

Logical Consistency Optimization for Few-Shot Weakly Supervised Video Anomaly Detection

Supplementary Materials

Hantao Zheng¹, Ning Han^{*2}, Yawen Zeng¹, Hegui Zhu³, and Hao Chen¹

¹Hunan University, China, {zhenghantao, yawenzeng11, chenhao}@hnu.edu.cn

²Xiangtan University, China, hanninginf@gmail.com

³Northeastern University, China, zhuhegui@mail.neu.edu.cn

This document provides supplementary material for our paper. Section A offers a detailed, step-by-step walkthrough of our entire pipeline, from initial visual parsing to final keyword extraction. We present the exact prompts used to interface with the language models at each stage, accompanied by concrete output examples from a single, coherent case study. Sections B and C provide additional experimental results that further support and validate the claims made in the main text.

A Detailed Method Walkthrough and Prompts

A.1 Visual Parsing

The pipeline begins with Visual Parsing. In this step, a Vision-Language Model (VLM) processes a video sample, selected from our few-shot set, to generate a textual description (D). The process is guided by the targeted prompt p_d shown below.

- Input $Video$ and p_d .



p_d = Briefly describe this video. It may contain frames of anomalous events. Do not include any unnecessary details or descriptions.

- Output video descriptions D .

D = The video depicts a scene of a road accident involving a white SUV. The vehicle is seen driving on the road, and suddenly, it crashes into a motorcycle, causing the motorcycle to fall onto the ground. The SUV's hood is crumpled, and the motorcycle is lying on the ground, indicating a severe impact.

A.2 Initial Guideline Generation

The resulting textual description (D) then serves as the input for Guideline Generation. We employ a sophisticated prompting strategy (p_{gen}) to guide a Large Language Model (LLM) in producing an initial set of detection guidelines

^{*}Corresponding author.

(G_{gen}). The prompt directs the LLM to analyze the description from three crucial facets: human activities, states, and the overall scene context.

- Input video descriptions D and p_{gen} .

```

 $p_{gen} = [$ 
  {"role": "system",
   "content": "I am an anomaly content analyst working with the {data_full_name} dataset. My job is to determine whether the description contains any anomaly-related information. If it does, I will generate relevant anomaly detection guidelines. If not, I will respond with 'no.'"},
  {"role": "user",
   "content": "Please check if there are any descriptions of exceptions related to the anomaly keywords in the context. If so, generate relevant guidelines. If not, reply with 'normal.'"},
  {"role": "assistant",
   "content": "
    **Guidelines for Anomaly Human Activities:
    1.
    **Guidelines for Anomaly Human States:
    1.
    **Guidelines for Anomaly Scenes:
    1. "},
  {"role": "user",
   "content": "You are provided with descriptions: {context}. Identify the Anomaly Guidelines from this information. Think through the process step by step. Begin with abstract concepts and then move towards specific activities states and scenes. List your findings using concise terms rather than full sentences." } ]

```

- Output guidelines as G_{gen} .

```

 $G_{gen} =$  **Anomaly Human Activities:**
1. Sudden veering of vehicle causing collision
2. Motorcycle rider falls onto road
3. Vehicle occupants visible after crash (possible ejection or exposure)
**Anomaly Human States:**
1. Motorcycle rider impacted, potential for injury
2. SUV occupants potentially unharmed post-crash
**Anomaly Scenes:**
1. Chaotic aftermath with debris scattered
2. Vehicle overturning and lying on its side
3. Wet road conditions indicating recent weather impact

```

A.3 Logical Consistency Optimization Framework (LCO)

The initial guidelines (G_{gen}) are then refined by our core **Logical Consistency Optimization (LCO)** framework. This module introduces a meta-reasoning loop to ensure the final guidelines are not only accurate but also logically sound. The three-step process is detailed below.

A.3.1 Logical Reasoning Generation

This module takes the video description D and the generated guideline G_{gen} as input, and outputs an explainable logical reasoning text L_{reason} , revealing the logical relationship between video content and the guideline.

- Input video description D , generated guideline G_{gen} and $p_{logical}$.

```

plogical = [
  {"role": "system",
   "content": "I am a logical reasoning analyst working with the {data_full_name} dataset. My task is to thoroughly analyze all provided materials and perform accurate logical reasoning."},
  {"role": "user",
   "content": "Based on the descriptions and existing guidelines, analyze whether the guidelines are correct and conduct faithful logical reasoning."},
  {"role": "assistant",
   "content": ""
   **Logical Reasoning:
   1. Identify the logical relationships within the context.
   2. Evaluate the accuracy of the existing guidelines in reflecting these relationships.
   3. Propose modifications or confirmations for each guideline based on the analysis."},
  {"role": "user",
   "content": "You are provided with descriptions: {context} and existing guidelines: {guidelines}. Based on these descriptions and existing guidelines, analyze whether the derived guidelines accurately reflect the descriptions. Conduct accurate logical reasoning step by step." ]

```

- Output logical reasoning text L_{reason} .

L_{reason} = To conduct a thorough analysis and logical reasoning based on the provided descriptions and guidelines, we need to examine whether the guidelines accurately encapsulate the scenarios depicted in the video descriptions. The guidelines effectively capture several crucial aspects of road accident scenarios and align with video descriptions. With adjustments, they can provide an even clearer framework for analysis.

A.3.2 Consistency Scoring

This module takes the video description D , the generated guideline G_{gen} , and the logical reasoning L_{reason} as input, and outputs a consistency score S_{belief} , indicating the alignment between the reasoning and the guideline.

- Input video description D , generated guideline G_{gen} , logical reasoning L_{reason} and p_{belief} .

```

pbelief = [
  {"role": "system",
   "content": "I am a belief score evaluator working with the {data_full_name} dataset. My task is to thoroughly analyze all provided materials and perform accurate score evaluations."},
  {"role": "user",
   "content": "Based on the descriptions, guidelines, and logical reasoning, conduct a faithful belief score evaluation. Scoring Guidelines: 'A' represents Confidently Correct (100%), 'B' indicates Correct with Moderate Confidence ([75%, 1]), 'C' signifies Incorrect with Some Confidence ([50%, 75%]), 'D' denotes Confidently Incorrect ([25%, 50%]), and 'E' reflects No Direct Answer ([0%, 25])."},

```

```

{"role": "assistant",
"content": "A"},
{"role": "user",
"content": "Given descriptions: {context}, guidelines: {guidelines}, and logical reasoning:
{logical_reasoning}, determine if the derived guidelines match the descriptions and assess the logical
consistency of the logical reasoning. Answer with a single letter from the scoring guidelines ('A', 'B', 'C',
'D', 'E')."} ]

```

- Output consistency score S_{belief} .

```

 $S_{belief} = B.$ 

```

A.3.3 Conditional Guideline Refinement

When the initial logical consistency score S_{belief} is below the threshold, the LCO module generates an optimized guide G_{refine} by fusing the initial guide G_{gen} with the logical reasoning L_{reason} . It is subsequently rescored for logical consistency $S_{rebelief}$. If $S_{rebelief} > S_{belief}$, the optimized guideline is retained; otherwise, the original guideline is maintained, thus ensuring that the final output of the anomaly detection guideline possesses higher consistency and confidence.

- Input video description D , generated guideline G_{gen} , logical reasoning L_{reason} and p_{refine} .

```

 $p_{refine} = [$ 
  {"role": "system",
  "content": "I am an anomaly guidelines refiner working with the {data_full_name} dataset. My task is
to evaluate whether the existing guidelines align with logical reasoning based on all provided descriptions.
If they align, I will output the existing guidelines. If not, I will generate new guidelines that faithfully
reflect the original descriptions and logical reasoning."},
  {"role": "user",
  "content": "Verify whether the existing guidelines are consistent with logical reasoning. If they align,
output the existing guidelines. If not, generate new guidelines that accurately reflect the original descrip-
tions and logical reasoning."},
  {"role": "assistant",
  "content": "
  **Guidelines for Anomaly Human Activities:
  1.
  **Guidelines for Anomaly Human States:
  1.
  **Guidelines for Anomaly Scenes:
  1."},
  {"role": "user",
  "content": "You are provided with original descriptions: {context}, existing guidelines: {guideline_init},
and logical reasoning: {guideline_logical_reasoning}. Determine if the existing guidelines are consistent
with the logical reasoning. Think through the process step by step. If they are consistent, list the existing
guidelines following the format: start from an abstract concept and then move to concrete activities, states,
and scenes. Use short terms rather than full sentences."} ]

```

- Output guidelines G_{refine} .

$G_{refine} =$ **Anomaly Human Activities:**

1. Unsafe driving leading to crashes.
2. Interaction with road medians and guardrails.
3. Collisions involving multiple vehicles.

Anomaly Human States:

1. Being thrown from vehicles during crashes.
2. Motorcyclists injured and lying on the ground.

Anomaly Scenes:

1. Vehicles flipping over.
2. Debris scattered on roads.
3. Vehicles in precarious positions.

A.4 Guideline Aggregation

The input to the guideline aggregation module is the logically optimized guideline set $G_{logical}$, which is composed of both the refined guidelines G_{refine} and the original generated guidelines G_{gen} . Leveraging the LLM as an aggregator, the model consolidates redundant or biased entries and outputs a refined guideline set G_{agg} , where frequently recurring, high-confidence guidelines are retained.

- Input guidelines $G_{logical}$ and p_{agg} .

```
 $p_{agg} = [$   
  {”role”: ”system”,  
    ”content”: ”I am an anomaly content analyst working with the {data_full_name} dataset. My task is to  
    create guidelines for detecting abnormal human activities, unusual human states, and irregular scenes.”},  
  {”role”: ”user”,  
    ”content”: ”You are provided with sets of guidelines for anomaly detection. From the organized  
    anomaly guidelines, extract and list all the provided anomaly guidelines.”},  
  {”role”: ”assistant”,  
    ”content”: ”  
    **Guidelines for Anomaly Human Activities:  
    1.  
    **Guidelines for Anomaly Human States:  
    1.  
    **Guidelines for Anomaly Scenes:  
    1. },  
  {”role”: ”user”,  
    ”content”: ”You are provided with a list of guidelines: {guidelines}. Identify the anomaly-related  
    guidelines from this list. Think through the process step by step and respond following the specified  
    format.”} ]
```

- Output guidelines G_{agg} .

$G_{agg} =$ **Anomaly Human Activities:**

1. Vehicle Collision with pedestrians or other vehicles.
2. Erratic driving behaviors, such as veering off-road or hit-and-run.
3. People rushing to provide aid or running away from an accident scene.
4. People lying on the ground post-collision.
5. Helping injured individuals.

****Anomaly Human Status:****

1. Individuals lying on the ground indicating possible injury.
2. People in visible shock, distress, or panic.
3. Injured pedestrians or vehicle occupants post-accident.
4. People scattered at the scene due to a collision.

****Anomaly Scenes:****

1. Chaotic aftermath of road accidents involving multiple vehicles or pedestrians.
2. Vehicles in states of disrepair post-collision, such as overturned cars.
3. High-impact, busy street scenes with traffic jams or vehicle pile-ups.
4. Presence of emergency response scenes with visible commotion.
5. Disarray involving obstructions to regular traffic conditions.

A.5 Keyword Extraction

The input to the keyword extraction module is the aggregated guideline set G_{agg} . Using LLM-based semantic analysis, representative keywords are extracted to form the final keyword set $K = \{keyword_0, keyword_1, \dots, keyword_m\}$, which serves as compact and high value cues for the detection of downstream anomalies.

- Input guidelines G_{agg} and $p_{keyword}$.

```

pkeyword = [
  {"role": "system",
   "content": "You are an abnormal content analyst working with the {data_full_name} dataset. Your task is to extract the most relevant words related to abnormal human activities, class states, and scenarios from a provided word list."},
  {"role": "user",
   "content": "You will receive a set of guidelines for detecting abnormal activities, states, and scenes. Extract the most relevant anomaly-related words from the following word list. Format the output as a list: ['anomaly word1', 'anomaly word2', ...]"},
  {"role": "assistant",
   "content": "['anomaly word1', 'anomaly word2', ... ]"},
  {"role": "user",
   "content": "From the provided word list ({cleaned_data}), extract the most relevant words that align with the guidelines for detecting anomalies related to violence, crime, danger, emergencies, destruction, and harm. Return the results in a list format: ['word1', 'word2', ...]"} ]

```

- Output keyword set K .

```

K = [collision, panic, injured, chaotic, overturned, disrepair, pile-ups, commotion, disarray, dangerous, damage, accidents].

```

B Experimental Details for Reproducibility

B.1 Few-Shot Sample Identifiers

To ensure full reproducibility of our few-shot learning experiments, Table 1 provides the complete list of video identifiers used for generating the initial guidelines. For each dataset, one video per anomaly category was randomly selected from the official training split.

Dataset	Category	Video Identifier / Filename
UCF-Crime	Abuse	Abuse023_x264
	Arrest	Arrest004_x264
	Arson	Arson024_x264
	Assault	Assault001_x264
	Burglary	Burglary004_x264
	Explosion	Explosion009_x264
	Fighting	Fighting013_x264
	Road Accidents	RoadAccidents085_x264
	Robbery	Robbery038_x264
	Shooting	Shooting036_x264
	Shoplifting	Shoplifting019_x264
	Stealing	Stealing020_x264
Vandalism	Vandalism020_x264	
XD-Violence	Abuse	Taken.2.UNRATED.EXTENDED.2012._#00-12-25_00-13-25_label_B5-0-0
	Explosion	Operation.Red.Sea.2018._#0-53-51_0-55-00_label_G-0-0
	Fighting	v=zzxg8qfqQPA_#1_label_B1-0-0
	Car Accident	v=0aOC63IZPTs_#00-04-00_00-05-19_label_B6-0-0
	Shooting	Deadpool.2.2018._#0-03-38_0-03-50_label_B2-0-0
	Riot	v=qtqQeJ315nw_#1_label_B4-0-0

Table 1: List of video samples and their corresponding categories used for the few-shot guideline generation phase.

C Additional Ablation Studies

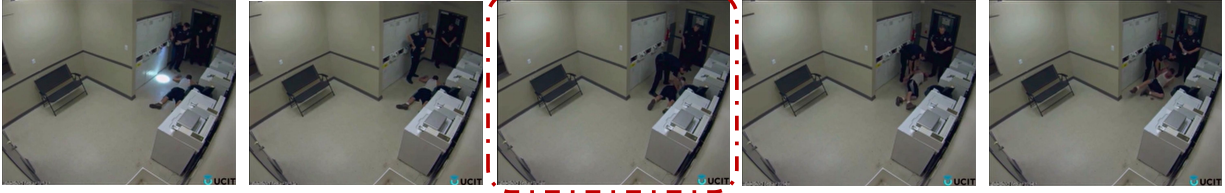
C.1 Qualitative Comparison of Vision-Language Models

Our framework is designed with modularity, allowing the visual parsing module to be replaced with different Vision-Language Models. To justify our choice of **LLaVA-NeXT-Video-7B-hf** in the main paper, we conducted a qualitative comparison against two other prominent models in a similar parameter class: **Qwen2-VL-7B-Instruct** and **CogVLM2-Llama3-Caption**. The goal was to assess which VLM provides the most informative and semantically rich descriptions suitable for our downstream logical reasoning task.

We selected representative frames from three distinct anomaly videos. The verbatim descriptions generated by each model, along with a case-by-case analysis, are presented from Table 2 to 4. In the generated descriptions within these tables, text highlighted in **green** denotes phrases that correctly identify the anomalous event or its key semantic components. Conversely, text highlighted in **red** denotes phrases that are either factually incorrect, too generic to capture the anomaly, or inconsistent with the context of the event.

Case 1: Arrest (Arrest007_x264, frame 1760)

In the “Arrest” scenario, Table 2, LLaVA-Next accurately identifies the core event: “seemingly being arrested by security personnel” and correctly identifies the actors and state “security personnel,” “person lying on the ground,” “restrained state”. In contrast, CogVLM2 misidentifies the location as a “public restroom” and focuses on irrelevant static objects like “a roll of toilet paper.” Qwen2-VL provides a highly generic description “man was leaning over a woman” that, while factually observing a person on the floor, completely misses the critical context of the arrest and the presence of security forces.



VLM Model	Generated Description
LLaVA-Next	This video appears to be a surveillance footage of an incident that occurred in a public place, possibly in a shopping mall or similar indoor facility. In the video, we see two people seemingly being arrested by security personnel . A person lying on the ground may be in a restrained state , while another person stands with the assistance of security personnel. Security personnel wearing uniforms indicate that they are members of the security personnel. The background is an indoor area with tiled flooring and benches. The lighting is artificial, and the environment indicates that this is a high ceiling public space. The video quality is not high, and the recording is a bit blurry, making it difficult to distinguish specific details. Details. Security personnel are focused on these individuals, and their actions indicate that they are being detained or restricted. The exact nature of the event cannot be determined solely from the video, but it appears to be a security related situation .
CogVLM2	Surveillance images show a series of incidents that occurred in a public restroom , with one person lying on the floor, and two others standing nearby, one holding a phone. The scene includes a bench, a mirror, and a sink, with security cameras displaying public or commercial environments. The presence of the "UCIT" watermark indicates that this video is from a security camera system. The video also includes a close-up of a roll of toilet paper .
Qwen2-VL	The video shows a man and a woman in a room with a white floor and white walls. The man was leaning over a woman lying on the floor, wearing a black shirt. There is a bench in the room, with a man wearing a black shirt and a woman wearing a black top and black shorts. There is a sink and a bench in the room, with a white wall in the background. The video seems to be recorded by a security camera.

Table 2: Qualitative comparison for **Arrest007_x264** (frame 1760).

Case 2: Shoplifting (Shoplifting033_x264, frame 688)

For the "Shoplifting" example, Table 3, LLaVA-Next is the only model that successfully identifies the anomalous action. It explicitly labels the actor as a "thief" and the event as an "obvious act of theft". This event-centric identification is critical. Conversely, Qwen2-VL fundamentally misinterprets the scene as a "typical shopping experience" and "considering purchasing". CogVLM2 provides a neutral, descriptive narrative "inspecting items", "casual and enticing" that completely fails to capture the "theft" context, rendering its description useless for anomaly detection.

Case 3: Robbery (Robbery102_x264, frame 1120)

In the "Robbery" scene, Table 4, LLaVA-Next successfully captures the confrontational nature of the event. It describes "physical altercations", identifies the specific struggle "restrained by a woman who grabbed the car door", and correctly contextualizes the scene as "moments of tension and conflict, which may be related to robbery attempts". CogVLM2 also identifies the "physical clashing" and "conflict", but provides less contextual detail about the specific actions. Qwen2-VL fails entirely to identify any conflict, providing a benign description of people "gathering" and "climb[ing] onto the car", which completely misses the anomalous event.

Concluding Analysis

The qualitative results, detailed from Table 2 to Table 4, clearly illustrate why **LLaVA-NeXT-Video-7B-hf** was chosen as the default VLM for our visual parsing module. For all three distinct anomaly categories, it consistently delivers a more event-centric and semantically rich narrative. It successfully identifies not only the actors and objects but also the core interaction and its critical context, such as "arrested by security personnel", "obvious act of theft" and "moments



VLM Model	Generated Description
LLaVA-Next	In the video, a man stands inside a clothing store surrounded by various items such as jackets, shirts, and other clothing. He seems to be browsing or possibly considering purchasing. The atmosphere of this store is relaxed and casual, with other customers in the background. The man was wearing a blue shirt and khaki shorts, seemingly looking at the items on display. The surveillance footage shows that there is a person in the store, possibly a thief, who was caught stealing. It can be seen that the person took a jacket off the shelf and took it away, which is an obvious act of theft. The video also shows other customers in the store, some of whom are looking at items. ”And others are browsing or walking around. Thieves are the main focus of the video, and their behavior is recorded in detail.
CogVLM2	At a modern retail store, a man wearing a blue shirt and khaki shorts initially placed a dark jacket on a table with neatly folded clothes. Then he checked a pair of gray pants surrounded by men’s clothing and accessories. His attention shifted to a table wearing a dark jacket, white shirt, and black pants, with neatly arranged clothes and bags inside. He held a dark green jacket, then a dark brown hat, and later a dark green scarf, surrounded by hangers and bags. The atmosphere of this store was casual and enticing, emphasizing style and comfort.
Qwen2-VL	The video shows a man standing in a clothing store. He is wearing a blue shirt and beige shorts, looking at a display stand. The store seems to be well lit, displaying a variety of clothing. The man seems to be browsing the store, possibly looking for something to buy. The store has a modern and tidy appearance, displaying various clothing and accessories. The man seems to be focused on the items on the display stand, and he seems to be considering purchasing. Overall, the video captures the typical shopping experience of a modern clothing store.

Table 3: Qualitative comparison for **Shoplifting033_x264** (frame 688).

of tension and conflict”. This level of semantic detail is essential for our downstream logical reasoning module to perform meaningful logical reasoning.

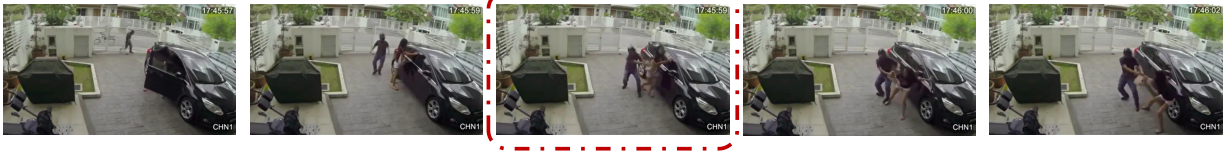
In contrast, **CogVLM2-Llama3-Caption** tends to overemphasize static or irrelevant objects like “a roll of toilet paper” or provide neutral descriptions that, while partially correct as in “conflict” lack the rich contextual details of the interaction. **Qwen2-VL-7B-Instruct** often provides overly generic or factually inaccurate descriptions that fail to capture the anomalous context, for instance describing a shoplifting scene as a “typical shopping experience”. Therefore, while the architecture of our framework supports replacing the VLM, the output quality of LLaVA-NeXT-Video-7B-hf provides the most reliable and informative foundation for our method.

C.2 Ablation on Confidence Bins

To justify our choice of a five-bin grading system for the logical consistency score (S_{belief}), as mentioned in the main paper, we provide a detailed explanation and empirical validation. Our approach to discretizing confidence scores is inspired by the widely adopted A–F letter grading system in education[1], which maps continuous performance into distinct, interpretable categories. This aligns with best practices in confidence calibration research, where such discretization enhances the interpretability of the certainty of a model [2].

To assess the robustness of our five-bin scheme, we conducted an ablation study comparing it with three-bin and seven-bin alternatives. The details of each scheme and their corresponding performance on the XD-Violence dataset are presented in Table 5.

The results clearly indicate that both the three-bin and seven-bin alternatives lead to a drop in performance compared to our chosen five-bin setup. This suggests that three bins may be too coarse to capture meaningful confidence levels, while seven bins may be too granular, potentially overfitting to the LLM’s scoring nuances. Therefore, the



VLM Model	Generated Description
LLaVA-Next	The video depicts a scene where a group of people are engaged in physical altercations . In the foreground, there seems to be a man opening the car door, possibly to enter and exit the vehicle. He was restrained by a woman who grabbed the car door, preventing him from opening it further . The woman was wearing a sleeveless top and shorts, and her facial expression indicated that she was trying to stop the man from opening the car door. In the background, there seem to be two other people observing the situation. One of them is a woman standing to watch the interaction, and the other is a man also observing the scene. "The background seems to be a residential area, with a house and a fence visible in the background. The timestamp in the video shows that it was taken at 11:11:11. This video captures moments of tension and conflict , which may be related to robbery attempts or similar situations.
CogVLM2	There are two people in the image, one wearing a dark blue shirt and the other wearing a light colored top, who are physically clashing with a black car in the lane. The scene includes a security camera timestamp indicating that the event is being recorded. As the conflict unfolded, one person was holding another person's arm , with a black bag lying on the ground and a white building with a balcony in the background. Conflict seems to be the central focus , with no other people or objects visible.
Qwen2-VL	The video shows a group of people gathering around a car in the lane. One person gets on the car while others stand beside it. The scene seems to take place in a residential area, with a house visible in the background. The people in the video seem to be engaged in some activity , possibly related to the car.

Table 4: Qualitative comparison for **Robbery102_x264** (frame 1120).

# of Bins	Binning Scheme Description	AUC (%)
3	A: Confidently Correct ([75, 100%]) B: Correct with Uncertainty ([50, 75%]) C: Incorrect or No Answer ([0, 50%])	70.14
5	A: Confidently Correct (100%) B: Correct with Moderate Confidence ([75, 100%]) C: Incorrect with Some Confidence ([50, 75%]) D: Confidently Incorrect ([25, 50%]) E: No Direct Answer ([0, 25%])	73.20
7	A: [90, 100%], B: [75, 90%), C: [60, 75%), D: [45, 60%), E: [30, 45%), F: [15, 30%), G: [0, 15%)	68.77

Table 5: Comparison of different confidence binning schemes on the XD-Violence dataset. The five-bin configuration, used in our final model, demonstrates superior performance.

five-bin scheme provides the optimal balance between descriptive granularity and robust detection performance.

References

- [1] N.E. Gronlund and C.K. Waugh. *Assessment of Student Achievement*. Pearson, 2009. ISBN 9780205597284. 9
- [2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. 9