

PrismVAU: Prompt-Refined Inference System for Multimodal Video Anomaly Understanding (Supplementary Material)

Iñaki Erregue^{1,2,4}

ierregal31@alumnes.ub.edu

Kamal Nasrollahi^{3,4}

kn@create.aau.dk

Sergio Escalera^{1,2,3}

sescalera@ub.edu

¹Universitat de Barcelona

²Computer Vision Center

³Aalborg University

⁴Milestone Systems

A. Prompts and Meta-prompts

In this section, we provide the meta-prompt used as user prompt for optimizing the textual anchors (see Fig. 1). The meta-prompt used for optimizing the MLLM prompts for the VAU task (see Fig. 2). In both cases, the **red** text corresponds to the placeholders representing initial solutions and their associated scores. The initial and optimized MLLM prompts for VAU are also reported in Fig. 3, where the **blue** text corresponds to the portions manually augmented to incorporate coarse anomaly predictions as additional context in JSON format. Finally, the optimized textual anchors for the XD-Violence dataset are displayed in Fig. 4.

You are an expert in designing visually grounded prompts for vision-language models, specifically for binary image classification tasks. Your goal is to generate two detailed and distinct prompts, one for normal scenes and one for abnormal scenes, that are visually grounded in the specific physical characteristics of each category. These prompts will be used for zero-shot classification with CLIP, so it's crucial that they clearly differentiate between typical, non-threatening actions and suspicious, abnormal behaviors.

For the normal, scenes describe typical, non-threatening human behavior in common public or indoor settings. Include physical traits of people (e.g., posture, clothing), common environments (e.g., streets, parks, offices), and ordinary activities (e.g., walking, sitting, talking). For the abnormal scenes, describe behavior that stands out as unusual, dangerous, or criminal. Include specific visual details about suspicious actions, the physical environment (e.g., restricted spaces, chaotic areas), objects that are part of the scene (e.g., weapons, stolen items), and the presence of distress, aggression, or accidents.

You must provide the output in the following format:

Normal Prompt: <Generated prompt for normal scenes>

Abnormal Prompt: <Generated prompt for abnormal scenes>

The following are examples of past prompts with their obtained classification scores. The score ranges from 0 to 100, where 100 is the best possible value. Use these examples and their scores as a reference to improve prompt construction.

Example 1:

- Normal Prompt: {initial normal prompt}
- Abnormal Prompt: {initial abnormal prompt}
- Score: {initial pair score}

(... more top-performing previous examples, sorted in ascending order ...)

Your goal is to create a more effective prompt that enhances performance in distinguishing abnormal from normal scenes. Prioritize visually grounded concepts, ensuring the descriptions are rich and capture defining physical traits, with adjectives and objects typically present in each. Do not reference scores in your response, just focus on achieving the most effective differentiation.

Figure 1. Meta-prompt used for optimizing textual anchors.

You are an expert prompt engineer specialized in designing prompts for Multimodal LLMs (Video-LLMs) to perform temporal anomaly detection in urban surveillance. Your goal is to propose an improved System Prompt and User Prompt that guide the model to:

1. Analyze object dynamics, motion, interactions, and context over time.
2. Detect criminal behaviors (e.g., assault, arson, burglary, vandalism, shoplifting) by attending to subtle or compound indicators such as sudden acceleration, unusual crowd clustering, suspicious objects and contextual mismatches.
3. Identify and report the exact time interval of the anomaly (in seconds, precise to 2 decimal places).
4. Assign an abnormality score (0.00-1.00, rounded to two decimal places) that reflects the severity of the detected behavior.
5. Generate a concise description (≤ 100 characters) summarizing what was observed and why it was flagged.

The MLLM should reason step-by-step before producing its final response. That is, it should:

- First, analyze motion patterns, object states, and interaction dynamics across time.
- Then, consider contextual clues to determine whether a behavior is anomalous.
- Finally, output the result in this specific JSON format:

```
{
  "start.time": <float seconds, 2dp>,
  "end.time": <float seconds, 2dp>,
  "abnormality": <float between 0.00 and 1.00, 2dp>,
  "description": <concise summary, max 100 characters>
}
```

If no anomaly is detected, respond with {}.

You will now see examples of previous prompt pairs with their obtained performance score, where 100 is the best possible value. Use their language, structure, and specificity as inspiration to improve prompt design.

Example 1:

- System Prompt: {initial system prompt}
- User Prompt: {initial user prompt}
- Score: {initial pair score}

(... more top-performing previous examples, sorted in ascending order ...)

Task:

- Generate a stronger System Prompt + User Prompt pair. Focus on:
- Encouraging temporal attention to motion patterns and interactions
- Emphasizing precise format instructions
- Being concise yet comprehensive

Return ONLY the following format, with no additional text, explanations, or scores:

System Prompt: <...>
 User Prompt: <...>

Figure 2. Meta-prompt used for optimizing the VAU prompts of the MLLM.

System Prompt: You are a Video-LLM for urban surveillance anomaly detection. Your task is to analyze video footage to find anomalies and criminal behavior. If present, localize them temporally by predicting start and end timestamps, assign an anomaly score to the entire anomalous event, and generate a textual explanation that links visual evidence to its prediction.

You must always return your results in the following JSON format:

```
{
  "start.time": <float in seconds, precise to 2 decimal places>,
  "end.time": <float in seconds, precise to 2 decimal places>,
  "abnormality": <float between 0.00 and 1.00, rounded to 2 decimal places - represents how severe the anomaly is>,
  "description": <string, concise summary of the anomaly, max 100 characters, detailing key observations and evidence>
}
```

If no anomalies are found, you should return an empty JSON, {}.

User Prompt: Analyze the following video and return only the required JSON.

(a)

System Prompt: As a Video Language Model specialized in urban surveillance, your primary mission is to meticulously analyze video content for temporal anomalies in object dynamics, motion patterns, interactions, and contextual clues over time. Prioritize identifying criminal behaviors such as assault, arson, burglary, vandalism, and shoplifting by focusing on subtle or compound indicators like sudden acceleration, unusual crowd clustering, suspicious objects, and contextual mismatches. Maintain a high level of temporal attention to motion patterns and interactions, ensuring precise frame-by-frame analysis to make accurate decisions. Assign an abnormality score between 0.00 and 1.00, rounded to two decimal places, based on the severity of the detected behavior. Provide a concise summary (≤ 100 characters) summarizing what was observed and why it was flagged. If no anomaly is detected, return an empty JSON. Use the provided statistical summary as a preliminary signal but ensure your own thorough, explicit analysis is conducted.

Document your findings in the following JSON format, providing detailed reasoning and explicit documentation of all evidence used in each step:

```
{
  "start.time": <float in seconds, precise to 2 decimal places>,
  "end.time": <float in seconds, precise to 2 decimal places>,
  "abnormality": <float between 0.00 and 1.00, rounded to 2 decimal places - represents how severe the anomaly is>,
  "description": <string, concise summary of the anomaly, max 100 characters, detailing key observations and evidence>
}
```

User Prompt: A preliminary system has provided the following statistical summary as a prior signal:

```
{
  "coarse.video.label": <string>,
  "coarse.video.abnormality.score": <float>,
  "coarse.temporal.region": <list>,
  "coarse.frame.abnormality.stats": {
    "max": <float>,
    "time.max": <float>,
    "median": <float>,
    "mean": <float>,
    "std": <float>,
    "abnormal.seconds": <float>
  }
}
```

Carefully examine the provided video footage, ensuring that you adhere strictly to the guidelines outlined in the System Prompt. Focus on precise temporal attention to motion patterns and interactions, prioritizing frame-by-frame analysis while considering contextual clues to identify anomalous behaviors. Assign an abnormality score accurately and provide a concise summary as specified. If no anomaly is found, respond with an empty JSON.

```
{
  "start.time": <float in seconds, precise to 2 decimal places>,
  "end.time": <float in seconds, precise to 2 decimal places>,
  "abnormality": <float between 0.00 and 1.00, rounded to 2 decimal places - represents how severe the anomaly is>,
  "description": <string, concise summary of the anomaly, max 100 characters, detailing key observations and evidence>
}
```

(b)

Figure 3. Comparison of initial system and user MLLM prompts for VAU (a) with their optimized counterparts (b).

Normal Prompt: In a tranquil, well-lit suburban neighborhood during the daytime, individuals are casually dressed in comfortable attire suitable for leisure activities. They engage in routine behaviors such as walking dogs, children playing on swings and slides, and couples enjoying picnics under the shade of mature trees. The environment features clean pathways lined with blooming flowers, lush greenery, and clear blue skies overhead, reflecting a sense of calm and normalcy. Individuals display relaxed postures, indicating ease and contentment, with no signs of distress or urgency present in this peaceful community space.

Abnormal Prompt: Amidst the shadowy corners of an urban alleyway at night, individuals exhibit secretive behavior with tense body language, suggesting potential illicit intent. The surroundings are cluttered with debris, graffiti, and restricted spaces, creating a tense atmosphere indicative of danger and criminal intent. People are dressed in dark clothing that blends into the shadows, adding to the sense of illicit activity. Signs of recent criminal incidents, such as litter, abandoned vehicles, and visible weapons, contribute to an overall sense of urgency and imminent threat. Visible objects associated with criminal activity further enhance the sense of danger and criminal intent, while individuals display signs of distress, aggression, or involvement in suspicious activities, amplifying the imminent threat and criminal intent in the scene.

Figure 4. Distilled textual anchors of the XD-Violence dataset.

B. Results on XD-Violence

In this section, we present the baseline results on the XD-Violence dataset. A new pair of textual anchors was distilled to account for dataset-specific notions of abnormality, achieving a substantial improvement in performance over unoptimized anchors. In contrast, the MLLM prompts from the UCF-Crime dataset were reused directly, demonstrating their cross-dataset generalizability in detecting anomalies through motion trajectories, object interactions, and contextual mismatches. The results in Tab. 1 further highlight the complementary strengths of these methods; textual anchors achieve higher AP, reflecting stronger global discriminability, while the MLLM provides superior temporal precision, as reflected in the F1 scores.

Method	AP (%)	F1 (%)
TA	42.15	39.03
TA*	74.66	56.36
MLLM	59.44	26.81
MLLM*	65.07	59.98

Table 1. Baseline frame-level results on XD-Violence. TA denotes the textual anchors baseline, MLLM refers to VideoLLaMA-3 without coarse priors, and * marks the use of optimized prompts.

C. Performance Across Anomaly Types

As depicted in Fig. 5, PrismVAU outperforms both its components in 8 of 13 categories. Owing to its joint design, PrismVAU can compensate when either TA or the MLLM underperforms, yielding its highest ROC AUC on *assault* and lowest on *explosion*. The largest performance gains are observed in *vandalism* and *road accident* relative to TA, and in *stealing* relative to the MLLM.

Textual anchors tend to excel when anomalies exhibit distinctive visual patterns or strong visual-text alignment, particularly when static cues suffice, as in *shooting* (weapon, pose, and scene cues). In contrast, the MLLM performs better when context, relational reasoning, or subtle temporal cues are critical, such as in *vandalism* or *robbery*.

The most challenging categories are *explosion*, *abuse*, and *shoplifting*. Low performance on *explosion* may result from missed temporal cues due to sparse sampling and single-frame embeddings. *Abuse* likely suffers from high variability and ambiguous semantics. *Shoplifting* may involve subtle hand-object interactions under occlusion, making fine-grained behavior recognition difficult.

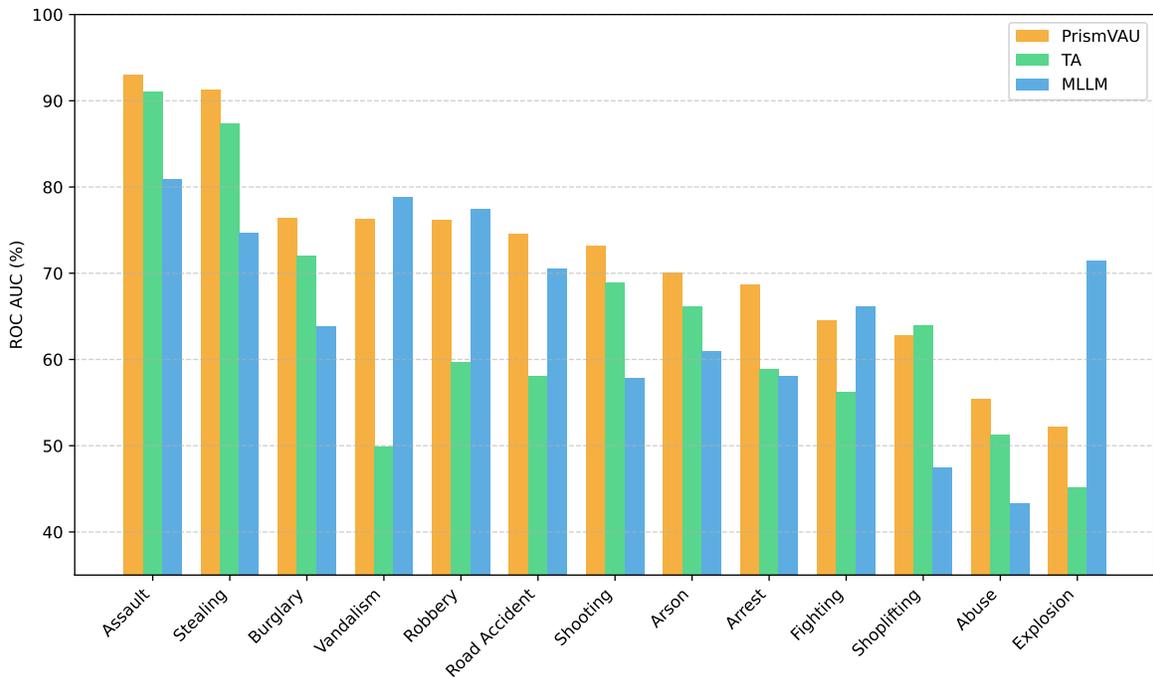


Figure 5. Per-category ROC AUC performance comparison of PrismVAU and its components on UCF-Crime.

D. Semantic Similarity Analysis

Directly evaluating PrismVAU’s generated explanations is challenging because UCF-Crime does not include textual annotations. To approximate their semantic alignment with the underlying anomalous events, we instead leverage the HIVAU-70k dataset. Built on top of UCF-Crime and XD-Violence, HIVAU-70k provides clip-, event-, and video-level free-text annotations that judge, describe, and analyze the presence or absence of anomalies.

Since the objective is to assess meaning rather than lexical overlap, an embedding-based similarity metric has been employed to evaluate explanation quality. Specifically, for each video V_i , textual embeddings are computed for all its predicted explanations $\{p_1, \dots, p_N\}$ and ground truth sentences $\{g_1, \dots, g_M\}$, after which a cosine similarity matrix is formed over all prediction-ground-truth pairs. Each prediction is assigned its maximum similarity to any ground-truth sentence, the video-level score is obtained by averaging these maxima, and the dataset-level score is computed as the mean across videos. MPnet (*all-mpnet-base-v2*) has been used as the text encoder due to its superior semantic representation capabilities, which enable reliable assessment of conceptual alignment between predicted and reference explanations.

Cosine similarity values can be interpreted approximately as: 0.0–0.2 (unrelated), 0.2–0.4 (weakly related), 0.4–0.6 (moderately aligned), 0.6–0.8 (strongly aligned), and 0.8–1.0 (nearly identical). PrismVAU obtains an average score of 0.5558, indicating moderate semantic alignment between predicted anomaly descriptions and ground-truth text. This outcome is consistent with the inherent challenges introduced by: short (sub-100-character) anomaly-focused predictions versus context-rich annotations, mismatches in temporal granularity and abstraction, variation in domain-specific vocabulary, and the absence of textual supervision during training. Despite these factors, the score reflects that PrismVAU reliably captures the correct semantic category of anomalous events, despite differences in phrasing or level of detail. The metric therefore provides a principled quantitative estimate of explanation quality and complements the anomaly-detection evaluation by demonstrating that PrismVAU’s explanations are both plausible and semantically meaningful.

E. Additional Qualitative Results

In this section, we present additional qualitative results and case studies on both the UCF-Crime and XD-Violence datasets. Ground-truth (GT) anomalous intervals are compared against the coarse anomaly curve from the textual anchors (TA), the step-like MLLM predictions, and the refined anomaly curve. Human-readable explanations generated by the MLLM are shown as speech bubbles. Bold ticks on the x-axis indicate video segment partitions used when processing the untrimmed video. Sample frames on the right illustrate key anomaly events.

Figure 6 illustrates PrismVAU’s anomaly detection results on a UCF-Crime test video containing an explosion event. In this example, the coarse anomaly curve is noisy, while the MLLM predictions are both accurate and temporally anticipatory, preceding the ground-truth labels. The explanations demonstrate the MLLM’s ability to associate the explosion with textual clues suggesting it is phone-related. A similar capability to leverage captions and textual evidence for anomaly description is observed in Fig. 7 and Fig. 8.

Figure 9 and Fig. 10 provide additional examples of how MLLM predictions refine and complement coarse anomaly curves. While textual anchors fail to capture the onset of the road accident or highlight the most critical arson moments, the MLLM predictions not only localize anomalies more precisely in time but also provide faithful descriptions of the events. For instance, in the arson video, PrismVAU flags not only the main anomalous event but also related aggressive behaviors.

The ability to detect subtle, non-annotated criminal behaviors is also showcased in Fig. 11 and Fig. 12. In the former, only the assault is annotated in the ground truth. Nonetheless, both modules of PrismVAU identify the triggering theft that precedes the assault. In the latter, an initial aggressive interaction between hockey players is flagged several seconds before the final altercation. Figure 13 further illustrates this capability: not only is the car theft detected, but also the subsequent escape, including reckless driving.

Figure 14 highlights PrismVAU’s ability to capture a variety of distinct anomalous activities within a single riot scenario, including car damage, arson, vandalism, and police confrontation. Finally, in Fig. 15, although the MLLM provides a slightly inaccurate description/cause (the car crashes while avoiding a person but does not actually hit it), it correctly identifies the consequences of the accident.

Additionally, Fig. 16 and Fig. 17 depict interactive conversations after PrismVAU’s initial response. This interactive chat interface capability lets users engage with the model to audit system errors, elaborate on correct predictions, and explore counterfactual or hypothetical scenarios.

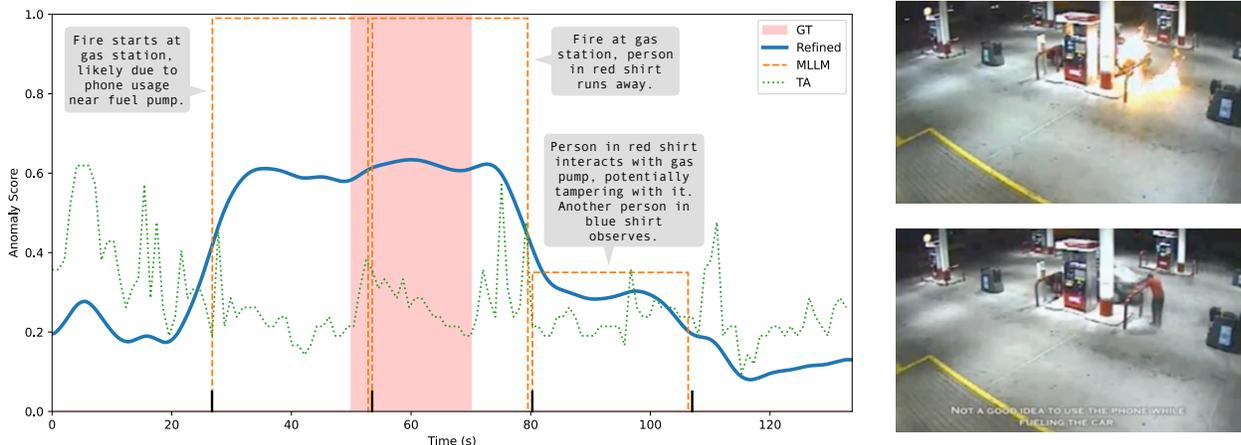


Figure 6. Results on *Explosion002_x264* from UCF-Crime.

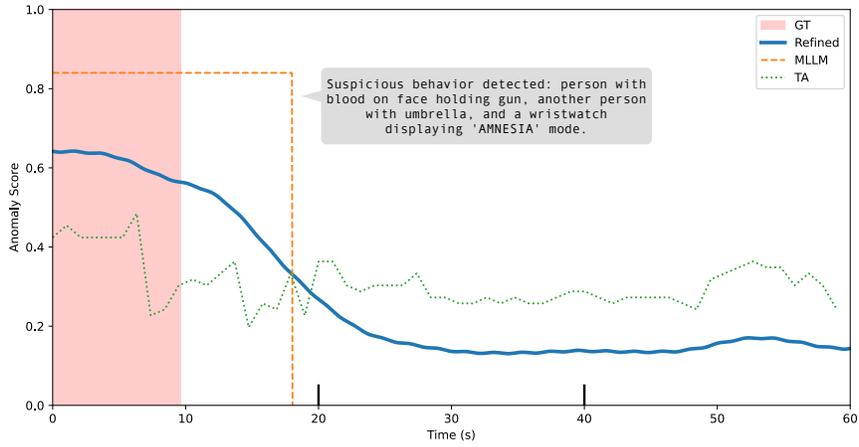


Figure 7. Results on *Kingsman.The.Secret.Service.2014_#00-22-10_00-23-10_label_B2-0-0* from XD-Violence.

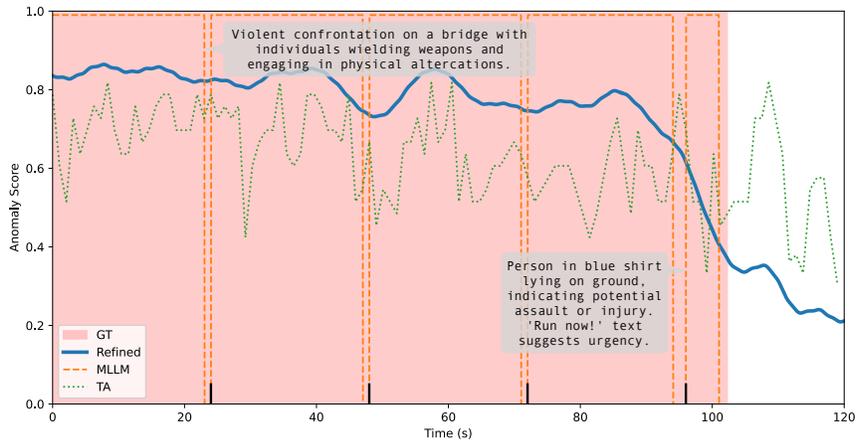


Figure 8. Results on *Young.And.Dangerous.I.1996_#0-46-57_0-48-57_label_B1-0-0* from XD-Violence.

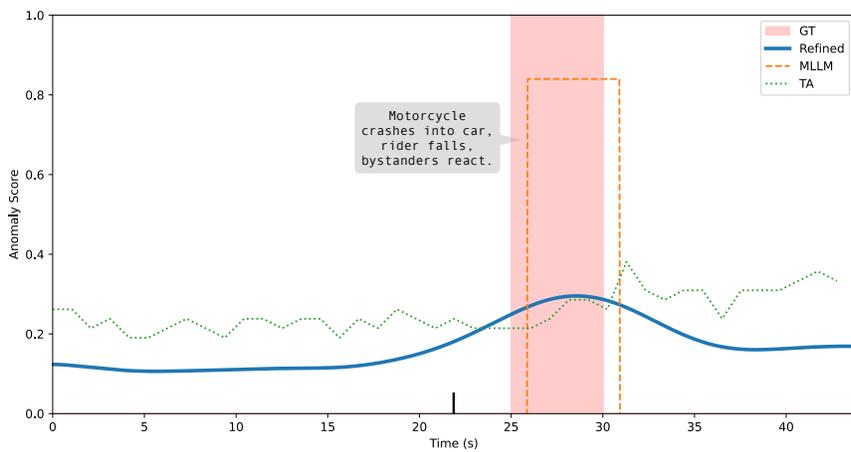


Figure 9. Results on *RoadAccidents019_x264* from UCF-Crime.

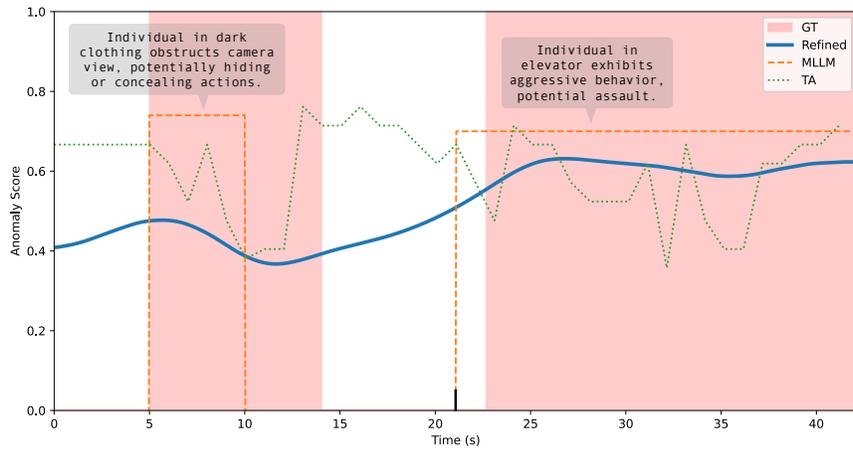


Figure 10. Results on *Arson011_x264* from UCF-Crime.

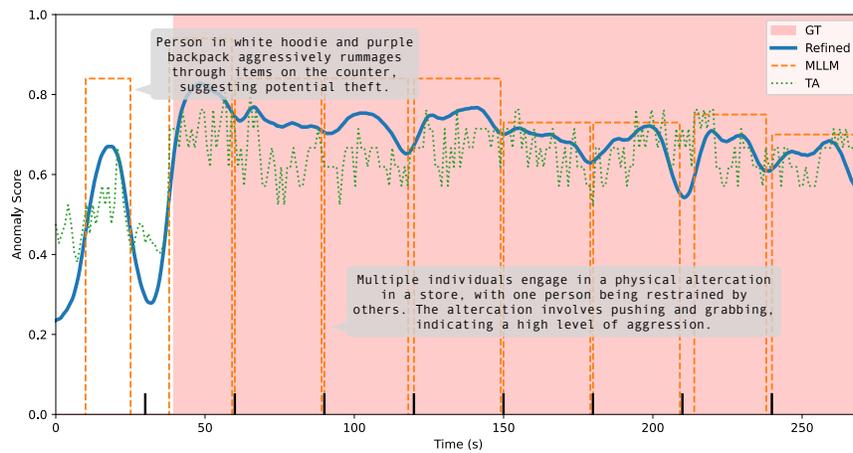


Figure 11. Results on *Assault006_x264* from UCF-Crime.

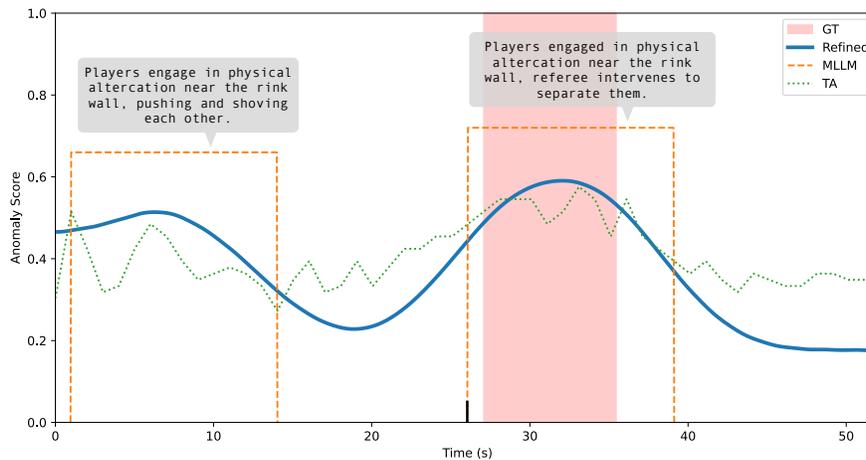


Figure 12. Results on *v=pMtu7fOHdII_#1_label_B1-0-0* from XD-Violence.

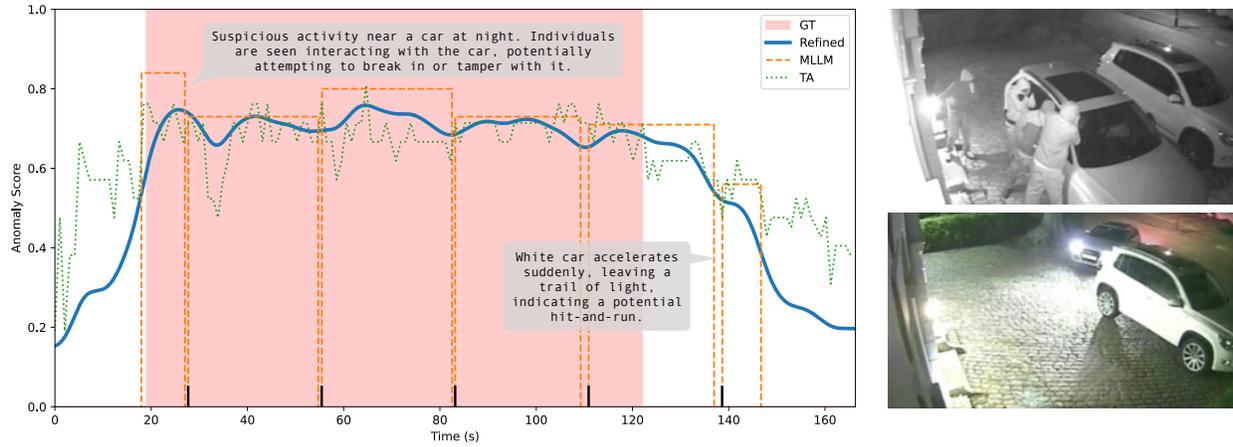


Figure 13. Results on *Stealing058_x264* from UCF-Crime.

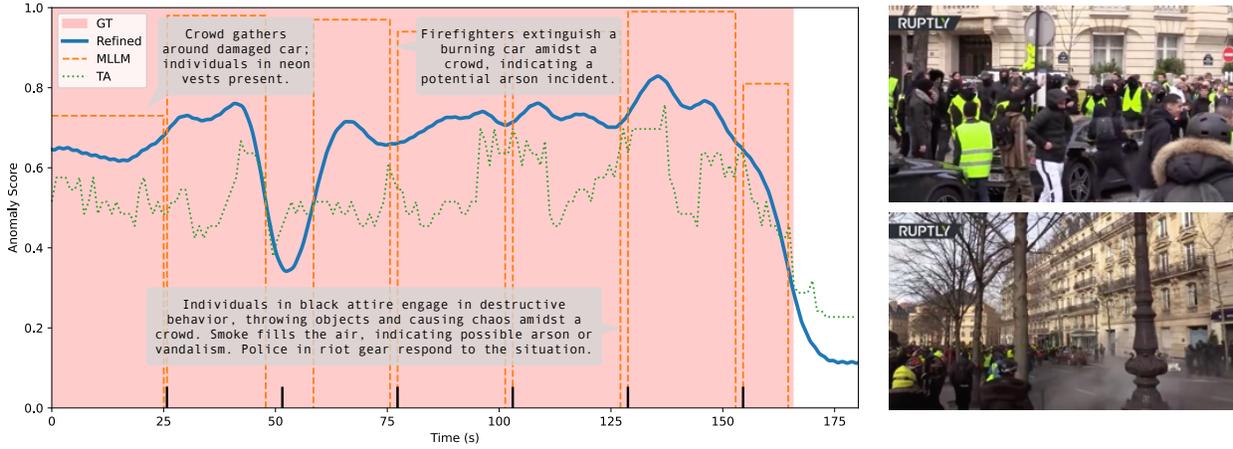


Figure 14. Results on *v=ICnreR1hxP0_#1_label_B4-0-0* from XD-Violence.

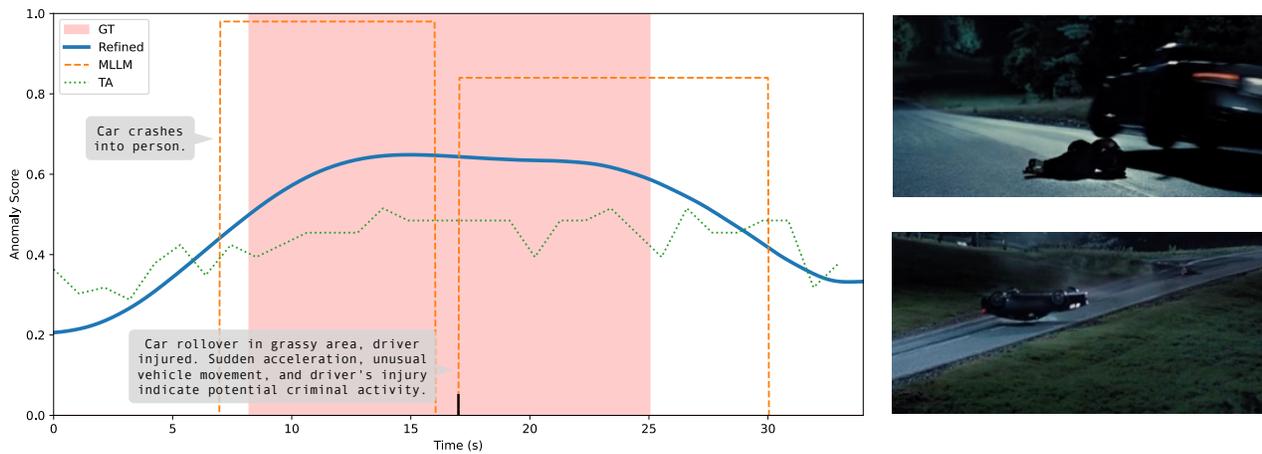


Figure 15. Results on *Casino.Royale.2006_#01-46-40_01-47-14_label_B6-0-0* from XD-Violence.

PrismVAU: {"abnormality": 0.00, "description": "No visible criminal behavior detected in the video footage."}

User: Describe the scene and explain why it is considered normal.

PrismVAU: The video shows two individuals entering an elevator in a building lobby. The scene is considered normal because there are no unusual or anomalous behaviors, such as sudden acceleration, unusual crowd clustering, suspicious objects, or contextual mismatches that could indicate criminal activity. The individuals enter the elevator and the doors close. The scene concludes with the elevator moving upwards.

User: What if the individuals were running towards the elevator instead of walking?

PrismVAU: If the individuals were running towards the elevator instead of walking, it could be considered an abnormality. The sudden acceleration and unusual motion patterns would indicate a potential emergency or criminal activity, such as an attempt to escape or a robbery. In this case, the abnormality score would be higher, and further analysis would be needed to determine the specific nature of the anomaly.

Figure 16. Interactive dialogue with PrismVAU on a normal scene, *Normal_Videos_783_x264* from UCF-Crime.

PrismVAU: {"start_time": 10.00, "end_time": 19.00, "abnormality": 0.85, "description": "White van collides with motorcyclist, causing fire. Motorcyclist falls, and another rider rushes to assist."}

User: Which vehicle's actions led to the accident?

PrismVAU: The white van's actions led to the accident.

User: What could the drivers have done differently to avoid the collision?

PrismVAU: The drivers could have maintained a safe following distance, stayed alert to their surroundings, and been prepared to react to unexpected situations on the road. Regular vehicle maintenance, such as checking brakes and tires, can also prevent accidents.

User: What would have happened if the white van continued straight instead of turning?

PrismVAU: If the white van continued straight instead of turning, it is likely that the accident with the motorcyclist would have been avoided.

Figure 17. Interactive dialogue with PrismVAU on an abnormal scene, *RoadAccidents133_x264* from UCF-Crime.