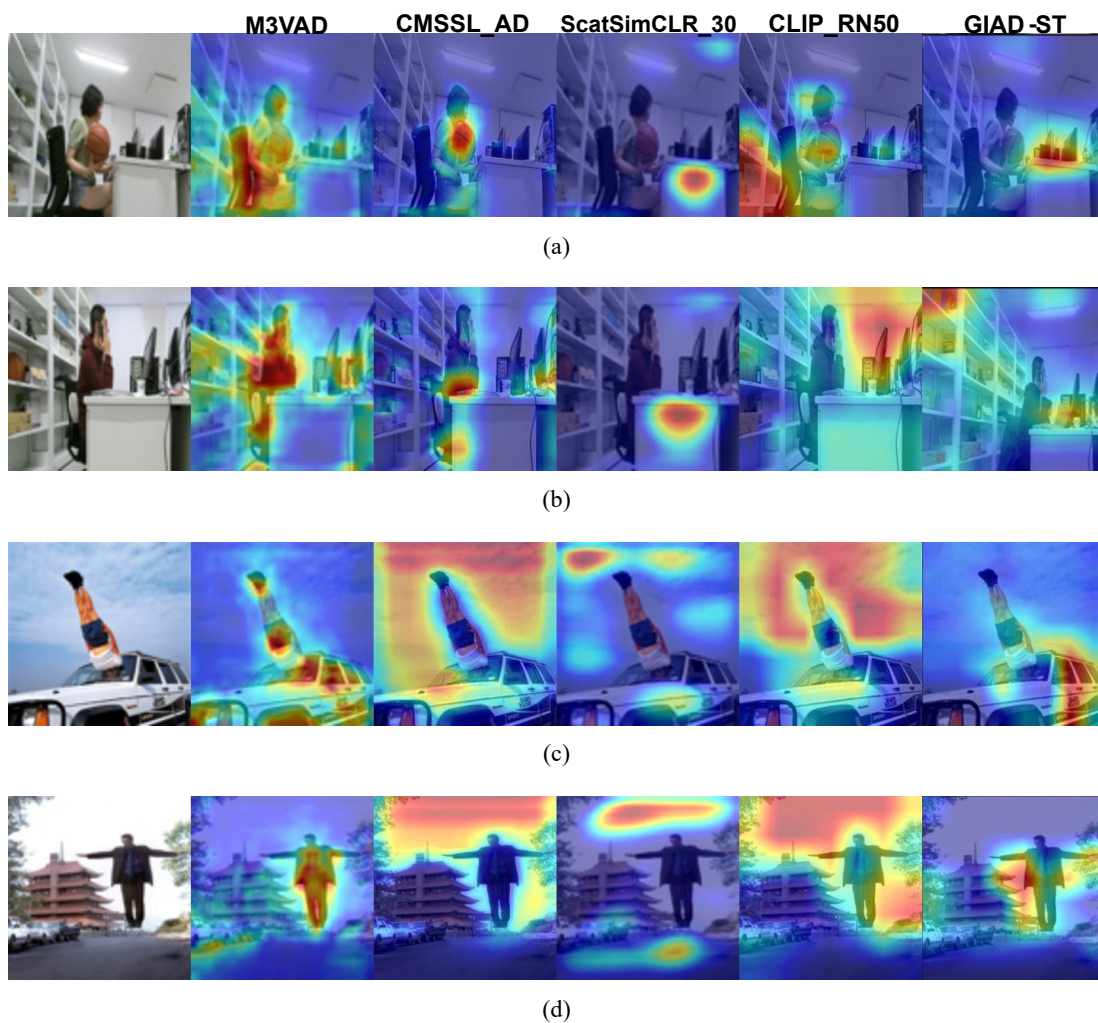


Figure 3.8: Comparison of the attention maps for abnormal images for (a) HAM, (b) BM, and (c, d) AnoVG.

is performing an unusual action, M3VAD focuses on the individual, while other methods focus on the sky or the vehicle. Similarly, in Figure 3.8.d, M3VAD detects the unusual levitating posture of the person, whereas other models highlight background structures instead of the actual anomaly. These visualizations provide evidence that M3VAD’s multimodal approach improves context-aware anomaly localization, reducing false positives and misclassifications.

3. M3VAD achieves the largest performance gap over the baselines in the AnoVG dataset, with improvements of 0.055 in ROC-AUC and 0.149 in AUPRC. These results demonstrate the model’s ability to detect anomalies effectively, even in small-scale datasets. This advantage is derived from M3VAD’s joint use of multimodal and multiview objectives, which enable it to learn more structured and meaningful feature representations with fewer data.

To better understand this improvement, Figures 3.7 and 3.8 provide a qualitative comparison of Grad-CAM [91] attention maps across different methods. These visualizations highlight key differences in the models’ attention mechanisms. As discussed in finding 2, M3VAD demonstrates superior anomaly localization by correctly focusing on human-related regions and their interactions with objects, while baseline methods often misplace attention on background elements or non-informative regions. However, this improvement is not limited to abnormal data, the same pattern is observed in normal samples, as shown in Figure 3.7. Here, M3VAD maintains well-structured attention, focusing on people and relevant objects, whereas the baseline methods tend to scatter attention across the background, exhibiting inconsistent attention patterns even in normal cases.

4. Our results indicate that ScatSimCLR_30, the largest ScatSimCLR variant, is the second-best performing model. While it achieves competitive performance on

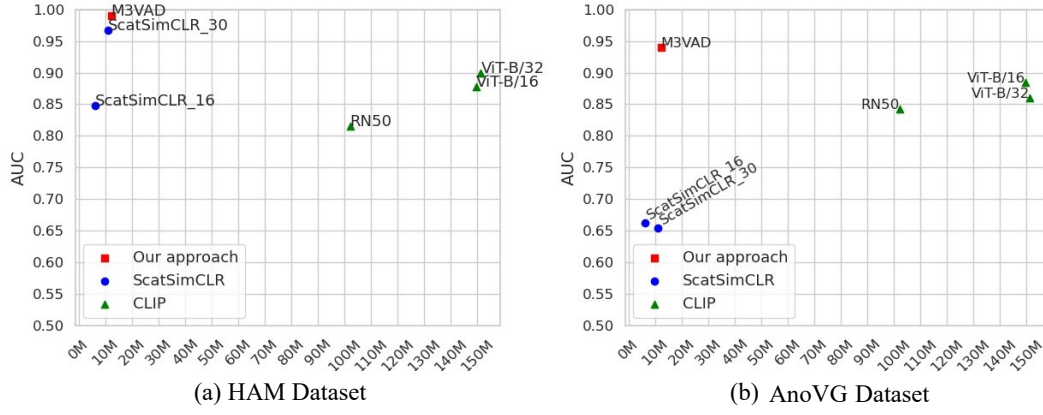


Figure 3.9: Comparison in terms of ROC-AUC and number of trainable parameters

the HAM and BM datasets, its performance degrades on the AnoVG and UCS-Dped2 datasets. One possible explanation is that ScatSimCLR_30 is a purely unimodal model which lacks the cross-modal supervision necessary to capture context well in small datasets. Without multimodal guidance, the model may learn feature representations that do not effectively separate anomalies from normal samples, particularly in limited data scenarios.

Figure 3.9 supports this observation, as it shows that ScatSimCLR_30 has a similar number of trainable parameters to M3VAD, yet its ROC-AUC score is significantly lower, especially on the AnoVG dataset. This fact suggests that increasing model complexity alone does not guarantee better anomaly detection performance when data is scarce. These results suggest that multimodal learning could be beneficial for improving generalization, particularly when data is scarce.

5. M3VAD consistently achieves the highest ROC-AUC and AUPRC scores across all datasets, demonstrating its superior anomaly detection capabilities. However,

its performance on UCSDped2 is lower compared to HAM and BM, achieving an ROC-AUC of 0.713, as shown in Table 3.2. While M3VAD still outperforms all baseline methods, the gap is narrower, with CLIP_ViT.B.16 and CLIP_ViT.L.14 achieving ROC-AUC scores of 0.671 and 0.650, respectively.

The primary reason for this difference lies in the nature of static image anomaly detection versus video-based anomaly detection. Unlike HAM and BM, where anomalies are identified through contextual interactions between humans and objects, anomalies in UCSDped2 often rely on motion patterns, such as skaters, bikers, or small carts moving through a pedestrian area. Since M3VAD does not incorporate temporal dependencies, it cannot leverage motion cues, making certain anomalies harder to distinguish from normal behaviors.

Despite this challenge, M3VAD still outperforms all baselines on UCSDped2, demonstrating that a well-structured multimodal feature space provides robust anomaly detection, even in scenarios where motion information is crucial. Notably, M3VAD maintains higher precision at high recall values, as illustrated in Figure 3.5.d, meaning it detects a significant portion of anomalies while controlling false positives. These results suggest that even without temporal cues, multimodal learning significantly enhances anomaly detection at the image level.

6. As discussed in finding 5, M3VAD achieves the highest performance among all baseline methods on UCSDped2, but its performance remains lower compared to HAM and BM. This difference is primarily due to the fundamental distinction between image-based and video-based anomaly detection. While the proposed method effectively captures contextual anomalies in static images using multimodal and multiview representations, UCSDped2 contains anomalies that depend on motion cues rather than object relations or scene context. Many



Figure 3.10: Examples of abnormal events from the UCSDped2 [9] dataset. A skater is seen on the pedestrian walkway, highlighted by a red arrow.

anomalous activities in this dataset, such as people skating or riding bicycles, are not always visually distinguishable from normal pedestrian behavior in a single frame. As shown in Figure 3.10, a skater captured in one frame appears similar to a walking pedestrian because the defining characteristic (the motion) is missing. This limitation is inherent in static image anomaly detection, whereas video-based approaches can utilize temporal information to track movement patterns across frames.

Despite the difficulty of detecting motion-based anomalies in static images, image anomaly detection remains an essential approach in human monitoring scenarios where processing single frames is more practical than analyzing video sequences. For instance, in fields such as biometric screening, medical imaging, and occupational safety monitoring, static image analysis is widely employed because motion information is not required [92].

Unlike video-based methods, which require significantly more computational resources to track motion over time, static image-based approaches are more efficient and applicable in scenarios where analyzing individual frames is sufficient. Video anomaly detection requires continuous frame processing, leading to substantial computational costs and storage demands, making it impractical for

real-time monitoring in many environments [93]. Additionally, the annotation process for video-based models is significantly more complex and expensive, further limiting the scalability of video anomaly detection systems [93]. The ability to analyze single frames also simplifies data annotation, making image-based methods more practical [94].

Our findings highlight the need for future work to integrate temporal representations into the existing multimodal framework, allowing the model to enhance anomaly detection in cases where movement patterns play a crucial role.

7. Another factor that may influence M3VAD’s performance on UCSDped2 is the accuracy of the automatically generated captions. The model relies on captions to capture semantic relationships between humans and objects, which aids in anomaly detection. However, the GIT captioning model [84], while highly effective for our primary target problem of image-level anomaly detection in human-object interactions, struggles in video-based scenarios where motion information is essential.

GIT was chosen for this study due to its strong performance in describing human-object interactions in single images, a crucial aspect of anomaly detection at the image level. Unlike region-based captioning models that may focus on isolated objects without capturing full contextual interactions, GIT generates scene-level descriptions. This capability aligns well with our target problem, where anomalies often stem from unusual contextual relationships between a person and surrounding objects, rather than from motion cues. The effectiveness of GIT in this regard is evidenced by M3VAD’s strong performance on HAM, BM, and AnoVG, where static images provide sufficient context for anomaly detection.

However, in UCSDped2, where many anomalies are motion-dependent, e.g.,



Figure 3.11: Examples of normal and abnormal events from the UCSDped2 [9] dataset along with their captions. On the left: a normal event with pedestrians on the walkway. In the middle: an abnormal event with a skater on the walkway. On the right: an abnormal event with a bicycle being ridden on the walkway.

skaters, cyclists, or small carts moving through pedestrian areas, the lack of temporal modeling in GIT becomes a limitation. Since GIT is designed for static image captioning, it does not incorporate motion cues, resulting in captions that fail to differentiate dynamic anomalies from normal pedestrian activity. This shortcoming is evident in Figure 3.11, where similar captions are generated for distinct activities, including walking, skating, and cycling. The inability to capture movement reduces the discriminative capacity of the multimodal features extracted from UCSDped2 images.

To address this limitation, our future work will explore integrating a video captioning method to capture temporal dependencies and motion cues, making M3VAD better suited for datasets like UCSDped2, where anomalies depend on movement rather than static attributes [95]. Despite the current limitations, the results indicate that M3VAD still achieves the highest overall performance among baseline methods, suggesting that multimodal learning can be effective even in cases where motion-based information is not explicitly modeled.

3.4.4 Component Analysis and Evaluation

This section evaluates the effectiveness of different components in M3VAD. Our objective is to assess how each part of the model contributes to performance and to analyze its robustness across different settings. We conduct three key experiments: (1) an analysis of how varying the amount of training data used for support sets impacts model stability, (2) an evaluation of the clustering quality, and (3) a study of M3VAD’s individual components. These experiments provide insights into the model’s effectiveness in handling small datasets and capturing meaningful anomaly representations.

3.4.4.1 Impact of Support Set Size on Stability

The effectiveness of M3VAD in anomaly detection is influenced by the composition of the support sets used in training. Given that one-class learning inherently lacks negative samples, conventional validation techniques become impractical. Instead, a statistical approach was adopted to assess the model’s robustness by systematically varying the proportion of training data allocated to create the support sets and evaluating the corresponding detection performance. To analyze the stability of the model, experiments were conducted by adjusting the support set size across a range of values: $\{10\%, 30\%, 50\%, 70\%, 90\%\}$. Each configuration was tested over ten independent trials to account for variations in sample selection, thereby providing a distribution of ROC-AUC scores that offers an insight into the model’s consistency.

Different from conventional settings where validation sets are used for model selection [96], one-class anomaly detection presents a fundamental challenge: the training data contain only normal (positive) examples, with anomalies appearing solely in the test phase [76]. This constraint makes it impractical to define a validation set that provides meaningful generalization feedback. Instead, our experimental methodology

relies on repeated trials and variance analysis, ensuring that performance stability can be assessed empirically. The results from Table 3.3 confirm that for datasets with sufficient training samples, such as HAM and BM, variance remains low, reinforcing the reliability of the learned representations.

For instance, as shown in Table 3.3, in the HAM dataset, the ROC-AUC score for the 90% support set reaches 0.978, with a variance of 1.24×10^{-5} , indicating highly stable performance. Similarly, the BM dataset achieves a mean ROC-AUC of 0.971 with a variance of 2.74×10^{-5} , demonstrating robustness in larger datasets. However, for smaller datasets such as AnoVG and UCSDped2, variance values are comparatively higher at 1.03×10^{-4} and 4.18×10^{-4} , respectively, reflecting the increased sensitivity to training data availability. These findings highlight the necessity of maximizing the support set size in practical deployments, particularly for datasets with limited training samples, where variability in representation learning could impact performance.

Figure 3.12 illustrates the ROC-AUC score value of M3VAD for the four datasets using boxplots, each of which displays the lower quartile, median, and upper quartile of the ROC-AUC scores in a box, and extends the box from the lowest to the highest scores by a line segment. The results, summarized in Figure 3.12, indicate that increasing the proportion of data used for support sets leads to an improvement in detection performance. This trend is evident in the reduction of standard deviation across trials, signifying enhanced model reliability when trained with a higher percentage of support data. As observed in Figure 3.12, the HAM and BM datasets maintain consistently low variability across different support set sizes, indicating that even with moderate data, stable feature representations can be learned. In contrast, the AnoVG and UCSDped2 datasets exhibit greater variability, particularly at lower support set sizes, which suggests that these datasets are more sensitive to training data availability due to their limited sample sizes.

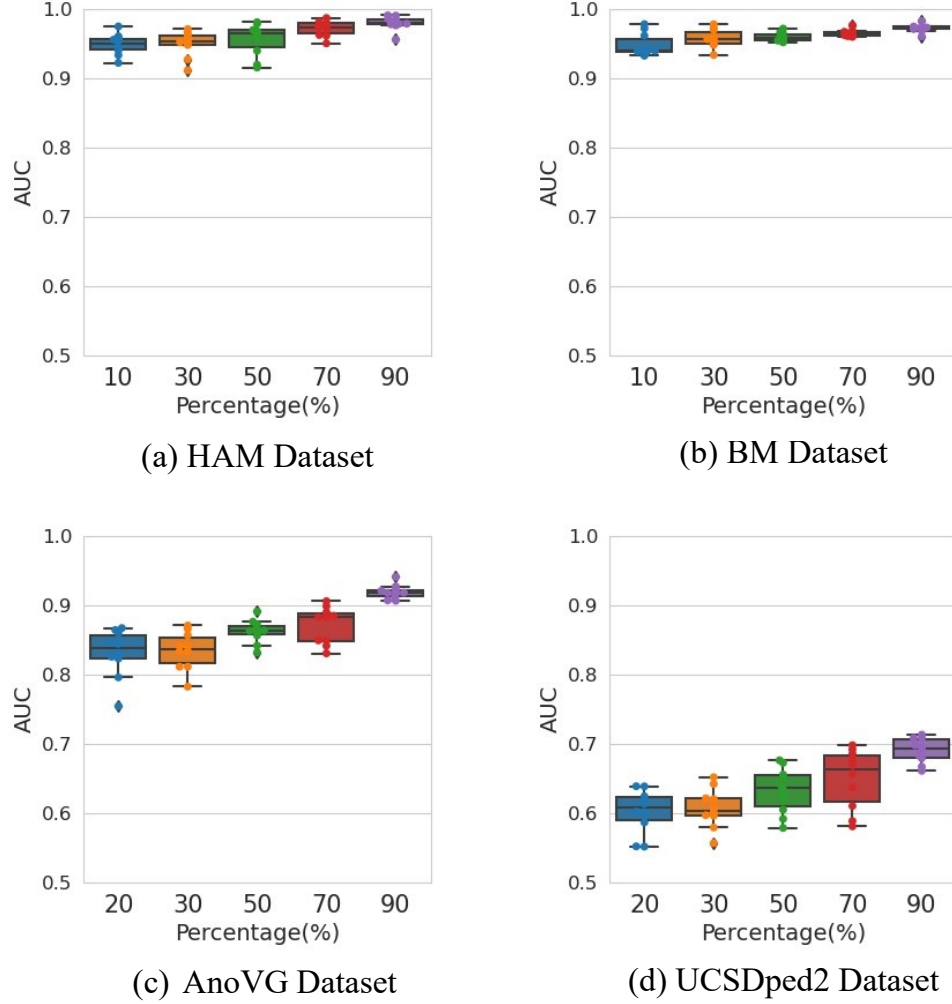


Figure 3.12: Distributions of the ROC-AUC score on the datasets for different support sets sizes.

Table 3.3: Mean and Variance of ROC-AUC Scores for 90% Support Set Across Datasets.

Dataset	Mean ROC-AUC	Variance
HAM	0.9789	1.24×10^{-5}
BM	0.9715	2.74×10^{-5}
AnoVG	0.9188	1.03×10^{-4}
UCSDped2	0.6915	4.18×10^{-4}

3.4.4.2 Effect of Clustering Refinement on Anomaly Detection

Our objective is to test whether our anomaly score, derived from the refined clustering algorithm using the prototypical loss, provides superior performance compared to well-established One-Class Learning (OCL) methods. To this end, we conducted a study comparing our approach against widely used OCL algorithms, including One-Class SVM (OCSVM) [97], Isolation Forest [98], and Local Outlier Factor (LOF) [99]. These methods rely on feature representations to distinguish between normal and anomalous instances. Given that M3VAD also produces feature representations, a direct comparison using the same feature space ensures a fair evaluation of the clustering refinement introduced in our approach.

To conduct this study, we used multimodal representation $\mathbf{r}$ defined in Equation (3.7) as multimodal features to train the OCL algorithms on the training set. The trained OCL models were then applied to test set features $\mathbf{r}^*$ for anomaly detection.

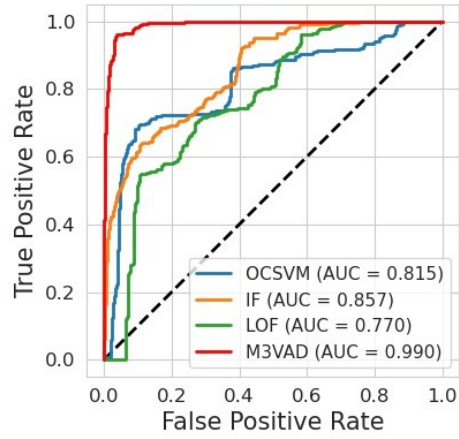
The results, summarized in Figure 3.13, demonstrate that M3VAD consistently achieves higher ROC-AUC scores than traditional OCL algorithms across all datasets. The second-best method varies depending on the dataset. For instance, in the HAM dataset, M3VAD achieves an ROC-AUC of 0.99, whereas Isolation Forest is the second-best with 0.85. In the BM dataset, M3VAD attains 0.98, with LOF ranking second at 0.79. Similarly, in the AnoVG dataset, M3VAD outperforms Isolation Forest by a notable margin (0.94 vs. 0.73). In the UCSDped2 dataset, the performance difference is smaller, with M3VAD reaching 0.71 and LOF achieving 0.70. These results underscore the robustness of our anomaly score in diverse anomaly detection scenarios.

The superior performance of M3VAD can be attributed, in part, to the refinement of the clustering algorithm through the prototypical loss. By minimizing $\mathcal{L}_{prot}$, each multimodal representation $\mathbf{r}_i$ is encouraged to align more closely with its correspond-

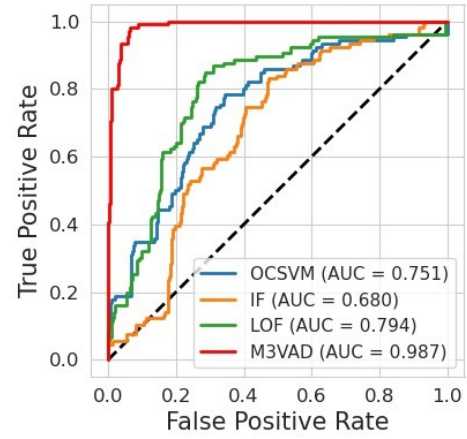
ing multimodal prototype $\mathbf{c}_i$, while maximizing the distance to other prototypes $\mathbf{c}_m$, as explained in Section 3.3.2.2. This refinement improves the structuring of the feature space, contributing to more effective anomaly detection. While the observed improvements suggest that prototypical loss enhances the clustering process, further analysis with alternative loss functions could provide additional insights into its specific impact. Our future work could explore variations in the clustering objective to further validate the role of $\mathcal{L}_{prot}$ in anomaly detection. The limitations of the OCL methods in this study can be explained by their sensitivity to the nature of multimodal features. OCSVM, for instance, struggles with high-dimensional spaces and is highly sensitive to the hyperparameter setting, leading to inconsistent results in an unsupervised setting [100]. Isolation Forest, which relies on axis-parallel data partitioning, is often ineffective in complex feature spaces where anomalies are not easily separable [98]. LOF, on the other hand, suffers from the curse of dimensionality, which makes distance-based anomaly detection unreliable in high-dimensional feature representations [99]. Given that M3VAD refines the clustering structure within the same feature space, its superior performance highlights the effectiveness of the prototypical loss in structuring meaningful clusters for anomaly detection.

3.4.4.3 Effectiveness of Multimodal and Unimodal Objectives

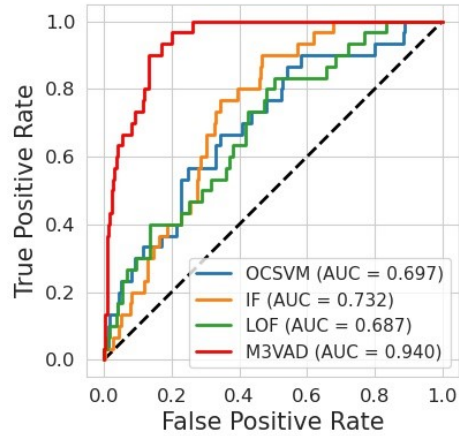
We further investigate the effectiveness of the components in our method, i.e., the impacts of jointly considering unimodal multiview and multimodal objectives in M3VAD to gain a better understanding of the contribution of each component. The goal of this study is to determine whether both types of objectives independently contribute to improving the model’s ability to detect anomalies. To ensure a fair evaluation, we followed the same experimental framework used for the baseline methods discussed in



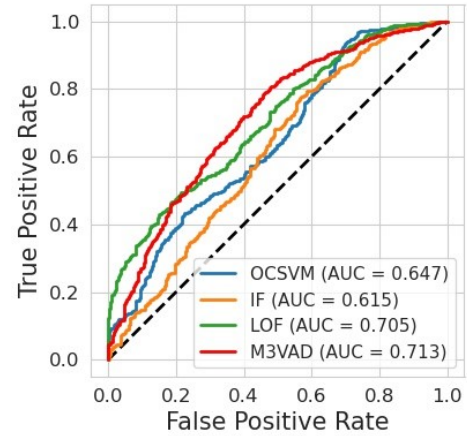
(a) HAM Dataset



(b) BM Dataset



(c) AnoVG Dataset



(d) UCSDped2 Dataset

Figure 3.13: Comparison with OCL methods in terms of ROC curves and ROC-AUC scores.

Section 3.4.2.3, applying the same feature extraction process and test-time anomaly detection approach.

The evaluation consists of two experimental setups. The first considers the impact of multimodal objectives, specifically the cross-modal alignment and prototype-based clustering, by considering only the multimodal loss functions while omitting unimodal objectives. We intentionally avoid handling these multimodal objectives separately for two reasons. First, the prototype objective serves as an auxiliary component, enhancing the model’s understanding of normality by refining the clusters. Second, considering only the cross-modal objective would yield results equivalent to Chapter 2, which were discussed in Section 3.4.3.

The second setup isolates the effect of unimodal objectives by training separate models using only image-based and text-based contrastive learning without multimodal components. These controlled experiments allow us to determine whether the combination of both multimodal and unimodal objectives leads to superior anomaly detection performance.

To investigate the role of multimodal learning, we implement a variation of M3VAD that relies exclusively on the multimodal clustering and alignment objectives (denoted as **MM_prototype**). Similarly, to analyze the importance of unimodal contrastive objectives, we implement variations trained using either image-based nearest-neighbor contrastive learning (**NN_image**) or text-based nearest-neighbor contrastive learning (**NN_text**). These variations ensure that each objective’s contribution is isolated while keeping all other settings consistent.

The results summarized in Table 3.4 confirm that the full M3VAD model outperforms all individual variations, demonstrating that integrating both multimodal and unimodal objectives is essential for improved performance. The multimodal prototype-based clustering significantly improves anomaly detection, as evidenced by the lower

Table 3.4: Ablation study of the key components of M3VAD.

Dataset Model	ROC-AUC				AUPRC			
	HAM	BM	AnoVG	UCSD Ped2	HAM	BM	AnoVG	UCSD Ped2
MM prototypes	0.923	<u>0.968</u>	0.864	0.636	0.692	<u>0.851</u>	0.408	0.887
NN_Text	0.843	0.878	<u>0.895</u>	0.573	0.685	0.577	<u>0.480</u>	0.861
NN_Image	<u>0.932</u>	0.945	0.751	<u>0.662</u>	<u>0.812</u>	0.670	0.300	<u>0.902</u>
M3VAD	0.99	0.987	0.940	0.713	0.960	0.905	0.662	0.907

performance of the models that rely solely on unimodal learning. Similarly, the NN_text model achieves the best performance among variations for the AnoVG dataset, highlighting the robustness of the textual features in certain scenarios.

The effectiveness of the textual features in providing semantic consistency is evident in situations where intra-object variations occur. However, the reliance on textual descriptions also introduces challenges when captions are inaccurate. As shown in Figure 3.14, the test sample’s caption does not precisely match the visual content, as it describes the woman as “holding” the teddy bear, whereas in the training image, the teddy bear is on the table. Despite this discrepancy, the caption still retains the core semantic elements of the scene, namely the woman and the teddy bear, ensuring a degree of consistency in the latent space. This property can be beneficial when captions provide correct but flexible representations of the scene, allowing the model to generalize across variations in pose, angle, or object placement. However, when captions are significantly inaccurate, they introduce misleading relationships between normal and anomalous instances, affecting the similarity-based anomaly score. This issue is particularly evident in the UCSDped2 dataset, where lower-quality captions, as shown in Figure 3.11, lead to incorrect feature alignments, reducing the effectiveness of the similarity component in the anomaly score. As a result, the distance score



a woman sitting in a chair,teddy bear on the table.



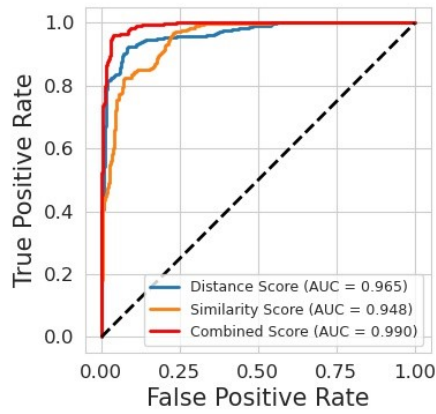
woman sitting in a chair, woman is holding a teddy bear.

Figure 3.14: Training and test images from the HAM dataset taken on different days. On the left: a training sample showing a woman with a teddy bear on the table. On the right: a test sample of the same woman with the teddy bear on the table. The test image’s caption is inaccurate but still relates to the woman and the teddy bear.

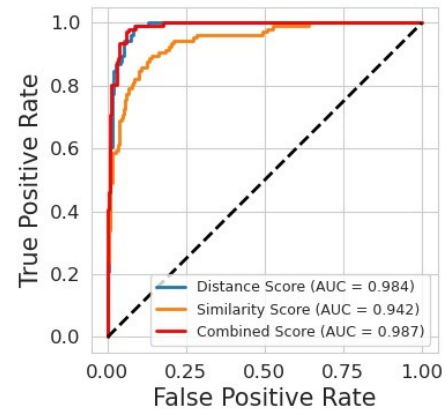
outperforms the combined similarity-distance score in that dataset, as seen in Figure 3.15. These findings emphasize that while language features improve robustness to intra-object variations, their effectiveness is dependent on caption accuracy.

Further considering the contribution of multimodal and unimodal objectives, we also analyze the effectiveness of our anomaly score function. Specifically, we examine the impact of using (1) the distance-based score, (2) the similarity-based score, and (3) the combined anomaly score in Equations (3.15), (3.13), and (3.14), respectively. Figure 3.15 illustrates the ROC-AUC scores across all datasets for each scoring method.

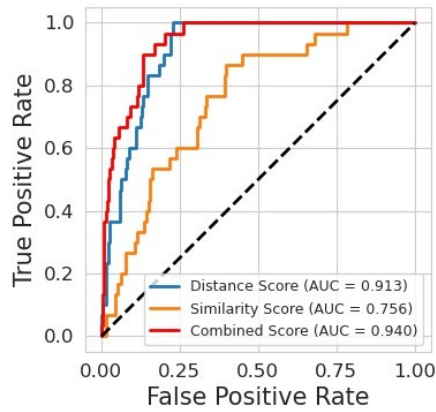
The results demonstrate that the combined anomaly score consistently achieves the highest ROC-AUC scores for the HAM, BM, and AnoVG datasets. Additionally, the ROC curve of our method almost always dominates those of the distance and similarity scores, confirming the robustness of our approach. These findings validate that the integration of both scores enhances detection performance by leveraging complementary aspects of multimodal embeddings.



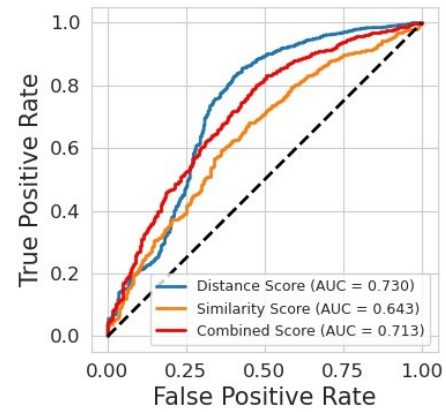
(a) HAM Dataset



(b) BM Dataset



(c) AnoVG Dataset



(d) UCSDped2 Dataset

Figure 3.15: Comparison of anomaly scores in terms of ROC curves and ROC-AUC scores.

3.5 Conclusion

In this chapter, we proposed an end-to-end contrastive VLM for anomaly detection to capture and extract abstract contextual information across modalities using multimodal and multiview objectives. We see that the combination of unimodal pre-trained features is insufficient in quality for human monitoring. Therefore, improving the quality of the feature space allows the cross-modal features to create well-defined groups. At the same time, we improved the groups in the latent space, learning the distribution of normality by a clustering algorithm and a contrastive objective based on prototypes. We also notice that the majority of contrastive VLM methods need large-scale datasets to learn high-quality representations. Therefore, using multiple views of the same data facilitates the learning of high-quality multimodal representations with small-scale data. Experiments on four real-world datasets demonstrate the superiority of our method in terms of detecting unseen anomalies which contain contextual information.

There are several directions which can be explored for future research, including (1) proposing a cross-modal attention mechanism to discard parts of the feature vectors in each modality that do not contribute to the detection, (2) determining the level of contribution of each part of an image and sentence to add an explanatory component to the detection method, (3) incorporating a video captioning method to improve captions on datasets that contain videos, and (4) extending our method to recognize anomalies in videos more accurately by including temporal relationships in the model.

Chapter 4

Leveraging Vision Language Models for Understanding and Detecting Violence in Videos

4.1 Overview

This Chapter builds on the vision language modeling approaches introduced in Chapters 2 and 3 and addresses the specific challenges of violence detection, focusing on its unique temporal and ethical complexities in video-based human monitoring scenarios.

As discussed in Section 1.3, existing approaches based on handcrafted features, supervised deep learning, or generative VLMs often fail to generalize across diverse environments and struggle with contextual bias, limited data, and lack of transparency.

To address these challenges, we introduce VIVID (Vision-Language Integration for Violence Identification and Detection), a generative VLM-based framework designed for interpretable and bias-aware violence detection. VIVID extends prior methods by

incorporating both temporal modeling and retrieval-augmented contextual reasoning. Inspired by the representation learning strategies used in CMSSL-AD and M3VAD, VIVID aligns visual and textual features in a shared semantic space and enhances decision-making by incorporating external definitions of violence.

The proposed framework consists of two stages: (1) key-frame selection based on motion analysis to reduce redundancy and focus on high-impact segments, and (2) Retrieval-Augmented Context Awareness, where VLMs generate multimodal embeddings of selected frames and compare them against violence-related definitions retrieved from an external knowledge base. This design allows the model to make classification decisions that are context-sensitive and semantically grounded, improving both robustness and transparency.

This Chapter introduces a generative VLM framework that enhances interpretability and bias awareness through semantic grounding in external knowledge. By combining key-frame selection with retrieval-augmented classification, the proposed method improves efficiency and robustness in data-scarce video-based monitoring. Through extensive evaluation, VIVID demonstrates its ability to generalize across diverse scenarios and provides a transparent decision-making process tailored to the ethical and contextual demands of violence detection. This Chapter concludes the thesis’s exploration of contrastive and generative VLM paradigms, illustrating how external knowledge and semantic alignment can overcome the limitations of conventional models in human monitoring.

4.2 Problem formulation

Our objective is to detect violence within video clips in the context of human monitoring. Given the scarcity, diversity, and limited availability of labeled data specifically

containing violent content, we approach this problem as a zero-shot violence detection scenario, i.e., without prior training on video clips from target datasets that include violence or non-violence classes.

Following this setting, the model receives as input a set of video clips $\{(\mathbf{C}_i)\}_{i=1}^n$. For $\mathbf{C}_i$, the model produces two outputs: an associated class prediction $\hat{y}_i \in \{0, 1\}$ and an explanatory text $\mathbf{e}_i$ detailing the classification decision. In this context, a class label of 0 represents non-violent content, while 1 denotes violent content.

Our focus lies in the specific detection of physical violence defined as an act attempting to cause, or resulting in, pain and/or physical injury, or damage to the state of something [101].

To assess the effectiveness of our method, we employ accuracy and F1-score. These evaluation measures collectively provide insights into the model’s performance, ensuring robustness and reliability in detecting violent content within video clips.

4.3 Proposed method

4.3.1 Overview

We propose VIVID: **V**ision-**L**anguage **I**ntegration for **V**iolence **I**dentification and **D**etection. This method leverages the strengths of Vision-Language Models (VLMs) and Language Models (LLMs) to provide a robust, interpretable solution for video violence detection.

The detailed steps of the VIVID algorithm are outlined in Algorithm 2. It involves two main processes: key-frame selection and violence detection.

In the first step, key-frame selection, the model utilizes optical flow [102] analysis to quantify motion within the video. For each frame $\mathbf{I}_t$ in the video $\mathbf{C}$, the magnitude

Algorithm 2: Violence Detection using VIVID for a single video clip **C**

Input: Video clip **C**, Knowledge base **KB**, Number of frames k

Output: Prediction $\hat{y}$ and Explanation **e**

```
1 // Key-frame Selection;
2 key_frames  $\leftarrow$  Top  $k$  frames with the greatest magnitude of motion;
3 // Modality Translator;
4 definitions  $\leftarrow$  [];
5 for each frame  $I_t \in \text{key\_frames}$  do
6      $V_t \leftarrow \text{Q\_Former}(I_t)$ ;
7      $V_s^t \leftarrow \text{mean}(V_t)$ ;
8      $D_s^t \leftarrow \text{retrieve\_definition}(V_s^t, \text{KB}, \tau_r)$  ; definitions.append( $D_s^t$ );
9 // Definition Retriever;
10  $D_s \leftarrow \text{select\_definition}(\text{definitions})$ ;
11  $V_t' \leftarrow \text{reorder\_queries}(\text{queries}, D_s)$ ;
12 // Modality Translator;
13  $T_v \leftarrow \text{FC\_layer}(V_t')$ ;
14 // Output Generator;
15  $Q_u \leftarrow$  "Is this video violent?";
16  $T_u \leftarrow \text{LLM\_tokenizer}(Q_u)$  ;  $T_d \leftarrow \text{LLM\_tokenizer}(D_s)$ ;
17 final_prompt  $\leftarrow \text{create\_prompt}(T_u, T_v, T_d)$ ;
18  $(\hat{y}, e_t) \leftarrow \text{LLM\_process}(\text{final\_prompt})$ ;
19 return  $(\hat{y}, e)$ 
```

of motion is computed and stored. The frames with the highest motion activity are selected as keyframes, capturing essential moments that potentially contain violent actions.

In the second step, violence detection, the model performs several sub-tasks to analyze the keyframes. It begins by translating each frame into visual queries $\mathbf{V}_t$ through the Q-Former [17], enabling direct comparison with text embeddings. The definitions related to each query $\mathbf{D}_t$ are then retrieved from the knowledge base **KB**. The most relevant definition for the video $\mathbf{D}_s$ is then selected from the list that contains all the definitions of each frame. The **append**(x, y) function is used to add the element y to the list x . In this context, it is utilized to add new items to lists **queries** and **definitions** during the processing steps.

$\mathbf{D}_s$ is used to reorder $\mathbf{V}_t$ to mitigate the ‘lost in the middle’ effect, which will be explained in detail in Section 4.3.3.2. The reordered queries $\mathbf{V}'_t$ are subsequently processed through a fully connected layer to create visual tokens. A prompt is created using the visual tokens, definition tokens, and user query tokens. The final prompt is then processed by the language model to produce a label indicating whether the content is violent or not, along with an explanatory text detailing the classification decision.

A summary of the architecture is shown in Figure 4.1.

4.3.2 Key-frame selection

As we explained in Section 1.2, violent content is scarce and has limited availability. This scarcity makes it challenging to train models that can generalize patterns associated with violent actions. To address this issue, we propose using optical flow to extract potential violent frames.

Optical flow [102] effectively captures dynamic movements in video sequences,

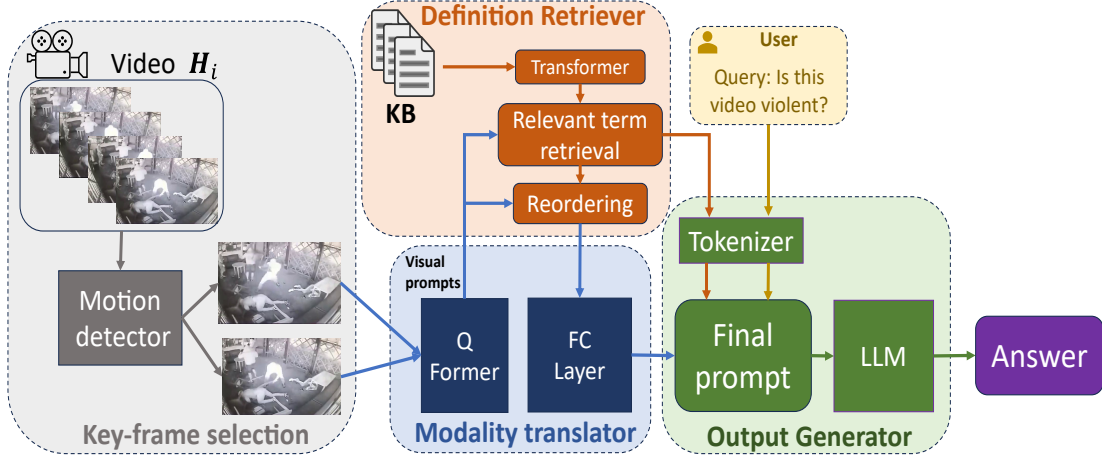


Figure 4.1: Architecture of the proposed method.

which are often indicative of violent behaviors. By semantic definition, a violent action involves forceful behaviors like hitting, kicking, or grappling, which inherently exhibit a greater degree of motion compared with other non-violent activities [103]. Consequently, we extract potential violent frames by calculating the norm of dense optical flow vectors, summarizing the overall motion magnitude.

Several alternatives to optical flow could be considered for frame extraction, including traditional approaches like Frame Differencing and Background Subtraction [104], as well as more advanced techniques such as Motion History Images (MHI) [102]. For instance, Frame Differencing calculates the difference between consecutive frames to detect changes in the scene. While simpler and computationally less expensive, it may miss subtle movements and fail to capture complex motion patterns associated with violent actions [104]. Similarly, Background Subtraction isolates moving objects by subtracting a background model from each frame. However, it is susceptible to lighting changes and background clutter, leading to false detections [104]. MHIs create a motion that represents a single image over time by accumulating motion information. However, they may blur detailed motion information and do not provide

designated frames in the video [105], making it challenging to identify and extract frames corresponding to specific violent actions.

Unlike these alternatives, optical flow offers a balanced approach that effectively captures dynamic motion. It quantifies the velocity and direction of motion at each pixel, providing detailed information about movements within the scene [104]. This granularity is essential for identifying violent actions, which typically involve rapid and forceful movements. Additionally, optical flow is computationally efficient and can be applied to both low-resolution and high-resolution video data.

The importance of using optical flow in violence detection scenarios has been emphasized in several works [15, 106–108]. These methods utilize optical flow to extract trajectories, build descriptors, or extract features from video data. However, they primarily use optical flow for training models to learn patterns from violent video content, whereas our approach focuses on selecting key-frames directly based on motion magnitude.

Consider video clip $\mathbf{C}$ that is decomposed into a set of frames, where a frame at time t , represented as $I(x, y, t)$, with (x, y) being the pixel coordinates. The dense optical flow relates the partial derivatives of image intensity with respect to spatial coordinates and time as follows:

$$I_x \cdot u_x + I_y \cdot u_y + I_t = 0, \quad (4.1)$$

where, I_x and I_y are the spatial gradients of image intensity in the x and y directions, respectively, and I_t is the temporal gradient. u_x and u_y represent the flow velocities in the x and y directions, respectively. The norm of optical flow for a single pixel is calculated as:

$$|u| = \sqrt{(u_x)^2 + (u_y)^2}. \quad (4.2)$$

By summing the norms across all pixels in the frame, we obtain the total motion magnitude for the frame as:

$$U_t = \sum_x \sum_y |u(x, y, t)|. \quad (4.3)$$

After calculating U_t for different frames along the video, we extract the top k frames with the highest values, which we assume correspond to the frames containing the highest activity. These frames serve as input for the subsequent analysis.

We acknowledge that extracting keyframes with the highest activity does not guarantee selecting all violent frames. For instance, high-speed motion, such as a passing car, might deceive the optical flow selector, causing it to neglect frames with slow-motion violence, such as a person threatening another with a knife. In such cases, VIVID may fail to detect the violent scenario. We leave this challenge to our future work, which can be addressed by incorporating the detection of potentially dangerous objects during the frame selection process.

4.3.3 Violence detection

In this step, we combine a Vision Language Model (VLM) [17] with a Large Language Model (LLM) to detect violence in videos. The detection process involves several key modules working together, as illustrated in Figure 4.1.

The Modality Translator component translates visual data from the video frames into sequence of tokens, which can be processed by the LLM. This process involves using the VLM’s Q-Former to generate visual queries $\mathbf{V}_t$. After reordering the queries $\mathbf{V}'_t$, they are processed by a fully connected (FC) layer that converts them into tokens $\mathbf{T}_v$ compatible with the LLM. This translation ensures that the visual information is appropriately represented for subsequent analysis.

Following this step, the Retrieval module plays a critical role. It retrieves relevant information by comparing $\mathbf{V}_t$ with the definitions in the knowledge base $\mathbf{KB}$, identifying the most pertinent violence-related definitions $\mathbf{D}_t$. From these steps, we extract the most relevant definition $\mathbf{D}_s$ for the entire video. To address the ‘lost in the middle’ effect often observed in LLMs, the visual queries are reordered to prioritize the most relevant information, ensuring the LLM receives a coherent and contextually appropriate set of inputs.

Finally, the Generator module combines $\mathbf{T}_v$, user query tokens $\mathbf{T}_u$, and the most relevant definition tokens $\mathbf{T}_d$ to create a comprehensive final prompt. This prompt is then processed by the LLM, which generates a response to the user’s query, determining whether the video contains violent content.

4.3.3.1 Modality translator

As shown in Figure 4.1, we selected an architecture based on BLIP-2 [17] for the VLM. This architecture includes a Q-Former, which provides visual queries containing visual representations $\mathbf{V}_t$ corresponding to the text, making them directly comparable with text embeddings in a common subspace. Additionally, the VLM has a fully-connected (FC) layer that linearly projects the reordered visual queries $\mathbf{V}'_t$ to match the input dimension of LLM. This design means that the FC layer acts as a visual tokenizer, converting the reordered visual queries into a set of tokens $\mathbf{T}_v$ that the LLM can process. The reordering of visual queries ensures that the most relevant visual information is emphasized, which is crucial for accurate analysis. This process will be explained in detail in Section 4.3.3.2.

4.3.3.2 Definition Retriever

In this Section, we explain the processes of retrieving the definitions to ensure their relevance and accuracy for violence detection. We break down this Section into three sub-parts: Creating KB, where we define and build the knowledge base of violence-related terms and definitions; Relevant Term Retrieval, which covers how visual queries are compared with the knowledge base to extract the most relevant definitions; and Reordering Visual Queries, where we discuss the reordering of visual queries to prioritize the most pertinent information, mitigating the ‘lost in the middle’ effect in LLMs.

Creating KB Retrieval Augmented Generation (RAG) [109] enhances LLMs by integrating external knowledge, such as databases, to improve performance in knowledge-intensive or domain-specific applications that require continually updating information [110]. RAG retrieves relevant documents based on the user’s query and combines them with the original prompt to generate a response.

Our proposed method addresses biases within LLMs by drawing inspiration from the RAG framework. However, instead of adding new information, it focuses on mitigating existing biases related to sensitive topics like violence. To achieve this goal, we leverage the VLM output, which provides informative visual representations $\mathbf{V}_t$ corresponding to the text. Instead of relying on user queries to retrieve documents, we compare these visual representations with a predefined set of violence-related terms and definitions, referred to as the knowledge base **KB**.

Since, to the best of our knowledge, there is no existing comprehensive knowledge base of violence-related terms, we propose one that is defined as:

$$\mathbf{KB} = \{(w_1, d_1), (w_2, d_2), \dots, (w_n, d_n)\}, \quad (4.4)$$

where w_i represents the i -th violence-related term (word) and d_i represents the corresponding definition.

We created the knowledge base (KB) of violence-related terms, comprising 48 entries, each explaining various aspects of physical violence. These aspects include direct physical contact, violent sports, collective violence, and broader terms denoting physical force [111–114].

First, direct physical contact encompasses terms such as “physical abuse”, “physical disputes”, “rape”, “battery”, and “fights”, describing incidents where physical harm is directly inflicted upon individuals.

Regarding terms that frequently escalate into physical confrontations, this includes “clash”, “assault”, “aggression”, and “confrontation”, highlighting situations that may start as verbal or non-physical conflicts but can quickly escalate to physical violence.

In terms of violent sports, these are activities where the objective is to subdue opponents through physical force, such as “boxing”, “wrestling”, “muay Thai”, and “karate”, where physical engagement is a structured and accepted part of the sport.

For collective violence, the terms include “riot”, “brawl”, “vandalism”, and “scuffle”, reflecting scenarios where multiple individuals are involved in violent acts, often leading to public disorder or damage.

Lastly, broader terms denoting physical force encompass words like “war”, “savagery”, “oppression”, “bullying”, and “genocide”, representing extensive and severe forms of violence, highlighting large-scale or systematic acts of harm and intimidation.

This diverse compilation ensures a comprehensive understanding of violence across different contexts. The definitions were obtained from well-known sources, including the Cambridge [113] and Oxford [114] dictionaries, Wikipedia [111], and Wex [112], a legal dictionary sponsored by the Legal Information Institute at Cornell Law School.

Relevant Term Retrieval By comparing $\mathbf{V}_t$ with the violence-related terms in the knowledge base, we identify the most relevant violence-related definition $\mathbf{D}_t$ associated with each content frame. Specifically, we calculate the similarity between the visual representations of each frame and the textual definitions.

Given a video frame $\mathbf{I}_t \in \mathbb{R}^{H \times W \times 3}$, the pre-trained Q-Former produces visual queries $\mathbf{V}_t \in \mathbb{R}^{N \times d_q}$. Here, N is the number of visual queries obtained by the Q-Former, and d_q is the dimension of the visual embedding of each query, respectively. Subsequently, we use the pre-trained text transformer from BLIP2 to obtain the definition embedding $\mathbf{D}_t \in \mathbb{R}^{M \times d_q}$. Here M is the number of text tokens, and the dimension of the embedding of each token are set to d_q . For each frame, we retrieve the definition from $\mathbf{KB}$ that has the smallest squared Euclidean (L2) distance to the visual embedding.

In BLIP2 [17], to calculate the contrastive loss, they propose to compare the [CLS] token from the text embedding and the visual query with the highest pairwise similarity. However, since our goal is not to align visual and text representations, but rather to select the most relevant violent definition for the video, we use mean pooling to compare the entire visual sentence with the entire definition sentence. Specifically, through the mean-pooling step, we obtain and compare the visual sentence $\mathbf{V}_s^t \in \mathbb{R}^{d_q}$ and definition sentence $\mathbf{D}_s^t \in \mathbb{R}^{d_q}$. After processing the top k frames, we select the definition with the highest pairwise similarity across all frames as most relevant for the video. We denote the definition sentence vector of the selected definition as $\mathbf{D}_s$.

To further ensure semantic relevance, we apply a retrieval threshold τ_r to filter out definitions whose similarity scores fall below a predefined value. In RAG systems, a retrieval threshold is commonly used to filter candidate documents based on their similarity to the input prompt. In our case, only definitions with similarity scores above this threshold are passed to the Reordering module. Less relevant candidates are discarded. This filtering process helps reduce the risk of incorporating unrelated or

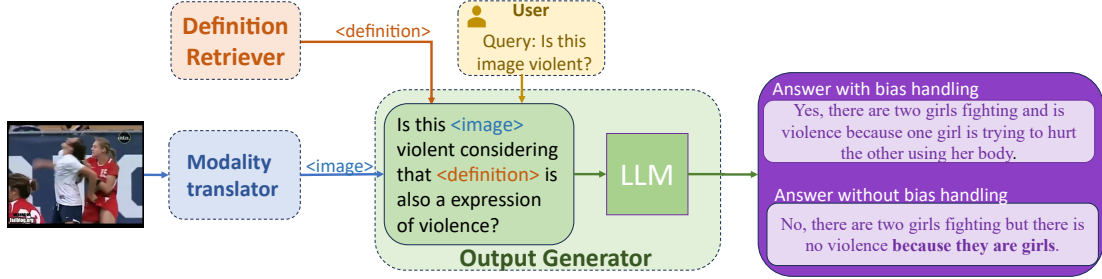


Figure 4.2: Example of handling bias.

misleading context, which in turn mitigates hallucinations [115].

Reordering Visual Queries The ‘lost in the middle’ effect is a phenomenon observed in LLMs where the model’s performance degrades when critical information is situated in the middle of a long input context [116]. This phenomenon occurs because LLMs tend to focus more on the input’s beginning and end, often neglecting the information in the middle. In our method, this effect poses a significant challenge.

To mitigate the ‘lost in the middle’ effect, we propose reordering the visual queries to prioritize those that are more relevant to the violence definition. By aligning the visual queries ($\mathbf{V}_t$) with the violence definition ($\mathbf{D}_s$), the LLM receives a coherent and contextually appropriate set of inputs. This alignment ensures that the most critical information is positioned at the beginning of each frame, where the LLM is more likely to focus its attention.

By doing so, the LLM can better understand whether the visual content is directly related to the definition of violence, enabling it to disregard irrelevant definitions when necessary. Overall, this approach not only addresses the ‘lost in the middle’ effect but also enhances the system’s robustness across diverse scenarios.

To reorder the visual queries, first, we compute a similarity vector $\mathbf{S}$ by calculating

the dot product between $\mathbf{V}_t$ and $\mathbf{D}_s$ as follows:

$$\mathbf{S} = (\mathbf{V}_1 \cdot \mathbf{D}_s, \dots, \mathbf{V}_t \cdot \mathbf{D}_s). \quad (4.5)$$

This similarity vector indicates how closely each visual query aligns with the definition. Next, we sort the indices of the visual queries based on their similarity scores in descending order. This sorting arranges the visual queries from the most to the least relevant according to their similarity to the definition.

Finally, the visual queries are reordered using these sorted indices, resulting in the reordered visual queries $\mathbf{V}'_t$. This process is repeated for each of the top k frames to ensure that the LLM receives a consistent and contextually relevant input.

4.3.3.3 Output Generator

To create the final prompt for the LLM, we extract the tokens from the user query $\mathbf{T}_u$ and the most relevant definition $\mathbf{T}_d$, obtained directly by the LLM tokenizer. The text embeddings are used solely for comparison purposes between visual and text data: first, to obtain the most relevant violence-related definition for the video, and second, to reorder visual queries according to this definition.

Finally, we receive from the Modality Translator module $\mathbf{T}_v$ for the corresponding frame. We then combine $\mathbf{T}_u$, $\mathbf{T}_v$, and $\mathbf{T}_d$ to form the **final_prompt**: “Is this video: $\langle \text{frame_1} \rangle, \dots, \langle \text{frame_k} \rangle$ violent, considering that $\langle \text{definition} \rangle$ is also an expression of violence?” This prompt enhances LLM’s understanding of the concept of violence, helping to mitigate biases and ensure robustness across diverse scenarios. Figure 4.2 illustrates an example using a single frame.

4.4 Experiments

4.4.1 Datasets

Table 4.1 summarizes the five datasets we used in the experiments.

Hockey Fight Detection Dataset [33] was specifically created to introduce a new video dataset for evaluating violence detection systems, where both normal and violent activities occur in similar, dynamic settings. It contains 1 000 clips of action from the National Hockey League (NHL) games. Each clip comprises 50 frames of 720×576 pixels and is manually labeled as “fight” or “non-fight”.

Movies Fight Detection Dataset [33] was constructed to explore the generalization capacity of learning fight patterns. It consists of 200 video clips obtained from action movies, of which 100 contain fight scenes, and 100 videos with non-fight scenes from football games, and other events. The dataset contains videos depicting a wide variety of scenes that were captured at different resolutions and are manually labeled as “fight” or “non-fight”.

Surveillance Camera Fight Dataset [41] was collected mostly from YouTube. It encompasses various fight scenarios, including kicks, fists, hitting with objects, and wrestling. The dataset features footage from different locations such as cafes, bars, streets, buses, and shops. In total, there are 300 videos in the dataset: 150 depict fighting sequences, while the remaining 150 show non-fighting sequences. These videos vary in resolution and frame numbers, and have an approximate duration of two seconds each.

Real-World Fighting-2000 Dataset (RWF-2000) [40] was collected from YouTube. It consists of 2 000 video clips captured from around 1 000 raw surveillance videos with extended footage. This dataset is split into two parts: the training set (80%) and the

Table 4.1: Dataset description.

<i>Dataset</i>	<i>Data Scale</i> (# clips)	<i>Length/Clip</i>	<i># Violent clips</i>	<i># Non-violent clips</i>	<i>Source of scenarios</i>
Movies	200	1.6-2 sec	100	100	Movies and sports
Surveillance	300	1.4-2 sec	150	150	CCTV and mobile cameras
Fight					
RWF-2000 ²	400	5 sec	200	200	CCTV and mobile cameras
Hockey	1000	1.6-1.96 sec	500	500	Ice hockey games
XD-Violence ²	800	0-10 min	500	300	Movies, sports, games, hand-held cameras, CCTV, car cameras, etc.

test set (20%), in which half depicts violent behaviors, while the other half represents non-violent activities. The videos have different resolutions and are uniformly cut into 5-second clips with a frame rate of 30 frames per second.

XD-Violence Dataset [117] is a large-scale and multi-scene dataset that was collected from movies and YouTube. The dataset includes a total of 91 movies, which were used to collect both violent and non-violent events. Additionally, in-the-wild videos were collected from YouTube. This dataset has a total duration of 217 hours, containing 4 754 untrimmed videos with audio signals and weak labels, from which 2 405 are classified as violent, while 2 349 are non-violent. The dataset is split into two parts: a training set containing 3 954 videos and a test set with 800 videos. The test set consists of 500 violent videos and 300 non-violent videos.

4.4.2 Implementation details

In the Key-frame selection step, we utilized the Gunnar Farneback technique, which is implemented in the Open Source Computer Vision Library (OpenCV)¹ to calculate dense optical flow. We set the parameter $k = 5$, meaning that five frames are selected to represent each video. This selection helps in reducing the computational load while maintaining essential motion information. We employed a divide-and-conquer strategy for longer videos, which we define as those containing more than 5 000 frames. This approach involves splitting the video into up to five parts, treating each part independently to manage the data more efficiently. A video is classified as violent if at least one of its parts contains violent content.

For the VLM (modality translator), the architecture includes a Q-Former with $N = 32$ visual queries, each having a dimension of $d_q = 768$ [17]. Additionally, the VLM has an FC layer that linearly projects each output visual query embedding to match the text embedding dimension of the LLM, resulting in a 2048-dimensional output layer. The retrieval threshold is set to $\tau_r = 1.45$.

For the LLM, we chose the instruction fine-tuned version of Google’s T5 [118], also known as the Text-to-Text Transfer Transformer Language Model (Flan-T5) [119]. Since we intend for VIVID to be used in detecting violence in real-time scenarios, we selected the Flan-T5-XL [119], which contains around 3 billion parameters. We set the generation parameters to temperature = 1.0, top-p = 0.9, repetition penalty = 1.0, and number of beams = 5. This model size allows for a nuanced understanding and processing of complex language inputs on a single NVIDIA GeForce RTX 3090 GPU, which has 24 GB of GDDR6X memory, and a boost clock of 1 700 MHz.

¹<https://opencv.org/>

²The Data Scale column lists video clips from the test split of the dataset, used for model comparison. We use only the test split to align our evaluation with other methods that measure performance using the test split.

4.4.3 Experimental results

4.4.3.1 Comparison with LLM-based methods

We compare VIVID with other LLM-based methods using the test split of video clips from each dataset as input, along with the user prompt: "Is this video violent?" This user query is consistently applied across all baseline methods to ensure a fair comparison.

The results of our experiments, as summarized in Table 4.2, reveal significant insights into the performance of various models on the five datasets. VIVID consistently outperforms other LLM-based models across most datasets, achieving improvements of 0.01–0.324, 0.007–0.187, 0.014–0.294, and 0.066–0.334 in terms of accuracy, and 0.01–0.492, 0.009–0.446, 0.011–0.277, and 0.062–0.537 in terms of F1-Score on the Movies, RWF-2000, Hockey, and XD-Violence datasets, respectively.

In the Movies dataset, VIVID significantly outperforms other LLM-based models by effectively capturing complex visual and contextual cues. It also leads in the RWF-2000 dataset, demonstrating its effectiveness in real-world fight detection. The Hockey dataset results show substantial improvements, highlighting VIVID’s capability to detect violent actions in fast-moving and complex interactions. In the XD-Violence dataset, VIVID’s robustness is validated by its ability to handle diverse and complex violent scenarios.

The effectiveness of our method lies in its ability to identify potential violent frames and manage various biases associated with LLMs: inherited bias, guardrails bias, and alignment bias. Figure 4.3 shows the classification made by VIVID over examples where our method successfully identifies the videos as violent. As illustrated in Figure

³BLIP2 [17] and InstructBLIP [18] do not support video input. We analyze violent content at the frame-level, extracting frames as in X-InstructBLIP [57]. If any frame contains violent content, the entire video is classified as violent.

Table 4.2: Comparison with LLM-based methods in terms of Accuracy and F1-Score.

Model	LLM Backbone	Movies		Surveillance Fight		RWF-2000		Hockey		XD-Violence	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BLIP2 ³	Flan-T5-XL (3B)	0.900	0.880	0.776	0.797	0.767	0.760	0.841	0.859	<u>0.760</u>	0.770
InstructBLIP ³	Vicuna-1.1 (7B)	0.810	0.765	0.796	0.784	0.742	0.685	0.932	0.932	0.535	0.436
X-InstructBLIP	Vicuna-1.1 (7B)	0.880	0.863	<u>0.793</u>	<u>0.800</u>	0.785	0.766	0.856	0.831	0.648	0.616
Video-LLaMA	LLaMA-2 (7B)	0.661	0.492	0.656	0.511	0.610	0.365	0.660	0.676	0.492	0.330
Video-LLaVA	Vicuna-1.5 (7B)	<u>0.975</u>	<u>0.974</u>	0.726	0.777	0.782	<u>0.802</u>	0.662	0.743	0.726	0.802
LLaMA-VID	Vicuna-1.5 (7B)	0.965	0.963	0.783	0.805	<u>0.790</u>	0.794	<u>0.940</u>	<u>0.942</u>	0.755	<u>0.805</u>
VIVID	Flan-T5-XL (3B)	0.985	0.984	0.736	0.781	0.797	0.811	0.954	0.953	0.826	0.867

4.4, incorporating a violence-related term in the prompt enables users to address the bias and influence the LLM’s judgment based on the definition of violence. The dataset [33] categorizes all martial arts as violent sports. Therefore, by including martial arts as violence-related terms in the **KB**, the LLM’s response aligns with the definition of Karate as a violent sport. This comprehensive approach ensures that our method remains robust and effective across various scenarios, providing a reliable tool for bias management in LLMs.

In the Surveillance Fight dataset, the “InstructBLIP” model shows the highest accuracy 0.796, while the “LLaMA-VID” model achieves the highest F1-Score 0.805. Our model performs well with an accuracy of 0.736 and an F1-Score of 0.781 but does not lead in this dataset. This fact suggests that while our model is effective, it may struggle with the specific challenges posed by surveillance footage, such as varying lighting conditions like insufficient illumination, and occlusions as shown in Figure 4.5. These challenges could be addressed by jointly fine-tuning the VLM and the LLM with a small amount of violent and non-violent videos under these conditions. How-

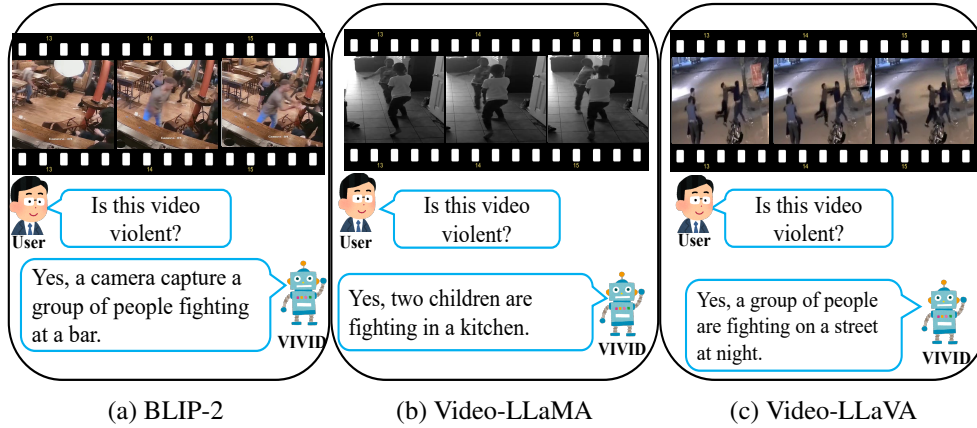


Figure 4.3: Example in which VIVID successfully handles the biases of the LLMs.

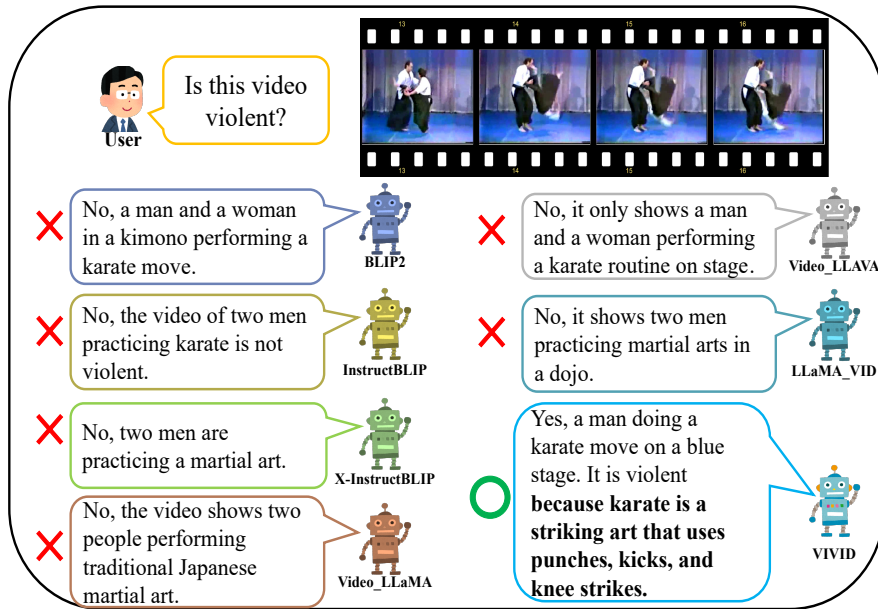


Figure 4.4: Example in which the inclusion of a violent-related definition improves performance compared to the baseline methods.

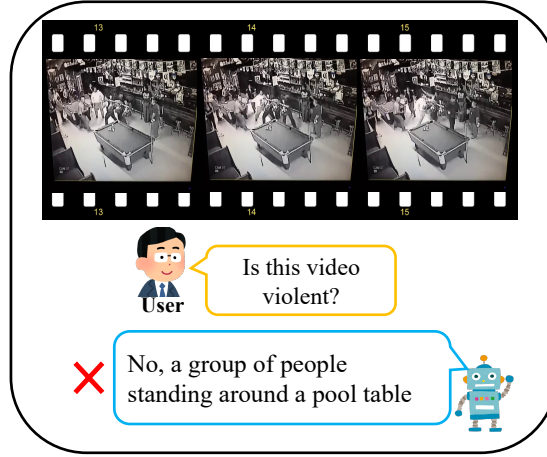


Figure 4.5: Example where occlusion caused by the pool table lights partially blocking the violent scene, along with lack of illumination, resulted in the LLM failing to detect the violent scene correctly.

ever, the primary objective of our research is to evaluate the capabilities of an LLM to classify violent content in a zero-shot manner. This approach allows us to assess the inherent strengths and limitations of the LLM without further modifications. Nonetheless, we recognize the potential for improvement through fine-tuning and consider it a promising direction for our future work.

4.4.3.2 Comparison with traditional hand-crafted features and deep learning-based methods

When comparing VIVID with traditional hand-crafted features and deep learning-based methods, as shown in Table 4.3, we notice a clear difference in performance. Traditional methods such as HOG + HIK [33], HOF + HIK [33], and MoSIFT + HIK [33] show lower accuracy across the datasets. For instance, MoSIFT + HIK achieves an accuracy of 0.895 in the Movies dataset and 0.909 in the Hockey dataset, which are significantly lower than our model’s performance. This fact indicates that traditional hand-crafted features are less effective in capturing complex patterns of

violence in videos. Deep learning-based methods, such as Xception + Bi-LSTM + attention [41], Flow Gated [40], and Conv3D [15], demonstrate high performance. For example, Xception + Bi-LSTM + attention achieves an accuracy of 1.0 in the Movies dataset and 0.98 in the Hockey dataset. However, these methods often require a considerable amount of training data to generalize. Due to the restricted amount of available training data, the resulting models are often ad-hoc and fail to generalize violent behavior across different scenarios.

In contrast, VIVID uses zero-shot classification, which leverages the pre-existing knowledge of the LLM along with bias handling to enhance performance. This approach not only effectively classifies violent content but also provides interpretability. The interpretability feature is particularly advantageous as it helps users understand why a video instance was classified as violent or not, making it a valuable tool for real-world applications. Even though VIVID does not always achieve the highest accuracy compared to some deep learning methods, its performance is competitive or comparable to these methods. Additionally, our method outperforms the Xception + Bi-LSTM + attention method in the Surveillance Fight dataset, achieving an accuracy of 0.736 compared to 0.720.

4.4.3.3 Ablation study

To investigate the effects of reordering visual queries, we conducted experiments using three distinct variants of our method. These variants share the common approach of comparing a visual sentence with text embedding to identify the most relevant violence-related definition for an image. However, in each variation, the text and visual embeddings interact differently.

Variation 1: Inspired by BLIP2 [17], we compare the [CLS] token from the text with

Table 4.3: Comparison with hand-crafted features and deep learning-based methods in terms of accuracy. A hyphen indicates that the results are not provided in the corresponding paper.

Type	Model	Movies	S. Fight	RWF 2000	Hockey
Hand	HOG + HIK	0.490	-	-	0.917
Crafted	HOF + HIK	0.590	-	-	0.886
Features	MoSIFT + HIK	0.895	-	-	0.909
DL Based	Xception				
	+ Bi-LSTM	1.000	0.720	-	0.980
	+ attention				
	Flow Gated	1.000	-	0.872	0.980
	Conv3D	1.000	-	-	0.981
LLM based	VIVID	0.985	0.736	0.797	0.954

Table 4.4: Comparison of our method’s variants in terms of Accuracy.

Model	Movies	S.	RWF	Hockey	XD
		Fight	2000		Violence
Variant 1	0.915	0.706	0.800	0.942	0.671
Variant 2	0.940	0.680	0.792	0.920	0.730
VIVID	0.985	0.736	0.797	0.954	0.826

the visual queries. The most relevant violent definition is identified as the one with the highest pairwise similarity.

Variant 2: Instead of using the [CLS] token, we use mean pooling. Specifically, we compare a visual sentence obtained by mean pooling all the visual queries with a definition sentence obtained by mean pooling all sentence tokens, as explained in Section 4.3.3.2.

VIVID: In addition to using the mean pooling approach from Variant 2, we also reorder the visual queries. By calculating the dot product between the visual sentence and each text token, we can prioritize visual queries that are more relevant to the violence definition, as explained in Section 4.3.3.2.

Table 4.4 shows the results of each variant on the five datasets. From these results, it is evident that VIVID demonstrates superior performance across most datasets. It achieves the highest accuracy in the Movies, Surveillance Fight, Hockey, and XD-Violence datasets. Although VIVID was not the best in the RWF-2000 dataset, the difference in accuracy was small enough to be negligible. Therefore, we consider VIVID to have achieved same performance to Variant 1 in this dataset. Overall, the robust performance of VIVID in four out of five datasets highlights its effectiveness in enhancing the model’s accuracy through the reordering of visual queries.

4.5 Conclusions

In this Chapter, we introduced VIVID, a novel method for detecting violent content in videos by integrating generative VLM with LLMs. Our method demonstrates superior performance across multiple datasets, effectively capturing complex visual and contextual cues.

VIVID’s zero-shot classification capability allows it to identify violent content without additional training, making it an alternative in scenarios with limited labeled data. By incorporating an external knowledge base, VIVID mitigates biases and provides reliable results. Its interpretability characteristics enhance user trust by explaining the reasons behind its classifications.

There are a number of directions that can be explored for future research, including (1) fine-tuning the generative VLM and LLM with a small amount of violent and non-violent videos to address specific challenges like varying lighting conditions and occlusions in surveillance footage, (2) including prompt engineering to have a broader range of violence-related terms and scenarios, and (3) enhancing the interpretability features to provide more detailed explanations for the model’s classifications. Overall, VIVID offers a balanced approach combining accuracy, interpretability, and generalizability, making it a robust solution for violent content classification in real-world applications.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis has explored how integrating visual and language modalities can improve the detection of anomalies and violent behaviors in human monitoring scenarios characterized by data scarcity, contextual ambiguity, and the need for interpretability. Through three methods built on the vision-language modeling paradigm, the thesis demonstrates that textual information can enhance the semantic understanding of visual data and support context-aware detection in complex settings.

Chapter 2 introduced CMSSL-AD, a contrastive VLM that applies cross-modal self-supervised learning to align visual and textual features. By maximizing mutual information between modalities, CMSSL-AD improves semantic clustering of normal behaviors and enhances anomaly detection in static image settings. The method outperforms conventional unimodal approaches by leveraging caption-guided representation learning but remains sensitive to the quality and relevance of textual inputs.

Chapter 3 extended this approach with M3VAD, a lightweight contrastive vision-language model that incorporates both multimodal alignment and unimodal multiview

learning objectives. These strategies improve the robustness and intra-modal consistency of learned features, enabling more compact clustering of normal behavior under small-scale training conditions. M3VAD also introduces a hybrid anomaly score that integrates prototype distances with cross-modal similarity, further improving performance in resource-limited environments.

Chapter 4 introduced VIVID, a generative VLM for violence detection in video. To address the challenges of limited training data and contextual bias, VIVID combines motion-based keyframe selection with retrieval-augmented context integration, grounding visual inputs in structured definitions of violence. This approach supports semantically grounded and bias-aware classification without task-specific training, improving generalization and interpretability in sensitive settings.

The contributions in this thesis demonstrate how contrastive and generative VLM paradigms can be adapted to context-sensitive human monitoring tasks, each tailored to specific requirements related to modality, temporal structure, and supervision. Together, they form a unified framework for robust and interpretable detection under real-world constraints.

5.2 Future Work

Building on the findings of this thesis, several directions can be explored to extend the proposed methods and further enhance their robustness and interpretability.

One promising direction is the incorporation of cross-modal attention mechanisms to better filter out modality-specific noise and refine the quality of shared representations. Unlike the fixed alignment strategies used here, attention-based integration could allow more flexible, fine-grained fusion of visual and language information.

Another important direction is improving interpretability through structured expla-

nation modules. While VIVID outputs language-based justifications, future systems could trace predictions back to specific image regions and semantic features, improving trust and accountability in safety-critical applications.

Improving the quality and reliability of textual input also remains an open challenge. Since caption quality significantly impacts representation learning and classification accuracy, future research could explore automatic refinement strategies using domain-adapted captioning models or feedback-driven correction mechanisms.

Temporal modeling could also be expanded. VIVID addresses motion through key-frame selection, but continuous temporal reasoning, such as sequence-aware contrastive learning or video captioning, could improve detection of subtle or evolving behaviors, particularly in long-term monitoring.

Finally, future work could investigate combining contrastive and generative vision-language paradigms into a unified self-supervised framework. Such integration may allow systems to benefit simultaneously from alignment-based learning and language-driven semantic grounding, opening new possibilities for context-sensitive detection in diverse real-world conditions.

Bibliography

- [1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015.
- [3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119. 2013.
- [4] Muhammad Fikko Fadjrimiratno, Yusuke Hatae, Tetsu Matsukawa, and Einoshin Suzuki. Detecting anomalies from human activities by an autonomous mobile robot based on “fast and slow” thinking. In *Proc. VISIGRAPP (Vol.5, VISAPP)*, pages 943–953, 2021.
- [5] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, 2016.
- [6] Ning Dong and Einoshin Suzuki. GIAD: Generative inpainting-based anomaly detection via self-supervised learning for human monitoring. In *Proc. PRICAI, Part II*, page 418–432, 2021.

- [7] Vitaliy Kinakh, Olga Taran, and Svyatoslav Voloshynovskiy. ScatSimCLR: Self-supervised contrastive learning with pretext task regularization for small-scale datasets. In *Proc. IEEE ICCV*, pages 1098–1106, 2021.
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763, 2021.
- [9] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2013.
- [10] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection. *ACM Computing Surveys*, 54(2):1–38, 2021.
- [11] Connor Esterwood and Lionel P. Robert. Personality in healthcare human robot interaction (H-HRI): A literature review and brief critique. In *Proc. HAI '20*, page 87–95, 2020.
- [12] Hao Lv, Pengfei Yi, Rui Liu, YingKun Hou, Dongsheng Zhou, Qiang Zhang, and Xiaopeng Wei. Asymmetric anomaly detection for human-robot interaction. In *Proc. IEEE CSCWD*, pages 372–377, 2021.
- [13] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.

- [14] Fath U Min Ullah, Mohammad S. Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. A Comprehensive Review on Vision-Based Violence Detection in Surveillance Videos. *ACM Comput. Surv.*, 55(10), 2023.
- [15] Jae-Hyuk Park, Mohamed Mahmoud, and Hyun-Soo Kang. Conv3D-Based Video Violence Detection Network Using Optical Flow and RGB Data. *Sensors*, 24(2):317, 2024.
- [16] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proc. ICML*, pages 19730–19742, 2023.
- [18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [19] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. *arXiv preprint arXiv:2311.17043*, 2023.
- [20] Nadia Mumtaz, Naveed Ejaz, Shabana Habib, Syed Muhammad Mohsin, Prayag Tiwari, Shahab S Band, and Neeraj Kumar. An Overview of Violence

- Detection Techniques: Current Challenges and Future Directions. *Artificial Intelligence Review*, 56(5):4641–4666, 2023.
- [21] Ferda Ofli, Firoj Alam, and Muhammad Imran. Analysis of social media data using multimodal deep learning for disaster response. *CoRR*, abs/2004.11838, 2020.
- [22] Akash Kumar Gautam, Luv Misra, Ajit Kumar, Kush Misra, Shashwat Aggarwal, and Rajiv Ratn Shah. Multimodal analysis of disaster tweets. In *Proc. IEEE BigMM*, pages 94–103, 2019.
- [23] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proc. CVPR*, pages 14679–14689, 2020.
- [24] Tongzhi Niu, Bin Li, Weifeng Li, Yuanhong Qiu, and Shuanlong Niu. Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders. *IEEE/ASME Transactions on Mechatronics*, pages 1–2, 2021.
- [25] John Taylor Jewell, Vahid Reza Khazaie, and Yalda Mohsenzadeh. OLED: One-class learned encoder-decoder network with adversarial context masking for novelty detection. In *Proc. IEEE WACV*, pages 3591–3601, 2022.
- [26] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. MOCCA: Multi-layer one-class classification for anomaly detection. *arXiv.20*, 12:12111, 2020.
- [27] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect

- anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proc. IEEE ICCV*, pages 1705–1714, 2019.
- [28] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *Proc. ACCV*, pages 622–637. Springer, 2019.
- [29] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.
- [30] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient GAN-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.
- [31] Yusuke Hatae, Qingpu Yang, Muhammad Fikko Fadjrimiratno, Yuanyuan Li, Tetsu Matsukawa, and Einoshin Suzuki. Detecting anomalous regions from an image based on deep captioning. In *Proc. VISIGRAPP (Vol.4, VISAPP)*, pages 326–335, 2020.
- [32] Fillipe DM De Souza, Guillermo C Chavez, Eduardo A do Valle Jr, and Arnaldo de A Araújo. Violence Detection in Video Using Spatio-Temporal Features. In *Proc. SIBGRAPI*, pages 224–230, 2010.
- [33] Enrique Nieves Bermejo, Oscar Suárez Déniz, Gloria García Bueno, and Rahul Sukthankar. Violence Detection in Video Using Computer Vision Techniques. In *Proc. CAIP, Part II 14*, pages 332–339, 2011.

- [34] Oscar Deniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim. Fast Violence Detection in Video. In *Proc. VISIGRAPP (VISAPP)*, volume 2, pages 478–485, 2014.
- [35] Tobias Senst, Volker Eiselein, Alexander Kuhn, and Thomas Sikora. Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation. *IEEE Transactions on Information Forensics and Security*, 12(12):2945–2956, 2017.
- [36] Sunanda Das, Amlan Sarker, and Tareq Mahmud. Violence Detection from Videos Using HOG Features. In *Proc. IEEE EICT*, pages 1–5, 2019.
- [37] IP Febin, K Jayasree, and Preetha Theresa Joy. Violence Detection in Videos for an Intelligent Surveillance System Using MoBSIFT and Movement Filtering Algorithm. *Pattern Analysis and Applications*, 23(2):611–623, 2020.
- [38] Al-Maamoon R Abdali and Rana F Al-Tuma. Robust Real-Time Violence Detection in Video Using CNN and LSTM. In *Proc. IEEE SCCS*, pages 104–108, 2019.
- [39] Swathikiran Sudhakaran and Oswald Lanz. Learning to Detect Violent Videos Using Convolutional Long Short-Term Memory. In *Proc. IEEE AVSS*, pages 1–6, 2017.
- [40] Ming Cheng, Kunjing Cai, and Ming Li. RWF-2000: An Open Large Scale Video Database for Violence Detection. In *Proc. ICPR*, pages 4183–4190, 2021.
- [41] Şeymanur Aktı, Gözde Ayşe Tataroğlu, and Hazım Kemal Ekenel. Vision-Based Fight Detection from Surveillance Cameras. In *Proc. IEEE IPTA*, pages 1–6, 2019.

- [42] Yuening Li, Ninghao Liu, Jundong Li, Mengnan Du, and Xia Hu. Deep structured cross-modal anomaly detection. In *Proc. IEEE IJCNN*, pages 1–8, 2019.
- [43] Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence GAN: Towards better anomaly detection. In *Proc. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 141–148, 2019.
- [44] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-Ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *Proc. 2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.
- [45] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021.
- [46] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Mohamed Ahzam Amanullah, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, and Muhammad Imran. Clustering-based real-time anomaly detection—a breakthrough in big data technologies. *Transactions on Emerging Telecommunications Technologies*, 33(8):e3647, 2022.
- [47] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *Gi/itg Workshop Mmbnet*, volume 7, 2007.
- [48] Kalle Burbeck and Simin Nadjm-Tehrani. ADWICE—anomaly detection with real-time incremental clustering. In *Proc. ICISC*, pages 407–424, 2005.
- [49] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, and Muhammad Imran. Real-time big

- data processing for anomaly detection: A survey. *International Journal of Information Management*, 45:289–307, 2019.
- [50] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.
- [51] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pages 1597–1607, 2020.
- [52] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proc. ICCV*, pages 9588–9597, 2021.
- [53] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive unsupervised representations for reinforcement learning. In *Proc. ICML*, pages 5639–5650, 2020.
- [54] Kang Zhang, Muhammad Fikro Fadjrimiratno, and Einoshin Suzuki. Region anomaly detection via spatial and semantic attributed graph in human monitoring. *Sensors*, 23(3), 2023.
- [55] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

- [56] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proc IEEE CVPR*, pages 6995–7004, 2021.
- [57] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-InstructBLIP: A Framework for Aligning X-Modal Instruction-Aware Representations to LLMs and Emergent Cross-Modal Reasoning. *arXiv preprint arXiv:2311.18799*, 2023.
- [58] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [59] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*, 2023.
- [60] Jose Alejandro Avellaneda, Tetsu Matsukawa, and Einoshin Suzuki. Cross-modal self-supervised feature extraction for anomaly detection in human monitoring. In *Proc. IEEE CASE*, pages 1–8, 2023.
- [61] Jose Alejandro Avellaneda Gonzalez, Tetsu Matsukawa, and Einoshin Suzuki. Detecting anomalies in human monitoring based on multimodal multiview self-supervised learning. *Machine Vision and Applications*, 36(3):1–26, 2025.
- [62] Jose Alejandro Avellaneda Gonzalez, Tetsu Matsukawa, and Einoshin Suzuki. Leveraging vision language models for understanding and detecting violence in videos. In *Proceedings of the International Joint Conference on Computer*

- Vision, Imaging and Computer Graphics Theory and Applications*, volume 2, pages 99–113, 2025.
- [63] Kang Zhang, Muhammad Fikko Fadjrimiratno, and Einoshin Suzuki. Context-based anomaly detection via spatial attributed graphs in human monitoring. In *Proc. ICONIP*, pages 450–463, 2021.
- [64] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *Advances in Neural Information Processing Systems*, volume 34, pages 11834–11845, 2021.
- [65] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
- [66] Kenji Kashima, Moe Watanabe, and Ichiro Maruta. Simultaneous identification of nonlinear dynamics and state distribution using jensen-shannon divergence. *IFAC-PapersOnLine*, 54(14):25–30, 2021.
- [67] Vittoria Bruni and Domenico Vitulano. Jensen Shannon Divergence as Reduced Reference Measure for Image Denoising. In *Advanced Concepts for Intelligent Vision Systems*, pages 311–323. Springer, 2016.
- [68] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [69] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. ACL (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [70] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *Proc. ICLR*, 2015.

- [71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, page 1097–1105, 2012.
- [72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.
- [73] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distil-BERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.
- [74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [76] Naeem Seliya, Azadeh Abdollah Zadeh, and Taghi M Khoshgoftaar. A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, 8:1–31, 2021.
- [77] Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2), 2016.
- [78] Jian Pei Jiawei Han, Micheline Kamber. *Data mining: Concepts and Techniques*, volume 10. 2006.

- [79] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.
- [80] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2nd edition, 2018.
- [81] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proc. ICLR*, 2021.
- [82] Jaeill Kim, Duhun Hwang, Eunjung Lee, Jangwon Suh, Jimyeong Kim, and Wonjong Rhee. Enhancing contrastive learning with efficient combinatorial positive pairing. *arXiv preprint arXiv:2401.05730*, 2024.
- [83] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.
- [84] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.
- [85] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.
- [86] Joan Bruna and Stephane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

- [87] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. AugGPT: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023.
- [88] Jingqian Wu and Shibiao Xu. From point to region: Accurate and efficient hierarchical small object detection in low-resolution remote sensing images. *Remote Sensing*, 13(13):2620, 2021.
- [89] Mahyar Najibi, Bharat Singh, and Larry S Davis. Fa-rpn: Floating region proposals for face detection. In *Proc. CVPR*, pages 7723–7732, 2019.
- [90] Kevin Faust, Quin Xie, Dominick Han, Kartikay Goyle, Zoya Volynskaya, Ugljesa Djuric, and Phedias Diamandis. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics*, 19:1–15, 2018.
- [91] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, pages 618–626, 2017.
- [92] Lakshmi Vara Prasad G, Murali A, Bhaskar Reddy Y V, Haribabu Marturi, and Valeru Vision Paul. Deep learning in medical imaging: Image processing - from augmenting accuracy to enhancing efficiency. In *Proc. DMIP*, page 101–106, 2025.
- [93] Sabah Abdulazeez Jebur, Khalid A Hussein, Haider Kadhim Hoomod, Laith Alzubaidi, and José Santamaría. Review on deep learning approaches for anomaly event detection in video surveillance. *Electronics*, 12(1):29, 2022.

- [94] Osama ElSahly and Akmal Abdelfatah. Developing a machine-learning-based automatic incident detection system for traffic safety: Promises and limitations. *Infrastructures*, 9(10):170, 2024.
- [95] Qixiang Chen, Lei Wang, Piotr Koniusz, and Tom Gedeon. Motion meets attention: Video motion prompts. In *The 16th Asian Conference on Machine Learning (Conference Track)*, 2024.
- [96] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.
- [97] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.
- [98] Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604, 2023.
- [99] Hanlin Wang and Zhijian Li. A Robust Outlier Detection Method in High-Dimensional Data Based on Mutual Information and Principal Component Analysis. In *International Conference on Intelligent Computing*, pages 270–281, 2024.
- [100] Albert Thomas, Vincent Feuillard, and Alexandre Gramfort. Calibration of One-Class SVM for MV set estimation. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9, 2015.

BIBLIOGRAPHY

- [101] American Psychological Association. Physical Abuse and Violence, 2024. Accessed on 2024/12/20 from: <https://www.apa.org/topics/physical-abuse-violence>.
- [102] Aaron F. Bobick and James W. Davis. The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
- [103] Todd I Herrenkohl. *Violence in Context: Current Evidence on Risk, Protection, and Prevention*. OUP USA, 2011.
- [104] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.
- [105] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human Action Recognition from Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3200–3225, 2022.
- [106] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In *Proc. ECCV, Part II*, pages 428–441, 2006.
- [107] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent Flows: Real-Time Detection of Violent Crowd Behavior. In *Proc. IEEE CVPR Workshops*, pages 1–6, 2012.
- [108] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision*, 103:60–79, 2013.

BIBLIOGRAPHY

- [109] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [110] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [111] Wikipedia contributors. Wikipedia, The Free Encyclopedia, 2025. [Online; accessed 16-June-2025].
- [112] Legal Information Institute. Wex: Legal Dictionary and Encyclopedia, 2025. [Online; accessed 16-June-2025].
- [113] Cambridge University Press. Cambridge Dictionary, 2025. [Online; accessed 16-June-2025].
- [114] Oxford University Press. Oxford English Dictionary, 2025. [Online; accessed 16-June-2025].
- [115] Wan Zhang and Jing Zhang. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics*, 13(5):856, 2025.
- [116] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

- [117] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not Only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In *Proc. ECCV*, pages 322–339, 2020.
- [118] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [119] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Published Papers

- 1 **Jose Alejandro Avellaneda Gonzalez**, Tetsu Matsukawa and Einoshin Suzuki.
Cross-Modal Self-Supervised Feature Extraction for Anomaly Detection in Human Monitoring. In *Proc. IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1-8, 2023.
- 2 **Jose Alejandro Avellaneda Gonzalez**, Tetsu Matsukawa and Einoshin Suzuki.
Detecting Anomalies in Human Monitoring Based on Multimodal Multiview Self-Supervised Learning. In *Machine Vision and Applications*, Vol. 36 (3), pages 1-26., 2025.
- 3 **Jose Alejandro Avellaneda Gonzalez**, Tetsu Matsukawa and Einoshin Suzuki.
Leveraging Vision Language Models for Understanding and Detecting Violence in Videos. *Proc. VISIGRAPP, (Vol. 2, VISAPP)*, pages 99 - 113, 2025.