



DEPARTMENT OF
COMPUTER SCIENCE

PEDRO ACÚRCIO DOMINGOS

BSc in Computer Science and Engineering

**VIDEO SURVEILLANCE METADATA
EXTRACTION WITH MULTIMODAL LLM
AGENTS THROUGH ITERATIVE DIALOGUE
EXPLORATION**

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon
September, 2025



VIDEO SURVEILLANCE METADATA EXTRACTION WITH MULTIMODAL LLM AGENTS THROUGH ITERATIVE DIALOGUE EXPLORATION

PEDRO ACÚRCIO DOMINGOS

BSc in Computer Science and Engineering

Advisers: David Fernandes Semedo

Assistant Professor, NOVA University Lisbon

João Carlos Raposo Neves

Associate Professor, University of Beira Interior

Co-adviser: Vasco Lopes

CEO, Deep Neuronic

Examination Committee

Chair: Jörg Matthias Knorr

Associate Professor, NOVA University Lisbon

Rapporteur: Catarina Helena Branco Simões da Silva

Associate Professor, University of Coimbra

Advisers: David Fernandes Semedo

Assistant Professor, NOVA University Lisbon

João Carlos Raposo Neves

Associate Professor, University of Beira Interior

Co-adviser: Vasco Lopes

CEO, Deep Neuronic

Video Surveillance Metadata Extraction with Multimodal LLM Agents through Iterative Dialogue Exploration

Copyright © Pedro Acúrcio Domingos, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

To all the people I have met along my journey: you have shaped me, either by good or bad experiences, and I am grateful to each of you. To the friends who are sadly no longer with us: your memory endures, and I hope this work, together with the consistency and discipline I maintained throughout its development, honours you.

Disclaimer: In this work, generative artificial intelligence tools were used, namely ChatGPT 4 and 5, for grammatical and punctuational revision. These tools were employed under the author's supervision, and all the generated content was verified for accuracy. Authorship and validation of the content remain entirely the author's responsibility, and the use of artificial intelligence complies with the institution's standards of academic integrity.

ACKNOWLEDGEMENTS

I want to start by expressing my gratitude to Professor David Semedo and Professor João Neves for all the guidance, support and availability throughout this project. They made this project possible by providing solution ideas, debating methods, and ensuring consistency in our thought process. Next, I would like to thank Vasco Lopes of DeepNeuronic, who actively contributed to methodological debates, providing expert insight into real-time surveillance needs and João Pereira, a Ph.D colleague, who collaborated by sharing his developments and results, enabling us to make better-informed decisions and avoid mistakes. Next, I want to thank NOVA LINCS for the opportunity to carry out this research project, to contribute to the development of the surveillance metadata extraction community, and to participate in new experiences. Finally, my heartfelt appreciation goes to all my loved ones for their unwavering support and motivation.

”

“Everything that you thought had meaning: every hope, dream, or moment of happiness. None of it matters as you lie bleeding out on the battlefield. None of it changes what a speeding rock does to a body—we all die. But does that mean our lives are meaningless? Does that mean that there was no point in our being born? Would you say that of our slain comrades? What about their lives? Were they meaningless? They were not! Their memory serves as an example to us all! The courageous fallen! The anguished fallen! Their lives have meaning because we, the living, refuse to forget them! And as we ride to certain death, we trust our successors to do the same for us.”

— **Erwin Smith**, Attack on Titan
(Commander of the Survey Corps)

ABSTRACT

Current surveillance methods are unsustainable, demanding constant human oversight and vast storage for millions of hours of footage. This reliance on manual monitoring is not only labour-intensive but also error-prone, owing both to the challenging characteristics of surveillance footage (low resolution, poor lighting, and unconventional camera angles) and to the sheer number of cameras that must be supervised. While Large Vision-Language Model (LVLM)s have shown promise in understanding generic videos, they often fail in this domain and tend to omit critical details, limiting their use for evidence extraction.

This thesis addresses surveillance metadata extraction by employing multi-agent approaches sharing a foundational principle: obtaining additional scene context through iterative question-answering dialogue exploration. We begin with static questioning schemes; progress to dynamic, class-aware question generation; investigate reasoning-based approaches; and finally develop a surveillance-fitted, reasoning-guided method. The developed approaches use a set of models to overcome major single-model and surveillance limitations, ultimately being capable of producing richer representations of the video. To evaluate the developed methods, we conduct quantitative experiments on surveillance anomaly detection and classification tasks using the UCF-Crime Annotation (UCA) [59] dataset, and present qualitative examples demonstrating their fragilities and extraction capabilities. One method improves anomaly detection by 21 percentage points (p.p.) over the baseline. We also curate a higher-quality subset of surveillance data to study resolution effects on the prominent developed approaches. On this subset, we are able to surpass the baseline in anomaly detection by 10 p.p. and in anomaly classification by 4 p.p., without retraining. Resolution impacts anomaly classification by up to 12 p.p.

This research contributes to the surveillance field by introducing alternative methods to perform zero-shot anomaly detection and classification, while pioneering the surveillance metadata extraction domain, and presenting insights into context extraction by employing multi-agent solutions.

Keywords: Surveillance Video Analysis, Exploratory Question-Answering, Multimodal Large Language Models

RESUMO

Os métodos atuais de vigilância são insustentáveis, exigindo supervisão humana constante e vasta capacidade de armazenamento para milhões de horas de vídeo. Esta dependência de monitorização é intensiva em mão de obra e propensa a erros, devido às características das filmagens (baixa resolução, fraca iluminação e ângulos de câmara inconventionais) e ao elevado número de câmaras para supervisionar. Embora os *Large Vision-Language Model (LVLN)*s consigam compreender vídeos genéricos, falham frequentemente neste domínio e tendem a omitir detalhes críticos, limitando a sua utilização na recolha de provas.

Esta tese aborda a extração de metadados de vigilância através de abordagens multiagente que partilham um princípio fundamental: obter contexto através de um diálogo de pergunta-resposta. Iniciamos com esquemas perguntas estáticas; progredimos para geração dinâmica de perguntas por class; investigamos abordagens baseadas em raciocínio; e desenvolvemos um método de raciocínio guiado, adaptado a vigilância. As abordagens utilizam um conjunto de modelos para superar limitações individuais e de vigilância, sendo no final capazes de produzir representações mais fiéis do vídeo. Para avaliar os métodos desenvolvidos, realizámos testes quantitativos em tarefas de deteção e classificação de anomalias utilizando o conjunto de dados UCF-Crime Annotation (UCA) [59], e apresentamos exemplos qualitativos para demonstrar tanto as suas fragilidades como as suas capacidades de extração. Um dos métodos melhora a deteção de anomalias em 21 pontos percentuais (p.p.) face ao modelo base. Seleccionámos ainda um subconjunto com maior resolução para estudar o seu impacto nas abordagens desenvolvidas. Neste subconjunto, conseguimos superar o modelo base em deteção por 10 p.p. e em classificação por 4 p.p. sem treino. A resolução impacta a classificação de anomalias até 12 p.p.

Esta investigação contribui para o domínio da vigilância, ao introduzir métodos alternativos para realizar deteção e classificação de anomalias em *zero-shot*, ao mesmo tempo que inaugura o domínio da extração de metadados em vigilância e apresenta *insights* sobre extração de contexto através de abordagens multiagente.

Palavras-chave: Análise de Vídeo de Vigilância, Pergunta-Resposta Exploratória, Modelos Grandes de Linguagem Multimodal,

CONTENTS

List of Figures	xi
Acronyms	xiv
1 Introduction	1
1.1 Context and Motivation	1
1.2 Problem Formulation	2
1.3 Challenges and Research Hypothesis	2
1.4 Main Contributions	3
1.5 Document Organization	4
2 Background	6
2.1 Transformers	6
2.1.1 Sequence Modelling	6
2.1.2 The Transformer Architecture	6
2.1.3 Advantages of Transformers	8
2.2 Transformers for Vision	8
2.2.1 Vision Transformers	8
2.3 Vision-Language Transformers	9
2.3.1 CLIP	9
2.4 Summary	10
3 Related Work	11
3.1 Visual Question Answering	11
3.1.1 VideoMME	11
3.2 Temporal and Multimodal Reasoning	12
3.2.1 TOMATO	13
3.2.2 VideoEspresso	14
3.3 Surveillance	15
3.3.1 Surveillance Datasets	16

3.3.2	VideoICL	16
3.3.3	VERA	17
3.4	Large Language Models	17
3.4.1	LLaMA	18
3.4.2	Qwen	18
3.5	Large Vision Language Models	19
3.5.1	LLaVA	19
3.5.2	LLaVA-Video	21
3.5.3	Qwen-VL	23
3.5.4	Qwen2-VL	23
3.6	Models as Judges	24
3.6.1	LLM Judges	24
3.6.2	LVLN Judges	26
3.7	Multi-Agent LLM Setting	26
3.7.1	Types of Agents	27
3.7.2	Important Discussion Aspects	28
3.7.3	Decision-Making Mechanisms	28
3.8	Reasoning Strategies	29
3.8.1	Chain-of-Thought	29
3.8.2	Tree-of-Thought	29
3.8.3	Iteration-of-Thought	30
3.8.4	Layered-Chain-of-Thought	30
3.8.5	Critical Summary	31
4	Problem Definition and Initial Framework	32
4.1	Problem Definition in Surveillance Video Metadata Extraction	32
4.1.1	Surveillance Footage Challenges	32
4.1.2	Metadata Extraction Challenges	33
4.2	Preliminary Analysis Observations	33
4.3	Agent-Based Question-Answer Framework	35
4.3.1	Motivation	35
4.3.2	Architecture	35
4.3.3	Error-resilience	35
4.3.4	Evaluation Tasks	36
5	Baseline and Base Multi-Agent Approaches: Static, Dynamic and Reasoning-Based	38
5.1	Single LLM-Based Baselines	38
5.2	Static Fixed-Set Questioning Approaches	39
5.3	Dynamic Top-K Questioning Approaches	41
5.4	Reasoning-Based Approaches	42

6	Extended Framework and Surveillance Layered Graph Strategy	47
6.1	Extended Q-A Framework	47
6.1.1	Motivation	47
6.1.2	Architecture	48
6.1.3	Reasoning Graph	48
6.1.4	Balancing Dynamism and Control	49
6.2	Surveillance Layered Graph Strategy	49
6.2.1	Lessons Learned from Previous Approaches	49
6.2.2	Layered Architecture for Metadata Extraction	51
6.2.3	Graph Operations and Reasoning Control	52
6.2.4	The Criminal Module	55
6.2.5	Final Overview	55
6.2.6	Goal and Intended Outcome	56
7	Experimental Evaluation	58
7.1	Evaluation Protocol	58
7.1.1	Datasets	58
7.1.2	Metrics	61
7.2	Baseline Performance	63
7.2.1	Setup	63
7.2.2	Results	63
7.2.3	Conclusion and Insights	64
7.3	Static Approaches Performance	64
7.3.1	Phase A - Static Variants on UCA	65
7.3.2	Phase A - Results	65
7.3.3	Phase B - Question Count Ablation on UCA-25	66
7.3.4	Phase B - Results	66
7.3.5	Phase C - M-VERA-Optimized Question Sets	68
7.3.6	Conclusion and Insights	70
7.4	Dynamic Approaches Performance	70
7.4.1	Setup	70
7.4.2	Results	70
7.4.3	Conclusion and Insights	73
7.5	Reasoning-Based Approaches Performance	74
7.5.1	Setup	74
7.5.2	Results	74
7.5.3	Conclusion and Insights	78
7.6	Surveillance Layered Graph Reasoning Strategy Performance	79
7.6.1	UCA-25 Setup	79
7.6.2	UCA-25 Results	79
7.6.3	Critical Overview of UCA Results	80

7.6.4	BUCA Setup	81
7.6.5	BUCA-HR Results	81
7.6.6	Resolution Impact	81
7.6.7	Conclusion and Insights	83
8	Conclusions and Future Work	86
8.1	Overview	86
8.2	Limitations	87
8.2.1	Data Limitations	87
8.2.2	Model Limitations	89
8.2.3	Solution Limitations	89
8.3	Future Work	90
	Bibliography	91
	Appendices	
A	Prompt Templates	99
A.1	Static Questions	101
A.2	Dynamic Top-K Questions	104
A.3	Reasoning-based Strategies	105
A.4	Surveillance Layered Graph Strategy	109
B	M-Vera Generated Questions	114

LIST OF FIGURES

2.1	The Transformer Architecture, illustrating the Encoder and Decoder components. Reproduced from [60].	7
2.2	The Vision Transformer (ViT) architecture [11]. Left: Illustrating the process of dividing an image into patches, embedding them with positional information, and processing them through the Transformer Encoder. The final classification is performed using a Multilayer Perceptron (MLP) head. Right: Overview of the operations within the Transformer Encoder.	9
2.3	The architecture of CLIP [38], illustrating its three phases: (1) contrastive pre-training to align text and image embeddings, (2) creating classifiers using label text, and (3) zero-shot prediction by comparing input image embeddings with label embeddings.	10
3.1	Video Multi-Modal Evaluation (Video-MME) video categories and subcategories [13].	12
3.2	<i>VideoEspresso</i> data sources [16].	14
3.3	<i>VideoEspresso's</i> question-answer pair construction pipeline [16], illustrating the frame-level captioning, redundancy removal, frame grouping, and Question Answering (QA) pair construction steps.	15
3.4	The architecture of <i>Large Language and Vision Assistant (LLaVA)</i> [30], illustrating the integration of a pre-trained visual encoder with a language model through a projection matrix.	19
3.5	AnyRes, dynamic high resolution scheme [28].	20
3.6	Representation of slowFast video in <i>LLaVA-Video</i> [62]. Key frames (F_{slow}) have more visual tokens assigned, while intermediate frames (F_{fast}) have fewer tokens allocated. This balance reduces the needed computational resources.	22
3.7	Hierarchical multi-level caption generation process [62]. The model produces descriptions at three different temporal levels to capture both global and local video context.	22
3.8	<i>Qwen-Vision-Language (VL)</i> 's training pipeline [2].	23

3.9	Comparison between standard Input-Output (IO) prompting and Chain-of-Thought (CoT) prompting [52]. While IO prompting reaches the final solution in one single step, Chain-of-Thought (CoT) introduces intermediate reasoning steps that help guide the model towards a better final solution.	29
3.10	Comparison of reasoning strategies [56]. Chain-of-Thought with Self-Consistency (CoT-SC) extends CoT prompting by exploring multiple paths in parallel and selecting the final answer through majority voting. Tree-of-Thought (ToT) generalises this idea by exploring different intermediate-step solutions and selecting the most promising paths at each stage.	30
3.11	Use case of Iteration-of-Thought (IoT) [39] prompting to count the occurrences of the letter "R" in the word <i>Strawberry</i>	31
4.1	Example of a <i>Video-MME</i> daily life scenario, illustrating the loss of information in a description when compared to directly questioning the existence of cars.	34
4.2	Agent-Based Q-A Framework. An interactive dialogue between the Questioner (LLM), which generates context-aware questions, and the Answerer (LVLM), which provides answers grounded in the video content.	36
5.1	Static approaches interaction diagram. The Answerer LVLM responds to a predefined set of questions and the Questioner Large Language Model (LLM) uses the resulting Q-A pairs to decide the final output.	40
5.2	Dynamic approaches interaction diagram. The Answerer LVLM iteratively refines the set of probable classes $\mathcal{K} \rightarrow \mathcal{K}'$ through the analysis of class-dependent-question-answer pairs, generated initially by the Questioner LLM and answered by the Answerer LVLM	42
5.3	Comparison between original Chain-of-Thought (CoT) reasoning and its adaptation to the Questioner–Answerer framework.	43
5.4	Comparison between original Tree-of-Thought (ToT) reasoning and its adaptation to the Questioner–Answerer framework.	44
5.5	Comparison between original Iterative-of-Thought (IoT) reasoning and its adaptation to the Questioner–Answerer framework.	45
5.6	Comparison between original Layered Chain-of-Thought (LCoT) reasoning and its adaptation to the Questioner–Answerer framework.	46
6.1	Extended Question-Answer Framework. The architecture extends the previous version by introducing a Controller persona and a reasoning graph to guide the interaction.	48
6.2	The proceed operation. It generates three pertinent follow-up questions and selects the one whose question-answer pair contributes most towards the main goal.	52
6.3	The refine operation. It rewrites the same question with a more developed context based on the previous answer.	53

6.4	The split operation. It divides the current knowledge context into multiple segmented branches.	54
6.5	The stop operation. It dictates that a certain node has already achieved its goal, excluding it from the exploration list.	54
6.6	Surveillance Layered Graph interaction diagram.	56
7.1	Comparison between original UCA test set and the sampled version UCA-25. Left: class distribution expressed as percentages. Right: video duration distribution expressed as percentages, with videos categorized as short (< 30s), medium (30–60s), and long (> 60s).	60
7.2	Comparison of per-class recall (%) between the baselines.	64
7.3	Multiclass confusion matrices (row-normalized in %) Top: Qwen2.5-VL baseline confusion matrix. Middle: Qwen2.5-VL-BiasQ confusion matrix. Bottom: Qwen2.5-VL-GenQ confusion matrix.	67
7.4	Comparison of per-class recall (%) between the baseline and the best static approaches.	69
7.5	Answer inconsistency in a <i>Fighting</i> scenario.	73
7.6	Comparison of per-class recall (%) between the baseline and the best dynamic approaches.	73
7.7	Qualitative metadata extraction example for <i>Arrest048</i> (UCA).	77
7.8	Comparison of per-class recall (%) between the baseline and the best reasoning approaches.	78
7.9	Resolution-induced prediction shifts on BUCA (HR – LR). Each panel shows the difference of row-normalised (%) confusion matrices for the indicated model.	84
7.10	Comparison of per-class recall (%) between the baseline and the best SLG approaches.	85

ACRONYMS

AI	Artificial Intelligence (<i>pp. 6, 56</i>)
BGE-M3	Beijing General Embedding Model - Multi-Functionality, Multi-Linguality, Multi-Granularity (<i>p. 14</i>)
BUCA	Better UCA (<i>pp. 60, 61, 81, 83, 87, 88</i>)
BUCA-HR	BUCA High Resolution (<i>pp. 60, 61</i>)
BUCA-LR	BUCA Low Resolution (<i>pp. 61, 87</i>)
CLIP	Contrastive Language-Image Pretraining (<i>pp. 9, 10, 19, 23</i>)
CNN	Convolutional Neural Network (<i>pp. 8, 9</i>)
CoT	Chain-of-Thought (<i>pp. xii, 13, 14, 25, 26, 29, 30, 43, 49, 50, 74, 77</i>)
CoT-SC	Chain-of-Thought with Self-Consistency (<i>pp. xii, 30</i>)
CV	Computer Vision (<i>p. 8</i>)
FFN	Feed Forward Network (<i>pp. 7, 8, 18</i>)
GPT	Generative Pre-training Transformer (<i>pp. 25, 26</i>)
GPT-4o	Generative Pre-training Transformer 4 Omni (<i>pp. 15, 17</i>)
ICL	In-Context Learning (<i>pp. 16, 17</i>)
IDA	Inner dialogue agent (<i>pp. 30, 44</i>)
InternVL2	Intern Vision-Language 2 (<i>p. 14</i>)
IO	Input-Output (<i>pp. xii, 29</i>)
IoT	Iteration-of-Thought (<i>pp. xii, 30, 31, 44, 50, 74</i>)
LCoT	Layered Chain-of-Thought (<i>pp. 30, 31, 44, 50, 51, 74</i>)
LLaMA	Large Language Model Meta AI (<i>pp. 18, 19, 21</i>)
LLaVA	Large Language and Vision Assistant (<i>pp. xi, 19–23</i>)

LLM	Large Language Model (<i>pp. xii, 3, 11, 15, 17–21, 23–26, 30, 35, 39–45, 79, 80, 86, 88, 89</i>)
LLMA	LLM agent (<i>pp. 30, 44</i>)
LSTM	Long Short Term Memory (<i>pp. 6, 8</i>)
LVLM	Large Vision-Language Model (<i>pp. v, vi, xii, 1–3, 11, 17, 19, 20, 24, 26, 32, 35, 38–42, 44, 45, 48, 66, 70, 80, 86, 88–90</i>)
M-RoPE	Multimodal Rotary Position Embedding (<i>p. 24</i>)
M-VERA	Multiclass VERA (<i>pp. 41, 55, 56, 65, 68, 70, 71, 87, 114</i>)
MEIDE	Metadata Extraction through Iterative Dialogue Exploration (<i>pp. 11, 31</i>)
MLP	Multilayer Perceptron (<i>pp. xi, 9, 20</i>)
NLP	Natural Language Processing (<i>pp. 6, 8, 10, 18</i>)
OCR	Optical Character Recognition (<i>pp. 21, 23</i>)
OOD	out-of-distribution (<i>pp. 16, 31</i>)
QA	Question Answering (<i>pp. xi, 15</i>)
RNN	Recursive Neural Network (<i>pp. 6, 8</i>)
RoPE	Rotatory Positional Embedding (<i>p. 21</i>)
SLG	Surveillance Layered Graph (<i>pp. 47, 55, 56, 79, 80, 83, 87, 90</i>)
SOTA	State-of-the-art (<i>pp. 1, 11, 22, 24, 86</i>)
TOMATO	Temporal Reasoning Multimodal Evaluation (<i>pp. 12–14, 24</i>)
ToT	Tree-of-Thought (<i>pp. xii, 29, 30, 44, 49, 50, 52, 74, 77</i>)
UCA	UCF-Crime Annotation (<i>pp. v, vi, xiii, 4, 16, 36, 58–61, 63–66, 68–72, 74, 75, 79–81, 83, 87–89</i>)
UCF-Crime	University of Central Florida Crime (<i>pp. 16, 17, 40, 47, 55, 58, 60, 88, 114</i>)
Video-MME	Video Multi-Modal Evaluation (<i>pp. xi, xii, 11, 12, 22, 24, 34</i>)
ViT	Vision Transformer (<i>pp. 8–10</i>)
VL	Vision-Language (<i>pp. xi, 23, 24</i>)
VQA	Visual Question Answering (<i>pp. 2, 4, 11, 12, 14, 19, 20, 23</i>)

INTRODUCTION

1.1 Context and Motivation

As society evolves, new measures are developed to ensure and maintain citizens' safety, aiming to reduce the overall violence and criminality while enhancing the lives of responsible citizens.

With the growing reality of a modern era where surveillance cameras monitor nearly every public space, new problems emerge. Usually, cameras require human oversight to effectively interrupt any criminal attempt in a reasonable amount of time, and in the hypothetical scenario of an intensely tracked civilisation this would not only demand excessive human labour but also immense storage capacity to save millions of hours of footage for future analysis and long-term statistics gathering. These problems could be theoretically solved by replacing human monitoring with an automated system capable of identifying criminal acts in real time, immediately flagging a human supervisor, and summarising surveillance video data in a compact way.

State-of-the-art (SOTA) Large Vision-Language Models (LVLMs) can already accurately analyse and understand common video events, presenting a potential replacement for the human monitoring but still lack competence in the surveillance domain and exploratory capabilities needed to capture non-prominent details that standard video benchmarks do not evaluate but could be relevant in this setting. Additionally, recent studies [4, 12, 26, 34, 50, 64, 66] demonstrate that using multiple models in complex tasks enhances performance, introducing new possibilities for frameworks that, upon receiving a surveillance video, could provide a compact, structured description with all the details needed.

The core motivation of this thesis is to leverage the capabilities of recent LVLMs within a multi-agent environment, aiming to address their weaknesses in the surveillance and metadata extraction domain, and adopt an iterative dialogue exploratory approach to extract surveillance meaningful aspects of the video.

This work was carried out under a national research project in collaboration with DeepNeuronic, and the resulting methods and insights are intended to support subsequent development.

1.2 Problem Formulation

Metadata extraction is commonly defined as the process of extracting information or features from a document, usually retrieving data such as the author, title, headers, and keywords. The surveillance domain footage usually contains critical but difficult-to-capture information, requiring manual human analysis, becoming impractical in large-scale monitoring scenarios. While employing automated systems seems reasonably better, current options often fail to extract meaningful aspects from the videos, particularly events occurring in spatially small regions, relevant sub-events and general background information, leading to inefficient surveillance video summarisation and reasoning.

In this thesis, metadata extraction refers to the process of compiling every necessary detail in a given video into a compact, structured description. This includes directly identifying the number of people in the scene, their clothing, their actions, and interactions with items, which could indicate a potential weapon or theft event, while indirectly analysing the surrounding environment as vehicles, location, time of day, and other relevant aspects that might be crucial for subsequent analysis of the footage.

With this type of structured description, it would be possible not only to reduce the storage capacity required to store surveillance footage but also to efficiently search for evidence of a crime by running queries to retrieve information about the actions and movements of suspicious individuals. Additionally, utilising a system capable of flagging suspicious events in a real-time video streams coming from the surveillance cameras would effectively reduce the human monitoring need and increase the overall security by allowing authorities to intervene almost immediately.

1.3 Challenges and Research Hypothesis

While widely used multimodal benchmarks evaluate tasks such as Visual Question Answering (VQA) and temporal reasoning they usually lack surveillance video specific features. Surveillance footage brings novel challenges, ranging from cluttered scenarios, low-quality video, strange camera angles, harsh lighting conditions and unusual events occurring in peripheral areas of the image. Additionally, the current state of surveillance lacks recent domain specific datasets, restricting the scientific community to use either weakly-annotated data that usually serves for anomaly detection tasks or low-resolution videos.

Even though recent models already demonstrate impressive results on common benchmarks, this is not the case in surveillance datasets, where only a limited subset of models are currently tested, and poor results demonstrate either insufficient video quality or model limitations.

This thesis builds on the observation that LVLMs tend to overlook details when prompted to describe a certain video, even if in-depth, yet can still perceive the events and accurately answer to direct questions about them (further explored in Section 4.1). We

believe these limitations come from the models' training. Like humans, when asked to describe a video, we tend to focus solely on the central actions and ignore all the middle ground events. While this is acceptable for most scenarios, surveillance heavily relies on these details as they can be crucial to identify a crime.

Building upon this observation, we propose developing and testing methods that advance the surveillance metadata extraction domain by surpassing single-model performance on anomaly detection and classification through enhanced, context-dependent reasoning. These methods share a common foundational principle: obtaining additional scene context through an iterative dialogue exploration. The dialogue is conducted between two agents: an Large Language Model (LLM), which generates pertinent follow-up questions based on the discovered context, and an LVLM, which answers according to the video content. By employing this dialogue in a multi-agent environment, we aim to mitigate both single-model limitations, such as superficial descriptions and omissions, and the ones introduced by the surveillance domain, such as hallucinations and normality bias introduced by the context and prompts.

1.4 Main Contributions

The main contributions of this thesis are:

- A novel multi-agent framework to perform iterative dialogue exploration, containing two agent roles: the Questioner, which poses pertinent follow-up questions based on the obtained context, and the Answerer, which answers the posed questions according to the video content.
- Development and study of static methods, i.e., approaches that utilise a fixed set of predefined questions to explore the environment. These questions can be generated by an LLM, assembled by human experts, or optimised based on the model's answers.
- Development and study of dynamic methods, i.e., approaches that iteratively reduce a set of probable criminal classes and pose questions regarding the remaining ones.
- Adaptation and study of reasoning-based methods, i.e., approaches that employ step-by-step reasoning strategies (*Chain of Thought*, *Tree of Thought*, *Iteration of Thought*, and *Layered Chain-of-Thought*).
- A novel multi-agent framework that extends the above (first contribution) multi-agent design with a new role, the Controller, a reasoning-planning agent that guides the exploration process by selecting operations defined by the user.
- Development and study of surveillance layered graph methods, utilising the extended multi-agent framework as well as the main strengths of the previous methods to perform surveillance-specific scenario exploration.

- An extensive comparison between the different methods on surveillance-specific tasks based on anomaly detection and classification. Additionally, qualitative examples of limitations and metadata extraction capabilities, making a fair comparison between reasoning-based approaches and static methods.
- A curated subset of higher-quality surveillance videos, comprising *UCF-Crime Annotation (UCA)* examples selected to be representative and reliably annotated to minimise discovered limitations, together with complementary high-resolution online videos.
- An analysis of the impact caused by video resolution on the prominent developed methods.
- This thesis initiates the research into the surveillance metadata extraction domain, establishing new methods to create rich representations of the video content. Additionally, it extends existing surveillance research by including multi-agent reasoning approaches, something unexplored to date.
- A conference paper titled "Chain-of-Anomaly Thoughts with Large Vision-Language Models" [10], accepted and presented at *RECPAD*, a Portuguese conference on pattern recognition, regarding the surveillance graph method.

1.5 Document Organization

The remainder of this document is organized in the following chapters:

- **Chapter 2 - Background:** This chapter introduces key concepts that provide a solid structure to understand the next advancements.
- **Chapter 3 - Related Work:** In this chapter, we explore relevant literature that directly aligns with the project's theme, examining the VQA and temporal reasoning tasks and their respective datasets, the current state of surveillance, recent updates in high-performing models, the use of models as judges, the multi-agent setting, and model reasoning strategies.
- **Chapter 4 - Problem Definition and Initial Framework:** In this chapter, we introduce the surveillance and metadata extraction challenges and define the initial iterative dialogue exploration framework.
- **Chapter 5 - Baseline and Base Multi-Agent Approaches: Static, Dynamic and Reasoning-Based:** This chapter defines the baselines used, as well as the methods that utilise the initial iterative dialogue exploration framework.
- **Chapter 6 - Extended Framework and Surveillance Layered Graph Strategy:** This chapter explores the key improvements made to the initial framework and defines

the final proposed strategy that leverages the combined strengths of the previously proposed methods within the extended framework.

- **Chapter 7 - Experimental Evaluation:** In this chapter, we discuss the results of applying the proposed methods to surveillance-specific tasks, evaluating their strengths and major limitations. Additionally, we test the impact of resolution on the prominent proposed methods.
- **Chapter 8 - Conclusions and Future Work:** In this chapter, we provide a complete overview of the progress and decisions made during the development process, state the major discovered data, model, and solution limitations, and propose future developments.

BACKGROUND

This chapter introduces key concepts and historical advancements that allowed recent models to work with visual and textual data. Firstly, sequence modelling is explained, followed by an in-depth look into the Transformers architecture, then examining vision transformers and concluding with *CLIP*, a Vision-Language Transformer that serves as a component in more recent models.

2.1 Transformers

Transformers, proposed in 2017 by Vaswani et al. [48], revolutionised sequence modelling, becoming the prominent approach for tasks like Natural Language Processing (NLP). This section introduces the concept of sequence modelling and the main components of the transformer architecture.

2.1.1 Sequence Modelling

Sequence modelling is a core concept in Artificial Intelligence (AI) that involves understanding the current data, by learning its underlying patterns, to predict its next elements. This capability enables models to not only understand but also generate and classify various types of structured data such as text and video.

The Transformer architecture marked a significant shift in sequence modelling because it overcomes the major limitations of previous methods. Older approaches, like Recursive Neural Networks (RNNs) and Long Short Term Memorys (LSTMs), process data sequentially, making them less efficient in general and less effective at modelling long-term relationships within data. Transformers address these limitations by introducing a more efficient and effective approach to sequence modelling.

2.1.2 The Transformer Architecture

The Transformer architecture [48] is composed of two main components: the Encoder and the Decoder (Figure 2.1). The Encoder builds contextual representations of the input

sequence, while the Decoder generates the output sequence using masked attention and the Encoder's context.

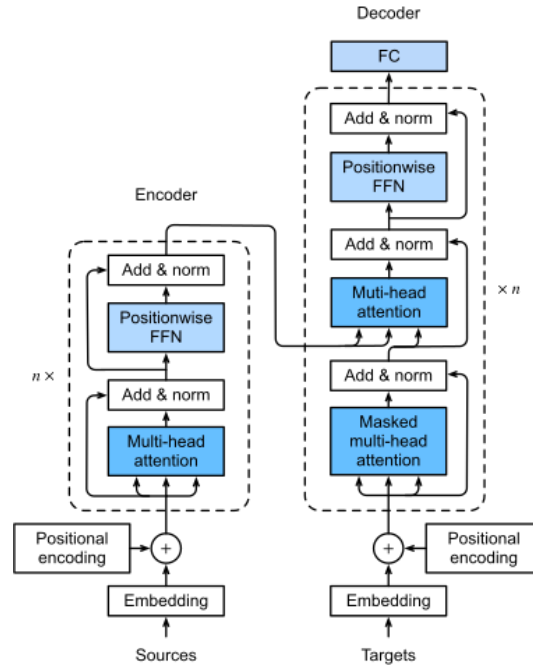


Figure 2.1: The Transformer Architecture, illustrating the Encoder and Decoder components. Reproduced from [60].

2.1.2.1 Encoder

The encoder creates a contextual representation of the input sequence that captures its meaning and positional relationships. This process begins by embedding the source sequence tokens into a new n -dimensional space format, where related tokens are positioned closer to one another, while non-related ones are further apart. Additionally, positional encoding is added to the embeddings, giving the model a better understanding of the sequence order.

Within the Encoder, the Self-Attention mechanism is applied. This mechanism evaluates how much focus each token should give to other tokens in the sequence. The Multi-Head Attention allows the model to separately process multiple attention layers in parallel, concatenating the outputs at the end. The result is then passed through a Feed Forward Network (FFN), refining the learned features. To address gradient stability problems, such as the vanishing gradients, the Encoder uses a residual connection and normalisation after each step. The residual connection mitigates the vanishing gradients by adding the input of each model to its output, while layer normalisation helps stabilise and accelerate training.

2.1.2.2 Decoder

The Decoder is used for language generation tasks, it predicts the subsequent token in a sequence based on the earlier context. Like the Encoder, it receives as input the embeddings of a sequence enriched by the positional encodings. This sequence consists of tokens generated by previous steps, with the initial token being a special <START> token.

The Masked Multi-Head Attention mechanism ensures that predictions only depends on the preceding tokens by masking every other token. The next step performs another Multi-Head Attention mechanism, but now integrates contexts coming from the Encoder, allowing the Decoder to align its outputs with the input sequence. The resulting representation is then passed through a FFN followed by a fully connected network. Like the Encoder, the Decoder uses residual connections and normalisation after each step. The Decoder is capable of calculating the probabilities for all tokens in the vocabulary and applies a selection strategy to determine the most appropriate output token. The generation process ends when a special <END> token is generated.

2.1.3 Advantages of Transformers

Transformers address the sequential processing problem of RNNs and LSTMs by using the Multi-Head attention mechanism, enabling parallelisation and significantly accelerating training and inference. Additionally, this new architecture is capable of discovering long-range dependencies without being constrained by distance, a fundamental limitation of earlier approaches.

2.2 Transformers for Vision

Transformers also impacted the field of Computer Vision (CV), which previously was dominated by Convolutional Neural Networks (CNNs), which effectively extract local features but usually struggle to capture global relationships or processing complex long-range dependencies. Transformers in the other hand, use the attention mechanism to effectively fill these gaps, offering an alternative to CNNs for vision tasks. This section covers the Vision Transformer (ViT) architecture, focusing on how it processes images.

2.2.1 Vision Transformers

ViTs [11] adapts the Transformer architecture to process image data. Unlike CNNs, which rely on convolutional layers to extract features, ViTs divide the image into patches and uses them as tokens, similar to words in NLP (Figure 2.2).

The processing starts by partitioning the image into patches, flattening them, and passing the result through a linear projection to generate patch embeddings. Positional encodings are then added to capture spatial details, along with an additional learnable embedding, usually known as the class embedding, representing the final classification

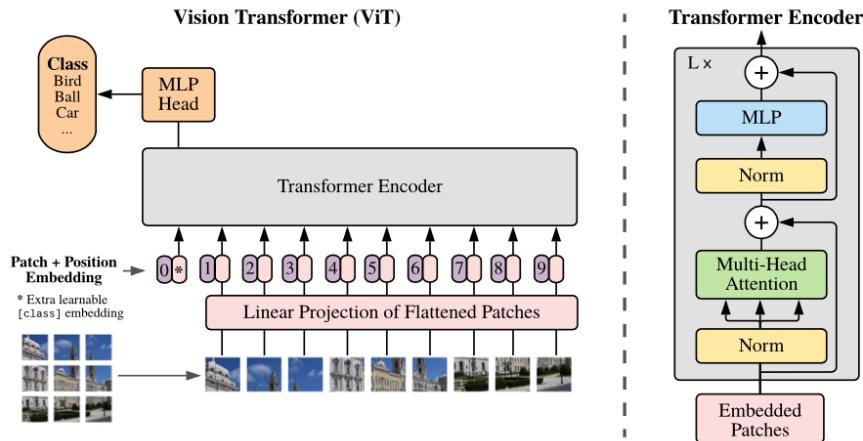


Figure 2.2: The Vision Transformer (ViT) architecture [11]. **Left:** Illustrating the process of dividing an image into patches, embedding them with positional information, and processing them through the Transformer Encoder. The final classification is performed using a Multilayer Perceptron (MLP) head. **Right:** Overview of the operations within the Transformer Encoder.

output. The embeddings are then used by the Transformer Encoder, which uses the attention mechanism to capture relationships between patches without being constrained by spatial distance, unlike CNNs. The encoder result is then passed through an MLP head that classifies the input based on the added class embedding.

2.3 Vision-Language Transformers

After the innovative works around transformers both in textual and vision fields, joining these two modalities became prominent within the scientific community. Integrating spatial information from visual inputs with context from textual details promised a more enriched understanding of the content.

A typical Vision-Language Transformer consists of three key components: a text encoder, a vision encoder, and a mechanism to join information from both modalities. These models are designed to align both textual and visual information within a shared embedding space. This section covers Contrastive Language-Image Pretraining (CLIP) [38], a Vision-Language Transformer that stood out for developing a novel technique to fuse the two modalities.

2.3.1 CLIP

CLIP [38] introduces a contrastive learning strategy to align textual and visual information within a shared embedding space. CLIP's architecture consists of two main components: a text encoder that converts textual tokens into embeddings and an image encoder, typically a ViT, that maps visual inputs into the same shared embeddings space.

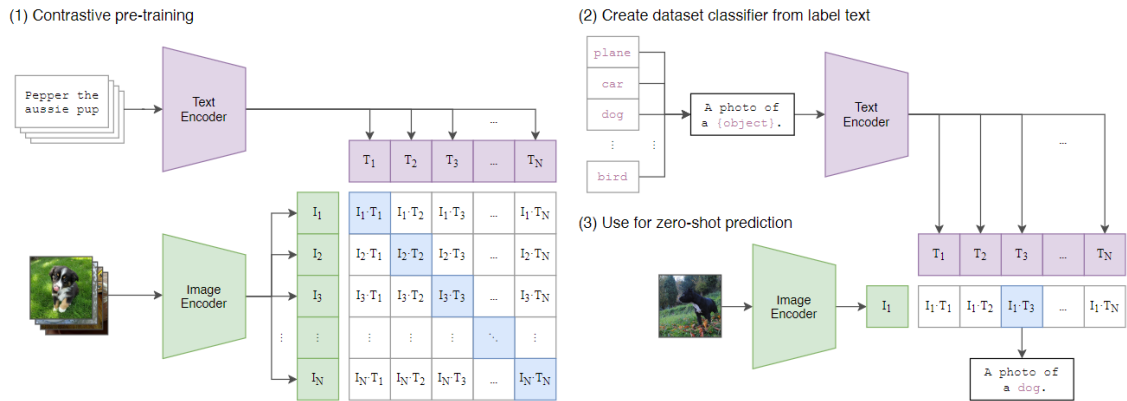


Figure 2.3: The architecture of CLIP [38], illustrating its three phases: (1) contrastive pre-training to align text and image embeddings, (2) creating classifiers using label text, and (3) zero-shot prediction by comparing input image embeddings with label embeddings.

During pre-training, *CLIP* learns to represent both modalities in the same embedding space by optimising the two encoders at the same time, using contrastive loss. This loss function guarantees that embeddings of matching image-caption pairs become closer, while embeddings of non-matching pairs are pushed further apart in the shared feature space. The similarity between these embeddings is measured using cosine distance, enabling the model to capture semantic alignment effectively.

CLIP's pre-training utilizes an extensive dataset of image-caption pairs, providing a solid generalisation across a wide range of tasks without requiring fine-tuning. *CLIP* achieves impressive zero-shot capabilities by using textual prompts describing the possible output classes, such as "a photo of a dog". These textual prompts are encoded by the text encoder and compared to the generated image embeddings. The class with the highest similarity score is selected as the prediction, enabling *CLIP* to perform tasks like image classification without additional training.

2.4 Summary

The Transformer architecture, initially developed for NLP tasks, ended up being a foundational basis for other models that impacted diverse areas, enabling a better understanding of structured data. ViTs introduce a novel way to treat images almost like tokens in NLP, enabling transformers to understand visual data. *CLIP* then builds upon the ViT innovations to propose a new mechanism for joining both the textual and visual modalities, enabling the development of more robust and complex models in the future.

RELATED WORK

This chapter discusses research work that aligns directly with the project’s theme and further shapes the proposed framework, *Metadata Extraction through Iterative Dialogue Exploration (MEIDE)*. Firstly, we explore two prominent visual tasks: visual question answering and temporal reasoning (both crucial for surveillance analysis), followed by an look into important surveillance datasets and recent works. Next, we present the historical evolution and major architectural updates of Large Language Models (LLMs) leading up to State-of-the-art (SOTA) Large Vision-Language Models (LVLMs). We then explore studies that use models as performance evaluators, and finally, we take an in-depth look at the taxonomy of multi-agent frameworks.

3.1 Visual Question Answering

Visual Question Answering (VQA) involves assessing a model’s ability to understand and reason over visual content to answer a specific question. These questions are usually either open-ended or have multiple choices to select from. This section highlights *Video Multi-Modal Evaluation (Video-MME)*, a prominent video benchmark designed to evaluate VQA performance across a wide range of multimodal reasoning tasks.

3.1.1 VideoMME

Video-MME [13] is a comprehensive video-only multi-modal benchmark designed to evaluate the capabilities of LVLM. This dataset ensures:

1. **Diversity in Video Types:** Featuring six primary categories and 30 subcategories, ensuring models handle a wide range of real-world video scenarios.
2. **Duration in the Temporal Dimension:** Containing three different levels of video length (short, medium and long) spanning videos with durations of seconds until one hour of length.

3. **Breadth in Data Modalities:** Containing not only video frames but also subtitles and audio tracks, enabling the evaluation of models' all-around multi-modal understanding.
4. **Quality in Annotations:** Providing high-quality, expert-annotated questions, effectively assessing the model's capabilities.

Video-MME consists of 900 videos, totalling 254 hours of content and 2,700 carefully curated question-answer pairs.



Figure 3.1: **Video-MME** video categories and subcategories [13].

Video-MME not only provides a solid testing base for evaluating VQA performance across categories like Life Record, which share challenges with surveillance scenarios, but also evaluates the model's ability to interpret normal day life events and deal with long-duration videos.

3.2 Temporal and Multimodal Reasoning

Temporal a multimodal reasoning involves determining the model's capabilities in interpreting complex, temporal or multimodal, relationships. Unlike static content, videos contain patterns that are only detectable by understanding underlying relationships across time, making temporal reasoning essential for tasks that rely on dynamic content analysis. This section highlights two notable works: *Temporal Reasoning Multimodal Evaluation*

(*TOMATO*), which addresses existing issues in common video datasets resulting in overestimated metric results, and *VideoEspresso*, a large-scale dataset designed with novel Chain-of-Thought (CoT) reasoning techniques and an interesting question-generation pipeline.

3.2.1 TOMATO

TOMATO [43] is a recent benchmark specifically design to assess models’ temporal reasoning capabilities in video understanding. Built from the observation that recent models achieve overestimated performance scores, not due to genuine reasoning capacity but rather due to video simplicity, *TOMATO* introduces three new metrics designed to measure how rigorously a benchmark targets visual temporal reasoning:

1. **Multi-Frame Gain:** This metric evaluates the performance gain κ when using multiple frames compared to using a single frame. The single frame is either randomly selected or hand-picked, while the multiple frames are uniformly sampled at equal intervals throughout the video.

$$\kappa = \frac{Acc(m \text{ frames})}{Acc(1 \text{ frame})}. \quad (3.1)$$

The lower the κ values, the easier it is for the model to answer accurately using a single frame, indicating a weak dependence on temporal information.

2. **Frame Order Sensitivity:** This metric measures the impact of temporal order by comparing the performance τ on correctly ordered frames versus the same frames presented in a shuffled order.

$$\tau = \frac{Acc(m \text{ frames})}{Acc(\text{shuffled } m \text{ frame})} - 1. \quad (3.2)$$

The lower the τ value, the easier it is for the model to answer accurately with unordered frames, indicating a weak reliance on frame ordering.

3. **Frame Information Disparity:** This metric measures the performance gain ρ when answering a question using a specifically hand-picked frame compared to a randomly selected one.

$$\rho = \frac{Acc(\text{hand-picked } 1 \text{ frame})}{Acc(\text{random-sampled } 1 \text{ frame})} - 1. \quad (3.3)$$

A higher ρ value indicates that certain frames carry more critical information for accurately answering the question, whereas a lower value suggests that importance is equally distributed over the frames.

TOMATO provides a different perspective on well-known benchmark results by introducing new metrics that highlight critical issues such as frame order sensitivity. These metrics emphasize the importance of evaluating temporal reasoning in models correctly and reveals inconsistencies in traditional benchmarks. These insights are particularly relevant for the project, where effective model selection is crucial, as temporal reasoning can lead to completely opposite interpretations. For example, the actions of picking up an item and returning it to the shelf are different in terms of reasoning, but frame-wise, they can be seen as merely a reordering of the frames.

3.2.2 VideoEspresso

VideoEspresso [16] is a recent large-scale dataset design to address challenges of fine-grained video reasoning and multimodal understanding. This benchmark introduces techniques to enhance CoT reasoning with annotations, guiding models to provide intermediate steps to build a final answer. Unlike traditional VQA datasets, which normally rely heavily on creation methods that are either costly and manual or underdeveloped and automatic, *VideoEspresso* uses an automated pipeline to construct high-quality QA pairs while preserving temporal coherence. The dataset features carefully selected videos from seven distinct data sources (Figure 3.2), ensuring rich temporal dynamics, logical sequences, and causal relationships.

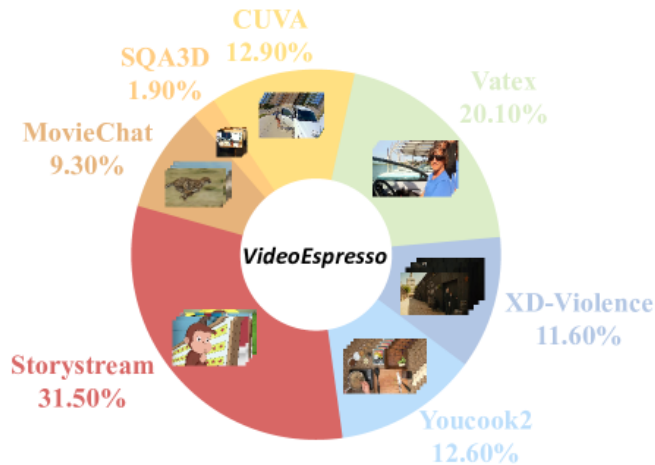


Figure 3.2: *VideoEspresso* data sources [16].

Taking a closer look at the automated pipeline for constructing Question-Answer pairs (Figure 3.3), it begins by generating frame-level captions using *Intern Vision-Language 2 (InternVL2)-8B* [9]. Next, It filters redundant frames through semantic filtering with the language retrieval model *Beijing General Embedding Model - Multi-Functionality, Multi-Linguality, Multi-Granularity (BGE-M3)* [7], which compares frame captions using cosine

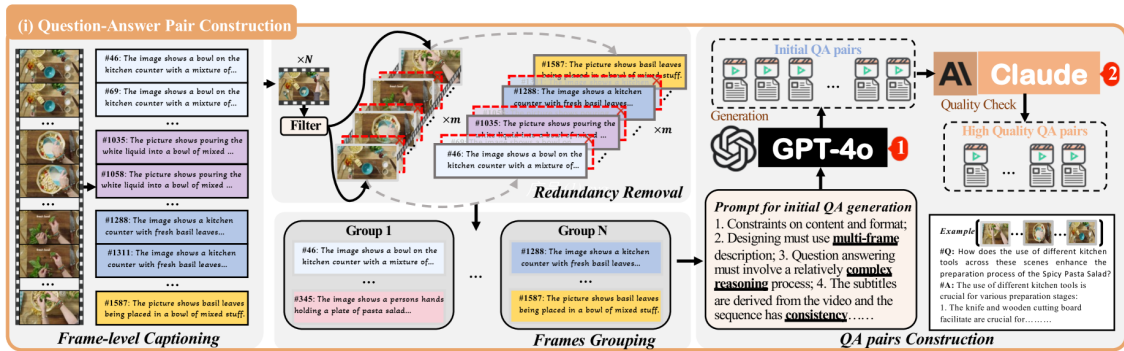


Figure 3.3: *VideoEspresso*'s question-answer pair construction pipeline [16], illustrating the frame-level captioning, redundancy removal, frame grouping, and Question Answering (QA) pair construction steps.

similarity and excludes frames exceeding a predefined threshold. Following this preprocessing step, frames are grouped to reduce the risk of hallucination and instruction-following failures caused by having an excessive number of tokens. Finally, the pipeline prompts *Generative Pre-training Transformer 4 Omni (GPT-4o)* [33] to generate question-answer pairs based on the multi-frame descriptions and ensures response quality with an additional LLM that removes hallucinations, verifies factual accuracy, and filters out subjective or difficult-to-evaluate questions. This last step occurs iteratively, refining the question-answer pairs.

VideoEspresso features a diverse collection of videos, including a curated selection from *XD-Violence*, a well-known benchmark focused on various types of violence videos that aligns directly with the surveillance theme.

3.3 Surveillance

The surveillance domain presents unique challenges, videos usually capture a wide range of actions in cluttered scenarios and anomalies occur in a matter of seconds. Building a comprehensive surveillance dataset can be particularly difficult due to privacy concerns, the limited availability of high-quality data and even the required human labour for relevant annotations. The current state of surveillance reveals not only a lack of complex datasets and task diversity, forcing researchers to rely on video models exclusively pre-trained on other settings and develop new strategies to slightly improve the scores in surveillance, but also the largely unexplored potential of newer models on existing surveillance benchmarks. This section covers some important datasets in this domain, followed by an examination of *VideoICL* and *VERA*, two of the most recent works addressing video analysis that cover surveillance.

3.3.1 Surveillance Datasets

Most surveillance datasets are designed for the anomaly detection task, some examples of interesting dataset choices include:

- **ShanghaiTech Campus Dataset** [31] is an anomaly detection dataset containing 130 abnormal events across 13 diverse scenes, totalling 317,398 frames, a significant increase in size compared with previous surveillance datasets. Additionally, the data introduces new types of anomalies involving sudden motion, such as chasing and brawling.
- **PoseLift** [40] is a pose-based anomaly detection dataset specifically design to detect shoplifting behaviours. The dataset contains anonymised pose data extracted from surveillance videos, meaning that only keypoints representing human body joints are shown, excluding the background video to address privacy concerns. The dataset contains 34 instances of shoplifting and 112 regular shopping events, making a total of 153 videos recorded across six different cameras in a local store. Additionally, *PoseLift* includes frame-wise annotations specifying bounding boxes and the coordinates of the keypoints. While *PoseLift* does not directly align with the type of input used in this project, it demonstrates a strategy to address privacy concerns.
- **University of Central Florida Crime (UCF-Crime)** [5] is a large-scale surveillance video dataset primarily designed for anomaly detection and classification tasks. The events are categorised as normal or one of 13 crime-related classes, such as robbery, arson, and vandalism. The videos are untrimmed, with anomalies usually occurring in short time span, simulating the real-world surveillance scenarios. The dataset contains a total of 128 hours of CCTV footage, providing a set of diverse and challenging environments. However, *UCF-Crime* consists mainly of low-resolution videos, limiting its applicability when testing requires fine-grained visual data.
- **UCF-Crime Annotation (UCA)** [59] extends the *UCF-Crime* dataset by introducing fine-grained annotations that describe events at each timestamp of the video. Unlike *UCF-Crime*, which focuses primarily on anomaly detection and classification tasks, *UCA* supports a wider range of multimodal tasks. The dataset contains a total of 23,000 sentence-level annotations, enriching *UCF-Crime* with semantic and temporal details. Additionally, *UCA* is the only publicly available dataset in the surveillance domain that contains scene-level description annotations, emphasising its importance not only for this project but also for the whole surveillance research community.

3.3.2 VideoICL

VideoICL [20] is one of the most recent video analysis works in the surveillance domain, it uses In-Context Learning (ICL) to improve performance on out-of-distribution (OOD)

tasks. ICL is a technique that revolves around sending meaningful contextual examples before the actual task, effectively guiding the model to achieve better performance. Kim et al. [20] go a step further and incorporate a similarity-base mechanism to select valuable examples from a wide range of videos stored in a database. The framework first selects a set of k examples that are similar to the current query and processes them in smaller batches of m examples ($with\ m < k$), ensuring the model’s context length is not exceeded. After each iteration, the answers’ confidence score (in this case, the lowest probability among the generated tokens) is calculated, and if the confidence level surpasses a defined threshold, the response is accepted as final. If, after multiple iterations, the confidence remains below the threshold, the most confident response is selected instead.

VideoICL is tested on the *UCF-Crime* dataset using three recent open source models, slightly increasing their performance but still not surpassing the proprietary model *GPT-4o*, which holds an accuracy score of 58%. While *VideoICL* demonstrates the potential of applying recent models on a surveillance benchmark and is capable of enhancing open-source model performance, the best scores are still below the 60% mark, emphasising again the largely unexplored opportunities. There is an urgent need to evaluate recent models in-depth and extend testing to more complex tasks within the surveillance setting.

3.3.3 VERA

VERA [57] is another example of using innovative approaches to tackle the inherent challenges of applying LVLMs to the surveillance domain. Given the fact that frozen models tend to underperform in this domain when prompted with manually written, generic questions, *VERA* treats a set of guiding questions as *learnable parameters* and optimises them through an interaction between two agents: a **learner**, which predicts anomalies given the current question–answer pairs, and an **optimiser**, which refines the set of questions based on the learner’s correctness. This setup allows *VERA* to discover a set of highly optimised questions for video anomaly detection suited for a specific LVM model.

On *UCF-Crime*, *VERA* achieves the highest score among explainable video anomaly detection methods, with an AUC of 86.55%, which is also comparable to the top-performing method in the non-explainable setting. This result highlights the effectiveness of searching for strategic questions to increase accuracy in tasks where models tend to underperform, without requiring any training.

3.4 Large Language Models

The evolution of LLMs changed the way we deal with text-related tasks, enabling systems to process, understand, and generate human language. Starting from the base Transformer architecture and implementing key innovations, LLMs achieved important milestones.

This section covers two fundamental LLM models: first, we explore key innovations introduced by *Large Language Model Meta AI (LLaMA)*, followed by the updates made in *Qwen*.

3.4.1 LLaMA

LLaMA [47] was introduced as a family of language models by Meta AI, designed to maximise performance under a given computational budget. By prioritising extensive pre-training over scaling model size, *LLaMA* offers great efficiency without needing proprietary datasets.

LLaMA is inspired by the original transformer 2.1.2 but adopts a decoder-only architecture, specifically design for autoregressive language tasks. *LLaMA* only generates the next token based on previously generated ones, unlike the usual transformer that receives an additional encoded sequence. Additionally *LLaMA* introduces key innovations that enhance performance and training efficiency:

Pre-Normalization: Root Mean Square Normalization [61] is applied to the input of each Transformer sub-layer instead of the output, enhancing training stability.

SwiGLU Activation: The standard ReLU activation function (within the Feed Forward Network (FFN)) is replaced with SwiGLU [44], leading to improved performance through smoother gradient flow.

Rotary Positional Embeddings (RoPE): RoPE [45] replaces absolute positional embeddings, allowing for more effective modelling of sequential data. This approach encodes relative positions rather than absolute ones, enabling the model to more effectively capture contextual dependencies within sequences.

3.4.2 Qwen

Qwen [1] was introduced by the Alibaba Group as a family of LLMs designed to perform a wide range of Natural Language Processing (NLP) tasks. Building upon the architecture and innovations of LLaMA 3.4.1, *Qwen* architectural updates further improve performance and versatility.

Embedding and Output Projection: *Qwen* separates the input embedding layer (used for token representation) from the output projection layer (used for generating token probabilities), allowing them to be optimised independently. While this decision increases the number of parameters by not sharing weights between the two layers, it enhances flexibility and performance.

Bias: *Qwen* removes biases from most layers to reduce computational cost and adds additional ones in the attention layers to enhance the model’s extrapolation ability.

3.5 Large Vision Language Models

The advancements made by *Contrastive Language-Image Pretraining (CLIP)* significantly impacted how the scientific community handles visual and textual data together. Earlier vision-language models were employed to analyse static image content, but the complexity of the real-world scenarios further demanded more advanced approaches that could process dynamic visual content. This section covers *Large Language and Vision Assistant (LLaVA)*, which laid a solid foundation for multimodal image-oriented models and its updates up until its latest version *LLaVA-Video*. Additionally, this section also covers *Qwen-VL* and its successor *Qwen2-VL*, an incredible model that exhibits impressive performances over diverse multimodal tasks.

3.5.1 LLaVA

LLaVA [30] is a foundational LVLM design to integrate textual and visual input seamlessly. Built upon a pre-trained vision encoder, such as *CLIP*, and a LLM like *Vicuna* (a finetuned version of *LLaMA*), *LLaVA* performs instruction tuning to align features from the visual encoder with the LLM’s word embeddings. This method enables the model to execute a variety of tasks on static images such as, image captioning, VQA and object recognition.

LLaVA [30] is a foundational LVLM design to integrate textual and visual modalities. Built upon a pre-trained vision encoder, such as *CLIP*, and a LLM like *Vicuna* (a fine-tuned version of *LLaMA*), *LLaVA* executes a variety of tasks on static images such as, image captioning and VQA.

The architecture of *LLaVA* uses a lightweight integration mechanism:

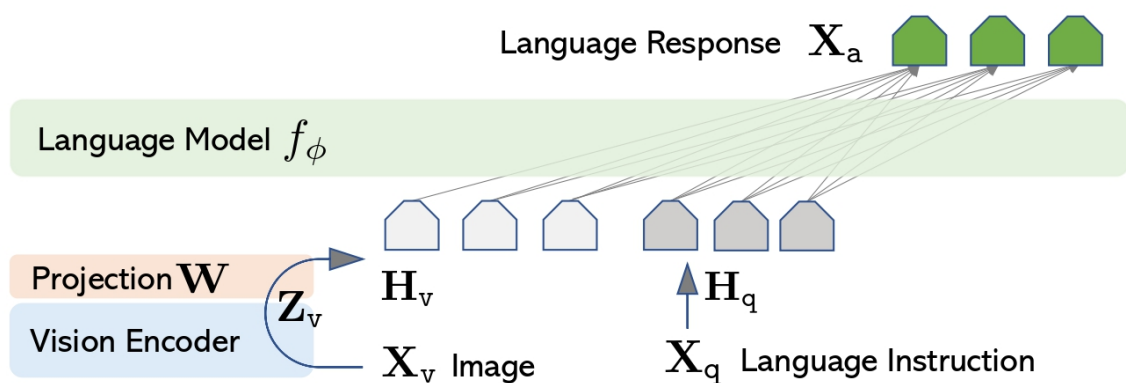


Figure 3.4: The architecture of *LLaVA* [30], illustrating the integration of a pre-trained visual encoder with a language model through a projection matrix.

The visual encoder processes the input image X_v to extract the visual features $Z_v = g(X_v)$, where g represents the visual encoder processing function. These features are transformed into the LLM's embedding space using a trainable projection matrix W , resulting in the matching visual embeddings $H_v = W \cdot Z_v$. This conversion ensures that the transformed visual features H_v are compatible with LLM's text embeddings H_q , enabling the multimodal reasoning (Figure 3.4).

LLaVA follows a two-step training methodology:

1. **Pre-training for Feature Alignment:** In the first stage, the visual encoder and LLM weights remain frozen, and only the projection matrix is trained. This step adjust the visual encoder's features to match the LLM's word embedding space, integrating visual and textual information.
2. **Fine-Tuning End-to-End:** In the second stage, the vision encoder weights remain frozen while updating both the projection matrix and LLM weights. This fine-tuning adapts the model to specific tasks, using instruction-following data to improve its capabilities.

While originally designed for static images analysis, *LLaVA* laid the foundation for more advanced LLMs, receiving various updates and variations to better adapt to the current scenario:

***LLaMA 2* based *LLaVA*:** The first major update in the *LLaVA* series was the substitution of the pre-trained base LLM from *LLaMA 1* to *LLaMA 2*. This transition enhanced the model's instruction-following capabilities, particularly for role-playing tasks, while only reducing its translation performance.

***LLaVA-1.5*:** The 1.5 version [28] then includes a Multilayer Perceptron (MLP) cross-modal connector and new academic-task-oriented VQA data. Additionally, the model implemented the AnyRes mechanism, allowing it to process images of any resolution by splitting them into grids and encoding them independently, while also providing a downsampled version of the complete image for global context (Figure 3.5).

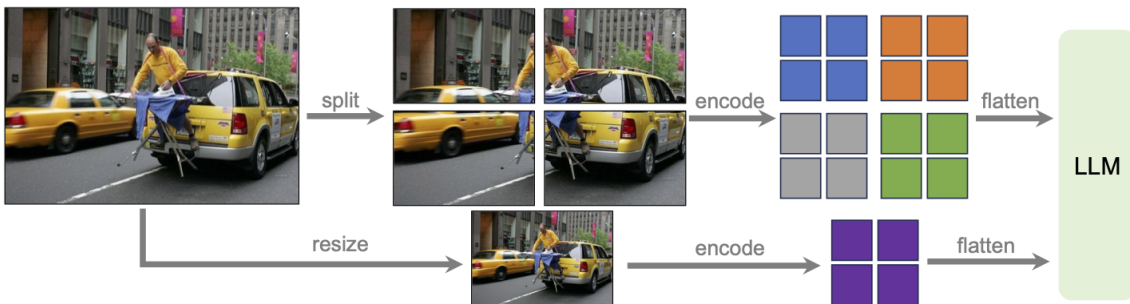


Figure 3.5: AnyRes, dynamic high resolution scheme [28].

LLaVA-NeXT: The development of *LLaVA-Next* [29] brought important updates. This iteration increased the base input image resolution by up to four times and introduced improved visual reasoning and Optical Character Recognition (OCR) capabilities through enhanced visual instruction tuning data. Additionally, it explored other LLM options rather than *Vicuna*. Another important advancement was the use of the AnyRes mechanism to handle videos as concatenated frames, a natural generalisation from the previous iteration. Inspired by Rotatory Positional Embedding (RoPE), the model also incorporated a linear scaling technique, increasing its maximum token capacity to handle longer video sequences, further extending its applicability.

LLaVA-Next-Strong: The release of *LLaVA-Next-Strong* [23] introduced two new LLM options, *LLaMA 3* and *Qwen-1.5*. These updates significantly increased both the model's size and multimodal capabilities. The authors also proposed newly developed datasets to evaluate the improved multimodal capabilities.

LLaVA-NeXT-Interleave: The major update released with *LLaVA-NeXT-Interleave* [25] was the interleaved image-text format, fixing it as a general data template for unifying diverse scenarios. This format alternates visual and textual elements in a sequence, enabling the model to establish contextual relationships between modalities. The model demonstrated improved generalisation across multiple settings, including multi-image, multi-frame (video), multi-view (3D environments), and multi-patch tasks. A new dataset was also proposed, specifically designed to align with the interleaved image-text format.

LLaVA-OneVision: The introduction of *LLaVA-OneVision* [24] brought an updated version of the AnyRes mechanism through the inclusion of bilinear interpolation. This update allowed the model to dynamically adjust the number of tokens per crop, providing a flexible visual representation framework adaptable to single-image, multi-image, and video representations. Additionally, the authors proposed a new collection of training data, combining three popular data sources, totalling a 99.8% reliance on synthetic data, highlighting the high costs and legal constraints involved in creating large-scale, high-quality datasets in the wild.

3.5.2 LLaVA-Video

LLaVA-Video [62] is the latest version of the *LLaVA* series. It develops even further the previous iterations bringing updates such as:

SlowFast Video Representation: *LLaVA-Video* adopts a SlowFast mechanism to efficiently balance temporal granularity and computational load. This technique uses more visual tokens to key frames (slow pathway) and fewer tokens to less critical frames (fast pathway), allowing the model to process larger videos without overwhelming computational resources (Figure 3.6).

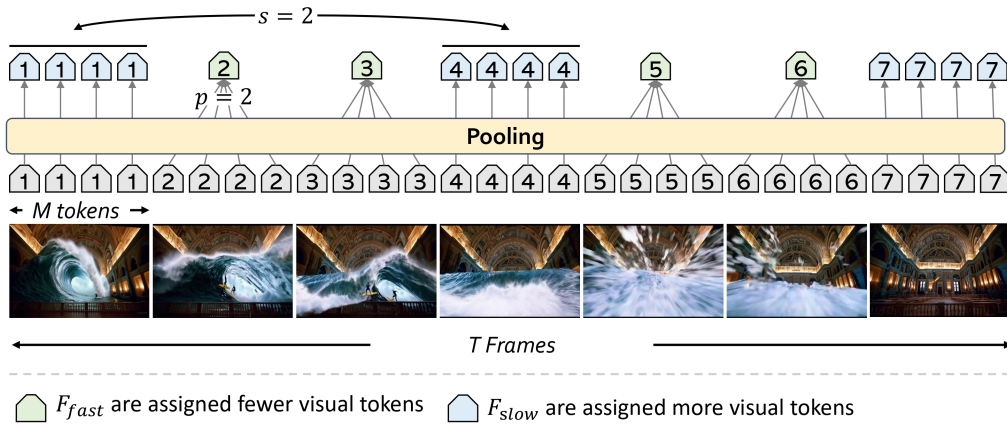


Figure 3.6: Representation of slowFast video in *LLaVA-Video* [62]. Key frames (F_{slow}) have more visual tokens assigned, while intermediate frames (F_{fast}) have fewer tokens allocated. This balance reduces the needed computational resources.

Recurrent Multi-level Captioning: *LLaVA-Video* utilises an hierarchical video description creation pipeline that produces captions at various time duration levels. This process allows both fine-grained and high-level representations of the video’s content, increasing the model’s ability to understand complex narratives (Figure 3.7).

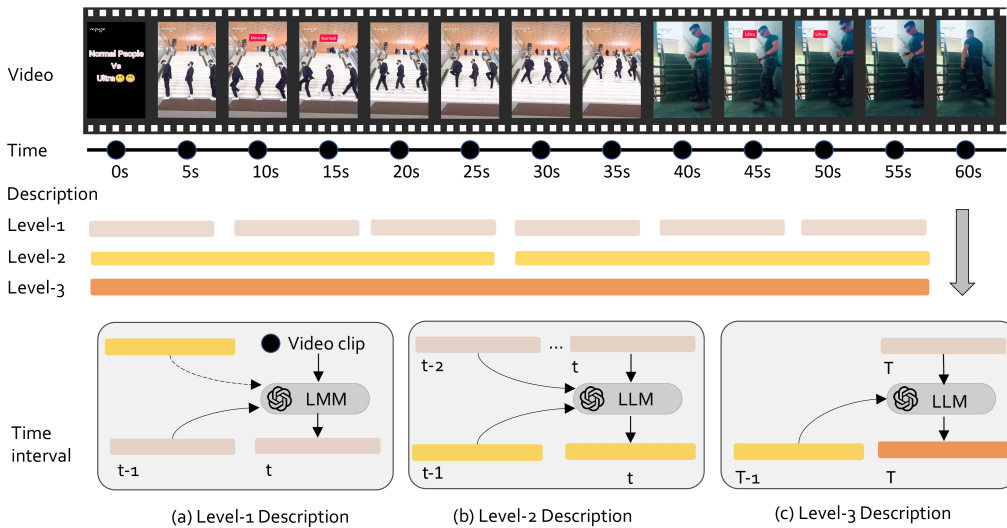


Figure 3.7: Hierarchical multi-level caption generation process [62]. The model produces descriptions at three different temporal levels to capture both global and local video context.

LLaVA-Video delivers competitive performance on various multimodal benchmarks such as *Video-MME*, while achieving SOTA results (in the open-source setting) on *VideoEspresso*’s benchmark. These results demonstrate *LLaVA-Video*’s robustness in the multi-modal setting, offering a viable solution for complex tasks that require temporal and multimodal reasoning.

3.5.3 Qwen-VL

Qwen-Vision-Language (VL) [2] represents the first iteration of VL models built upon the LLM *Qwen*. Similar to *LLaVA 3.5.1*, *Qwen-VL* incorporates an LLM (*Qwen*), a vision encoder (an updated version of *CLIP*), and a position-aware vision-language adapter that replaces *LLaVA*'s projection matrix (Figure 3.8). This adapter is a single-layer cross-attention module designed to compress the visual feature sequence from the vision encoder into a fixed-length representation. Additionally, two-dimensional absolute positional encodings are incorporated into the cross-attention mechanism, mitigating potential loss of positional details during the compression process.

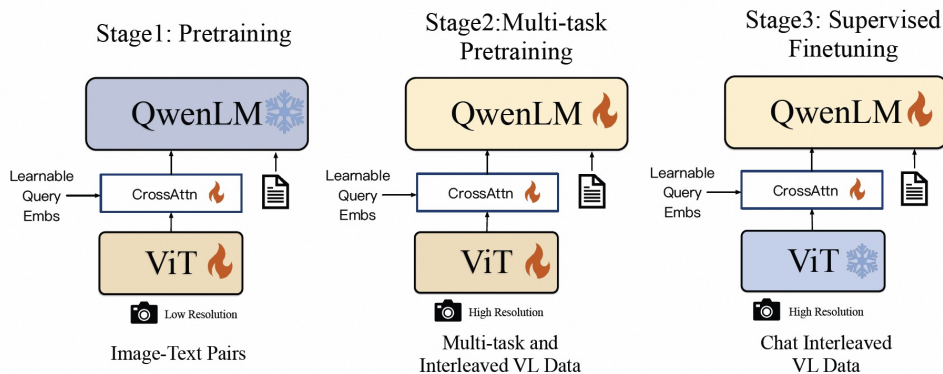


Figure 3: The training pipeline of the Qwen-VL series.

Figure 3.8: *Qwen-VL*'s training pipeline [2].

Qwen-VL follows a three-step training methodology:

- 1. Pre-training:** In the first stage, the vision encoder and adapter are trained using a large-scale image-text pairs dataset while the LLM's weights remain frozen. This process aligns the visual features with the LLM's embedding space.
- 2. Multi-task Pre-training:** In the second stage, introduces high-resolution inputs and interleaved image-text data, training the model in tasks such as captioning, VQA, OCR and grounding while maintaining the LLM's abilities with in-house collected corpus samples. During this stage, all components of the model are trained at the same time.
- 3. Supervised Fine-tuning:** In the final stage, the visual encoder weights are frozen while the LLM and adapter are fine-tuned through instruction tuning. This step enhances the model's instruction-following and dialogue capabilities.

3.5.4 Qwen2-VL

Qwen2-VL [49] introduces significant architectural innovations over its predecessor, designed to enhance multimodal understanding. The core improvements include:

Naive Dynamic Resolution Mechanism: *Qwen2-VL* dynamically adapts to the input’s resolution preserving its high-resolution details.

Multimodal Rotary Position Embedding (M-RoPE): While traditional models rely on one-dimensional positional encodings, *Qwen2-VL* decomposes positional encodings into temporal, height, and width components, significantly improving the model’s capability to reason over dynamic content.

Unified Image and Video Processing: *Qwen2-VL* adopts a unified framework for handling both images and videos. Video inputs are processed using three-dimensional convolutions and dynamically sampled frames, enabling temporal reasoning.

Qwen2-VL achieves competitive performances over various multimodal benchmarks and SOTA results (in the open-source setting) on *Video-MME* [13] and *TOMATO* [43]. These results demonstrate that *Qwen2-VL* is a great candidate for the role of LLM in the framework.

3.6 Models as Judges

As models and frameworks evolve, the need for high-quality evaluation becomes increasingly critical. Obtaining human labour to perform these tasks is costly, time consuming and in some cases challenging due to the growing complexity of models and tasks. This complexity often leads to more difficult metrics and exhausting comparison criteria. Using models as judges offers a scalable solution to this problem by allocating additional computational resources rather than relying only on human experts. This section explores insights from related works on this topic, presenting advantages and limitations of this type of approaches across both textual and visual modalities.

3.6.1 LLM Judges

As the models evolve, open-ended questions are getting more popular, in general they hold a lot more usability in different tasks but come with the negative aspect of being difficult to evaluate. Zheng et al. [63] propose three LLM-as-a-judge variations:

- **Pairwise Comparison:** The LLM judge receives a question along with two answers and determines which response is better, or if both are equally good.
- **Single Answer Grading:** The LLM judge evaluates a single answer by grading it directly based on a predefined criteria.
- **Reference-Guided Grading:** The LLM judge receives a reference answer in addition to the answers being evaluated, allowing comparisons between the answers and the ground truth for a more informed evaluation.

The authors identify as main advantages the scalability and explainability while as main limitations and respective solutions:

- **Position Bias:** This occurs when the judge is more likely to favour a certain position in the order of the presented answers. A possible solution is to evaluate the samples twice, swapping the appearance order. If the different evaluations are inconsistent, it is considered a tie.
- **Verbosity bias:** This happens when the judge favours longer, verbose responses, even if they are not as good. This issue can be mitigated by selecting a model that is less affected by this factor.
- **Self-enhancement bias:** This occurs when the judge model favours responses generated by its own model over others. While the problem is clearly detected, the authors could not provide a clear answer no how to mitigate this effect.
- **Limited Capability in Grading Math and Reasoning Questions:** Sometimes the judge struggles to grade questions involving math or reasoning. The authors propose two possible solutions for this problem:
 1. Prompt the judge model to solve the problem first and then evaluate the responses, encouraging CoT reasoning.
 2. Use a two-step approach: in the first step, ask the judge model to generate a solution; in the second step, use this solution as a reference for evaluating the responses.

Additionally the authors evaluate the judges performance in two different setups:

1. **MT-bench:** Authors used the proposed dataset *MT-bench*, containing 80 high-quality multi-turn questions that evaluate the LLM’s capabilities on the most common tasks. They answer the benchmark questions with six different models and compare the LLM judges evaluation with the judgement made by 58 human experts.
2. **Chatbot Arena:** Authors released a crowdsourcing benchmark platform that allows users to pose a question and receive two different anonymous answers from which they should elect the best one.

As a metric, the authors defined *agreement* as the probability that two different judges agree on the answer to a randomly selected question. These agreements are calculated in relation model-to-model, model-to-human and model-to-human-majority. The results demonstrate that advanced LLM judges, such as Generative Pre-training Transformer (GPT)-4, can achieve up to 85% agreement in pair-comparison grading (w/o tie) with human evaluators, exceeding even the agreement between human experts (81%).

LLM-as-a-judge highlights the potential of using another model to evaluate the performance instead of requiring manual human expert labour. However, this approach not

only uses more computational resources but also introduces novel challenges, particularly in selecting an appropriate judge model.

3.6.2 LVLM Judges

LVLMs bring not only significant advancements but also a new set of challenges in evaluation. Unlike LLMs, which focus only on text, these models require a deeper understanding of the visual data. This means that not only human evaluators take more time to analyse videos and evaluate model responses, but even model-based judges may face difficulties in evaluating the outputs.

Chen et al. [6] reveal that LVLMs demonstrate alignment with human judgements in the pair comparison setting but struggle with scoring evaluations and batch ranking (ordering a batch of responses by quality). In testing, scoring was measured using the previously defined *agreement* 3.6.1, pair comparison was evaluated with both F1-score [15] and recall [15], and batch ranking was assessed using normalized Levenshtein distance [22].

The best-performing judge was *GPT-4V*, demonstrating a clear performance advantage over the other models tested by achieving 70%, 80%, and 62% agreement in score, pair comparison and batch evaluation, respectively. Similarly to *LLM-as-a-judge* [63], the authors identified biases such as position bias, verbosity bias, and self-enhancement bias. Batch ranking showed a higher frequency of hallucinations, leading the authors to test the top two models, *GPT-4V* and *Gemini*, with CoT reasoning. While this approach effectively reduced hallucinations, it failed to improve the average agreement, even degrading the overall performance.

Additionally the authors propose the *MLLM-as-a-Judge* benchmark, curated with human annotations to assess the judging capabilities of the LVLMs. They also introduced two new datasets: *MLLM-as-a-Judge-HQ*, which contains high-quality questions, and *MLLM-as-a-Judge-HARD* that adds hallucination instances.

3.7 Multi-Agent LLM Setting

Nowadays single models already achieve impressive results in a diverse range of tasks, but sometimes using a set of specialized models to solve simpler tasks and discuss solutions can outperform one-model usage. This section summarises the main concepts introduced by Becker [4], a study that covers 20 multi-agent setting projects. Although the studies cover mostly text-based scenarios, extending this paradigm to include the vision modality is a straightforward process, making these insights pretty relevant for the framework development.

3.7.1 Types of Agents

An agent is defined as a node in the dialogue structure, any system that contributes to the ongoing discussion either by participating or ensuring its smooth progression. Agents (table 3.1) are usually categorised into two main categories: **moderators**, which play a neutral role and manage the interaction between other agents, and **participants**, which actively contribute to the dialogue by defending their perspective or enhancing the current solutions proposed.

Table 3.1: Types of Agents

Category	Classification	Description [Projects that use]
Moderator	Draft Proposer	Summarizes the already proposed ideas. [51]
Moderator	Turn Manager	Decides which models interact next. [46]
Moderator	Policy Feedback	Gives feedback to other agents, like remembering personas and guidelines. [14, 19]
Moderator	Task Specifier	Specifies the complex task into simpler steps for the other agents. [26, 66]
Moderator	State Classifier	Decides if the ongoing discussion should be ended. [14]
Participant	LLM	Generates constructive feedback to other agents and improves the current dialogue. [8, 27]
Participant	Persona	Executes the role of a domain expert. [19, 46, 51, 54]
Participant	Memory	Controls the current dialogue memory of each other agent. [34]
Participant	Tools	Provide external tools for problem solving. [65]

3.7.2 Important Discussion Aspects

While agents interact and form new solutions, certain discussion aspects (table 3.2) must be considered to create a stable structure for the ongoing problem-solving task. These aspects are divided into two categories: **paradigm**, which handles the communication protocols, and **prompting**, which defines agent-related specifications.

Table 3.2: Important Discussion Aspects

Category	Classification	Description [Projects that use]
Paradigm	Turn Order	Sequence of interactions between agents. [58]
Paradigm	Summarizer	Reducing the extensive dialogue. [12]
Paradigm	Visibility	Agent dialogue information access. [58]
Paradigm	Agent Number	Number of agents communicating, can be static or flexible. [12, 51]
Prompting	Context Length	Balance between shorter but efficient or high-quality but detailed prompts.[12]
Prompting	Persona Assigner	Generating suitable personas for a specific task. [51]

3.7.3 Decision-Making Mechanisms

At some point during the discussion, selecting a solution that satisfies the agents will be required. Decision mechanisms (table 3.3) can be split among three key categories: **voting**, where each agent votes in the solution(s) that best fit its interests; **consensus**, which selects a solution without directly interacting with the agents; and **tree search**, where a solution tree is built as solutions are proposed and then employs exploration techniques to select the best solution.

Table 3.3: Decision-Making Mechanisms

Category	Classification	Description [Projects that use]
Voting	Ranked	Each agent ranks solutions from best to worst. [55]
Voting	Cumulative	Each agent distributes a fixed number of points among the solutions. [55]
Voting	Approval	Each agent approves a number of solutions. [55]
Consensus	Consistency	Selects the solution that is most similar to all other solutions. [50]
Consensus	Iterative	A single solution is iteratively improved until it satisfies most agents. [8, 42, 51]
Tree Search	Critic	Uses an external evaluator to find errors and assess solutions. [17, 26]
Tree Search	Heuristic	Uses heuristic algorithms to traverse the solution tree. [21, 53, 64]

3.8 Reasoning Strategies

Solving non-trivial problems with a single interaction between the user and a model is fairly rare, as the vast majority of complex tasks require some high-level strategy. Instructing a model to solve a problem iteratively, by building a step-by-step solution, often results in more accurate final answers. This is a consequence of encouraging the model to generate intermediate reasoning tokens, sentences that simulate coherent thought. These steps help the model navigate through latent reasoning states that would otherwise be inaccessible.

3.8.1 Chain-of-Thought

This simple step-by-step request prompting technique was named CoT [52] (Figure 3.9) and marked a new valuable strategy, instead of solely developing more capable models why not change the way a model approaches the problem.

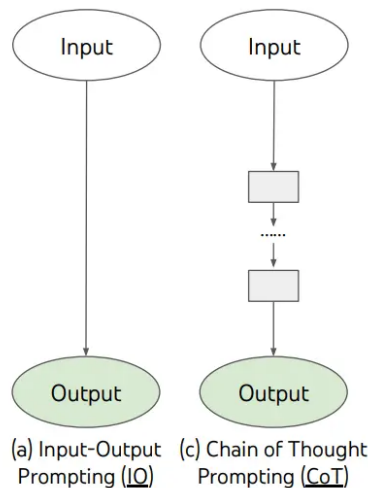


Figure 3.9: Comparison between standard Input-Output (IO) prompting and Chain-of-Thought (CoT) prompting [52]. While IO prompting reaches the final solution in one single step, CoT introduces intermediate reasoning steps that help guide the model towards a better final solution.

3.8.2 Tree-of-Thought

On top of this idea, new implementations tested multiple-path solutions, i.e., solving the problem step-by-step multiple times to verify the final answer. Yao et al. [56] go even further and developed Tree-of-Thought (ToT) (Figure 3.10), an explorative approach that suggests different solutions for each step and then selects the ones that perform best. This method constructs a tree structure where each node represents a thought, and where, iteratively, a new solution for a step is selected. This enables a more error-resilient development process, as a single-step error that would otherwise impact the final answer

can be replaced by an alternative step solution that better guides the problem-solving reasoning process.

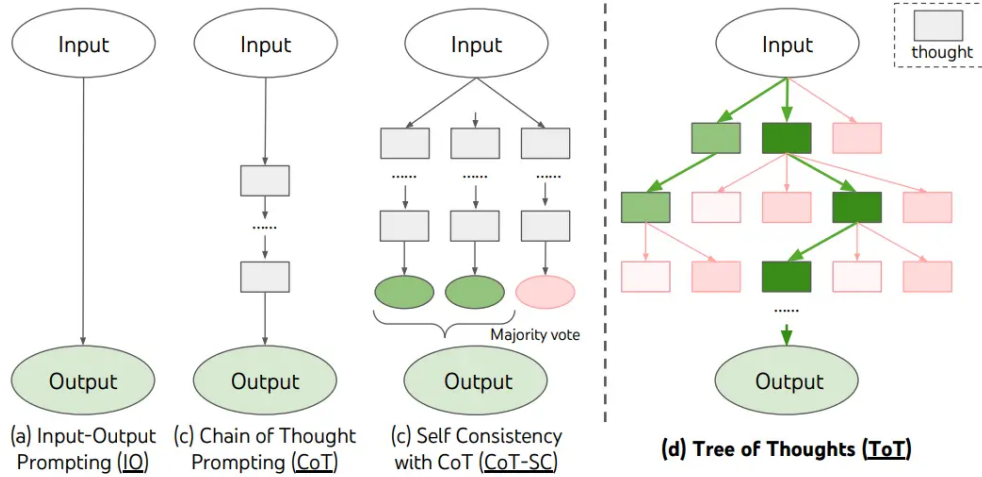


Figure 3.10: Comparison of reasoning strategies [56]. Chain-of-Thought with Self-Consistency (CoT-SC) extends CoT prompting by exploring multiple paths in parallel and selecting the final answer through majority voting. ToT generalises this idea by exploring different intermediate-step solutions and selecting the most promising paths at each stage.

3.8.3 Iteration-of-Thought

Iteration-of-Thought (IoT) [39] introduces a more dynamic and adaptive alternative to the ToT structure by utilising two LLM agents (Figure 3.11):

- **Inner dialogue agent (IDA):** The IDA functions as a "guide" that dynamically generates context-sensitive prompts based on the initial task defined and the previous answer. This way, IDA can provide refined prompts that utilise the details captured in the previous iteration.
- **LLM agent (LLMA):** The LLMA performs the core reasoning needed, producing the answers for each prompt generated by the IDA.

This strategy provides more context-dependent prompts which improve performance in highly flexible tasks.

3.8.4 Layered-Chain-of-Thought

To tackle the mid-reasoning errors found in CoT, a new layered strategy was developed. Layered Chain-of-Thought (LCoT) [41] starts by dividing the initial problem into partial objectives and verifies the sub-problem conclusions between layers using external sources (e.g., domain-specific databases, search engines, or human input). In case of detecting

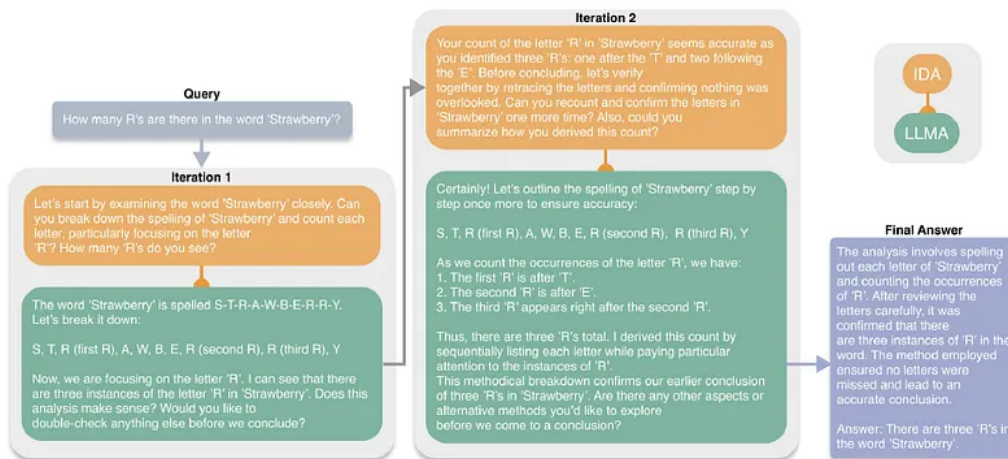


Figure 3.11: Use case of IoT [39] prompting to count the occurrences of the letter "R" in the word *Strawberry*.

inconsistencies during the verification process, the corresponding layer can update its partial chain-of-thought to reach a concordant state. LCoT is considered a safer and more reliable approach to conduct step-by-step reasoning.

3.8.5 Critical Summary

In summary, even though the current state of surveillance is far from optimal (due to the lack of novel large-scale open-source complex datasets, most tasks are directly related to anomaly detection and baseline models remain largely unexplored in terms of surveillance-specific training) the minimum requirements are met to proceed with surveillance video analysis tasks. Recent models demonstrate impressive results in common benchmarks but struggle with OOD tasks, forcing researchers to adapt and develop new methods. Recently developed works employ models as evaluators to reduce for human labour needs as comparing answers is becoming increasingly more difficult due to the complexity of the tasks, requiring extensive criteria and often a deep understanding of the problem context. Additionally, some works, particularly in the text-only modality, are utilising compositions of models or step-by-step reasoning approaches to solve complex problems more efficiently, as dividing the problem into simpler steps and sequentially solve or reason about them usually generates better answers. Finally, these literature works further emphasize the strong foundations of the *MEIDE* framework by combining a set of specialised models and dividing the surveillance video summarisation task in simpler question-answering pairs to iteratively explore and reason about the environment.

PROBLEM DEFINITION AND INITIAL FRAMEWORK

This chapter covers the problem definition and the initial framework developed. It starts with an analysis of the inherent challenges of the surveillance domain and the metadata extraction task, followed by an overview of how these initial challenges manifested during a preliminary testing phase, and concludes with an in-depth description of the base framework and the evaluation tasks used.

4.1 Problem Definition in Surveillance Video Metadata Extraction

This section explores the most prominent challenges in metadata extraction and surveillance footage. Explaining firstly how these limitations affect general video models and concluding with meaningful insights drawn from the first set of qualitative preliminary experiences.

4.1.1 Surveillance Footage Challenges

The surveillance domain presents unique challenges, as the data is not only heavily restricted due to legal and privacy concerns but also highly complex to analyse, mostly consisting of low-resolution videos filmed from a large distance, using unusual camera angles, sometimes containing harsh lighting conditions, and recorded within cluttered environments where multiple people perform actions simultaneously. Additionally, as a consequence of the scarce availability of high-quality surveillance videos and the underrepresentation of surveillance within the broader video-modelling landscape, LVLMS are predominantly trained on general-purpose video domains rather than surveillance-specific contexts. This misalignment introduces critical difficulties into the video analysis process, as these models lack contextualisation of criminal intent and dangerous situations. As a result, they tend to soften the interpretation of criminal acts unless these are completely explicit or the prompt introduces strong criminal biases.

4.1.2 Metadata Extraction Challenges

One of the main challenges in metadata extraction is the difficulty of obtaining large amounts of contextual information about a certain video with a single model call (i.e., in a unique generation process). Models usually overlook important information when prompted to describe a general scenario, focusing primarily on the central actions and entities while ignoring all the middle ground events. While this is sufficient for some use cases, metadata extraction aims to generate a dense representation of the video’s content, which should include many of these ignored aspects. Additionally, surveillance is also heavily dependent on these subtle details as they can often be crucial for identifying criminal activities. As an example, Figure 4.1 shows a simple daily-life video and the corresponding answers to different prompts.

4.2 Preliminary Analysis Observations

To better understand how impactful these initial challenges manifest in practice, we designed a preliminary experimental setup. In this setup, the baseline video model was prompted with increasingly fine-grained questions about the video environment, while providing the correct answers from the previous prompts as context. The setup conducted tests across different model sizes, frame rates, video durations, and cropped video variants. Building on this setup, we drew the following conclusions:

- **Resolution:** Low-resolution inputs consistently caused the model to hallucinate when questioned about more fine-grained information. Even high-resolution videos showed this issue when asked directly about something that is small in scene (e.g., a specific object).
- **Frame-rate:** Increasing the frame rate above one frame per second did not significantly change the results.
- **Clutter:** Cropping background noise did not eliminate hallucinations, further emphasising that resolution is the main bottleneck for good performance.
- **Video duration:** Shorter clips may give less context, making reasoning more difficult, while longer videos may capture more unnecessary information for answering a specific question. Additionally by handling really large video durations sampling becomes a prominent topic as selecting a smaller percentage of the total frames can significant impact on the overall content representation.

Daily Life Example Scenario

Visual Input: VideoMME Daily Life



Description: The video depicts a student making his way into the school building passing by a parking lot containing cars and trees. Additionally some captions give further context.

Textual Input and Qwen2-VL:72BQ4 Model Responses:

Prompt: Describe this video.

Answer: The video shows a boy walking down the street and passing by a building. The boy is wearing a suit and carrying a bag.

Prompt: Describe this video in detail.

Answer: The video shows a school with a **parking lot** and a building with large windows. A boy wearing a suit is walking down the street, carrying a book. The video also features a man walking on the sidewalk in front of the building. The school has been around for about 400 years and has approximately 1000 students divided into seven age groups, all of whom are boys. The school day starts at 8:30 am and consists of two lessons until 10:15 am, followed by a 15-minute break. After that, there are two more lessons until lunchtime at 12:40 pm.

Prompt: Are there any cars in the video?

Answer: **Yes, there are several cars parked outside the school building** in the video. The cars are parked on the street and in front of the school building. They appear to be parked in designated parking spots.

Figure 4.1: Example of a *Video-MME* daily life scenario, illustrating the loss of information in a description when compared to directly questioning the existence of cars.

Analysis:

The model can easily understand the video's purpose, identifying the main focus and respective actions, but does not mention the presence of vehicles until directly asked. This

highlights the need for targeted questions to extract more information from the scene, details often get overlooked and omitted.

4.3 Agent-Based Question-Answer Framework

Accounting for the core challenges introduced by the video metadata extraction task and the surveillance domain footage, the initial framework concept revolved around using two agents: one LLM agent capable of generating pertinent follow-up questions, and one LVLM agent to answer them based on the video's content. This section covers the base framework's architecture, advantages, and evaluation tasks.

4.3.1 Motivation

As previously identified, the core challenge introduced by the metadata extraction task is the inability of, in a single model execution, obtaining all the information needed to produce a rich representation of the video's content. This can, however, be tackled by performing consecutive calls to answer pertinent follow-up questions that explore the depicted scenario. The questions would start with simple general searches and increasingly dive into the main aspects needed, aiming to explore the video's environment step by step and ultimately provide a rich representation of everything contained in the video.

4.3.2 Architecture

The architecture enables an interactive dialogue between its two agent parts: **The Questioner LLM**, instructed to generate pertinent questions taking into account the previous question-answers pairs, which provide valuable context allowing the model to explore different themes and understand underlying relations between environment elements, and **the Answerer LVLM**, that answers the generated questions based on the video's depicted content (Figure 4.2).

4.3.3 Error-resilience

The Q-A framework reduces the total omissions, as the model is not forced to describe the complete scenario in a single answer, which considerably reduces the amount of information retrieved about each subject. Confusions caused by cluttering are also more easily managed, as the Questioner agent has to identify the points of interest by their description, allowing the Answerer agent to focus on a single entity at a time. Additionally, this base approach has a slight advantage in handling hallucinations, as it can detect them by searching for inconsistent responses throughout the dialogue

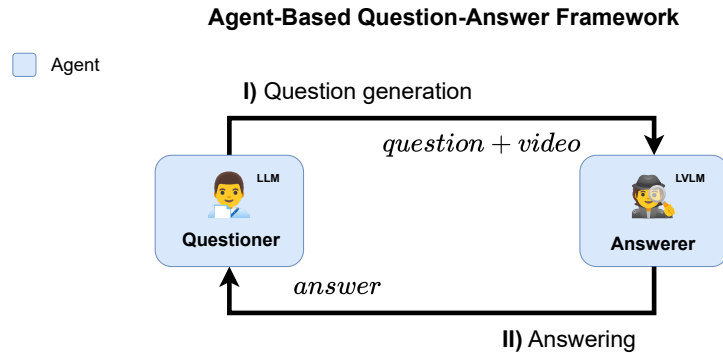


Figure 4.2: **Agent-Based Q-A Framework.** An interactive dialogue between the Questioner (LLM), which generates context-aware questions, and the Answerer (LVLM), which provides answers grounded in the video content.

4.3.4 Evaluation Tasks

While the main focus of this thesis is to perform metadata extraction in surveillance footage, there is no established benchmark for that purpose in the surveillance domain. In order to test and compare the developed solutions quantitatively, we use the *UCA* [59] dataset to perform anomaly detection and classification tasks, aiming to use these comparative results to discover critical fragilities of the base framework and measure the performance improvements over single input-output model executions.

The tasks utilised to evaluate the solution performance initially were:

- **Abnormal vs Normal** - The solution is questioned about the normality of a certain video.
- **Belongs vs Not Belongs** - The approach is questioned, for each class iteratively, whether a certain video belongs to that specific class.
- **Multiclass Classification with Normal** - The solution is questioned about which class the video belongs to.
- **Multiclass Classification without Normal** - The approach is questioned about which class the video belongs to, but this time the Normal class is not a valid option and the Normal examples are not processed.

With this initial testing setup, we aimed to distinguish solution criminal-based capabilities and clarify potential directions for future strategies. **Abnormal vs Normal** is an obvious choice for two main reasons: firstly, any surveillance solution must be able to recognise crime, even if it struggles to detect its specific criminal category, and secondly, due to its natural adaptation through the already existing surveillance data (discussed in Chapter 3), requiring mostly just class relabelling to be performed. **Belongs vs Not**

Belongs is a binary classification task, somewhat similar to Abnormal vs Normal but with a different objective. With this task, we aim to explore the capabilities of the base model to decide if a certain video belongs to a specific class individually. This enables the detection of similar classes that could cause model confusion, as the model can now select them both as belonging individually. Additionally, such conclusions could enlighten a future solution scheme where, by developing the contextual knowledge, the architecture would be capable of reducing the total number of probable categories and eventually presenting its certainty about the video belonging to each remaining individual category. **Multiclass Classification** is the direct implementation over the already existing benchmarks, allowing models to compete in correctness when classifying different types of crimes, and its without-normal variant is meant to test the solutions while erasing the normality bias present in the models, i.e., reducing the model's tendency to say that everything is normal.

BASELINE AND BASE MULTI-AGENT APPROACHES: STATIC, DYNAMIC AND REASONING-BASED

The Questioner-Answerer framework enables different strategies to perform either anomaly detection/classification or metadata extraction tasks. This chapter highlights the baselines used as well as the methods developed: static approaches, which fix a set of predefined questions to use; dynamic top-k approaches, that utilise criminal-class-based question generation; and finally, reasoning-based approaches, that explore the video’s environment through iterative, context-aware questions.

5.1 Single LLM-Based Baselines

We define baselines as direct, single-turn inferences in which the LVLM is queried to produce the target task output given a video, without auxiliary context or iterative questioning (prompt available in Appendix A.1). The selected baselines used were *Qwen2.5-VL* [3] (7B), and *LLaVA-Video* [62] (7B), recent representatives of the *Qwen-VL*, and *LLaVA* families, respectively. This selection was motivated by their competitive results across various benchmarks (as highlighted in Chapter 3), as well as their open-source availability, making them suitable for extensive experimentation and easy replication.

Qwen2.5-VL *Qwen2.5-VL* was selected as one of the core candidates to take the LVLM roles in the upcoming strategies. In comparison with its predecessor *Qwen2-VL* (introduced in Section 3.5.4), this new version incorporates key architectural refinements, such as absolute time encoding, enabling it to process videos of extended durations with second-level event localization, thereby improving its temporal reasoning and making it more suitable for longer surveillance video context relationships.

LLaVA-Video *LLaVA-Video* (introduced in Section 3.5.2) was selected as an alternative LVLM candidate to face *Qwen2.5-VL* in surveillance footage analysis. However, as

discussed in Section 7.3, our baseline and initial strategy evaluation revealed that the *LLaVA-Video* baseline and its variants were being outperformed by the *Qwen2.5-VL* variants. For this reason, subsequent experiments focused primarily on developing a solution within the *Qwen2.5-VL* variants domain.

Model Roles in Comparisons The defined LVLMs will perform either the single-turn inference, in baseline solutions, or the role of answerer when performing metadata extraction through iterative dialogue. We fix *Qwen2.5* [37] (7B) to perform the reasoning and question-generation tasks in the upcoming strategies. Its extended context length, improved training, and enhanced reasoning ability allow greater sustainability during the iterative dialogue cycles and stronger foundations for common sense and domain knowledge, making it a reliable Questioner to explore the video’s environment. Selecting a good LVLM is key to achieving desirable results in both anomaly detection/classification and metadata extraction, as the models need to understand the subjects, their attributes, and the events presented in the video, representing the main bottleneck of the whole system. On the other hand, the LLM selection is more flexible, as generating pertinent questions based on previously discovered context is much more natural than understanding rough criminal footage and translating it into textual answers. The only key requirement for the LLM is having sufficient knowledge to relate previous events and predict human intents, thereby enabling a better exploration of criminal activity based on the given descriptions.

5.2 Static Fixed-Set Questioning Approaches

The first proposed method covers static questioning approaches, strategies where the same exact set of questions is used in order to decide the task’s output.

Interaction Diagram The interaction process begins with the Answerer LVLM responding individually to each question $q \in \mathcal{Q} = \{q_1, q_2, \dots, q_Q\}$, a predefined set of pertinent questions. Then, the Questioner LLM agent utilises the produced question-answer pairs $\mathcal{P} = \{(q_1, a_1), (q_2, a_2), \dots, (q_Q, a_Q)\}$ to decide the final output for the anomaly detection and classification tasks (Figure 5.1).

Variants This proposed method includes four different variants where the set of static questions used is selected in varying ways. These variants select a total of five pertinent questions per category and choose the final answer based on the LLM’s thoughts about the question-answer pairs provided. The number of questions performed was selected with the intuition that having fewer questions would reduce the likelihood of achieving a clear separation between classes (which was identified as a recurrent problem in Chapter 8), whereas having a larger number of pertinent questions per class would not significantly improve the results, as after extensively analysing the possible similarities between actions

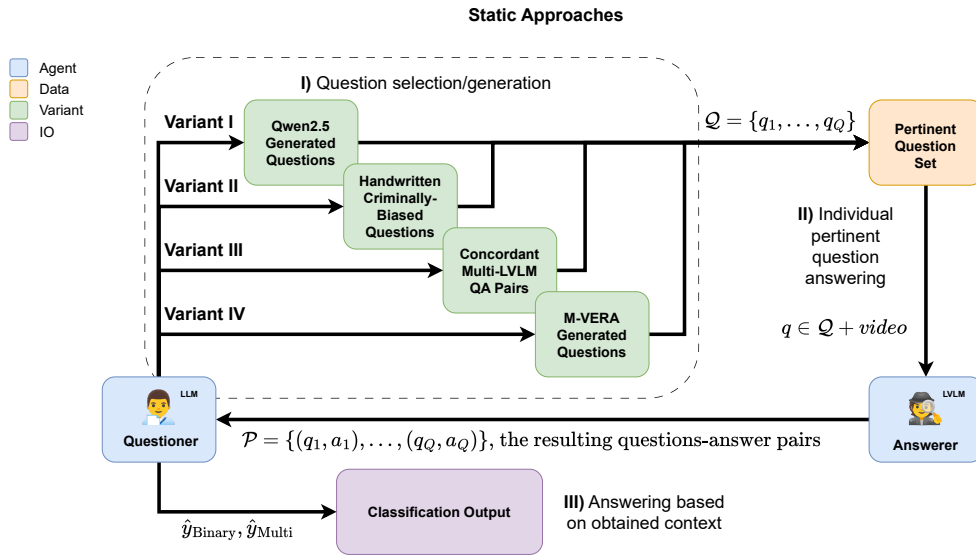


Figure 5.1: **Static approaches interaction diagram.** The Answerer LVLM responds to a predefined set of questions and the Questioner LLM uses the resulting Q-A pairs to decide the final output.

contained in UCF-Crime [5] classes, we concluded that they are usually confusable with at most four other classes.

Of the four variants designed, three modify the questions posed to the answering LVLM, while the other filters the remaining question-answer pairs, thereby ensuring consistency before using them in the final evaluation call:

AI-generated Questions The first variant utilises a set of questions completely generated by *Qwen2.5* (prompt available in Appendix A.1). This strategy allows to not only evaluate the capabilities of *Qwen2.5* in generating pertinent questions without any prior context, aiming to identify if a certain video belongs to that class, but also its criminal knowledge to separate categories based on their, sometimes subtle, inherent differences.

Handwritten Criminal-Biased Questions The second variant utilised handwritten questions produced by the development team, aiming to clearly separate each class from its most similar classes (e.g., to distinguish an abuse scenario from a fighting one, a pertinent question might be: "Does the video depict one-sided violence where one individual is overpowering another without retaliation?"). These questions are, by consequence, much more criminally biased, seeking correctness by reducing the normality bias through directly questioning criminal scenarios and intents.

Cross-Checked Multi-LVLM Approach The third variant takes advantage of both tested LVLMs. It utilises the questions generated in the previously designed variants, but now

it only uses the question-answer pairs in the final evaluation call if the answers given by both tested LVLMs are consistent (prompt available in Appendix A.1). This variant aims to reduce single-model failures due to hallucination over the already predefined question-generation schemes.

M-VERA-Generated Questions The final variant, Multiclass VERA (M-VERA), makes use of the *VERA* [57] project’s concept (explored in Section 3.3.3) to generate pertinent class-dependent questions. Unlike *VERA*, which optimises a set of questions to separate abnormal scenarios from normal ones, we filtered the dataset videos by category and optimised a set of pertinent questions for each individual class. The produced sets of questions (displayed in B) were impressively precise, taking into account the model had no additional context beyond the generic abnormal label. These results highlight both the model’s ability to capture event cues and the effectiveness of prompt-optimising strategies in improving question quality.

5.3 Dynamic Top-K Questioning Approaches

The analysis of the baseline and static approaches (Section 7.2 and Section 7.3) revealed that baseline methods achieve unexpectedly strong performance. These findings motivated the design of a second proposed method, one that could utilise this advantage in its favour. It starts with the LVLM providing its point of view on the possible criminal categories contained in the video, aiming to surpass baseline performance in both Abnormal vs Normal and Multiclass Classification by better adjusting its final answer using additional context extracted from the question-answer pairs.

Interaction Diagram The interaction process begins with the Answerer LVLM selecting an initial set of K most probable classes that could be depicted in the surveillance video, denoted as \mathcal{K} , without any additional context beyond the video footage, class names, and respective class descriptions (mimicking the baseline direct selection, but this time selecting a set of classes rather than a single class, prompt available in Appendix A.2). For each selected class c in \mathcal{K} , the Questioner LLM generates a set $\mathcal{Q}_c = \{q_1^c, q_2^c, \dots, q_Q^c\}$ of Q questions, that are answered individually by the Answerer LVLM, aiming to achieve more criminal context from the video. Then, the Answerer model is again prompted to reduce the number of current probable classes, producing a new smaller set \mathcal{K}' (with $|\mathcal{K}'| < |\mathcal{K}|$) and then repeating the loop, creating new pertinent questions for this new smaller set and consequently erasing more classes contained in the current set \mathcal{K}' . When the set reaches $|\mathcal{K}'| = 1$, meaning there is only one probable class left, it is selected as the final answer for both Abnormal vs Normal and the Multiclass Classification tasks (Figure 5.2).

Variants This proposed method includes variants that change the starting number of classes K , allowing more steps in the reduction process; variations in the number of

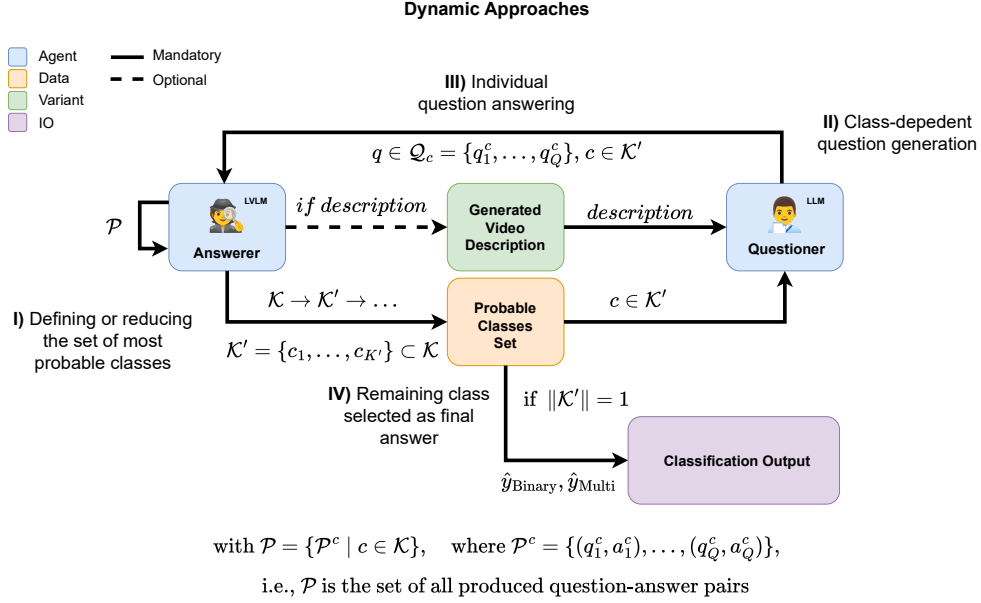


Figure 5.2: **Dynamic approaches interaction diagram.** The Answerer LVLMM iteratively refines the set of probable classes $\mathcal{K} \rightarrow \mathcal{K}'$ through the analysis of class-dependent-question-answer pairs, generated initially by the Questioner LLM and answered by the Answerer LVLMM

questions generated per probable class (Q); and, finally, a version where the Questioner LLM receives a brief description of the video’s content, produced initially by the LVLMM, aiming to generate more environment-context-dependent questions.

5.4 Reasoning-Based Approaches

While the previous approaches focused on achieving better performances than the currently established baselines on simpler tasks such as anomaly detection and classification, the reasoning-based approaches place more in-depth focus on surveillance aspects. We are not only interested in detecting criminal activity but also in understanding the environment, participants, objects of interest, and respective attributes. As the evaluation of dynamic approaches revealed (Section 7.4.2), there is the possibility, that improvements come not from solution-related capabilities but merely from using a more appropriate distribution of randomly selected choices over the possible outputs due to the increased criminal bias. Considering this, the reasoning-based approaches aim not only to surpass baselines on the defined tasks but also to go beyond first-order detections, progressing through the environment via evidence-grounded, hypothesis-driven reasoning in search of plausible predictions and extractable aspects, thereby demonstrating deeper video understanding and ensuring that the results reflect inherent model capabilities rather than artefacts of bias.

Motivation Exchanging previous approaches that generated pertinent questions based on categorical criminal descriptions with strategies that explore the video’s environment with meaningful inquiries about the location, subjects, events, and relations presents a new set of challenges. While we were previously constrained to generating criminal-dependent questions, the strategy is now free to iteratively explore the points of interest detected throughout the dialogue. This new paradigm supports proper metadata extraction and ensures that these approaches surpass the baseline due to their surveillance video understanding capabilities. Solutions need to be capable of starting with no bias towards criminal activity and gradually building suspicion about the content depicted in the video, rather than being directly biased by criminal statements that may reduce normality bias but risk leading to hallucinations without having proper evidence-based hypotheses.

We have developed innovative versions of the reasoning strategies explored in Section 3.8 to match the Questioner-Answerer framework roles, as their previous iterations were designed for single LLM model usage only. The four reasoning baselines explored and their respective changes are:

Chain of Thought In base CoT [52] the model is prompted to think step by step, which is not directly applicable in this scenario, as the reasoning process is developed with an iterative dialogue between the two framework agents (prompts available in Appendix A.3). To address this, we restructured CoT by making the Questioner LLM initiate the reasoning process with a single question, and adjusting the next steps depending on the answer received (Figure 5.3).

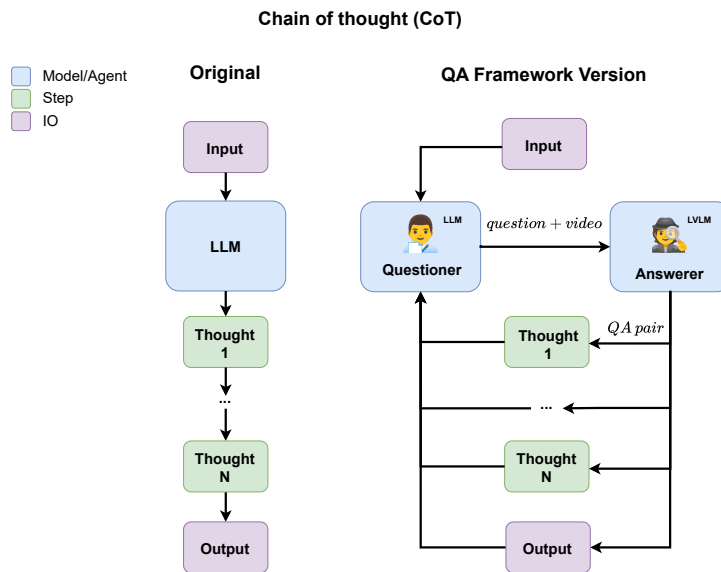


Figure 5.3: Comparison between original Chain-of-Thought (CoT) reasoning and its adaptation to the Questioner–Answerer framework.

Tree of Thought ToT [56] fundamentally changes by splitting the question-generation and answering processes into two different agents, whereas previously all these operations were managed iteratively by a single model. The Questioner LLM generates a set of questions at each step, the Answerer LVLM answers them individually, and the best step solution is then selected (prompt available in Appendix A.3). This latter selection evaluates question-answer pairs based on the relevance of the information revealed and their usefulness toward the main final goal, taking into account all the information discovered (Figure 5.4).

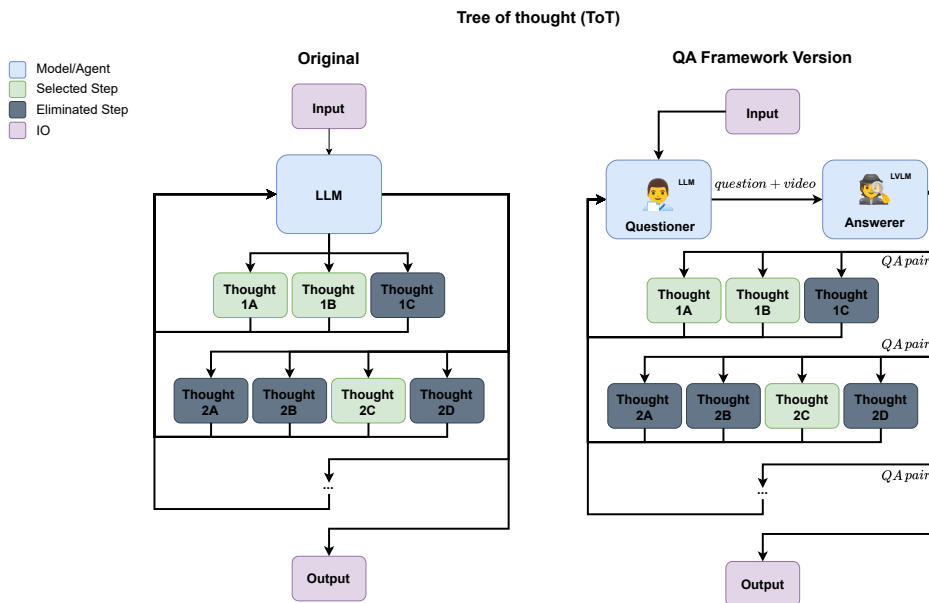


Figure 5.4: Comparison between original Tree-of-Thought (ToT) reasoning and its adaptation to the Questioner–Answerer framework.

Iteration of Thought IoT [39] can be directly translated into our framework, as it already uses two agents (see Section 3.8.3, available in Appendix A.3). IDA generates the context-sensitive prompts, similarly to how the Questioner LLM performs the question-generation process based on the discovered context. LLMA executes the core reasoning to answer IDA’s generated prompts, which is now performed by the Answerer LVLM agent, as we need to analyse the video to uncover new details about its content (Figure 5.5).

Layered Chain-of-Thought The LCoT [41] first divides the problem into logical layers and solves them iteratively, allowing a more structured approach compared with the default step-by-step reasoning and ultimately allowing a more reliable detection of inconsistencies (prompt available in Appendix A.3). While the base strategy uses external tools to verify each layer’s conclusion, we do not have any source capable of performing such verification for our scenario, and thus we restricted LCoT by skipping these mid-layer correctness

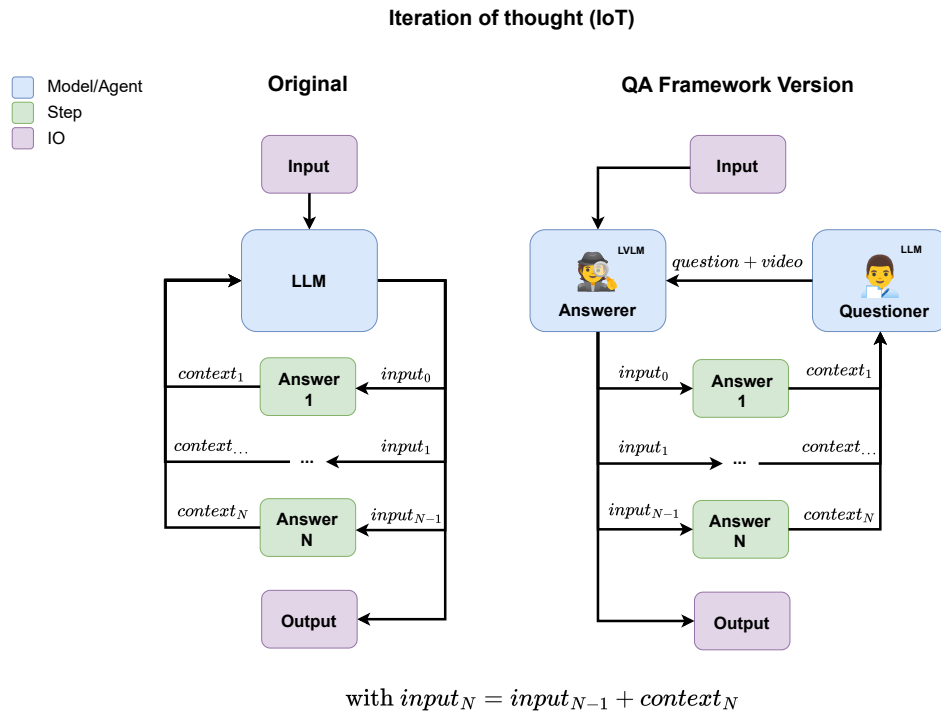


Figure 5.5: Comparison between original Iterative-of-Thought (IoT) reasoning and its adaptation to the Questioner–Answerer framework.

assessments. Our version utilises the Questioner LLM to manage the layer and question-generation processes and uses the Answerer LVLML to simply answer the generated inquiries (Figure 5.6).

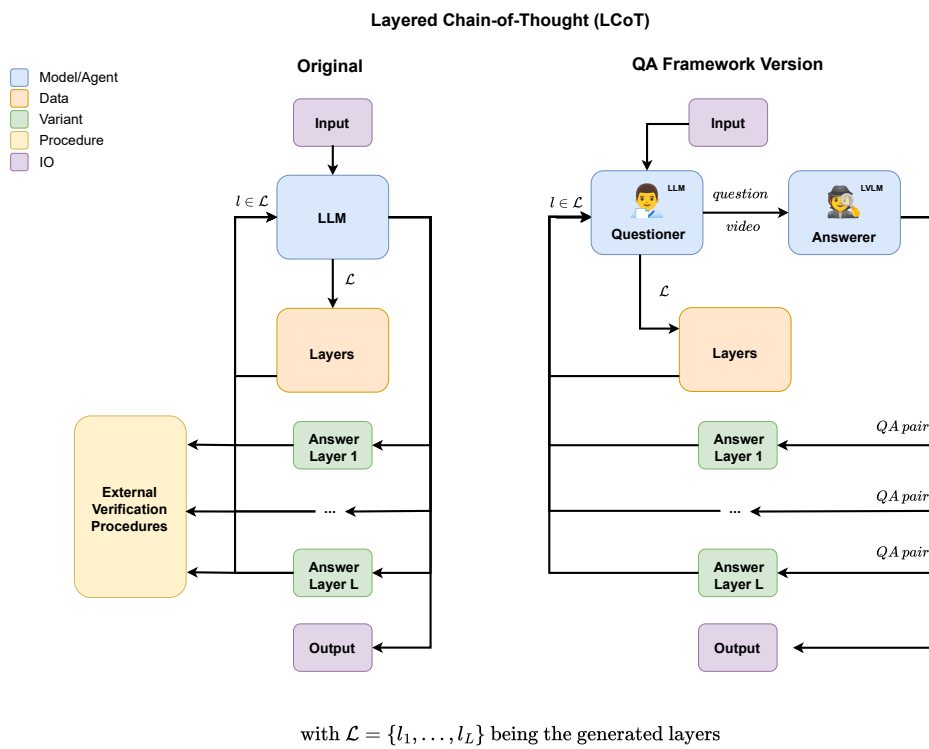


Figure 5.6: Comparison between original Layered Chain-of-Thought (LCoT) reasoning and its adaptation to the Questioner–Answerer framework.

EXTENDED FRAMEWORK AND SURVEILLANCE LAYERED GRAPH STRATEGY

This chapter builds upon the foundations established in the baseline, static, dynamic, and reasoning approaches, presenting the improvements that ultimately shaped the final solution, Surveillance Layered Graph (SLG). It starts by covering the key enhancements to the Q-A framework, allowing it to more flexibly perform surveillance-specific tasks and metadata extraction, followed by an in-depth exploration of the final proposed method, SLG.

6.1 Extended Q-A Framework

This section covers the fundamental thought process and key innovations presented in the extended version of the question–answer framework. It begins with an introduction to the core motivation, followed by an analysis of the architectural changes and the new data structure employed, and concludes with a discussion on balancing dynamism and control.

6.1.1 Motivation

Upon the developed strategy studies, achieving a fair balance between strategies was still missing. Even though static and dynamic strategies were, in theory, developing pertinent questions, these were still forced to be criminally dependent on the UCF-Crime [5] defined classes, resulting in good exploration toward anomaly detection and classification but struggling with more broader information extraction. On the other hand, reasoning strategies were more flexible in acquiring general context, as they could explore most of the normal video aspects independently, but lacked in clear objectives, often developing themes for too long, forgetting important topics, and even ignoring criminal aspects completely. This exploration process uncovers most of the important information about the video’s environment, making it a prominent approach for metadata extraction, but it is still incapable of independently diving into the criminal aspects. These problems mainly

reside in the question-generation process: LVLMs have intrinsic limitations (Section 8.2.2) that need to be surpassed by utilising better questions rather than model retraining, due to the existing data limitations (Section 8.2.1). The extended framework aims to balance these two conflicting needs, allowing a more flexible question-generation process without ignoring the criminal aspects completely.

6.1.2 Architecture

While still maintaining the previously declared personas, Questioner and Answerer, this new architecture introduces a new role, the Controller, along with a reasoning graph data structure. The interaction is now performed by the three agents in sequence: firstly, the Controller agent acts as a lightweight reasoning-planning agent, selecting an operation to perform based on the current state of the reasoning graph; then the Questioner agent is prompted to generate different types of inquiries depending on the operation selected; and finally, the Answerer model receives these questions and answers them individually, updating the reasoning graph state and repeating this cycle until the reasoning process is complete.

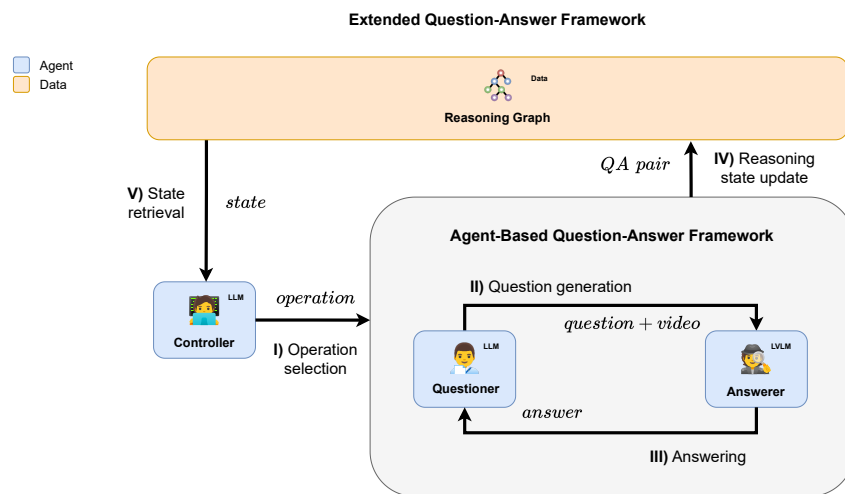


Figure 6.1: **Extended Question-Answer Framework.** The architecture extends the previous version by introducing a Controller persona and a reasoning graph to guide the interaction.

6.1.3 Reasoning Graph

The introduced reasoning graph data structure acts as persistent memory of the exploration process. Each node represents a question-answer pair performed within a sequence derived from the upcoming path nodes, meaning that each branch develops a single topic. Edges represent operations that can be defined by the framework users; these might include, for example, proceeding along the current reasoning path, halting, or performing some form of external or internal verification. This structure allows the Controller agent to

guide the reasoning process by providing a quick and simple view of the current reasoning state, preventing degeneration.

6.1.4 Balancing Dynamism and Control

While the previous framework lacked either question-generation dynamism, in static and dynamic approaches, or control, in reasoning-based solutions, these new improvements enable a more refined thinking strategy. With a set of strictly defined operations, the Controller agent is able to evaluate the state of the reasoning process and decide whether a topic is already overdeveloped or underdeveloped, whether there is a need to introduce criminal biases in case of clear individual suspicion, or even resolve previous hallucinations by demanding the Questioner agent to generate questions in that sense.

6.2 Surveillance Layered Graph Strategy

The extended Q-A framework provides a solid foundation for solutions aiming to perform both anomaly detection/classification and metadata extraction. This section covers the combined analysis obtained from previous experiences that jointly led to the development of the final proposed solution. It begins with the lessons learned from the previous reasoning-based approaches, followed by an in-depth explanation of the components of the final solution, and concludes with the main intended goal.

6.2.1 Lessons Learned from Previous Approaches

The first step to building a solid and coherent solution is to analyse and truly understand its base predecessors' fundamentals. While many key takeaways derive from the previous evaluation results (discussed in depth in Chapter 7), the base reasoning strategies developed in Section 3.8 already share important lessons that, when combined, create a solid foundational knowledge of the complex domain of solving problems with reasoning approaches. Table 6.1 summarises these lessons, and we briefly elaborate below.

Basic CoT [52] teaches that solving a complex problem is more effective when using a step-by-step approach, which is the main reason why reasoning strategies are fundamentally better for this type of complex tasks.

ToT [56] introduces a novel aspect into the equation by reflecting on the fragility of taking a wrong step during the reasoning process. Since it is common to deviate from the optimal solution when solving complex problems, thereby sacrificing all the previously developed work, using a multi-path solution enables a more error-resilient approach. Even if the algorithm takes a wrong step, it retains a backup path to proceed. This idea directly aligns with one central key question in this thesis development: "What is a good question to ask?" Sometimes, a good question is one that finds the solution efficiently, but

Table 6.1: Main takeaways extracted from each reasoning-based strategy.

Strategy	Main Takeaways
CoT	Solving step by step provides a more effective approach than direct solving for complex problems.
ToT	Exploring multiple sub-steps increases problem-solving capabilities. The best question is the one whose answer contributes most toward the main goal.
IoT	Refining the same question with additional, previously obtained, context helps capture hallucinations. Enables extraction of more fine-grained details.
LCoT	Anchoring discussion topics leads to a more structured exploration. Cross-validating information between layers makes detecting hallucinations easier.

in metadata extraction it is not that simple. There is no quick path to understanding the whole depicted scenario and extracting the important information, making it difficult to define a fixed set of questions that works in every single environment. However, by using a strategy like ToT, we can generate more than one question at each step and select the one that contributes most toward the main goal. Since question quality is generally a poor metric for evaluating the generated alternatives, we will instead evaluate question-answer pairs, which provide a much more reliable view of the subsequent reasoning states, with the obvious downside of having to answer every generated question, regardless of using them afterwards.

IoT [39] distances itself from the previous strategies by utilising a different question-generation pattern that aligns with one of our previous expectations regarding exploration limitations (detailed in Section 4.1). Instead of further developing a theme like the previous strategies, it refactors the last question with more context extracted from the previous answer, allowing the model to revise itself in case of hallucinations and enabling the exploration of more fine-grained details.

Finally, LCoT [41] establishes a new segmenting procedure to divide the reasoning process task into multiple layers. These layers provide two fundamental lessons: firstly, anchoring certain discussion topics is a solid technique to ensure all the needed themes are debated during exploration, addressing overlooked details and overall short, uninformative responses; and secondly, having the problem divided into multiple smaller debates helps detect single-layer hallucinations. As models are not directly prompted to follow an already defined reasoning path, they are more likely to create inconsistencies that can be easily detected.

6.2.2 Layered Architecture for Metadata Extraction

As explained in Section 6.2.1, LCoT mindfully introduces problem segmentation into reasoning approaches. Building on this key improvement, and on the conclusions drawn from the previous experiments in Chapter 7, we identified the most important aspects to search for in a surveillance video. The defined layers aim not only to perform general metadata extraction but also to contribute to overall video understanding, enabling the solution to classify the current footage as normal or abnormal. The layers were further divided into sublayers, aiming to simplify each theme and ensure that every necessary subtopic is debated. The layers and their respective sublayers are presented in the Table 6.2.

Table 6.2: Layers and respective sublayers.

Layer	Sublayers
Scenario Understanding	Location Type Time of Day Scene Objects
Entity Extraction	Person Count/Grouping Demographics Clothing
Social Context	Proxemics Gestures/Expressions Social Roles
Event Understanding	Events/ Actions Spatio-temporal Information Causality Links Suspicion Potential

Scenario understanding is a crucial topic, as it not only provides valuable context for metadata extraction but also defines a more focused scope for possible criminal classes (e.g., if the depicted scenario is within a forest, it will definitely not be a shoplifting scenario). Additionally, it provides meaningful context for the next layers, as it is much easier to generate questions about entities if you already know the surrounding environment. Entities are the main object of study, representing our biggest concern in terms of protection but also of suspicion, as they are the principal cause of most criminal scenarios. It is important to understand and identify people and their respective attributes. Social context is not as important as the previous ones but contributes by identifying second-level information, further developing the scenario knowledge through strategic approaches (e.g., understanding that an individual holding a gun is performing the role of a policeman and that everyone closely behind him are probably scared civilians, while the suspect is further away in front of him). Event understanding represents the most important topic in identifying crimes, discovering what is happening across the video's duration, understanding causality between actions, and finally raising suspicion about committed acts.

6.2.3 Graph Operations and Reasoning Control

Operations play a central role in the reasoning-guiding task, as they allow the Controller agent to manipulate the current direction of the reasoning process and take steps toward the main objective. Operations need to provide functionalities for the Controller to act appropriately in every possible situation, meaning that they must not be single-case specific and should allow generalisation in metadata extraction and free exploration, without forcefully eradicating meaningful discoveries but also without allowing the permanent development of already overexplored themes. The list of defined operations for this approach is as follows:

Proceed Every reasoning strategy needs to develop its thoughts, aiming to reach the final objective. Proceed, mainly inspired by ToT, creates a set of three new pertinent questions, which are then answered individually and evaluated according to their contribution toward the main goal, while also considering novelty from the already known intra-layer information (prompt available in Appendix A.4). The most prominent candidate is selected as the next step's question. The number of new questions created and selected derives from the conclusions drawn in Section 7.5. Proceed acts as the natural development of the reasoning exploration process (Figure 6.2).

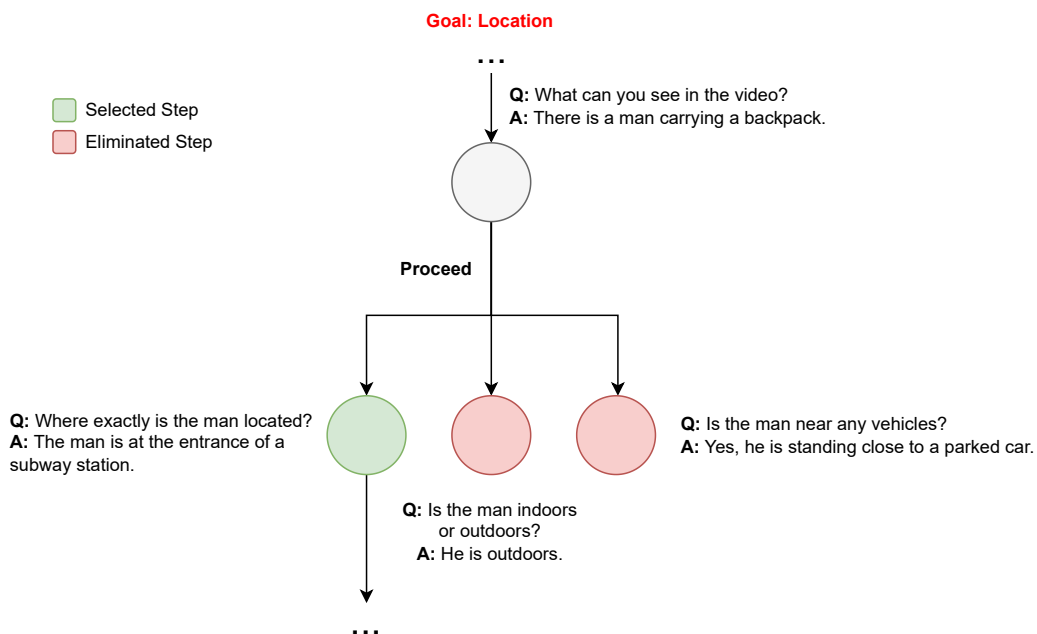


Figure 6.2: **The proceed operation.** It generates three pertinent follow-up questions and selects the one whose question-answer pair contributes most towards the main goal.

Refine Refine rewrites the same question with a more developed context based on the previous answer. This enables a more refined search, useful when details are initially

overlooked and needed for more fine-grained aspects. This operation is meant to be used wisely, as its repetition could lead to looped reasoning when the new context is already part of the prompt and no new aspects are detected, or to dangerous hallucinations, imagining details that do not correspond to the reality depicted in the video (Figure 6.3, and prompt available in Appendix A.4).

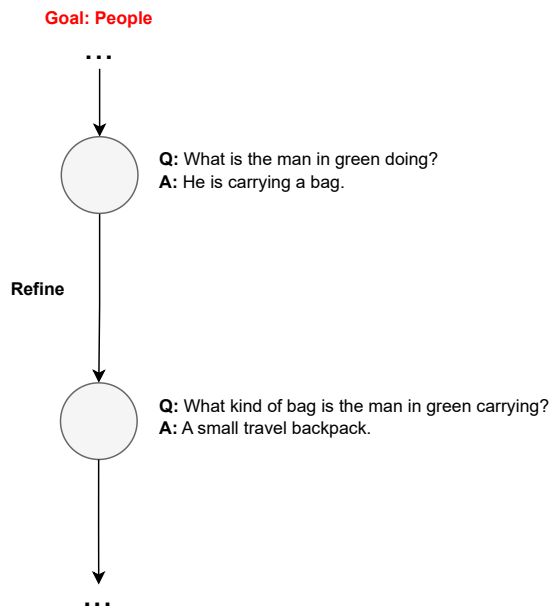


Figure 6.3: **The refine operation.** It rewrites the same question with a more developed context based on the previous answer.

Split Taking care of previously envisioned problems, split divides the current knowledge context into multiple segmented branches, allowing an individual exploration of the separated aspects and aiming to achieve better representations (prompt available in Appendix A.4). Its main use case is scenarios where the current context contains many similarly difficult subjects (e.g., a list of all the persons or objects), which are more easily explored by dividing the whole context into smaller sub-branches, thereby reducing the risk of shallower answers with omitted details (Figure 6.4).

Stop While developing the reasoning is crucial, solutions need mechanisms to end the reasoning paths. Stop dictates that a certain node has already achieved its goal, either by actually completing all the requirements or by demonstrating incapacity to obtain the needed information, and excludes it from the exploration list. This not only helps reduce the total execution time, but also contributes to finding a more stable balance between efficiency and effectiveness (Figure 6.5).

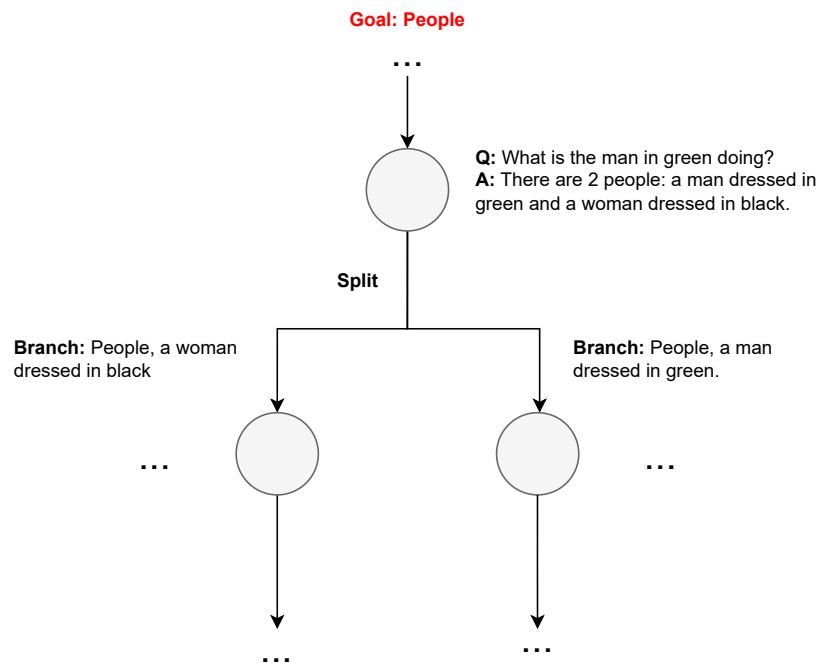


Figure 6.4: **The split operation.** It divides the current knowledge context into multiple segmented branches.

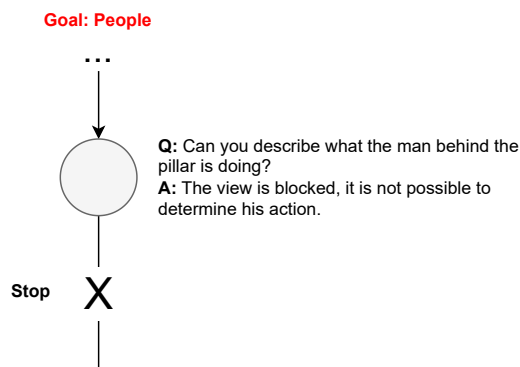


Figure 6.5: **The stop operation.** It dictates that a certain node has already achieved its goal, excluding it from the exploration list.

With this set of operations, we aim to enable the Controller agent to dynamically manage the reasoning process, focusing mainly on discovering new aspects while also providing more restricted search operations for fine-grained objectives.

6.2.4 The Criminal Module

The layered approach, combined with the set of dynamic operations to manage the reasoning-guiding process, enables the solution to start from an empty context and uncover most of the important general video aspects, successfully completing the metadata extraction task. However, as discussed in Section 6.1.1, Section 7.5, and Section 8.2.2, utilising a set of generated non-criminally biased questions results in responses with high normality bias, making the solution unable to detect any criminal activity. The criminal module is a conceptual attachment designed to act as the classification tool, providing a more focused scope on criminal detections and intents. Similar to the classification layer in deep learning classifiers, the criminal module utilises the already pre-processed input results to determine the final answer. In this case, the module makes use of the previously discovered context in order to execute its anomaly detection and classification task. This module can be implemented in various ways; one of the most prominent tested modules added a new final exploration layer that divided all the UCF-Crime crimes into their own branches and performed the optimised sets of questions discovered with M-VERA (discussed in Section 5.2). After answering all the questions posed, the classification module selected the most prominent criminal category based on the collection of summaries produced with each layer's context. With this addition, the solution is not only comparable to the previously developed approaches in terms of anomaly detection and classification but also capable of performing metadata extraction, creating a rich representation of the depicted scenario.

6.2.5 Final Overview

The SLG method utilises the three core agents defined in the extended Q-A framework (Section 6.1) to perform a complete exploration of the surveillance video environment. This exploration is divided into two stages: an initial investigation of non-criminal aspects, where the core agents collaborate to extract meaningful information about each of the layers defined in Section 6.2.2; and a second stage in which the criminal module concept is applied. In our case, the criminal module adds a set of additional exploration layers, each concerning one specific criminal category and employing the already produced sets of optimised questions (see Section 5.2). The interaction cycle between the core agents starts with the Controller agent receiving the reasoning-graph status, which specifies the ongoing layer and sublayer exploration objectives, and the current node's context (i.e., which aspects have already been extracted in that specific reasoning path). Based on this status, one of the defined operations is selected (see Section 6.2.3) and processed by the question-answering dialogue agents. The resulting question-answer pair is then used to update the current state of the reasoning graph, and this cycle repeats until every node belonging to the non-criminal exploration layers has been fully explored. Once the first-stage exploration is complete, the agents behave differently: mimicking the static solution approaches (Section 5.2), the Controller is now constrained to a single operation that

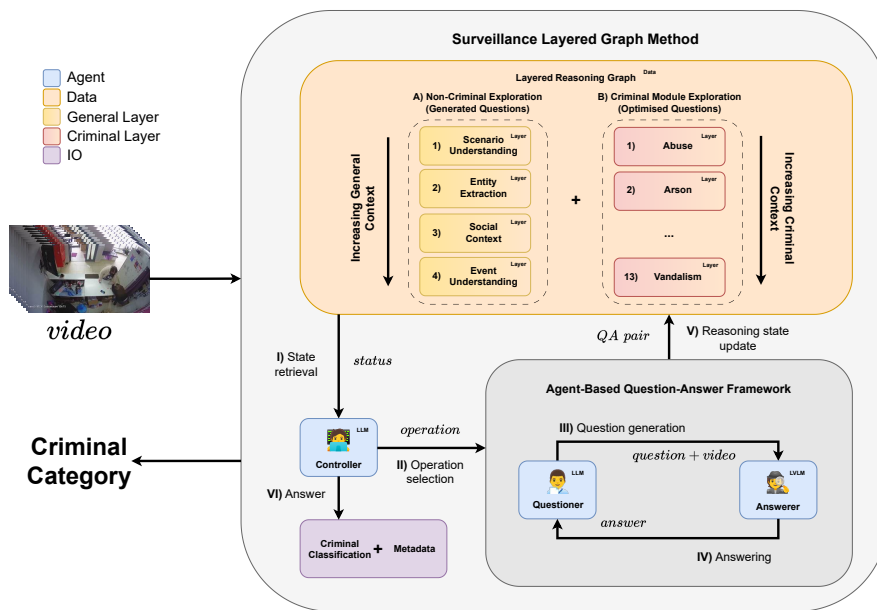


Figure 6.6: Surveillance Layered Graph interaction diagram.

instructs the dialogue agents to address the predefined M-VERA optimised question sets. Upon completion of this second stage of exploration, the solution has sufficient knowledge about criminal and non-criminal aspects to build a structured description comprising all the extracted content. It then uses this description to make an informed prediction of the final criminal classification (Figure 6.6, prompts available in Appendix A.4).

6.2.6 Goal and Intended Outcome

Initially, static and dynamic solutions were aimed at performing criminal detection and classification tasks, reducing their need for model reasoning and overall applicability in comparison with other Artificial Intelligence (AI) approaches, such as surveillance-specific trained classifiers, which directly map visual data into a final answer without requiring the whole time-consuming questioning process. However, this thesis focuses on developing a tool capable of creating rich representations of the surveillance video's depicted content, requiring an exploration strategy capable of debating meaningful topics to extract information. Afterwards, the reasoning baselines addressed the metadata extraction challenges more suitably but lacked efficient criminal exploration due to their flexibility and absence of criminal bias. Our final approach, SLG, aims to merge the advantages of the previous proposed methods while reducing their fragilities. It focuses on having a solid architecture capable of performing an initial reasoning-based scenario exploration, creating value in metadata extraction and providing an in-depth contextual knowledge of the scene that is then used by a criminal-restricted module capable of detecting and classifying crimes. While this represents a substantial leap compared with previous proposed methods, it is not an arbitrary design but rather a composed solution derived directly from the qualitative and quantitative insights obtained in the previous

evaluation stages described in Chapter 7.

EXPERIMENTAL EVALUATION

In this chapter, we discuss the results of applying the proposed methods for performing metadata extraction through iterative dialogue exploration (Chapters 5 and 6) to the surveillance-specific tasks defined in Section 4.3.4. We start by describing the datasets and evaluation metrics used, then sequentially evaluate each proposed method and select its best variants to compete with the subsequent solutions. Finally, we evaluate the impact of resolution on the final proposed approach.

7.1 Evaluation Protocol

This section covers the datasets and metrics used throughout the evaluation process. It begins by presenting the datasets and their developed variants and follows with an overview of the metrics and their relative importance, considering the main project use case.

7.1.1 Datasets

In order to evaluate how the developed solutions perform, we require a dataset that suits the evaluation tasks (described in Section 4.3.4). Among the surveillance-oriented options presented in Section 3.3.1, *UCA* [59] represents the most appropriate choice, as it includes not only a wide collection of criminal videos classified into 14 labels, but also event-level descriptions, which are useful for comparing model exploration capabilities within the metadata extraction domain. Taking this into consideration, we utilised the base *UCA* dataset, as well as some derived variants, to test our approaches.

7.1.1.1 UCA

UCA, which extends UCF-Crime with 23,000 event-level descriptions, has its class counts, video duration, and resolution distributed as shown in Table 7.1. Overall, the dataset is roughly balanced between *Normal* and *Abnormal* examples, containing a total of 944 videos with anomalies and another 910 without. For our evaluation scenarios, we will utilise the defined test set, containing a total of 206 criminal videos and 104 normal ones.

Table 7.1: Class distribution, average video duration, and resolution across the training, test, and validation splits.

Class	Train	Test	Val	Total
Abuse	36	6	8	50
Arrest	42	7	1	50
Arson	18	19	13	50
Assault	25	14	9	48
Burglary	64	23	13	100
Explosion	7	23	20	50
Fighting	37	6	7	50
RoadAccidents	106	35	7	148
Robbery	84	28	37	149
Shooting	31	12	7	50
Shoplifting	30	8	12	50
Stealing	66	13	21	100
Vandalism	30	12	7	49
Normal	589	104	217	910
Total Videos	1165	310	379	1854
Avg. Duration (s)	233.40	292.99	201.12	236.77
Avg. Resolution	320×240			

This selection takes into account our final objective: we aim to firstly extract metadata from surveillance footage but, to some extent, also classify the different types of crimes; having more examples of criminal videos provides more robust testing. Additionally, while maintaining the average resolution of 320×240px, the defined test set presents the highest average video duration, again ensuring more consistency in longer surveillance videos, which is important to replicate as much as possible the expected untrimmed video slots that a deployed solution would have to deal with.

UCA-25 (representative test subset). While solutions relied on a single set of generated questions per class, we opted to test using the whole *UCA* test set, allowing a clear identification of baseline and initial static variants’ trade-offs and limitations (explored in depth in Section 7.2 and Section 7.3). However, with the introduction of dynamic approaches, testing times exceeded our budgeted expectations, requiring significantly more time to run and, as a consequence of the increased architectural complexity, also to ensure quality quantitatively through example analysis. We therefore sampled a representative subset comprising 25% of the original test set (which we named *UCA-25*). To ensure representativeness in class and duration distribution, we produced Figure 7.1, which displays the similarities between the sampled set and the original test set. With this new sampled set, we balanced reasonable testing times for the upcoming complex methods of solutions while ensuring representativeness over the different classes and video durations.

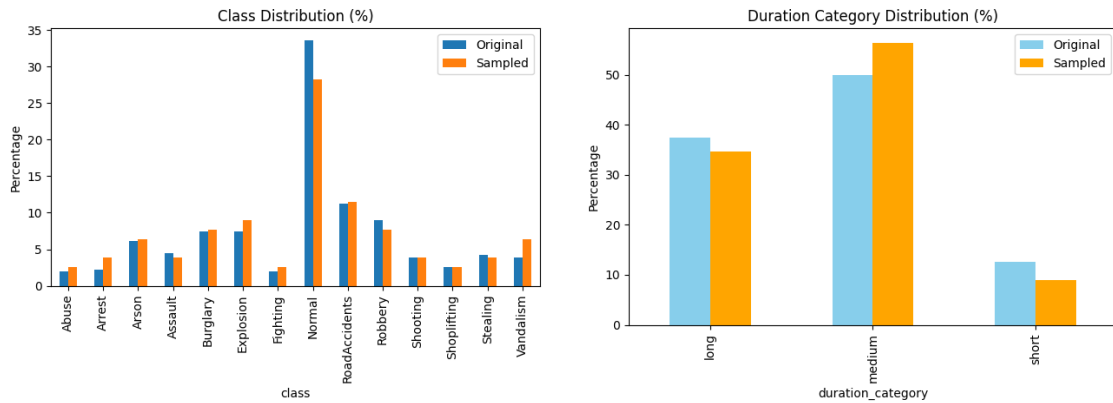


Figure 7.1: Comparison between original UCA test set and the sampled version UCA-25. **Left:** class distribution expressed as percentages. **Right:** video duration distribution expressed as percentages, with videos categorized as short (< 30s), medium (30–60s), and long (> 60s).

7.1.1.2 BUCA

Taking into account *UCA*'s data limitations (explored in detail in Section 8.2.1), we gathered a small subset of pertinent surveillance videos. This collection was built using native UCF-Crime footage while adding videos from the website itemfix [18], aiming to extract the best examples already contained in the base data source while also adding new videos from the online source. During this selection process, the development team focused on four main aspects:

- **Single-Label Correctness:** Videos contained a single crime; for example, scenarios where a fight is followed by an arrest were excluded.
- **Absence of Hints:** Videos had no overlaid text explaining what is depicted in the footage, as this provides unfair additional context. This is commonly found in *UCA*-native examples, such as TV-based footage or edited videos.
- **Criminal Clarity:** Videos explicitly demonstrated the criminal actions, without substantial environmental clutter that obstructs the view during key interactions.
- **Resolution:** We ensured the highest possible quality by selecting *UCA* original footage based on criminal perception and preferring higher-resolution options from the online source.

Better UCA (BUCA) has its average duration, average resolution, and class and respective source counts shown in Table 7.2.

Resolution variants (BUCA-HR and BUCA-LR). To measure the impact of resolution in the final approaches, we stored two variants of BUCA that share the exact same videos while differing in their resolution. BUCA High Resolution (BUCA-HR) comprises the

Table 7.2: Comparison of class distributions, average durations, and resolutions for online vs. UCA video sources.

Class	Online Source	Original Dataset	Total
Abuse	1	3	4
Arrest	2	3	5
Arson	2	3	5
Assault	1	3	4
Burglary	1	4	5
Explosion	2	3	5
Fighting	2	3	5
RoadAccidents	2	3	5
Robbery	1	3	4
Shooting	1	3	4
Shoplifting	2	3	5
Stealing	1	3	4
Vandalism	2	3	5
Normal	2	3	5
Total Videos	22	43	65
Avg. Duration (s)	90.99	118.19	108.99
Avg. Resolution	1343×769	320×240	666×419

Table 7.3: Average resolutions in BUCA-HR and BUCA-LR.

Subset	Average Resolution		
	Native	Online	Total
BUCA-HR	320×240	1343×769	666×419
BUCA-LR	320×240	320×186	320×222

curated clips with their native resolution. BUCA Low Resolution (BUCA-LR) is obtained by downscaling each BUCA video to a standard width of 320 pixels (to match the *UCA* native resolution), while preserving the aspect ratio. The differences in average resolutions are displayed in Table 7.3.

7.1.2 Metrics

The metrics selected to evaluate the tasks defined in Section 4.3.4 fundamentally depend on the surveillance domain, our project use cases, and dataset imbalance. We are especially focused on evaluating the correctness of binary selection (Abnormal vs Normal and Belongs vs Not Belongs) and multiple-choice selection (Multiclass Classification with and without Normal).

Accuracy. Providing a common and intuitive comparison metric, comparing models based on both true positive and true negative examples.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \quad (7.1)$$

However, this base accuracy formula has well-known limitations in imbalanced scenarios; for instance, when one class has substantially more examples than the other, a system that blindly predicts the majority class can appear to perform well. Given that our data exhibits imbalance in Abnormal vs Normal, Belongs vs Not Belongs, and especially in Multiclass Classification (where Normal has many more examples than any single crime class), we therefore use this balanced version of accuracy to evaluate these tasks.

$$Accuracy = \frac{TP}{(TP + FN)} \times 0.5 + \frac{TN}{(TN + FP)} \times 0.5. \quad (7.2)$$

With this balanced formula, the calculated accuracy ensures that the total number of positive examples contributes the same as the total number of negative ones, regardless of their actual counts.

Precision. Measuring how often the predicted crimes are correct, representing an important metric for evaluating cases where the solution thought that a crime occurred.

$$Precision = \frac{TP}{TP + FP}. \quad (7.3)$$

Recall. A critical metric in anomaly detection and classification, as identifying scenarios where the model was expected to detect crime but did not is key to understanding the main architectural fragilities and improving current techniques.

$$Recall = \frac{TP}{TP + FN}. \quad (7.4)$$

F1-score. The harmonic mean of Precision and Recall, representing a fair balance between the capability of detecting true examples and not missing positive ones. It is the main comparison metric used in the upcoming results, as it matches our use case expectations, ensuring the solution is capable of detecting criminal scenarios.

$$F1Score = 2 * \frac{Precision \times Recall}{Precision + Recall}. \quad (7.5)$$

During evaluation, metrics received different relative importance aligned with the project's use case. We are mostly interested in detecting crime, meaning that true negative examples are not as important as true positive and false negative ones, since these represent correctly detecting a crime and mistakenly classifying an anomalous video as normal, respectively. Taking this into account, Accuracy is less important than the other three metrics because it benefits from true negatives. F1-score is the primary metric for comparative analysis, given its ability to balance recall and precision; it is followed by Recall, as underestimating a crime (a false negative) is more relevant than overestimating a normal video (a false positive), then Precision, and finally ending with Accuracy. While we

have this clear preference, all metrics are reported throughout the evaluation, providing a transparent view of the trade-offs in each approach. Additionally, all the presented metrics were initially calculated per class and then averaged with class-frequency weights.

7.2 Baseline Performance

The baseline testing and evaluation provides not only key insights into major model capabilities and their limitations, but also sets a numerical reference for baseline performance, defining the project’s performance goal to surpass. This section covers the baseline testing setup, produced results, and respective insights.

7.2.1 Setup

In this baseline evaluation, we used the complete *UCA* test set to evaluate the two defined video-model baselines, *Qwen2.5-VL* and *LLaVA-Video*, on the binary tasks (Abnormal vs Normal; Belongs vs Not Belongs) and the multiclass tasks (Multiclass Classification with and without Normal). This evaluation measures performance when directly inquiring the baseline models with the target task (explained in detail in Section 5).

7.2.2 Results

As Tables 7.4 and 7.5 show, *Qwen2.5-VL* holds the majority of the highest values across all evaluated metrics, only losing to *LLaVA-Video* on the Multiclass Classification with Normal task. While the highest F1-scores achieved in the binary classification tasks are not particularly high, the multiclass tasks present substantially better results than completely random approaches, emphasising the baseline strength and capabilities for more complex collaborative approaches.

Table 7.4: Baseline performance on binary tasks (Abnormal vs Normal; Belongs vs Not Belongs) on the complete *UCA* test set. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column.

Model	Abormal vs Normal				Belongs vs Not Belongs			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	42.47	32.58	66.45	65.16	43.61	70.65	34.58	71.08
LLaVA-Video	36.98	27.10	66.45	60.32	39.08	69.68	28.60	65.98

To better visualise per-class recall, we compare the baselines (Figure 7.2). *Qwen2.5-VL* clearly presents advantages, as it loses to *LLaVA-Video* only in *Burglary*, *RoadAccidents*, and *Fighting*, while winning in *Stealing*, *Robbery*, *Abuse*, *Arrest*, *Assault*, and *Shoplifting*.

Table 7.5: Baseline performance on multiclass tasks (with and without Normal) on the complete UCA test set. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column.

Model	Multiclass w/ Nor.				Multiclass w/o Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	48.87	55.16	52.93	71.19	43.40	47.09	48.58	70.85
LLaVA-Video	48.96	53.54	54.89	69.42	42.50	42.23	48.48	68.62

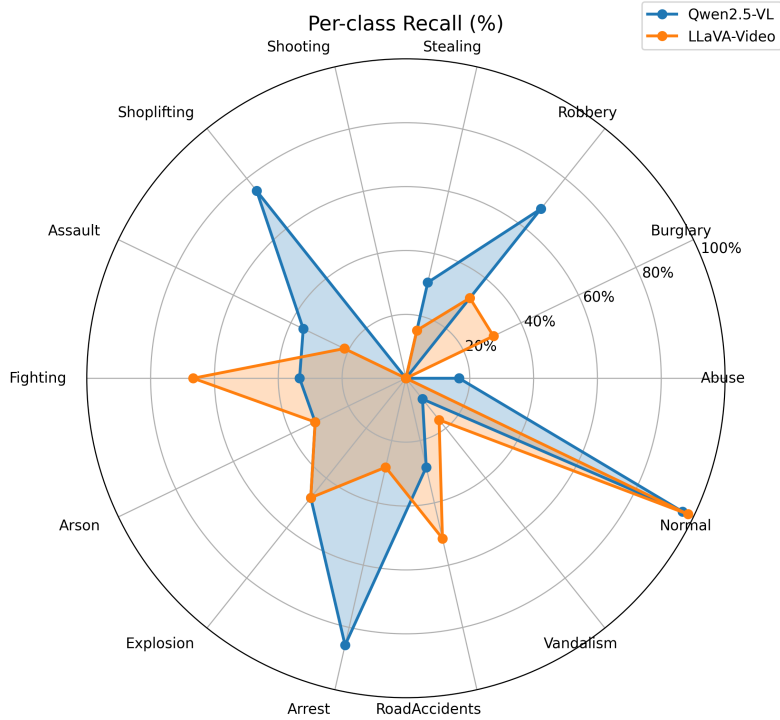


Figure 7.2: Comparison of per-class recall (%) between the baselines.

7.2.3 Conclusion and Insights

The *Qwen2.5-VL* baseline demonstrates greater base capabilities than the *LLaVA-Video* baseline across most of the defined tasks. While this is an interesting insight, it still requires further testing to verify its increased adaptability and to study possible multi-agent collaborations between the two video baselines.

7.3 Static Approaches Performance

Static approaches represent the simplest extension of the video-model baseline, employing a fixed set of questions to decide the final output. We further test the adaptability of *Qwen2.5-VL* and *LLaVA-Video* to different strategies. This section is organised into three phases: Phase A evaluates the initial static variants on *UCA*; Phase B investigates whether increasing the number of generated questions improves results; and Phase C analyses the

final variant, M-VERA, on the UCA-25 subset.

7.3.1 Phase A - Static Variants on UCA

To initially evaluate the first three developed static approaches, we utilised the complete UCA test set. These approaches utilise a collection of predefined questions obtained in varying ways: GenQ (Section 5.2) utilises Qwen2.5 to generate a set of five questions per criminal class; BiasQ (Section 5.2) utilises a set of handwritten questions designed by the development team, aiming to make a clear separation between the most similar classes; finally, Coop-GenQ and Coop-BiasQ (Section 5.2) analyse answer consistency between the two tested baselines to filter out non-concordant question-answer pairs before the final classification (explained in detail in Section 5.2).

7.3.2 Phase A - Results

As Tables 7.6 and 7.7 show, *Qwen2.5-VL* variants hold the majority of the highest F1-score values across all the defined tasks, with an increased performance gap in the multiclass tasks, where *LLaVA-Video* variants performed very poorly compared to their baseline evaluation results. Additionally, the concordant variants did not surpass single-model performance, highlighting that single-model fragilities compromise the performance of the cooperative solution. Due to the discovered fragilities in *LLaVA-Video* variants, we opted to exclude it as a possible baseline in our testing procedures, creating more room for discovery and testing with upcoming *Qwen2.5-VL* approaches and respective variants. **Bold values indicate the highest values in each column, while underlined values indicate the best scores within each section;** i.e., if two comparable solutions appear within a multi-row section, the highest metric values in that block are underlined (if they are not already bold).

Table 7.6: Static approaches performance on binary tasks (Abnormal vs Normal; Belongs vs Not Belongs) on the complete UCA test set. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Belongs vs Not Belongs			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	<u>42.47</u>	32.58	66.45	<u>65.16</u>	43.61	70.65	<u>34.58</u>	71.08
LLaVA-Video	36.98	27.10	66.45	60.32	39.08	69.68	28.60	65.98
Qwen2.5-VL-GenQ	43.41	34.19	66.45	60.00	<u>38.49</u>	44.19	42.71	<u>65.24</u>
LLaVA-Video-GenQ	<u>49.38</u>	<u>44.52</u>	64.52	78.06	32.72	<u>51.94</u>	35.40	62.13
Qwen2.5-VL-BiasQ	<u>48.82</u>	<u>40.97</u>	66.45	<u>71.94</u>	<u>36.23</u>	<u>69.35</u>	<u>27.97</u>	<u>66.11</u>
LLaVA-Video-BiasQ	59.54	55.80	66.45	55.80	13.63	39.03	10.98	56.25
Coop-GenQ	35.88	27.74	64.52	61.29	33.63	48.71	40.98	61.96
Coop-BiasQ	32.19	23.55	66.45	55.48	31.05	58.71	25.60	57.72

Table 7.7: Static approaches performance on multiclass tasks (with and without Normal) on the complete *UCA* test set. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Multiclass w/ Nor.				Multiclass w/o Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	48.87	55.16	52.93	71.19	43.40	47.09	<u>48.58</u>	70.85
LLaVA-Video	48.96	53.54	54.89	69.42	42.50	42.23	48.48	68.62
Qwen2.5-VL-GenQ	<u>44.44</u>	<u>47.74</u>	<u>52.03</u>	<u>67.33</u>	<u>34.75</u>	<u>30.58</u>	<u>49.00</u>	<u>63.70</u>
LLaVA-Video-GenQ	<u>18.43</u>	<u>18.39</u>	<u>41.92</u>	<u>56.26</u>	<u>11.18</u>	<u>14.56</u>	<u>20.95</u>	<u>53.22</u>
Qwen2.5-VL-BiasQ	<u>41.46</u>	<u>40.97</u>	<u>51.55</u>	<u>66.17</u>	<u>37.82</u>	<u>36.41</u>	52.19	<u>65.65</u>
LLaVA-Video-BiasQ	<u>7.37</u>	<u>12.26</u>	<u>7.83</u>	<u>52.74</u>	<u>12.51</u>	<u>18.45</u>	<u>14.55</u>	<u>53.93</u>
Coop-GenQ	39.25	42.90	46.25	65.58	22.63	19.90	40.11	57.84
Coop-BiasQ	28.09	27.42	45.98	58.86	21.53	22.33	41.26	57.77

It is also clear, as displayed in Figure 7.3, that the best-tested approaches all reveal a significant normality bias, identified by the intensively filled predicted-normal column, meaning that for most criminal scenarios the developed solutions struggle to identify criminal acts, thereby classifying them as *Normal*. Additionally, the solution with the least normality bias is currently Qwen2.5-VL-GenQ (right-side matrix), as it presents, on average, the lowest values in the predicted-normal column. While this is a great sign toward fighting the normality bias inherent to LVLMS, it does not directly translate into better classification in the multiclass task, as the Baseline still holds the highest metric scores. This shows that, even though these new approaches clearly classify more videos as crime, they are still incapable of identifying the correct criminal class.

7.3.3 Phase B - Question Count Ablation on UCA-25

To test whether generating more questions would significantly impact the final results, we used the UCA-25 subset (explained in Section 7.1.1.1). In this ablation, we evaluated different *Qwen2.5-VL* variants that generated between 1 and 30 questions per criminal class in the binary and multiclass tasks.

7.3.4 Phase B - Results

As Tables 7.8 and 7.9 show, the most prominent number of generated questions lies between 10 and 30, as this interval contains the majority of the highest metric values (considering only the ablation variants). Additionally, while the Belongs vs Not Belongs and Multiclass without Normal tasks were initially developed to understand class confusion and the impact of normality bias, these tasks were redundant, providing basically the same insights as the other tasks and were only loosely aligned with the project scope and standard anomaly detection and classification. Consequently, we opted to focus on the

7.3. STATIC APPROACHES PERFORMANCE

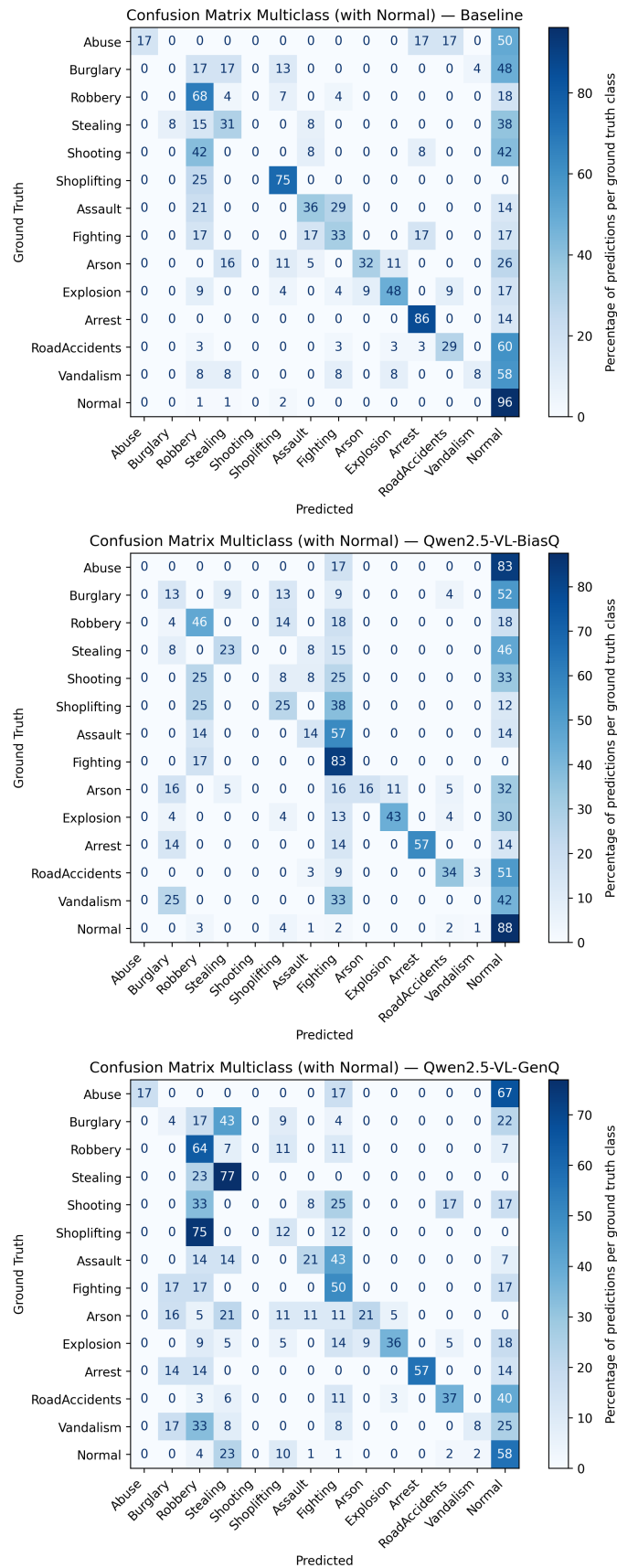


Figure 7.3: Multiclass confusion matrices (row-normalized in %) **Top:** Qwen2.5-VL baseline confusion matrix. **Middle:** Qwen2.5-VL-BiasQ confusion matrix. **Bottom:** Qwen2.5-VL-GenQ confusion matrix.

Table 7.8: Ablation on the number of generated questions per class for Qwen2.5-VL on the UCA-25 subset (see Section 7.1.1.1), evaluated on binary tasks (Abnormal vs Normal; Belongs vs Not Belongs). Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Belongs vs Not Belongs			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	40.28	32.05	62.82	58.97	40.76	67.95	31.92	71.24
GenQ1	31.22	23.08	55.13	51.28	31.74	50.00	27.38	60.68
GenQ3	33.01	25.64	55.13	53.85	30.94	47.44	29.43	61.16
GenQ5	44.11	35.90	65.38	55.13	31.12	38.46	36.17	61.56
GenQ10	49.43	41.03	65.38	64.10	<u>38.18</u>	47.44	34.99	<u>67.85</u>
GenQ20	50.65	43.59	69.23	67.95	35.89	48.72	34.39	64.86
GenQ30	47.13	39.74	62.82	67.95	37.33	<u>53.85</u>	35.37	66.01

Table 7.9: Ablation over the number of generated questions for multiclass classification (with and without the Normal class) on the UCA-25 subset. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Multiclass w/ Nor.				Multiclass w/o Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	48.33	56.41	52.14	72.47	46.28	50.00	55.27	72.90
GenQ1	32.32	39.74	35.82	62.05	22.51	19.64	35.54	58.11
GenQ3	39.46	<u>46.15</u>	43.72	66.33	31.04	30.36	33.66	63.43
GenQ5	38.40	42.31	45.19	64.26	32.87	30.36	43.49	63.76
GenQ10	<u>40.25</u>	44.87	43.52	<u>67.61</u>	30.68	28.57	<u>52.04</u>	62.45
GenQ20	37.39	43.59	<u>46.53</u>	64.68	33.99	32.14	42.30	64.66
GenQ30	33.96	41.03	<u>44.00</u>	63.10	<u>38.59</u>	<u>35.71</u>	49.82	<u>66.45</u>

Abnormal vs Normal and Multiclass with Normal tasks only.

7.3.5 Phase C - M-VERA-Optimized Question Sets

Motivated by the upcoming findings in Section 7.4, we later introduced a new variant developed around the VERA’s project (described in Section 5.2), aiming to test its performance and usability in comparison with our predefined sets of questions. We utilised M-VERA to optimise sets of pertinent questions for each individual criminal class. The summarised results for the static approach solutions and the M-VERA variant are displayed in Table 7.10. Even though M-VERA’s approach does not surpass any of the previous highest metric results, it still achieves the most prominent metric values within the static approaches for the multiclass task. Upon deeper analysis, we concluded that M-VERA’s success is mainly due to inherent model prompt sensitivity (further explored in Section 7.4). The biased and generated question approaches focus heavily on the criminal

Table 7.10: Baseline and prominent static solutions on Abnormal vs Normal and Multiclass (with Normal) tasks on the UCA-25 subset. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Multiclass w/ Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	40.28	32.05	62.82	58.97	48.33	56.41	52.14	72.47
GenQ20	50.65	43.59	69.23	67.95	37.39	43.59	46.53	64.68
BiasQ	51.24	44.87	65.38	70.51	29.32	30.77	35.58	61.30
M-VERA	50.32	42.31	65.38	65.38	<u>40.38</u>	<u>46.15</u>	<u>48.39</u>	<u>66.35</u>

class descriptions, becoming more subject to the use of aggressive terms, whereas VERA optimises the wording based on the model’s output, developing very interesting questions that apply only a soft criminal bias to the prompts. The full optimised question sets are provided in Appendix B.

To better visualise per-class recall, we compare the baseline with the best-performing static approaches (Figure 7.4). Although M-VERA achieves the best average F1-score among the static approaches in the multiclass setting, it surpasses the baseline only in *Burglary* and *RoadAccidents*. BiasQ, which achieves the best performance in the binary setting, surpasses the baseline only in *Burglary* and *Assault*.

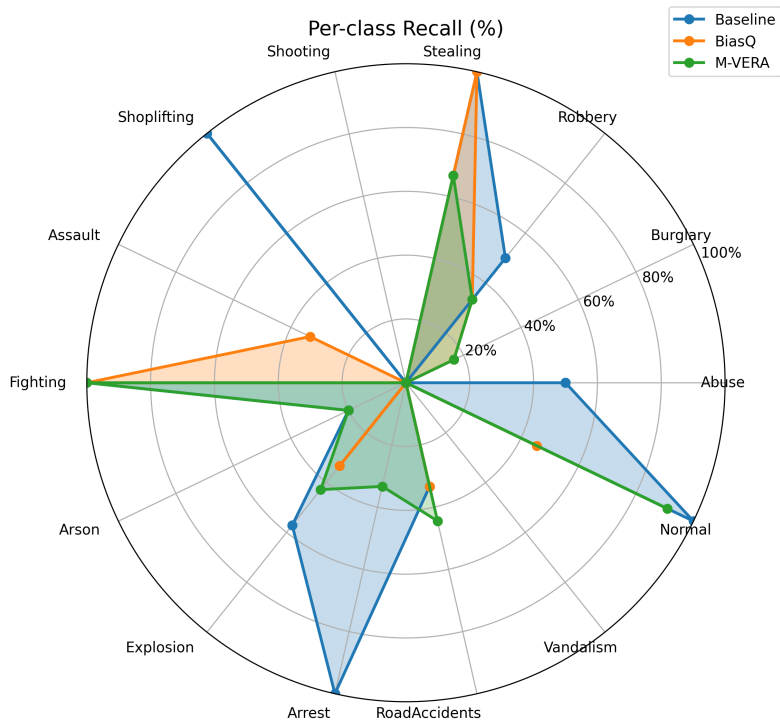


Figure 7.4: Comparison of per-class recall (%) between the baseline and the best static approaches.

7.3.6 Conclusion and Insights

Static approaches provide a simple and effective performance boost on Abnormal vs Normal, but fail to surpass the baseline on Multiclass classification. GenQ20 surpasses the previous anomaly detection record with an increase of 10.37 average F1-score, while M-VERA holds the best multiclass metric scores across the developed static solutions. Results reveal persistent normality bias, struggle to identify the correct criminal class and strong sensitivity to prompts (further explored in Section 7.4). An ablation study reveals that generating between 10 to 30 questions enables the best performance in this basic criminal-description-based-generative approaches. Based on the project scope and domain alignment, we decided to drop Belongs vs Not Belongs and Multiclass without Normal. Due to the poor results with *LLaVA-Video* variants in multiclass evaluation tasks, we opted for excluding it from the tested baselines, carrying on with only *Qwen2.5-VL* variants.

7.4 Dynamic Approaches Performance

Aiming to simultaneously surpass baseline performance in both Abnormal vs Normal and Multiclass Classification, dynamic approaches start their classification process by obtaining feedback from the LVLM on which crimes are likely to be depicted in the current video (simulating how the baseline works), aiming to refine the selection by iteratively reducing the number of probable classes using the obtained context. This section covers the setup, results, and main limitations found, and concludes with insights drawn from the dynamic approaches.

7.4.1 Setup

The approaches were tested in UCA-25 subset and vary in terms of K , the number of initial probable classes; Q , the number of questions generated per pertinent class; and whether the approach utilises a general description of the scenario, produced initially by the LVLM, when generating the questions for each class. The names of the experiments directly align with these terms: K5Q1-ND denotes a dynamic solution where the initial number of probable classes is five, the number of questions generated per probable class is one, and it does not utilise a description (ND = no description), while K5Q1-D differs only by using the description when generating the questions. More detailed information about the implementation can be found in Section 5.3.

7.4.2 Results

As Table 7.11 shows, K10Q5-ND achieved an all-time high average F1-score on the Abnormal vs Normal task, surpassing the baseline by 21 p.p., and K5Q1-ND surpasses the previously defined static solutions in the multiclass setting. Given that the dynamic

pipeline begins with a baseline-derived answer that is then refined with additional context, we expected to perform at least as well as the baseline in the multiclass setting. Instead, we observed the opposite: settings that used fewer contextual resources (fewer questions and no description) actually registered higher performance scores, while changing the starting number of probable classes had inconsistent effects across the developed variants. To understand these discrepancies, we conducted an example-level error analysis to address our main doubts:

Why larger K did not enhanced the results? We initially expected solutions that used a larger $K = 10$ to perform at least as well as their smaller- K counterparts. However, upon conducting an in-depth analysis of cases misclassified by $K = 10$ solutions but correctly classified by the corresponding counterparts, we found that larger- K approaches often either (i) failed to include the correct class at the beginning or (ii) inadvertently pruned it during execution due to criminal confusion, single-crime data inconsistencies, or occasional model misbehaviour (e.g., performing multiple reduction steps at once or maintaining the exact same set of classes).

Regarding (i), we hypothesise that allowing the baseline to select a larger number of classes introduces meaningful noise that could lead to selection biases. For (ii), we confirmed that criminal confusion is a base video-model limitation, as these models struggle to distinguish between similar criminal classes; single-crime data inconsistencies are a *UCA* dataset limitation, as videos often depict more than one criminal action (e.g., an individual steals from a shop, ends up in a fight with the owner, and is later arrested); and model misbehaviour generally forced the algorithm to make a non-contextual, deterministic selection among the remaining classes. More details on model and dataset limitations can be found in Section 8.2.

Why adding context degraded the results? We initially thought that providing more context would logically enhance the results, as the generated questions would not be based only on criminal descriptions but also on environmental contextual details. During the iterative process, the solution obtains context from two sources: a base scenario description, which provides details about the depicted scene and aims to generate more context-dependent questions; and question-answer pairs, which provide additional information before the reduction process. However, we identified that generating questions with additional context actually degraded the results by increasing the normality bias, softening the terms used. As an example of how these tiny adjustments can make an enormous difference in performance, we present a real scenario (Figure 7.5).

This demonstrated sensitivity is further corroborated by M-VERA’s success in the static approaches, as it takes advantage of these inherent textual aspects to achieve better performance in detecting crimes.

Table 7.11: Dynamic variants and prominent previous solutions on Abnormal vs Normal and Multiclass (with Normal) tasks on the UCA-25 subset. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Multiclass w/ Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	40.28	32.05	62.82	58.97	48.33	56.41	52.14	72.47
GenQ20	50.65	43.59	<u>69.23</u>	67.95	37.39	43.59	46.53	64.68
BiasQ	<u>51.24</u>	<u>44.87</u>	<u>65.38</u>	<u>70.51</u>	29.32	30.77	35.58	61.30
M-VERA	<u>50.32</u>	<u>42.31</u>	65.38	65.38	<u>40.38</u>	<u>46.15</u>	<u>48.39</u>	<u>66.35</u>
K5Q1-ND	48.16	41.03	65.38	69.23	<u>42.46</u>	<u>51.28</u>	45.08	<u>69.04</u>
K5Q1-D	38.74	30.77	62.82	58.97	30.69	38.46	36.70	60.63
K5Q3-ND	49.06	41.03	65.38	67.95	39.62	46.15	40.90	66.36
K5Q3-D	42.68	34.62	62.82	62.82	33.94	42.31	38.45	63.24
K5Q5-ND	46.55	38.46	65.38	66.67	40.59	47.44	41.39	66.56
K5Q5-D	44.46	35.90	65.38	64.10	39.10	46.15	43.65	65.53
K10Q1-ND	61.49	55.13	71.79	75.64	41.77	44.87	<u>49.55</u>	68.26
K10Q1-D	57.83	51.28	69.23	73.08	37.75	41.03	43.08	65.55
K10Q3-ND	60.33	55.13	69.23	76.92	34.89	38.46	39.65	64.73
K10Q3-D	46.06	38.46	62.82	66.67	36.17	43.59	40.49	64.61
K10Q5-ND	61.46	56.41	69.23	78.21	41.85	43.59	46.35	67.54
K10Q5-D	58.18	52.56	69.23	74.36	37.07	41.03	43.12	65.94

Fighting Scenario Answer Inconsistency

Category: Fighting

This example illustrates clear contradictory answers induced by slight prompt modifications.

Solution with additional context

Generated Question: Is there a scene in the video where two or more individuals are **physically attacking** each other?

Generated Answer: **No**, there is no scene in the video where two or more individuals are physically attacking each other.

Solution without additional context

Generated Question: Is there a **physical altercation** between two or more individuals in the video?

Generated Answer: **Yes**, there is a physical altercation between two or more individuals in the video.

Observation: The model drastically changes its final answer under minor wording variations, emphasising a high sensitivity to the words used in the prompt.

Figure 7.5: Answer inconsistency in a *Fighting* scenario.

To better visualise per-class recall, we compare the baseline with the best-performing dynamic approaches (Figure 7.6). Although K5Q1-ND achieves the best average F1-score among the dynamic approaches in the multiclass setting, it surpasses the baseline only in *Stealing*, *Robbery*, and *Explosion*. K10Q1-ND, which achieves the best performance in the binary setting, surpasses the baseline in *Stealing*, *Robbery*, *Burglary*, and *Explosion*.

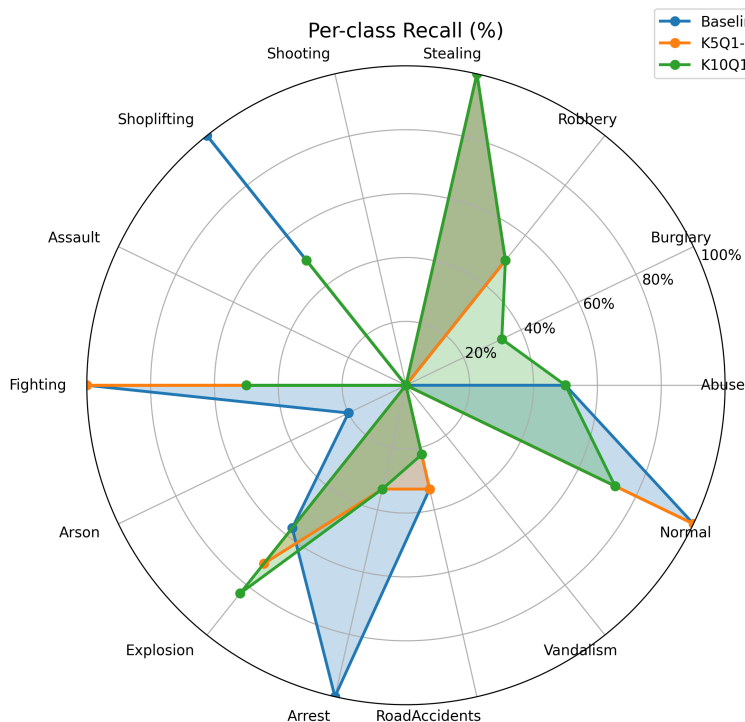


Figure 7.6: Comparison of per-class recall (%) between the baseline and the best dynamic approaches.

7.4.3 Conclusion and Insights

Dynamic approaches provide an effective way to outperform the static methods. K10Q5-ND surpasses the baseline by 21.18 p.p. in the average F1-score for the Abnormal vs Normal task, and K5Q1-ND holds the second-highest average F1-score in the multiclass setting (behind the baseline). While these are favourable improvements, we expected increased gains from variants that utilised additional context and larger initial class sets. Instead, these approaches revealed substantial limitations due to increased normality bias introduced by the additional context, and struggled to distinguish similar criminal categories. Additionally, a deeper analysis revealed intrinsic video-model and dataset

limitations (detailed in Section 8.2).

It is now important to fundamentally understand why the developed solutions are not capable of surpassing the baseline results in the multiclass setting, because there is a chance that the models do not truly understand the overall scenario and instead are simply better adjusted to the dataset’s criminal distributions. Therefore, we aim to test solutions oriented toward surveillance-footage understanding that use this additional context to surpass the previous baseline results.

7.5 Reasoning-Based Approaches Performance

Reasoning-based strategies place more in-depth focus on surveillance video analysis rather than classical anomaly detection and classification, providing a dense contextualisation of the depicted scenario. This section covers their testing setup, respective results, and the conclusions drawn from their analysis.

7.5.1 Setup

The approaches were tested on the UCA-25 subset and aimed to produce a maximum of 70 reasoning steps, balancing the amount of context with the previous $14 \times 5 = 70$ total generated questions in the initial static solutions. The approaches include: CoT (Section 5.4); ToT (Section 5.4), with three different tree compositions that modify the maximum depth M , branching factor B , and number of selected answers K ; IoT (Section 5.4); and LCoT (Section 5.4), with variants of four and six generated layers.

7.5.2 Results

As Table 7.12 shows, the reasoning-based approaches bring no novelty in terms of the highest metric scores, being unable to surpass the dynamic approaches in both binary and multiclass settings. While these advancements are not as meaningful as the previous ones, we need to take into consideration that these solutions enable a much deeper understanding of the non-criminal aspects of the scenario, representing a fair option as context producers for other classification methods. Even though their main advantage is clearly undervalued in these tasks, due to its non-criminal nature and the previously detected normality-bias limitations, the reasoning-based approaches were still capable of maintaining metric scores comparable to those obtained by the static solutions while producing much more valuable context. Considering only the developed reasoning-based approaches, LCoT-LGen6 demonstrates stronger capabilities in the binary task, while ToT-M4B5K1 dominates the multiclass setting.

To demonstrate the increased information-extraction capabilities of the reasoning-based approaches compared with the previous methods, we present a qualitative example showing the summarised information obtained from the question–answer pairs built during the exploration process for the BiasQ, CoT, and ToT-M4B5K1 strategies.

Table 7.12: Reasoning-based approaches and prominent previous solutions on Abnormal vs Normal and Multiclass (with Normal) tasks on the UCA-25 subset. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Multiclass w/ Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	40.28	32.05	62.82	58.97	48.33	56.41	52.14	72.47
GenQ20	50.65	43.59	<u>69.23</u>	67.95	37.39	43.59	46.53	64.68
BiasQ	<u>51.24</u>	<u>44.87</u>	<u>65.38</u>	<u>70.51</u>	29.32	30.77	35.58	61.30
M-VERA	50.32	42.31	65.38	65.38	<u>40.38</u>	<u>46.15</u>	<u>48.39</u>	<u>66.35</u>
K5Q1-ND	48.16	41.03	65.38	69.23	<u>42.46</u>	<u>51.28</u>	45.08	<u>69.04</u>
K10Q1-ND	61.49	55.13	71.79	75.64	41.77	44.87	<u>49.55</u>	68.26
CoT	41.63	34.62	58.97	55.13	34.39	38.46	<u>44.35</u>	60.92
ToT-M4B3K3	50.13	42.86	<u>69.84</u>	61.90	27.49	34.92	28.75	60.88
ToT-M4B5K1	<u>50.56</u>	<u>43.59</u>	<u>66.67</u>	<u>67.95</u>	<u>35.60</u>	<u>43.59</u>	<u>39.60</u>	<u>65.24</u>
ToT-M16B2K1	42.80	33.33	69.23	57.69	30.29	38.46	35.51	60.84
IoT	43.97	34.62	65.38	56.41	27.44	34.62	30.03	58.88
LCoT-LGen4	44.49	35.90	<u>66.67</u>	58.97	29.34	<u>37.18</u>	<u>35.40</u>	60.46
LCoT-LGen6	<u>51.73</u>	<u>44.87</u>	<u>69.23</u>	<u>65.38</u>	<u>29.39</u>	35.90	27.87	<u>61.06</u>

Qualitative Example of Metadata Extraction

The results displayed below are ChatGPT-5-summarised versions of the information obtained by reading all question-answer pairs generated during the respective strategies (prompt available in Appendix A.3). The example demonstrates how reasoning-based approaches extract more meaningful aspects of the video than the previous criminally oriented methods. Highlighted text indicates missing information; red symbolises meaningful information, while yellow indicates less meaningful aspects.

Visual Input: Arrest048 from UCA



Strategy: BiasQ (70 question–answer pairs)

Generated Summary: The video takes place inside a convenience store. Present are two men, one wearing a red shirt, and several police officers in uniform. The store interior appears intact, with no signs of forced entry, theft, vandalism, fire, or structural damage. No children, elderly people, animals, vehicles, or external infrastructure are visible in the footage. The individuals are positioned inside the store, with officers confronting and restraining the men.

The scene shows a sudden and violent physical altercation involving multiple individuals. The man in the red shirt is seen pushing, striking, and struggling with others before being subdued by police officers, who use physical restraint and handcuffing procedures. Several people end up on the ground, showing signs of struggle and possible minor injuries, with defensive postures noted. The altercation is limited to physical contact; no weapons, gunfire, fire, or property damage are present. The incident culminates in police officers restraining and arresting the individuals involved.

Strategy: CoT (4 question–answer pairs)

Generated Summary: The video takes place inside a well-lit convenience store with shelves of products and a tiled floor. Present are a man in a red shirt, a woman in a blue jacket, a man in a blue jacket, a man in a black shirt, two police officers, and several other customers and employees. One officer is visibly armed with a gun. Customers and staff are positioned near the counter or moving away from the confrontation, watching with concern.

The man in the red shirt runs toward the door and back into the store before being confronted by an officer. A physical struggle follows, involving the officer and the man in the red shirt, later joined by the woman in the blue jacket and the man in the black shirt. Additional officers enter to assist, and one person is restrained on the floor. Meanwhile, other individuals remain stationary or withdraw, observing the ongoing restraint and movement within the store.

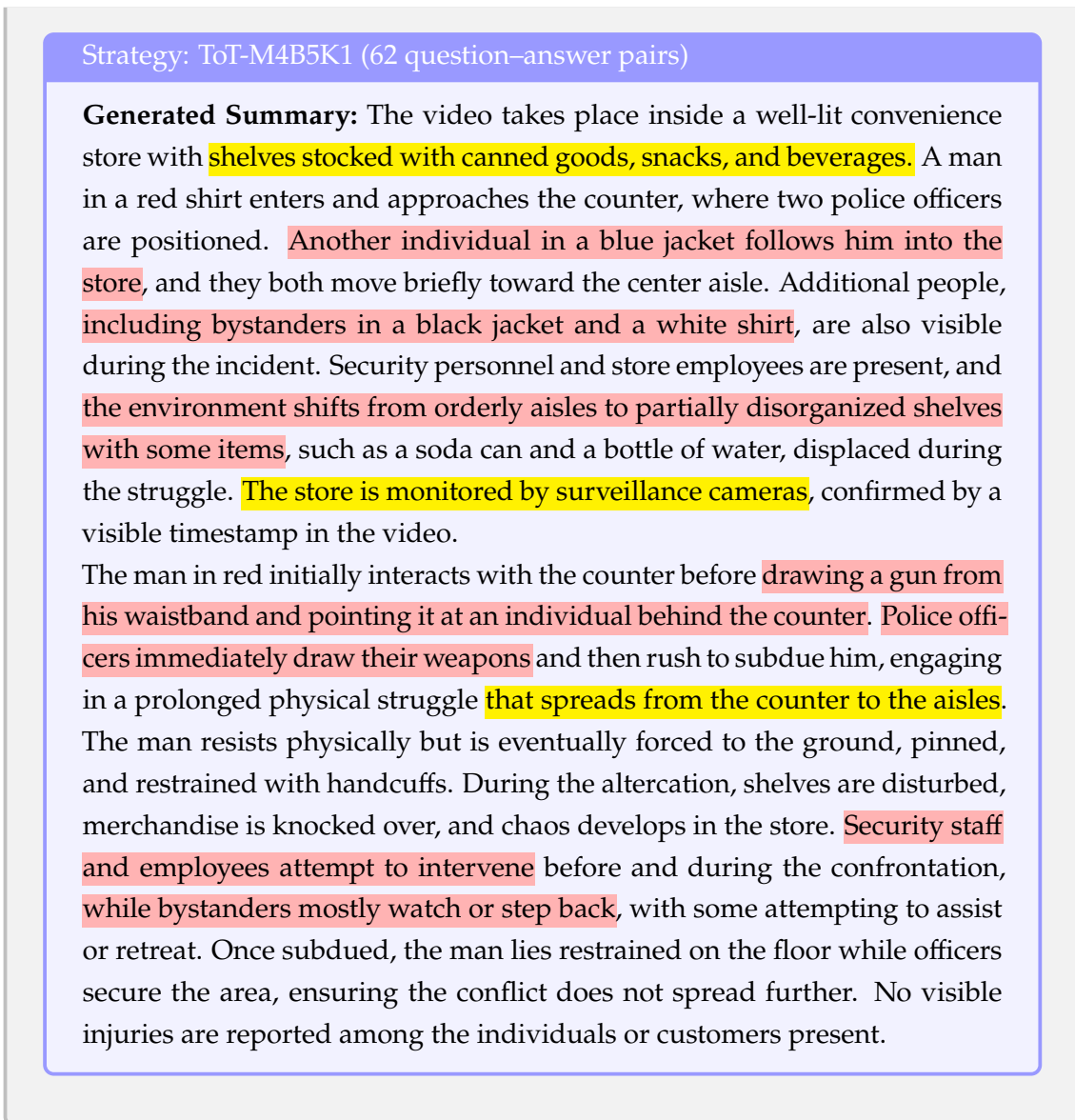


Figure 7.7: Qualitative metadata extraction example for *Arrest048* (UCA).

As shown in Figure 7.7, the reasoning-based approaches provide richer environmental context of the video content, making them more effective for metadata extraction. Additionally, the qualitative example highlights omissions of great importance in red and omissions of lower importance in yellow. The highlights clearly show the first summary with large video-understanding gaps, failing to extract information on central individuals and key interactions. It is also worth noting that the first summary was built from the extraction produced by the complete set of biased questions, containing 70 criminal-oriented questions, while the second summary was built using only four sequential reasoning steps of CoT, and the third with 62 steps of ToT, again emphasizing their performance in metadata extraction.

To better visualise per-class recall, we compare the baseline with the best-performing

reasoning approaches (Figure 7.8). Although ToT-M4B5K1 achieves the best average F1-score among the reasoning approaches in the multiclass setting, it surpasses the baseline only in *RoadAccidents*. LCoT-LGen6, which achieves the best performance in the binary setting (among the reasoning approaches), surpasses the baseline only in *RoadAccidents* and *Shoplifting*.

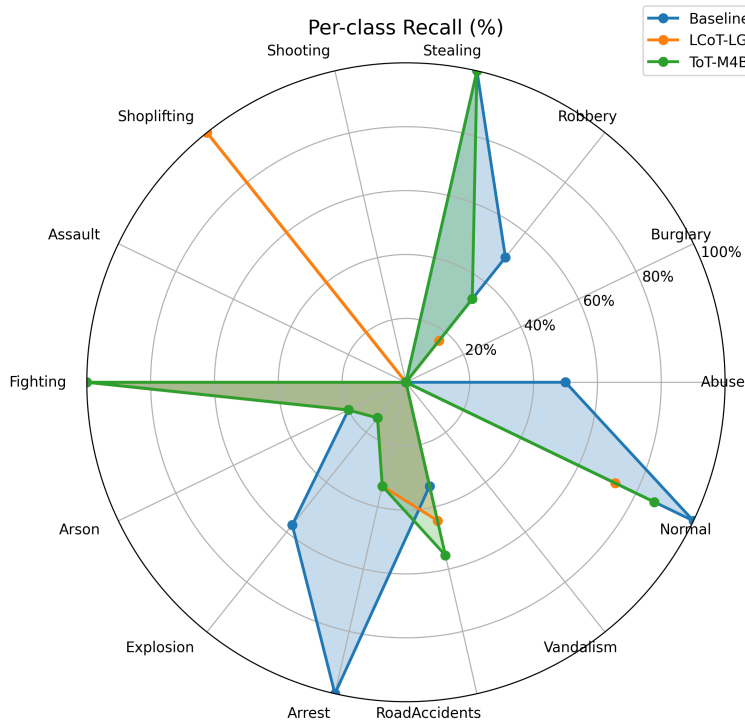


Figure 7.8: Comparison of per-class recall (%) between the baseline and the best reasoning approaches.

7.5.3 Conclusion and Insights

Reasoning-based methods were not able to surpass the most prominent dynamic approaches in the Abnormal vs Normal and Multiclass Classification tasks. However, these newly developed solutions are capable of extracting more valuable context from the depicted scenarios, becoming increasingly important for the final metadata extraction solution and providing key context for the final answer (as highlighted in Figure 7.7). Among the developed reasoning-based variants, LCoT-LGen6 demonstrates stronger capabilities in the binary task, while ToT-M4B5K1 dominates the multiclass setting. While these approaches struggle to correctly detect and classify crimes, they still achieve results comparable to static approaches. Considering these insights, it is clear that, to surpass previous results, we must employ a hybrid pipeline: starting with an extensive non-criminal scenario analysis to extract meaningful information and then proceeding to a criminally oriented inquiry in order to correctly classify the crime.

7.6 Surveillance Layered Graph Reasoning Strategy Performance

Our final reasoning strategy, SLG, mixes concepts from various reasoning-based approaches studied. It starts by performing a non-criminal, layered exploration to extract meaningful context about the analysed video, and then answers a set of criminally biased questions. The extracted context aids the exploration by providing information on what is depicted, allowing the Questioner LLM to produce more context-pertinent questions and to make the final criminal classification. The strategy is explained in detail in Chapter 6.2. This section covers the UCA-25 testing setup and the results obtained, followed by a comparison of performance between high- and low-resolution videos, and, finally, the conclusions and insights.

7.6.1 UCA-25 Setup

A first set of tests was performed in UCA-25, aiming to compare the final solution variants with the previous best approaches. The developed variants vary in the layers utilised during the metadata extraction phase and include: SLG-Layer1, utilising only the scenario understanding layer; SLG-Layer2, focusing solely on entity extraction aspects; SLG-Layer3, only exploring the social context; SLG-Layer4, performing solely event understanding; and SLG-Joint, which includes all the non-criminal exploration layers. We decided to test single-layer variants due to the limitations of introducing normality-bias by adding non-criminal context (explored in Section 7.4.2), and to compare layer importance towards the final answer.

7.6.2 UCA-25 Results

As Table 7.13 shows, baseline performance in Multiclass Classification is still unsurpassed by the developed approaches. Layer 4, the event understanding layer, demonstrates increased importance by providing the highest metric scores in both the binary and multiclass settings compared with the other single-layer approaches. This is logically coherent, since classifying criminal videos mostly relies on detecting events. Additionally, the effect of adding non-criminal context to the final decision is clear again: SLG-Joint is outperformed by every single-layer solution except SLG-Layer3 in the binary setting, and it still underperforms compared with SLG-Layer4 in the multiclass setting. If adding context did not increase the normality bias, thereby degrading the results, it would be counterintuitive to expect that combining information from different reasoning exploration layers would lead to a worse solution.

Table 7.13: SLG and prominent previous solutions on Abnormal vs Normal and Multiclass tasks. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Multiclass w/ Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	40.28	32.05	62.82	58.97	48.33	56.41	52.14	72.47
GenQ20	50.65	43.59	<u>69.23</u>	67.95	37.39	43.59	46.53	64.68
BiasQ	<u>51.24</u>	<u>44.87</u>	<u>65.38</u>	<u>70.51</u>	29.32	30.77	35.58	61.30
VERA	50.32	42.31	65.38	65.38	<u>40.38</u>	<u>46.15</u>	<u>48.39</u>	<u>66.35</u>
K5Q1-ND	48.16	41.03	65.38	69.23	<u>42.46</u>	<u>51.28</u>	45.08	<u>69.04</u>
K10Q1-ND	61.49	55.13	71.79	75.64	41.77	44.87	<u>49.55</u>	68.26
CoT-70	41.63	34.62	58.97	55.13	34.39	38.46	<u>44.35</u>	60.92
ToT-M4B5K1	50.56	43.59	66.67	<u>67.95</u>	<u>35.60</u>	<u>43.59</u>	39.60	<u>65.24</u>
IoT-70	43.97	34.62	65.38	56.41	27.44	34.62	30.03	58.88
LCoT-LGen6	<u>51.73</u>	<u>44.87</u>	<u>69.23</u>	65.38	29.39	35.90	27.87	61.06
SLG-Layer1	50.93	<u>43.59</u>	<u>69.23</u>	<u>69.23</u>	34.47	39.74	37.70	63.42
SLG-Layer2	51.31	<u>43.59</u>	<u>69.23</u>	67.95	36.45	39.74	43.98	63.43
SLG-Layer3	45.01	37.18	66.67	61.54	31.95	35.90	41.14	60.34
SLG-Layer4	<u>52.08</u>	<u>43.59</u>	<u>69.23</u>	<u>69.23</u>	<u>42.68</u>	<u>46.15</u>	<u>50.61</u>	<u>66.91</u>
SLG-Joint	50.71	42.31	<u>69.23</u>	67.95	39.28	43.59	46.28	65.26

7.6.3 Critical Overview of UCA Results

We previously hypothesised that these solutions were unable to fully comprehend the different crimes due to a lack of overall scene context, since the final answer comes from the reasoning LLM agent rather than the LVLM model (as in the baseline). During the development of the previous experiments, we gathered evidence that **introducing non-criminal context increases normality bias and degrades the final criminal-classification answers**. In light of this, we are now convinced that the **main bottleneck to surpassing the baseline in multiclass setting lies primarily in UCA’s dataset limitations**. After all, how could a solution that directly outputs a class with no additional reasoning outperform one that first extracts video-inherent details and then asks optimised, criminally focused questions to probe the candidate classes? This counterintuitive outcome is consistent with a faulty evaluation set that (i) **contains really subtle differences between classes**, (ii) **comprises videos with multiple crimes while using single-crime labels**, (iii) **provides additional textual hints that are often advantageous only for the baseline LVLM**, and (iv) **contains low-resolution videos, increasing the chance of hallucinations, which has a larger adverse impact on reasoning approaches**, as they explore based on the obtained answers, more easily propagating errors and degrading the final answer (see Section 8.2.1).

7.6.4 BUCA Setup

Considering this, we curated BUCA, a subset comprising the best native *UCA* examples together with additional footage from online sources. BUCA’s selection guidelines and analysis are presented in Section 7.1.1.2.

7.6.5 BUCA-HR Results

As Table 7.14 shows, there is not only a general increase in metric values, but also, for the first time, a solution that surpasses the baseline in every setting. SLG-Layer2 obtains the highest metric values in the binary setting, registering an increase of 9.96 p.p. in average F1-score, while SLG-Layer4 achieves an increase of 3.78 p.p. in average F1-score, both relative to the baseline.

Table 7.14: Baseline and static solutions performance on Abnormal vs Normal and Multiclass tasks. Metrics: F1-score (F1), Recall (Rec), and Precision (Pre) are support-weighted (by class frequency) averages of per-class values; Accuracy (Acc) is balanced accuracy. All the metrics are in (%). Bold = best in column; underline = best in section/block (if not already bold).

Approach	Abormal vs Normal				Multiclass w/ Nor.			
	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
Qwen2.5-VL	78.88	70.77	92.31	78.46	48.11	49.23	59.35	72.68
SLG-Layer1	86.25	81.54	92.31	89.23	45.76	46.15	58.33	71.07
SLG-Layer2	88.84	86.15	92.31	93.85	44.34	44.62	51.29	70.31
SLG-Layer3	87.99	84.62	92.31	92.31	41.79	44.62	45.60	70.29
SLG-Layer4	87.77	84.62	92.31	92.31	51.89	52.31	62.78	74.39
SLG-Joint	87.60	84.75	91.53	93.22	39.10	42.37	43.51	69.01

7.6.6 Resolution Impact

To understand how impactful resolution can be for the developed solutions, we downscaled BUCA’s videos to match the standard *UCA* native width of 320 pixels while preserving the video’s original aspect ratios (Table 7.3). Tables 7.15 and 7.16 show the resolution impact for each calculated metric.

Table 7.15: Resolution impact on each metric for Abnormal vs Normal. HR = native-resolution BUCA; LR = 320px width (aspect ratio preserved). $\Delta = LR - HR$. All the metrics are in (%).

Approach	Abnormal vs Normal											
	F1			Rec			Pre			Acc		
	LR	HR	Δ	LR	HR	Δ	LR	HR	Δ	LR	HR	Δ
Qwen2.5-VL	75.68	78.88	+3.2	67.69	70.77	+3.1	92.31	92.31	0.0	75.38	78.46	+3.1
SLG-Layer1	87.11	86.25	-0.9	83.08	81.54	-1.5	92.31	92.31	0.0	90.77	89.23	-1.5
SLG-Layer2	88.82	88.84	0.0	86.15	86.15	0.0	92.31	92.31	0.0	93.85	93.85	0.0
SLG-Layer3	85.18	87.99	+2.8	80.00	84.62	+4.6	92.31	92.31	0.0	87.69	92.31	+4.6
SLG-Layer4	87.99	87.77	-0.2	84.62	84.62	0.0	92.31	92.31	0.0	92.31	92.31	0.0
SLG-Joint	85.72	87.60	+1.9	81.36	84.75	+3.4	91.53	91.53	0.0	88.14	93.22	+5.1

Table 7.16: Resolution impact on each metric for Multiclass (with Normal). HR = native-resolution BUCA; LR = 320px width (aspect ratio preserved). $\Delta = LR - HR$. All the metrics are in (%).

Approach	Multiclass w/ Nor.											
	F1			Rec			Pre			Acc		
	LR	HR	Δ	LR	HR	Δ	LR	HR	Δ	LR	HR	Δ
Qwen2.5-VL	52.00	48.11	-3.9	53.85	49.23	-4.6	68.82	59.35	-9.5	75.15	72.68	-2.5
SLG-Layer1	34.29	45.76	+11.5	36.92	46.15	+9.2	40.67	58.33	+17.7	66.09	71.07	+5.0
SLG-Layer2	35.85	44.34	+8.5	38.46	44.62	+6.2	38.88	51.29	+12.4	66.98	70.31	+3.3
SLG-Layer3	38.57	41.79	+3.2	40.00	44.62	+4.6	47.06	45.60	-1.5	67.77	70.29	+2.5
SLG-Layer4	41.32	51.89	+10.6	44.62	52.31	+7.7	46.68	62.78	+16.1	70.23	74.39	+4.2
SLG-Joint	41.73	39.10	-2.6	42.37	42.37	0.0	52.15	43.51	-8.6	69.09	69.01	-0.1

As Tables 7.15 and 7.16 show, resolution clearly affects the performance of solutions, presenting only insignificant metric changes for the SLG-Layer2 in the Abnormal vs Normal task. This variant maintains strong performance across both resolutions in that task, retaining the F1-score lead among the variants. The largest improvement occurs in SLG-Layer1 (+11.5 p.p.), while in the multiclass setting SLG-Layer4 now holds the highest F1-score, improving by 10.6 p.p. over the low-resolution setting and surpassing the baseline by 3.78 p.p.

To aid the visibility of within-class prediction changes, we produced difference-of row-normalised (%) confusion matrices. These show how each class’s predictions evolved when changing from low resolution to high resolution (Figure 7.9). In these matrices, positive numbers on the diagonal represent straight improvements (predictions shifted from some other class to the correct classification), while negative numbers on the diagonal indicate degradation. Off-diagonal changes (negative or positive) represent redistributions among the incorrect classes. Figure 7.9 clearly displays a more favourable improvement for our reasoning-based approaches. Looking at the baseline prediction-shift matrix, we see that resolution negatively impacted *Shooting*, *Shoplifting*, *Assault*, *Fighting*, and *Arrest*, while improving *Arson* and *Vandalism*. Considering the layers that improved the most, SLG-L1 and SLG-L4 (with improvements above 10 p.p.), we can visualise that both improved *Abuse*, *Robbery*, *Arson*, *Explosion*, *Arrest*, and *RoadAccidents*.

To better visualise per-class recall, we compare the baseline with the best-performing SLG approaches (Figure 7.10). SLG-Layer4 clearly outperforms the baseline overall but still underperforms in *Stealing*, *Arrest*, and *Fighting*. SLG-Layer2, which achieves the best performance in the binary setting, surpasses the baseline in *Robbery*, *Burglary*, *Vandalism*, and *RoadAccidents*.

7.6.7 Conclusion and Insights

Our final strategy effectively surpasses the previously tested reasoning-based approaches. While SLG-Layer4, which develops event understanding, stands out as the best single-layer variant, it is still unable to surpass the baseline in the multiclass setting, despite achieving the best scores across metrics in both tasks. SLG-Joint, the variant that merges all reasoning layers, underperforms relative to several single-layer variants, emphasising again that adding non-criminal context negatively affects performance. Although *UCA* is widely used for anomaly detection and classification, **it presents crucial limitations that stall progress and confer an unfair advantage to the baseline (see Section 8.2.1). We obtained substantially better results on BUCA and observed clear benefits from higher resolution, especially in our reasoning-based approaches.**

CHAPTER 7. EXPERIMENTAL EVALUATION

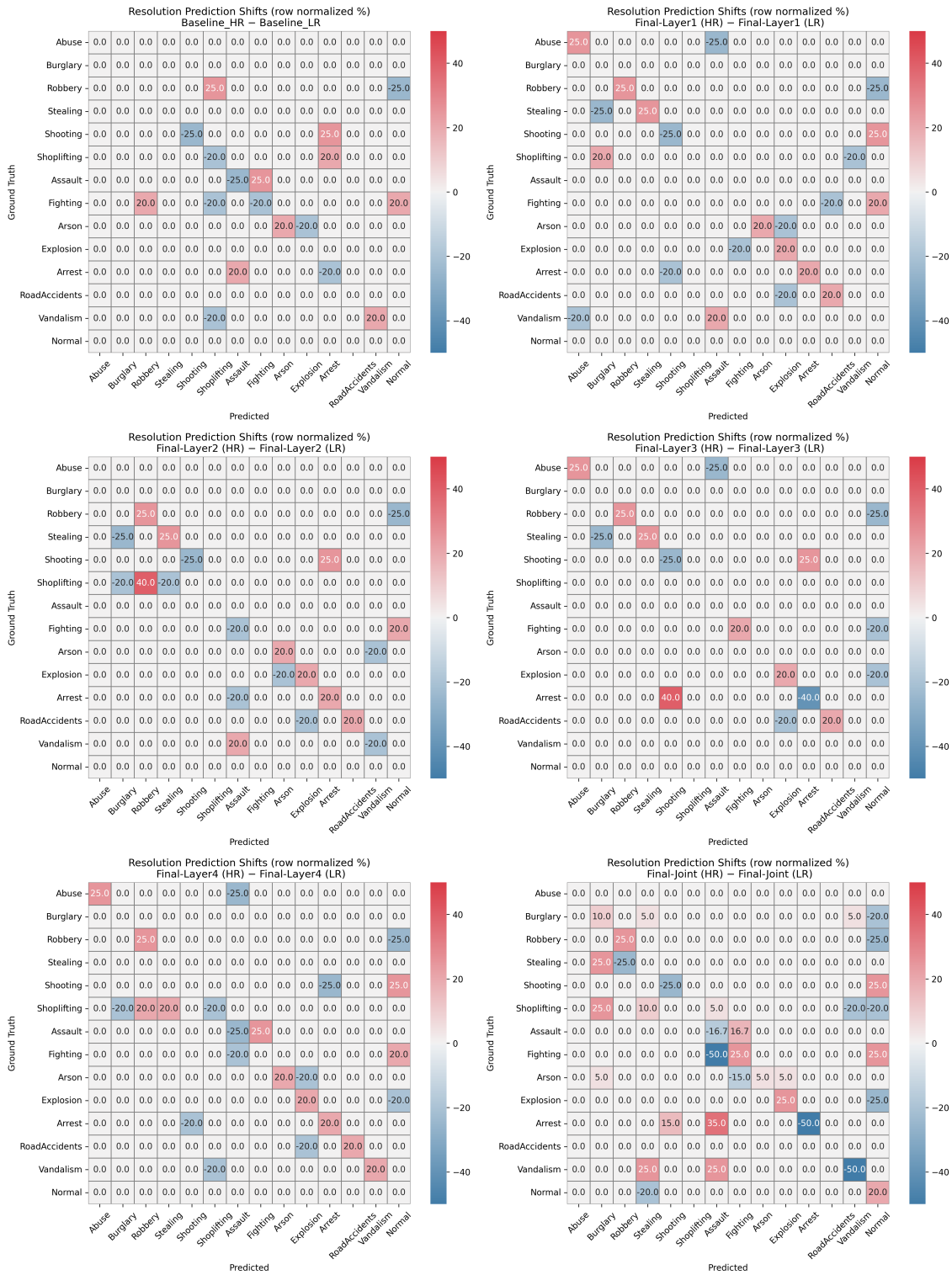


Figure 7.9: Resolution-induced prediction shifts on BUCA (HR – LR). Each panel shows the difference of row-normalised (%) confusion matrices for the indicated model.

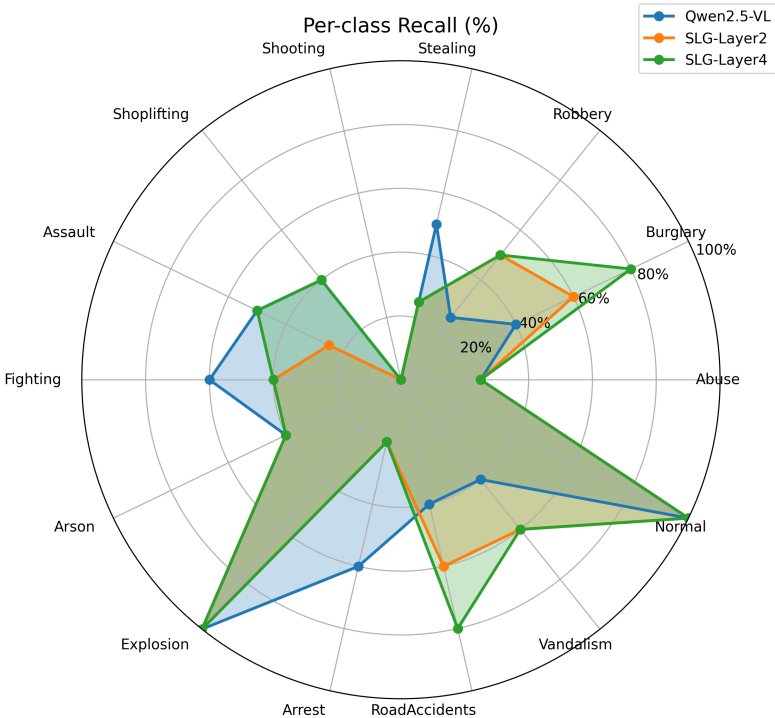


Figure 7.10: Comparison of per-class recall (%) between the baseline and the best SLG approaches.

CONCLUSIONS AND FUTURE WORK

8.1 Overview

In this thesis, we investigate and develop methods to perform metadata extraction in surveillance environments through iterative dialogue exploration, which is an unexplored approach to generating rich representations of the video content.

We started with an evaluation of current SOTA LVLMs to establish target performance metrics in surveillance-specific tasks such as Abnormal vs Normal and Multiclass Classification (with Normal), which represent the common anomaly detection and anomaly classification tasks. This evaluation showed favourable performance compared to randomly selecting answers and emphasised models' capacity to generally understand criminal videos.

This was then followed by the development of static approaches which, unlike the previous baseline solutions, now answer with a reasoning LLM rather than the vision model, revealing fundamental insights into how normality bias and criminal bias impact the final decision. Upon comparing the previous results, static approaches revealed fragilities mainly due to having a fixed set of questions that do not adapt to all the different scenarios. Additionally, these approaches revealed valuable insights into how optimising a set of questions could lead to better performance through the intelligent use of soft criminal biases, rather than basing the generated questions on tough criminal descriptions.

Since baselines demonstrated unexpectedly strong performance, we developed dynamic solutions. They start by querying the LVLm on which classes could be depicted in the video and restrict the criminal exploration process to these classes, iteratively reducing the probable set of classes and generating more pertinent questions. The results demonstrated that, while this new method can achieve up to 21.18 p.p. average F1-score over the baseline in Abnormal vs Normal, it still greatly underperformed in the multiclass setting. These results demonstrated a strict correlation between adding non-criminal context and degrading performance on common surveillance tasks.

To understand whether the previous performances resulted from genuine surveillance understanding rather than a simple better adjustment of "randomly selected" answers

and the extraction of non-criminal aspects of the videos, we developed reasoning-based strategies. These adapt existing text-only, step-by-step reasoning methods. The results demonstrated that, while these approaches were able to extract greater amounts of valuable information about the videos, they underperformed in surveillance-specific tasks due to the increased normality bias introduced by the exploration of non-criminal aspects.

Aiming to surpass the baseline on surveillance-specific tasks and extracted content, the final SLG approaches intelligently merge the fundamental aspects of previous methods. To better fit the surveillance metadata extraction domain, they start by performing a structured exploration of the scenario, entities, social context, and events. During this process, the obtained context helps to generate pertinent follow-up questions to explore the scenario and extract its meaningful aspects. Afterwards, a criminally biased module is applied to extract the final video classification. This criminal module first uses the optimised sets of questions present in M-VERA to criminally explore the environment and then selects, based on the context from both explorations, the final criminal classification. The obtained results for the multiclass setting still underperformed compared with the baseline, leading us to conduct an in-depth analysis and merge the limitations found in UCA’s dataset. These limitations represented meaningful progress barriers and an unfair comparison between approaches (see Section 8.2.1). Considering this, we curated BUCA, a small dataset that includes the best UCA native examples and adds new videos from online sources, aiming to work around the detected limitations and build a fair comparison environment. With this new dataset, the SLG approaches display increased performance and are able to surpass the baseline on surveillance-specific tasks (Table 7.14). To ensure these new scores were not caused by video simplicity but rather by higher resolution and label fairness, we tested the approaches on BUCA-LR, a resolution-reduced variant of the BUCA dataset. The results demonstrate similar multiclass scores in low-resolution videos and a clear improvement, for our developed solutions, in higher-resolution examples, emphasising the impact of resolution on these methods (Table 7.16). The SLG approaches ultimately demonstrate quantitatively superior capabilities in surveillance-specific tasks while ensuring a more appropriate surveillance metadata extraction procedure.

8.2 Limitations

Analysing the limitations found not only provides great insights into the next steps to take, but also gives broader context on some of the decisions taken during the development process. In this section, we focus on exploring the limitations regarding data, model, and produced solutions.

8.2.1 Data Limitations

Since our main objective is to develop a solution to perform metadata extraction strictly for surveillance environments, we need to utilise data that correctly resemble the usual

criminal activities performed. While acquiring data in other domains is generally easy, surveillance data presents increased difficulties due to privacy concerns. Although there are open-source options available in the surveillance domain, they mostly represent outdated sources, presenting videos with low resolution or undeveloped labelling, such as single-word labels and bounding boxes. *UCA* [59], which extends the base UCF-Crime [5] dataset, is an initially valid option for this project's scope but quickly reveals itself to be a barrier to progression, stalling the overall progress due to four main reasons:

- **Resolution** - *UCA* videos are restricted to 320×240 px, a small resolution that, when combined with the inherent challenges of surveillance footage, such as unusual camera angles, harsh lighting conditions, long camera-subject distance, and increased visual clutter, create challenging video environments to analyse even for the human eye.
- **Multi Crime** - It is generally rare to detect only a single crime in criminal footage. *UCA* videos often depict scenarios that contain more than one crime. The most frequent instances occur when perpetrators are confronted, either by other individuals or by the police. A scenario that starts with shoplifting can evolve into a fight when the owner catches the suspect red-handed; it may then escalate further through weapon use, police intervention, or immediate retaliatory acts carried out in revenge, such as arson and vandalism when leaving the scene. The main problem here is that *UCA* classifies these videos with a single criminal label, creating unfair examples for solutions that perform in-depth scene analysis and are more likely to detect multiple distinct crimes.
- **Classes Defined** - While all the categories defined in the *UCA* dataset make sense to exist, criminally speaking, they often represent very similar events and actions. Taking into consideration the previous two observations makes this problem even harder to manage, making zero-shot solutions struggle to understand the subtle differences.
- **Textual Hints** - *UCA* videos often contain textual hints in the form of journal openings, video titles, and additional editing. This does not correspond to the reality of surveillance camera footage, which contains no textual elements apart from the timestamp. Additionally, these textual hints can provide valuable information to the video models, presenting an unfair advantage for the baselines, which directly output the criminal class with the LVM, whereas the iterative dialogue methods require the LLM to hypothesise the idea first and then directly ask about it.

Taking this into account, and aiming to further test our approaches, we curated BUCA (Section 7.1.1.2), but this data source still represents a small collection, containing only 67 videos, which limits diversity.

8.2.2 Model Limitations

Model limitations represent direct progression barriers in any task and domain. While LVLMs are generally effective tools for video analysis, surveillance footage introduces novel challenges that make it more difficult for the models to comprehend the depicted scenarios (see Section 4.1.1). LVLMs are usually trained on broader video data, which is inevitably different from a distant, top-down view of the scene.

Additionally, standard LVLMs are trained to follow instructions that typically require only short- to medium-length descriptions, whereas in surveillance metadata extraction we expect to obtain much denser information about the video (see Section 4.1.2). This creates the need to use workarounds when more fine-grained, descriptive answers are required.

8.2.3 Solution Limitations

While our developments present meaningful advancements in the surveillance metadata extraction domain, measuring how good the solution’s performance is requires metrics that capture a richer, task-specific complexity than purely textual ones, which focus solely on n-gram overlap between the ground truth and the model’s answer. A metric for metadata extraction must compare semantic correctness, temporal reasoning, and event granularity. Recent work [35, 36] already utilises an LLM-as-Judge to evaluate performance on detected events in surveillance videos while demonstrating alignment with human raters. For our setting, we would adapt such a metric to also cover broader environmental aspects, while maintaining alignment with human experts and ensuring that evaluations are not biased by verbosity, answer position in comparisons, or own-model answers (see Section 3.6), thereby enabling a fair classification of the extracted content. At present, we provide qualitative illustrations showing that the reasoning approaches extract more meaningful information than static ones (Figure 7.7), and, given the model limitations discussed in Section 8.2.2, this strongly suggests that baseline performance on metadata extraction would be easily surpassed by our final, surveillance-fitted, solutions. We leave the integration of a judge-based metric as future work.

Measuring the time it takes to process one video example gives a clear view of optimisation needs, and while the developed solutions aimed to surpass the baseline in metric performance rather than runtime performance, it is interesting to examine the final time-per-example differences. Table 8.1 displays the average time it took to process one variant for each developed method and dataset. Note that these values are merely indicative and depend directly on the different tasks executed; i.e., comparisons across rows and columns can be misleading, as some values include tasks and variants that were later excluded. The exchange from the complete *UCA* test set for the smaller *UCA-25* subset is mainly due to the increased test times of the static methods in the *UCA* test set, which reached 47 hours to complete a single variant.

Table 8.1: **Average run times by method and dataset.** Times per example are expressed as MM:SS, while total times are expressed in HH:MM:SS.

Method	Run Time per Example			Total Run Time		
	UCA test (310)	UCA-25 (78)	BUCA (67)	UCA test (310)	UCA-25 (78)	BUCA (67)
Baseline	02:52	03:15	02:14	14:50:41	04:14:36	02:30:26
Static	09:06	07:32	–	47:05:35	09:48:07	–
Dynamic	–	04:05	–	–	05:18:37	–
Reasoning	–	05:58	–	–	07:46:34	–
SLG	–	21:25	25:26	–	27:51:47	28:25:01

8.3 Future Work

The proposed methods reveal crucial limitations in the surveillance metadata extraction domain (as described in Section 8.2). Future developments on this theme could include:

- **Dataset and Benchmarking** - curating a high-resolution surveillance video dataset that takes into account the footage limitations identified and provides meaningful descriptions of the events taking place in the video. Additionally, creating an unbiased and human expert aligned metadata extraction metric, by employing a model-as-judge, to perform quantitative evaluation of context extraction tasks.
- **Reasoning and Efficiency** - optimising reasoning strategies to balance the amount of context extracted and runtime. This can be done in different ways, such as modifying the first stage of exploration for the SLG method; developing more refined sets of operations; experimenting with new criminal module concepts that, throughout the exploration, prune criminal categories based on probability; integrating new agents or features that aid exploration; and testing external validity procedures to detect hallucinations.
- **Model Enhancements** - while this thesis focuses on developing zero-shot methods, having agents specialised in their roles could further improve results. A surveillance metadata extraction fine-tuned LVLM to answer the pertinent questions could lead to quicker convergence of themes and allow more hallucination-free exploration. Additionally, vision models would be a better fit for surveillance analysis if they could receive multiple cameras' footage and geospatially link entities and events.
- **System Application** - once solutions capable of performing metadata extraction effectively and efficiently are available, test these approaches on real-world use cases such as real-time surveillance monitoring systems, providing immediate detection of suspicious events; and forensic search engines, generating retrospective event descriptions that can be queried to retrieve particular metadata.

BIBLIOGRAPHY

- [1] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. *Qwen Technical Report*. 2023. arXiv: [2309.16609](https://arxiv.org/abs/2309.16609) [cs.CL]. URL: <https://arxiv.org/abs/2309.16609> (cit. on p. 18).
- [2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. 2023. arXiv: [2308.12966](https://arxiv.org/abs/2308.12966) [cs.CV]. URL: <https://arxiv.org/abs/2308.12966> (cit. on p. 23).
- [3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. *Qwen2.5-VL Technical Report*. 2025. arXiv: [2502.13923](https://arxiv.org/abs/2502.13923) [cs.CV]. URL: <https://arxiv.org/abs/2502.13923> (cit. on p. 38).
- [4] J. Becker. *Multi-Agent Large Language Models for Conversational Task-Solving*. 2024. arXiv: [2410.22932](https://arxiv.org/abs/2410.22932) [cs.CL]. URL: <https://arxiv.org/abs/2410.22932> (cit. on pp. 1, 26).
- [5] U. o. C. F. Center for Research in Computer Vision. *Real-World Anomaly Detection in Surveillance Videos*. Accessed: 2024-07-15. 2024. URL: <https://www.crcv.ucf.edu/projects/real-world/> (cit. on pp. 16, 40, 47, 88).
- [6] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun. *MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark*. 2024. arXiv: [2402.04788](https://arxiv.org/abs/2402.04788) [cs.CL]. URL: <https://arxiv.org/abs/2402.04788> (cit. on p. 26).

- [7] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 2024. arXiv: [2402.03216](https://arxiv.org/abs/2402.03216) [cs.CL]. URL: <https://arxiv.org/abs/2402.03216> (cit. on p. 14).
- [8] J. C.-Y. Chen, S. Saha, and M. Bansal. *ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs*. 2024. arXiv: [2309.13007](https://arxiv.org/abs/2309.13007) [cs.CL]. URL: <https://arxiv.org/abs/2309.13007> (cit. on pp. 27, 28).
- [9] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. *InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks*. 2024. arXiv: [2312.14238](https://arxiv.org/abs/2312.14238) [cs.CV]. URL: <https://arxiv.org/abs/2312.14238> (cit. on p. 14).
- [10] P. Domingos, J. Pereira, V. Lopes, J. Neves, and D. Semedo. *Chain-of-Anomaly Thoughts with Large Vision-Language Models*. 2025. arXiv: [2512.20417](https://arxiv.org/abs/2512.20417) [cs.CV]. URL: <https://arxiv.org/abs/2512.20417> (cit. on p. 4).
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: [2010.11929](https://arxiv.org/abs/2010.11929) [cs.CV]. URL: <https://arxiv.org/abs/2010.11929> (cit. on pp. 8, 9).
- [12] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. 2023. arXiv: [2305.14325](https://arxiv.org/abs/2305.14325) [cs.CL]. URL: <https://arxiv.org/abs/2305.14325> (cit. on pp. 1, 28).
- [13] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, C. Shan, R. He, and X. Sun. *Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis*. 2025. arXiv: [2405.21075](https://arxiv.org/abs/2405.21075) [cs.CV]. URL: <https://arxiv.org/abs/2405.21075> (cit. on pp. 11, 12, 24).
- [14] Y. Fu, H. Peng, T. Khot, and M. Lapata. *Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback*. 2023. arXiv: [2305.10142](https://arxiv.org/abs/2305.10142) [cs.CL]. URL: <https://arxiv.org/abs/2305.10142> (cit. on p. 27).
- [15] C. Goutte and E. Gaussier. “A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation”. In: vol. 3408. 2005-04, pp. 345–359. ISBN: 978-3-540-25295-5. DOI: [10.1007/978-3-540-31865-1_25](https://doi.org/10.1007/978-3-540-31865-1_25) (cit. on p. 26).
- [16] S. Han, W. Huang, H. Shi, L. Zhuo, X. Su, S. Zhang, X. Zhou, X. Qi, Y. Liao, and S. Liu. *VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection*. 2024. arXiv: [2411.14794](https://arxiv.org/abs/2411.14794) [cs.CV]. URL: <https://arxiv.org/abs/2411.14794> (cit. on pp. 14, 15).

-
- [17] M. Hu, Y. Mu, X. Yu, M. Ding, S. Wu, W. Shao, Q. Chen, B. Wang, Y. Qiao, and P. Luo. *Tree-Planner: Efficient Close-loop Task Planning with Large Language Models*. 2024. arXiv: [2310.08582](https://arxiv.org/abs/2310.08582) [cs.CL]. URL: <https://arxiv.org/abs/2310.08582> (cit. on p. 28).
- [18] ItemFix. *ItemFix – Videos, Images, and Audio*. <https://www.itemfix.com/>. Accessed: 2025-09-14 (cit. on p. 60).
- [19] S. Jinxin, Z. Jiabao, W. Yilei, W. Xingjiao, L. Jiawen, and H. Liang. *CGMI: Configurable General Multi-Agent Interaction Framework*. 2023. arXiv: [2308.12503](https://arxiv.org/abs/2308.12503) [cs.AI]. URL: <https://arxiv.org/abs/2308.12503> (cit. on p. 27).
- [20] K. Kim, G. Park, Y. Lee, W. Yeo, and S. J. Hwang. *VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding*. 2024. arXiv: [2412.02186](https://arxiv.org/abs/2412.02186) [cs.CV]. URL: <https://arxiv.org/abs/2412.02186> (cit. on pp. 16, 17).
- [21] J. Y. Koh, S. McAleer, D. Fried, and R. Salakhutdinov. *Tree Search for Language Model Agents*. 2024. arXiv: [2407.01476](https://arxiv.org/abs/2407.01476) [cs.AI]. URL: <https://arxiv.org/abs/2407.01476> (cit. on p. 28).
- [22] V. I. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics. Doklady* 10 (1965), pp. 707–710. URL: <https://api.semanticscholar.org/CorpusID:60827152> (cit. on p. 26).
- [23] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li. *LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild*. 2024-05. URL: <https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/> (cit. on p. 21).
- [24] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. *LLaVA-OneVision: Easy Visual Task Transfer*. 2024. arXiv: [2408.03326](https://arxiv.org/abs/2408.03326) [cs.CV]. URL: <https://arxiv.org/abs/2408.03326> (cit. on p. 21).
- [25] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li. *LLaVA-NeXT: Tackling Multi-image, Video, and 3D in Large Multimodal Models*. 2024-06. URL: <https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/> (cit. on p. 21).
- [26] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. *CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society*. 2023. arXiv: [2303.17760](https://arxiv.org/abs/2303.17760) [cs.AI]. URL: <https://arxiv.org/abs/2303.17760> (cit. on pp. 1, 27, 28).
- [27] H. Li, Y. Chong, S. Stepputtis, J. Campbell, D. Hughes, C. Lewis, and K. Sycara. “Theory of Mind for Multi-Agent Collaboration via Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. DOI: [10.18653/v1/2023.emnlp-](https://doi.org/10.18653/v1/2023.emnlp-)

- main.13. URL: <http://dx.doi.org/10.18653/v1/2023.emnlp-main.13> (cit. on p. 27).
- [28] H. Liu, C. Li, Y. Li, and Y. J. Lee. *Improved Baselines with Visual Instruction Tuning*. 2024. arXiv: 2310.03744 [cs.CV]. URL: <https://arxiv.org/abs/2310.03744> (cit. on p. 20).
- [29] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*. 2024-01. URL: <https://llava-vl.github.io/blog/2024-01-30-llava-next/> (cit. on p. 21).
- [30] H. Liu, C. Li, Q. Wu, and Y. J. Lee. *Visual Instruction Tuning*. 2023. arXiv: 2304.08485 [cs.CV]. URL: <https://arxiv.org/abs/2304.08485> (cit. on p. 19).
- [31] W. Liu, D. L. W. Luo, and S. Gao. “Future Frame Prediction for Anomaly Detection – A New Baseline”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on p. 16).
- [32] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User’s Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf> (cit. on p. i).
- [33] OpenAI. *GPT-4o System Card*. Available at <https://cdn.openai.com/gpt-4o-system-card.pdf>. 2024. URL: <https://cdn.openai.com/gpt-4o-system-card.pdf> (cit. on p. 15).
- [34] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. 2023. arXiv: 2304.03442 [cs.HC]. URL: <https://arxiv.org/abs/2304.03442> (cit. on pp. 1, 27).
- [35] J. Pereira, V. Lopes, J. Neves, and D. Semedo. “Towards Event-Driven Evaluation of Surveillance Video Understanding Using Natural Language”. In: 2025-07, pp. 43–55. ISBN: 978-3-031-99567-5. DOI: 10.1007/978-3-031-99568-2_4 (cit. on p. 89).
- [36] J. A. C. Pereira, V. Lopes, J. C. Neves, and D. Semedo. “FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding”. In: *The Fortieth AAAI Conference on Artificial Intelligence*. 2026 (cit. on p. 89).
- [37] Qwen, : A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. *Qwen2.5 Technical Report*. 2025. arXiv: 2412.15115 [cs.CL]. URL: <https://arxiv.org/abs/2412.15115> (cit. on p. 39).
- [38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: <https://arxiv.org/abs/2103.00020> (cit. on pp. 9, 10).

- [39] S. K. Radha, Y. N. Jelyani, A. Ghukasyan, and O. Goktas. *Iteration of Thought: Leveraging Inner Dialogue for Autonomous Large Language Model Reasoning*. 2024. arXiv: [2409.12618](https://arxiv.org/abs/2409.12618) [cs.CL]. URL: <https://arxiv.org/abs/2409.12618> (cit. on pp. 30, 31, 44, 50).
- [40] N. Rashvand, G. A. Noghre, A. D. Pazho, S. Yao, and H. Tabkhi. *Exploring Pose-Based Anomaly Detection for Retail Security: A Real-World Shoplifting Dataset and Benchmark*. 2025. arXiv: [2501.06591](https://arxiv.org/abs/2501.06591) [cs.CV]. URL: <https://arxiv.org/abs/2501.06591> (cit. on p. 16).
- [41] M. Sanwal. *Layered Chain-of-Thought Prompting for Multi-Agent LLM Systems: A Comprehensive Approach to Explainable Large Language Models*. 2025. arXiv: [2501.18645](https://arxiv.org/abs/2501.18645) [cs.CL]. URL: <https://arxiv.org/abs/2501.18645> (cit. on pp. 30, 44, 50).
- [42] T. Schick, J. Dwivedi-Yu, Z. Jiang, F. Petroni, P. Lewis, G. Izacard, Q. You, C. Nalmpantis, E. Grave, and S. Riedel. *PEER: A Collaborative Language Model*. 2022. arXiv: [2208.11663](https://arxiv.org/abs/2208.11663) [cs.CL]. URL: <https://arxiv.org/abs/2208.11663> (cit. on p. 28).
- [43] Z. Shanguan, C. Li, Y. Ding, Y. Zheng, Y. Zhao, T. Fitzgerald, and A. Cohan. *TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models*. 2024. arXiv: [2410.23266](https://arxiv.org/abs/2410.23266) [cs.CV]. URL: <https://arxiv.org/abs/2410.23266> (cit. on pp. 13, 24).
- [44] N. Shazeer. *GLU Variants Improve Transformer*. 2020. arXiv: [2002.05202](https://arxiv.org/abs/2002.05202) [cs.LG]. URL: <https://arxiv.org/abs/2002.05202> (cit. on p. 18).
- [45] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. 2023. arXiv: [2104.09864](https://arxiv.org/abs/2104.09864) [cs.CL]. URL: <https://arxiv.org/abs/2104.09864> (cit. on p. 18).
- [46] M. Suzgun and A. T. Kalai. *Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding*. 2024. arXiv: [2401.12954](https://arxiv.org/abs/2401.12954) [cs.CL]. URL: <https://arxiv.org/abs/2401.12954> (cit. on p. 27).
- [47] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: [2302.13971](https://arxiv.org/abs/2302.13971) [cs.CL]. URL: <https://arxiv.org/abs/2302.13971> (cit. on p. 18).
- [48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2023. arXiv: [1706.03762](https://arxiv.org/abs/1706.03762) [cs.CL]. URL: <https://arxiv.org/abs/1706.03762> (cit. on p. 6).

- [49] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. *Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution*. 2024. arXiv: 2409.12191 [cs.CV]. URL: <https://arxiv.org/abs/2409.12191> (cit. on p. 23).
- [50] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. 2023. arXiv: 2203.11171 [cs.CL]. URL: <https://arxiv.org/abs/2203.11171> (cit. on pp. 1, 28).
- [51] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji. *Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration*. 2024. arXiv: 2307.05300 [cs.AI]. URL: <https://arxiv.org/abs/2307.05300> (cit. on pp. 27, 28).
- [52] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. URL: <https://arxiv.org/abs/2201.11903> (cit. on pp. 29, 43, 49).
- [53] Y. Xie, K. Kawaguchi, Y. Zhao, J. X. Zhao, M.-Y. Kan, J. He, and M. Xie. "Self-Evaluation Guided Beam Search for Reasoning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 41618–41650. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/81fde95c4dc79188a69ce5b24d63010b-Paper-Conference.pdf (cit. on p. 28).
- [54] B. Xu, A. Yang, J. Lin, Q. Wang, C. Zhou, Y. Zhang, and Z. Mao. *Expert Prompting: Instructing Large Language Models to be Distinguished Experts*. 2023. arXiv: 2305.14688 [cs.CL]. URL: <https://arxiv.org/abs/2305.14688> (cit. on p. 27).
- [55] J. C. Yang, D. Dailisan, M. Korecki, C. I. Hausladen, and D. Helbing. "LLM Voting: Human Choices and AI Collective Decision-Making". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7 (2024-10)*, pp. 1696–1708. ISSN: 3065-8365. DOI: 10.1609/aies.v7i1.31758. URL: <http://dx.doi.org/10.1609/aies.v7i1.31758> (cit. on p. 28).
- [56] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. 2023. arXiv: 2305.10601 [cs.CL]. URL: <https://arxiv.org/abs/2305.10601> (cit. on pp. 29, 30, 44, 49).
- [57] M. Ye, W. Liu, and P. He. *VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models*. 2025. arXiv: 2412.01095 [cs.AI]. URL: <https://arxiv.org/abs/2412.01095> (cit. on pp. 17, 41).

- [58] Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, and X. Qiu. “Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, 2023-12, pp. 15135–15153. DOI: [10.18653/v1/2023.emnlp-main.936](https://doi.org/10.18653/v1/2023.emnlp-main.936). URL: <https://aclanthology.org/2023.emnlp-main.936/> (cit. on p. 28).
- [59] T. Yuan, X. Zhang, K. Liu, B. Liu, C. Chen, J. Jin, and Z. Jiao. *Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges*. 2023. arXiv: [2309.13925](https://arxiv.org/abs/2309.13925) [cs.CV]. URL: <https://arxiv.org/abs/2309.13925> (cit. on pp. v, vi, 16, 36, 58, 88).
- [60] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning*. <https://D2L.ai>. Cambridge University Press, 2023 (cit. on p. 7).
- [61] B. Zhang and R. Sennrich. *Root Mean Square Layer Normalization*. 2019. arXiv: [1910.07467](https://arxiv.org/abs/1910.07467) [cs.LG]. URL: <https://arxiv.org/abs/1910.07467> (cit. on p. 18).
- [62] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. *Video Instruction Tuning With Synthetic Data*. 2024. arXiv: [2410.02713](https://arxiv.org/abs/2410.02713) [cs.CV]. URL: <https://arxiv.org/abs/2410.02713> (cit. on pp. 21, 22, 38).
- [63] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: [2306.05685](https://arxiv.org/abs/2306.05685) [cs.CL]. URL: <https://arxiv.org/abs/2306.05685> (cit. on pp. 24, 26).
- [64] A. Zhou, K. Yan, M. Shlapentokh-Rothman, H. Wang, and Y.-X. Wang. *Language Agent Tree Search Unifies Reasoning Acting and Planning in Language Models*. 2024. arXiv: [2310.04406](https://arxiv.org/abs/2310.04406) [cs.AI]. URL: <https://arxiv.org/abs/2310.04406> (cit. on pp. 1, 28).
- [65] Y. Zhuang, Y. Yu, K. Wang, H. Sun, and C. Zhang. “ToolQA: A Dataset for LLM Question Answering with External Tools”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 50117–50143. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/9cb2a7495900f8b602cb10159246a016-Paper-Datasets_and_Benchmarks.pdf (cit. on p. 27).
- [66] M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordás, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, L. Kirsch, B. Li, G. Li, S. Liu, J. Mai, P. Piękos, A. Ramesh, I. Schlag, W. Shi, A. Stanić, W. Wang, Y. Wang, M. Xu, D.-P. Fan, B. Ghanem, and J. Schmidhuber. *Mindstorms in Natural Language-Based Societies of Mind*. 2023. arXiv: [2305.17066](https://arxiv.org/abs/2305.17066) [cs.AI]. URL: <https://arxiv.org/abs/2305.17066> (cit. on pp. 1, 27).

BIBLIOGRAPHY

PROMPT TEMPLATES

Here, we present the templates used to generate dynamic prompts sent to the models in each solution variant. The prompts shown omit the instructions regarding expected output format (always a JSON object) and additional rule enforcement (applied when the models were not cooperating properly) for better readability. Some prompts present more than one variant, depending on the availability of specific variables. These variants are clearly marked in the templates using placeholders such as {VARIABLE}, and distinct colour boxes. Additionally, to avoid redundancy, prompts that have already been introduced in earlier sections of this appendix are not repeated when subsequent solutions make use of them; instead, those solutions are assumed to rely on the previously presented templates.

Firstly, and common to almost every other prompt, there is a snippet containing the list of classes and their respective descriptions:

Category List Used in Prompts

The following categories are available. Each category name is in quotes, followed by its description.

- **"Abuse"**: This category includes acts of cruelty or violence against children, elderly people, animals, or women.
- **"Burglary"**: This category involves individuals unlawfully entering a building with the intent to commit theft.
- **"Robbery"**: This category captures instances where thieves take money or valuables by force or threat.
- **"Stealing"**: This category covers theft of property or money without permission.
- **"Shooting"**: This category focuses on incidents where a person shoots a firearm, regardless of whether they hit someone.
- **"Shoplifting"**: This category involves individuals stealing items from a store while posing as a shopper.
- **"Assault"**: This category includes sudden or violent physical attacks on individuals.
- **"Fighting"**: This category includes physical altercations between two or more individuals attacking each other.
- **"Arson"**: This category involves individuals deliberately setting fire to property.
- **"Explosion"**: This category covers incidents of something accidentally blowing up.
- **"Arrest"**: This category covers police arresting individuals.
- **"RoadAccidents"**: This category includes traffic accidents involving vehicles, pedestrians, or cyclists.
- **"Vandalism"**: This category involves deliberate destruction of or damage to public or private property. The term includes property damage, such as graffiti and defacement directed towards any property without permission of the owner.
- **"Normal"**: This category includes videos where no crime occurred, capturing both indoor and outdoor scenes in different lighting conditions.

A.1 Static Questions

Prompt to generate questions for each class

Context:

{CLASSES AND RESPECTIVE DESCRIPTIONS}

If a description is used:

Brief description of the video:

{DESCRIPTION}

If there is no previous dialogue

When creating the questions take into account:

- **1 Separability**: The set of question(s) should be able to differentiate between the class "{CLASS NAME}" and the other classes.
- **2 Relevance**: The questions should be mostly relevant to the class "{CLASS NAME}" and should help in identifying the class of the video.
- **3 Diversity**: The questions should be diverse and cover different aspects of the class "{CLASS NAME}".

If there is already some previous dialogue

These are the previous questions and respective answers provided by the Vision-Language Model, which should be taken into account when generating the new questions either to avoid repetition or to build upon the previous findings

Previous Dialog:

{PREVIOUS DIALOGUE}

Task:

Create a set of {NUMBER OF QUESTIONS} pertinent question(s) that, when asked about a specific surveillance video, can help determine if the video belongs to the class "{CLASS NAME}".

Prompt to answer question based on the video's content

Answer the following question regarding the video:

{QUESTION}

Prompt for Abnormal vs Normal classification

****Context**:**

Categories:

- ****Normal****: This category includes videos where no crime occurred, capturing both indoor and outdoor scenes in different lighting conditions.
- ****Abnormal****: This category encompasses videos depicting criminal or anomalous activities, such as Abuse, Burglary, Robbery, Stealing, Shooting, Shoplifting, Assault, Fighting, Arson, Explosion, Arrest, RoadAccidents, and Vandalism. These events may occur in both indoor and outdoor settings, under varied lighting and environmental conditions, and often involve individuals engaging in harmful, illegal, or disruptive actions.

If there are question-answer pairs

A set of questions were asked to a Vision-Language Model regarding a video to determine whether this video belongs either to the category "Normal" or the category "Abnormal".

Below are the provided questions and respective answers:

****Questions and Answers**:**

{QUESTION-ANSWER PAIRS}

****Task**:**

Based on the answers provided, identify if the video belongs to the category "Normal" or the category "Abnormal".

If there are no question-answer pairs

****Task**:**

Identify if the video belongs to the category "Normal" or the category "Abnormal".

Prompt for Multiclass classification

****Context**:**

{CLASSES AND RESPECTIVE DESCRIPTIONS}

If there are question-answer pairs

A set of questions were asked to a Vision-Language Model regarding a video to determine the category of the video.

Below are the provided questions and answers:

****Questions and Answers**:**

{QUESTION-ANSWER PAIRS}

****Task**:**

Based on the answers provided, identify the category of the video from one of the options above.

If there are no question-answer pairs

****Task**:**

Identify the category of the video from one of the options above.

Prompt to check agreement

****Context**:**

A question was answered by two different video models.

****Question**:**

{QUESTION}

****Answer 1**:**

{ANSWER 1}

****Answer 2**:**

{ANSWER 2}

****Task**:**

Determine if the two answers are in agreement or not.

A.2 Dynamic Top-K Questions

Prompt to Reduce Set K of Probable Classes

****Context**:**

{REMAINING CLASSES AND RESPECTIVE DESCRIPTIONS}

If there are question-answer pairs

A set of questions were asked to a Vision-Language Model regarding a video to determine the category of the video.

Below are the provided questions and answers:

****Questions and Answers**:**

{QUESTION-ANSWER PAIRS}

****Task**:**

Based on the question-answer pairs and the video provided, identify the top {TOP-K} categories of the video from the ones presented above.

If there are no question-answer pairs

****Task**:**

Identify the top {TOP-K} categories of the video from the ones presented above.

Prompt to generate video description

****Context**:**

A surveillance video was provided.

****Task**:**

Describe the video in detail, including the actions of the people involved, the environment, and any other relevant information.

A.3 Reasoning-based Strategies

Prompt to perform Chain-of-Thought reasoning

Your task is to generate pertinent step-by-step questions in order to complete the objective:

"Describe the scenario, entities, and events taking place in the video."

These questions will be answered by another agent that can observe the video.

If it is the first thought

Generate the first question to begin the investigation.

If is not the first thought

****Questions and Answers**:**
{QUESTION-ANSWER PAIRS}

If you believe that no further questions are needed and the scenario is sufficiently understood, respond with just the word "Finished". Otherwise, generate the next most pertinent question to further develop the analysis.

Prompt for Tree-of-Thought Reasoning

Your task is to generate pertinent step-by-step questions in order to complete the objective:

"Describe the scenario, entities, and events taking place in the video."

These questions will be answered by another agent that can observe the video.

If it is the first thought

Generate the first {BRANCHING FACTOR} questions to begin the investigation.

If it is not the first thought

****Questions and Answers**:**
{QUESTIONS AND ANSWERS}

If you believe that no further questions are needed and the scenario is sufficiently understood, respond with just the word "Finished". Otherwise, generate {BRANCHING FACTOR} different pertinent follow-up questions to explore different lines of inquiry.

Prompt to Evaluate Question-Answer Pair in Tree-of-Thought

You are evaluating follow-up questions in the context of a surveillance video analysis task.

Given the prior dialog context, a candidate follow-up question, and the corresponding answer provided by a visual-language model (LVLM), rate the overall utility of the question-answer pair in advancing the investigation. Focus on how much new, relevant, and concrete information the answer provides.

Rate on a scale from 1 to 10, where:

1-3 = The answer is irrelevant, vague, uninformative or repeated.

4-6 = The answer provides moderately useful or partially redundant information.

7-9 = The answer is relevant, specific, and meaningfully enhances the understanding.

10 = The answer is highly informative and significantly advances the scenario analysis.

****Context (previous questions and respective answers):****

{CONTEXT}

****Candidate Question:****

{QUESTION}

****Answer:****

{ANSWER}

Prompt to Generate Layered Chain-of-Thought Layers

Your task is to generate a series of layers for a layered chain-of-thought reasoning process.

Each layer should focus on a specific aspect of the final objective:

"Describe the scenario, entities, and events taking place in the video."

Generate {NUM LAYERS} layers, describing each layer's focus and the type of questions that should be asked to explore that layer.

Prompt for Layered Chain-of-Thought Reasoning

You are performing a layered analysis of video to extract scene understanding. Each layer focuses on a different sub-aspect of the video.

Current Layer {LAYER ID}: {LAYER DESCRIPTION}

If it is the first layer

Based on the video and layer objective, generate a focused question for this layer.

If it is not the first layer

****Context from previous layers:****

{QUESTIONS AND ANSWERS}

Based on the video, previous context, and layer objective, generate a focused question for this layer.

Prompt to generate summary based on built question-answer pairs

You are given a set of question–answer pairs obtained during the exploration of a surveillance video. Your task is to create a highly detailed summary that integrates all the meaningful information from these pairs.

Do:

- Write exactly two paragraphs:
 - First paragraph: describe the environment and the people (appearance, clothing, objects, positions).
 - Second paragraph: describe the actions or events taking place.
- Be factual, concise, and consistent.
- Use simple, direct sentences that maximize information density.

Don't:

- Do not add speculation or assumptions.
- Do not repeat information.
- Do not use narrative, stylistic, or inconsistent text schemes.
- Do not exceed two paragraphs.

question-answer pairs:

{QUESTION-ANSWER PAIRS}

A.4 Surveillance Layered Graph Strategy

In these prompts, the previous content is constructed by a function that traverses from the current node up to the root, extracting every question–answer pair produced and merging them into a single structured string.

Operation Descriptions For Selection (Proceed and Split)

Proceed

Continue developing the current line of reasoning by asking a new question that builds on the previous context.

Use this to expand the scope of the investigation or uncover new aspects related to the current conversational theme.

This operation is suitable when the goal is to explore further, not necessarily to clarify previous information.

Split

Decompose the current answer into multiple distinct research lines ****only when it contains several different elements****, topics, or facts that warrant focused, individual exploration.

Use this operation when separating these components can lead to clearer or more detailed insights relevant to the current sub-layer objective — for example, different people, actions, or locations described together.

Do not use Split if only one meaningful research line can be extracted — in that case, use other operation instead.

Avoid splitting when the identified elements are too similar, repetitive, uninformative, immobile, or irrelevant to the sub-layer objective (e.g., background furniture in an emotional expression task).

Do not repeat previously created research lines — each split must introduce new and relevant directions. Split should only be selected when multiple, clearly distinguishable and semantically useful paths are present for continued reasoning.

Operation Descriptions For Selection (Refine and Stop)

Refine

Reformulate the same question using additional details from the previous answer only if the answer was vague, ambiguous, or incomplete.

Do not choose this operation if the previous answer is clearly negative, conclusive or demonstrated lack of capabilities from the answering side (e.g., “No, there is no such object in the scene”, “The video does not provide specific details about...”).

This operation is suitable when more precision or elaboration on existing evidence is likely to yield richer insight (e.g., If you already know there’s a person in the scene, why not ask about their clothing or position?)

Stop

End the current reasoning path. Select this when the previous answer fully addresses the sub-layer objective or if the latest answers show no further insight can be gained from continuing the line of questioning.

Use this especially after clear, conclusive answers (e.g., a confirmed absence or a fully detailed explanation).

Sometimes, some objectives can not be answered by lack of capabilities from the answering side or the video content, so this operation is useful to stop the reasoning process.

Prompt for Operation Selection

You are an Operation Selection Agent. Your task is to select the most appropriate operation based on the current context and previous question–answer dialogue.

Context:

- Current Reasoning Layer: {LAYER}
- Current Sub-layer Objective: {SUB-LAYER OBJECTIVE}
- Previous Dialogue: {PREVIOUS DIALOGUE}

Answer with one of the following operation options:

{AVAILABLE OPERATIONS}

Prompt for Proceed Operation

You are a Follow-up Question Generation Agent. Your task is to create multiple detailed, context-sensitive follow-up questions that build upon already observed evidence in a surveillance video.

Context:

- Reasoning Layer: {LAYER}
- Sub-layer Objective: {SUB-LAYER OBJECTIVE}
- Previous Observations: {PREVIOUS CONTEXT}

Instructions:

- Based on the previous observations, generate 3 focused follow-up questions.
- Each question should aim to clarify, or specify aspects of what has already been observed.
- Questions should remain grounded in the visual domain.
- You may ask about specific types, attributes, relationships, or implications of observed elements.

Generate 3 questions that deepen the exploration of the current sub-layer based on the prior context.

Prompt for Split Operation

You are a reasoning assistant working with a knowledge graph extracted from a video. The current node contains a detailed but dense description, which may include multiple ideas, observations, or facts.

Your task is to split this node's content into distinct research lines, each corresponding to an aspect that could be investigated further in a separate branch of the reasoning process.

Each research line should:

- Be given a name that summarises the main question or topic it focuses on.
- Include a paragraph that explains what is currently known based on the context and why this aspect might require further exploration.

****Context to split:****

{CONTEXT}

Prompt for Refine Operation

You are refining a line of questioning that aims to extract detailed information from a surveillance video.

****Current Focus:**** - Layer: {LAYER}

- Sub-layer: {SUB-LAYER}

****Objective:****

Continue to improve the original question by increasing its granularity and specificity, while maintaining alignment with the current layer and sub-layer goals.

****Guidelines:****

- Incorporate key insights from the previous answer to build a more targeted and informative follow-up question.
- Maintain the same thematic focus as the original question.
- Avoid shifting to unrelated aspects of the scene.
- Do not simply repeat the previous question.

****Context:****

{PREVIOUS CONTEXT}

Now, write a refined follow-up question based on the context above.

M-VERA GENERATED QUESTIONS

Here, we present the optimised sets of questions generated by *M-VERA*'s approach (explained in Section 3.3.3 and later developed in Section 5.2). Each set was optimised to distinguish a specific individual criminal class from the *Normal* class. For transparency and reproducibility, all generated questions, along their respective *M-VERA*-reported accuracies (on the *UCF-Crime* test set), are included in the lists below.

Abuse | 0.86 acc

1. Are there any individuals who appear to be moving in a way that is inconsistent with their usual behavior?
2. Are there any objects that have been moved or rearranged in a way that seems unusual?
3. Is there any change in the overall scene composition that suggests something out of the ordinary?
4. Are there any individuals who are interacting in a manner that is not typical for the setting?
5. Is there any indication of a sudden shift in the environment, such as a change in lighting or background elements?

Arrest | 0.90 acc

1. Are there any sudden changes in pedestrian behavior or actions that seem out of place?
2. Is there any unusual interaction between pedestrians or objects?
3. Are there any abrupt changes in the environment or background?
4. Is there any evidence of a vehicle entering or exiting the scene in an irregular manner?
5. Are there any objects being moved or handled in an unusual manner?

Arson | 1.00 acc

1. Does the video contain any sudden or unexpected events like explosions or fires?
2. Are there any significant changes in lighting or color that suggest an anomaly?
3. Do the objects in the scene exhibit unusual movements or positions?
4. Is there any evidence of damage or destruction in the environment?
5. Are there any abrupt changes in the background or surroundings?

Assault | 0.75 acc

1. Are there any pedestrians or vehicles moving in directions that are inconsistent with the flow of traffic?
2. Do the individuals or objects exhibit any unusual postures or behaviors that deviate from the norm?
3. Are there any sudden changes in the density or distribution of people and vehicles on the street?
4. Does the video contain any abrupt changes in lighting or shadows that are not typical for the scene?
5. Are there any objects or people that appear out of place or misplaced in the environment?

APPENDIX B. M-VERA GENERATED QUESTIONS

Burglary | 0.69 acc

1. Are there any individuals who appear to be moving unusually or in a manner inconsistent with typical shopping behavior?
2. Have any objects been moved or repositioned in a way that suggests tampering or theft?
3. Are there any individuals who seem to be lingering or behaving suspiciously?
4. Has the layout of the store or the arrangement of items changed unexpectedly?
5. Are there any signs of forced entry or damage to the store's security systems?

Explosion | 0.83 acc

1. Are there any sudden movements or actions that deviate from the usual behavior of the individuals in the scene?
2. Is there any noticeable change in the environment, such as the appearance of new objects or the displacement of existing ones?
3. Are there any individuals who appear to be in a hurry or exhibiting unusual behavior?
4. Is there any indication of a sudden increase or decrease in the number of people in the scene?
5. Are there any signs of urgency or distress among the individuals present?

Fighting | 0.80 acc

1. Are there any sudden changes in the movement patterns of individuals?
2. Do individuals appear to be moving unusually fast or slow?
3. Are there any abrupt changes in the number of people present?
4. Are there any signs of urgency or distress among the individuals?
5. Are there any unusual interactions between individuals, such as physical altercations or prolonged conversations?

Road Accidents | 0.75 acc

1. Are there any sudden movements of pedestrians that disrupt the flow of traffic?
2. Are there any changes in the positioning of vehicles that suggest a collision or near-miss?
3. Is there any indication of a sudden change in the number of pedestrians present?
4. Are there any abrupt changes in the positioning of objects like cones or barriers?
5. Is there any indication of a sudden change in the activity level of the area, such as a crowd gathering?

Robbery | 0.85 acc

1. Has a person suddenly left the frame in a way that seems out of place?
2. Are there any sudden changes in the number of people in the scene?
3. Is there a noticeable change in the overall activity level in the scene?
4. Are there any sudden changes in the background or environment that could indicate tampering?
5. Are there any individuals who appear to be in a hurry or behaving unusually?

Shooting | 0.62 acc

1. Are there any significant changes in the positions of individuals within the scene?
2. Do the individuals exhibit any unusual behaviors or actions?
3. Are there any sudden changes in the number of people present in the scene?
4. Does the environment show any unexpected changes in lighting or shadows?
5. Are there any abrupt changes in the activities or interactions between the individuals?

Stealing | 0.83 acc

1. Are there any sudden movements or actions that deviate from the typical behavior of the individuals or objects in the scene?
2. Is there any noticeable change in the interaction between objects or individuals that seems unusual?
3. Are there any objects that appear to be moving unexpectedly or in a way that contradicts their expected behavior?
4. Is there any evidence of a sudden change in the environment, such as the appearance of new objects or the disappearance of existing ones?
5. Are there any abrupt changes in the posture or position of individuals that suggest an unexpected event?

Vandalism | 1.0 acc

1. Are there any sudden or unexpected changes in the number of individuals in the scene?
2. Do the individuals exhibit any irregular or non-repetitive patterns in their movements?
3. Are there any significant changes in the lighting conditions or shadows that are not typical for the setting?
4. Is there any unusual interaction between the individuals and the objects around them that deviates from the norm?
5. Are there any abrupt changes in the background elements or environment that are not typical for the setting?



2025 Video Surveillance Metadata Extraction with Multimodal LLM Agents through Iterative Dialogue Exploration Pedro Domingo

