The Road Less Seen: Segment Exploration for Weakly Supervised Video Anomaly Detection

Supplementary Material

Contents

1. List of Symbols

Table 1. List of symbols and their meanings

Symbol	Meaning
q	Quantile threshold for similarity distribution
τ_s	Threshold for adjacent frame similarity
τ_a	Quantile threshold for anomaly score
τ_u	Threshold for uncertainty
τ_s^m	Threshold for memory similarity
AM	Abnormal Memory
γ	Weighting parameter in uncertainty loss
λ	Weighting parameter in knowledge fusion
α	Desired False Positive Rate
MCC	Misclassification Cost
W	Weight to False Negatives in misclassification cost
S	Cosine similarity between two features
N	Number of segments
C	Cluster assignment of each segment in a video
p_j	Anomaly score of segment j
$\mathcal{K}$	Set of selected segments for MIL loss
$\mathcal{U}$	Uncertainty set
AP	Average Precision
$AUROC$	Area Under ROC Curve

2. Algorithms

In this section, we present pseudocode for Temporal Clustering-based selection 1 and Uncertainty-based exploration 2.

Algorithm 1 Temporal Clustering of Video Segments

Require: Segment features $\{\mathbf{x}_1^+, \ldots, \mathbf{x}_N^+\}$, similarity threshold τ_s.
Initialize set of clusters $\mathcal{C} \leftarrow \emptyset$
Create first cluster $C_1 \leftarrow \{1\}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{C_1\}$, set $m \leftarrow 1$
for $i = 2$ to N **do**
  Compute similarity $S(\mathbf{x}_i^+, \mathbf{x}_m^c)$
  if $S(\mathbf{x}_i^+, \mathbf{x}_m^c) \geq \tau_s$ **then**
    Assign i to C_m: $C_m \leftarrow C_m \cup \{i\}$
  else
    Create new cluster: $C_{m+1} \leftarrow \{i\}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{C_{m+1}\}$
    Update $m \leftarrow m + 1$
  end if
end for
return $\mathcal{C}$

Algorithm 2 Uncertainty-Based Exploration of Video Segments

Require: $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\{u_1, \ldots, u_N\}$, τ_u, τ_s^m, abnormal memory AM
Ensure: A Set of selected uncertain segments $\mathcal{U}$
Initialize $\mathcal{U} \leftarrow \emptyset$
for $i = 1$ to N **do**
  Compute: $s_i = \frac{1}{|AM|} \sum_{j \in AM} S(\mathbf{x}_i^+, \mathbf{x}_j^{AM})$
  if $u_i > \tau_u$ **and** $s_i > \tau_s^m$ **then**
    Add i to $\mathcal{U}$: $\mathcal{U} \leftarrow \mathcal{U} \cup \{i\}$
  end if
end for
return $\mathcal{U}$

3. Implementation Details

The hyperparameters q and τ_u are fine-tuned over the ranges $[0.01,\ 0.1]$ (with a step size of 0.01) and $[0.1,\ 0.4]$ (with a step size of 0.05), respectively. We performed a grid search for the fusion weight parameter λ over the interval [0, 1] to identify the value that maximizes the AP score. The best-performing hyperparameters for the UCF-Crime dataset are: $q = 0.03$, $\tau_a = 0.955$, $\tau_u = 0.2$, $\tau_s^m = 0.7$, $\gamma = 1$, and $\lambda = 0.45$, While we can also set the same hyperparameter setting for XD-Violence, we get the best performance on $q = 0.03$, $\tau_a = 0.95$, $\tau_u = 0.3$, $\tau_s^m = 0.7$, $\gamma = 0.1$, and $\lambda = 0.55$ All other features, training settings, and baseline model hyperparameters are the same as in the UR-DMU model [16], for example, a batch size of 64, a learning rate of 0.001, 60 memory slots, and so on.

For training-free inference with the InternVL3-14B model, we experiment with temporal windows of 48, 144, and 288 frames, and select a 48-frame window to capture short-clip anomalies, passing 8 frames at a time for inference. To estimate the model's confidence, we average the results over 5 runs, which is denoted by y_{VLM}.

For uncertainty calculation in the dual exploration, three models are trained. Although they can be run in parallel on three GPUs, we execute them sequentially on a single NVIDIA RTX A6000 with 50 GB of memory. With reduced testing frequency, evaluating every 100 steps instead of every 10, training takes only about one and a half hours for the UCF-Crime dataset.

Computational Cost: Since InternVL fusion is only performed during testing, our method has 19.47 million training parameters, and inference runs at 97 frames per second (FPS) for dual exploration model.

4. Baseline Comparision

Since AUROC is the primary evaluation metric for the UCF-Crime dataset, Table 2 presents a comparison of AU-

Table 2. Results on UCF-Crime dataset including AUROC Score

Method	Feature	AUROC (%)	AP (%)
MIL [9]	I3D	76.21	25.03*
RTFM [10]	I3D	84.30	29.46*
BNSVP [8]	I3D	83.39	30.68*
MGFN[3]	I3D	80.21*	18.88*
GS-MOE[1]	I3D	91.58	34.08*
UR-DMU [16]	I3D	86.97	35.48*
UMIL [5]	CLIP	86.75	–
TSA[4]	CLIP	87.58	–
TPWNG [12]	CLIP	87.79	–
PEMIL[2]	I3D+Text	86.83	–
PEL4VAD [7]	I3D+Text	86.76	33.99
VadCLIP [16]	CLIP	88.02	33.55
DSANet [14]	CLIP	89.44	37.41
InternVL3-14B [17]	Training Free	79.61	29.50
Ours (Dual Exploration)	I3D	85.63	**36.42**
Ours (Dual Exploration + InternVL)	I3D + InternVL	87.80	**38.33**

ROC and AP scores. As baseline models do not report AP for UCF-Crime, we retrained them to evaluate this metric. Among single-modality models, URDMU achieves the highest AUROC, while for multi-modality models, VadCLIP shows the best performance. We have omitted recent baselines such as [6] because neither their code nor their model is publicly available. This prevents us from computing their AP score, therefore, only best best-performing, open-source models are reported in Tables 1 and 2 of the main paper. The link to our source code is available at https://github.com/Anushaacharya607/DualExplore.

5. Effect of Hyperparameters

5.1. Effect of τ_a

We study the effect of varying the threshold τ_a for the anomaly score in Table 4 and 3. For both datasets, we find that a threshold of 0.99 is too high for the model and misses a significant ratio of abnormal events, indicated by the low recall values. As we decrease the threshold, we find that around $\tau_a = 0.95/0.96$ results in a good balance between precision and recall. A lower threshold would increase recall but compromise precision.

5.2. Effect of τ_u

In Table 6 and 5, we study the effect of varying the uncertainty threshold τ_u. Note that uncertainty is the standard deviation of the model's predicted score over an ensemble of 3 models. A low threshold ($\tau_u = 0.10/0.15$) would render most predictions as uncertain, thereby enabling the model to explore all such events, causing over-exploration. This would not prioritize the truly uncertain cases with high standard deviation, and thus lead to a low AP score. Similarly, a high threshold would result in an under-exploration. The tables show that a good balance occurs around $\tau_u = 0.20/0.25$ for UCF-Crime, and around $\tau_u = 0.30/0.35$ for XD-Violence datasets. For real-world deployment, τ_u can be initialized as the third quartile (Q3) of the uncertainty distribution after a few initial epochs.

5.3. Effect of fusion parameter λ

Table 7 shows the effect of varying the fusion parameter λ on anomaly detection performance, measured using AP. For both datasets, combining the dual-exploration prediction and the VLM prediction equally (i.e., $\lambda = 0.5$) yields better performance than relying solely on either one. We also observe that the AP score is relatively insensitive to the choice of λ within the range 0.4–0.6. In this stable region, λ can be further adjusted based on the overall uncertainty of the model. The average prediction uncertainty (std) of the dual-exploration model is approximately 0.07 for UCF-Crime and 0.02 for XD-Violence dataset. Since the model exhibits higher uncertainty on UCF-Crime, it is reasonable to trust the VLM prediction more by choosing a λ slightly lower than 0.5, and vice versa for XD-Violence. Accordingly, we select $\lambda = 0.45$ for UCF-Crime and $\lambda = 0.55$ for XD-Violence, which correspond to the best AP scores of 38.33 and 84.58, respectively.

Table 3. Recall and precision at different thresholds for varying values of τ_a on the UCF-Crime dataset.

τ_a	Recall									Precision									AP
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
0.99	76.1	67.7	58.4	50.8	44.6	39.8	33.3	28.4	23.1	22.1	24.7	27.0	28.6	30.0	31.9	32.4	33.5	35.7	32.8
0.98	84.9	77.7	71.4	63.8	56.4	47.8	42.8	35.9	29.7	19.9	22.1	24.0	25.8	27.6	28.6	30.0	30.7	33.0	33.1
0.97	79.6	70.9	61.1	51.2	44.0	36.5	24.5	19.8	15.3	21.3	24.4	26.8	28.8	31.3	33.8	35.0	40.4	50.6	34.5
0.96	82.7	76.9	70.3	68.8	56.7	47.9	40.9	34.8	29.0	20.0	22.7	24.6	26.0	28.0	29.1	30.6	32.3	35.5	**34.7**
0.95	84.2	76.6	70.0	63.2	56.6	49.8	43.5	36.7	29.6	20.0	22.0	23.7	24.9	26.2	28.2	29.5	30.8	32.8	34.6
0.94	86.8	80.0	74.3	67.1	58.2	51.6	45.0	37.3	31.6	19.4	21.6	23.1	24.7	26.1	27.7	29.2	30.2	33.8	34.2

Table 4. Recall and precision at different thresholds for varying values of τ_a on the XD-Violence dataset.

τ_a	Recall									Precision									AP
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
0.99	84.3	78.7	73.9	68.1	61.9	54.7	48.2	41.6	34.1	64.8	68.3	71.5	74.6	77.2	79.9	82.3	84.8	87.7	78.68
0.98	87.5	82.4	76.8	70.3	62.9	55.7	47.5	38.7	29.6	62.7	66.7	70.0	73.3	76.3	79.7	82.9	86.8	89.8	79.13
0.97	87.6	84.2	80.0	76.0	71.0	65.3	59.1	53.8	47.3	62.0	65.4	67.6	70.0	72.1	74.5	76.5	78.8	81.5	78.81
0.96	87.5	83.5	79.1	75.0	69.6	64.0	56.9	51.1	43.8	61.8	65.2	67.9	70.4	73.2	75.9	78.7	81.0	83.9	79.32
0.95	90.5	87.2	83.7	79.9	75.2	70.2	63.4	57.4	50.0	60.0	63.4	66.2	68.7	71.4	74.1	77.3	79.8	82.7	**80.88**
0.94	92.1	88.8	85.0	81.0	75.6	96.9	63.1	56.5	47.8	56.6	60.5	63.5	66.4	69.0	72.0	75.4	78.8	82.5	79.29

Table 5. Recall and precision at different thresholds for varying values of τ_u on the UCF-Crime dataset.

τ_u	Recall									Precision									AP
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
0.10	83.7	76.7	69.7	63.8	57.5	49.0	41.8	35.3	29.5	21.0	23.3	24.6	26.4	28.5	29.3	30.2	31.1	34.7	35.78
0.15	83.0	75.9	70.0	61.1	54.1	48.6	42.0	35.9	29.5	20.3	22.8	25.3	26.2	27.8	29.8	31.0	32.3	34.0	35.20
0.20	88.4	81.9	73.4	67.2	61.9	55.1	48.3	39.8	31.6	18.2	20.7	22.2	24.0	26.3	28.5	31.0	32.8	36.3	36.42
0.25	84.4	79.1	72.3	64.7	59.3	51.8	45.1	39.3	32.1	19.7	22.5	24.4	25.8	27.9	29.6	30.8	32.9	34.9	**36.47**
0.30	85.9	79.4	71.5	65.2	59.3	53.4	44.8	36.6	29.2	19.6	22.0	23.6	25.3	27.2	29.3	31.2	32.8	36.7	35.92
0.35	84.7	79.8	72.7	66.1	59.5	52.0	44.7	38.7	31.5	19.7	22.2	23.5	24.6	26.8	28.2	29.7	31.3	33.2	35.39
0.40	83.4	75.3	67.2	61.8	54.8	47.8	41.6	36.3	29.6	20.4	22.6	24.0	26.6	28.6	29.9	30.7	32.2	34.8	35.00

Table 6. Recall and precision at different thresholds for varying values of τ_u on the XD-Violence dataset.

τ_u	Recall									Precision									AP
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
0.10	91.5	88.1	84.3	80.4	75.1	69.6	62.2	55.3	46.6	58.9	62.0	64.8	67.3	69.9	73.0	75.7	78.4	81.3	78.70
0.15	92.2	89.1	85.8	82.5	77.90	72.6	65.7	60.0	52.1	57.3	60.4	63.4	65.9	68.8	71.1	74.3	77.1	80.2	79.70
0.20	94.1	91.8	88.3	85.1	80.6	75.7	69.0	62.7	54.4	54.4	58.1	61.2	64.1	66.8	69.6	72.7	76.2	79.9	79.13
0.25	91.8	88.6	85.4	81.7	76.6	71.5	64.3	58.1	49.8	56.5	60.2	63.2	65.7	68.5	71.1	74.2	77.8	81.6	78.95
0.30	90.5	87.2	83.7	79.9	75.2	70.2	63.4	57.4	50.0	60.0	63.4	66.2	68.7	71.4	74.1	77.3	79.8	82.7	**80.88**
0.35	60.1	86.3	82.1	78.1	72.8	67.3	60.2	54.5	47.0	59.1	62.7	65.9	68.7	71.7	74.8	77.9	80.6	83.7	79.57
0.40	90.4	86.2	81.8	77.3	71.7	65.7	58.9	52.2	43.4	59.3	63.4	66.4	69.2	72.1	74.8	77.9	81.1	84.0	79.53

5.4. Effect of Temporal Window and Frame Selection

As shown in Table 8, we observe only minimal performance differences across different temporal windows and frame selection strategies. Therefore, we adopt the smallest temporal window, which provides a more fine-grained frame-level analysis without sacrificing overall performance.

5.5. Sensitivity to Prompt

To identify the best-performing prompt, we experiment with different prompting settings (Table 9) using only abnormal videos from the test set of the UCF-Crime dataset. Since the evaluation is performed on abnormal videos, we report AUROC as a preliminary metric for these experiments, as it is less inflated than when normal videos are included. Binary prompt with no anomaly prior indicates that VLM should generate only 1 or 0 output without any context of abnormal events (refer to example 1 below). The 'Definition' prompt defines abnormal and normal events with examples of different types of abnormal events per dataset as shown in example 2, with the removal of the specific anomaly type. While in the 'Likelihood' prompt, the output is a value between 0 and 100. Since we cannot leverage the exact class type of abnormal event during training,

Table 7. Effect of the fusion parameter λ on detection performance measured using Average Precision (AP) on UCF-Crime and XD-Violence dataset.

λ	UCF-Crime	XD-Violence
0.0	29.27	69.66
0.1	37.11	82.60
0.2	37.51	83.23
0.3	38.08	83.87
0.4	38.33	84.37
0.5	38.26	84.57
0.6	37.91	84.47
0.7	37.28	83.99
0.8	36.95	83.29
0.9	34.58	82.42
1.0	36.43	80.83

Table 8. Effect of Temporal Window and Frame Selection on InternVL.

Frame Selection	**Temporal Settings**		**AUROC (Ab)**
	Window	**Frames**	
Uniform	48	8	70.00
Uniform	144	8	70.84
Uniform	288	8	70.78
Uniform	144	16	70.54
Uniform	288	16	70.62
FPS	288	8	70.48

we select a likelihood definition prompt to compare against different models. We further observe that a likelihood-based output performs slightly better than a binary one. This may be because the VLM engages in implicit chain-of-thought reasoning when estimating likelihood.

Table 9. Sensitivity to Prompt on InternVL3-14B Model

Output	**Anomaly Prior**	**AUROC (Ab)**
Binary	None	68.67
Binary	Definition	70.00
Likelihood	Definition	70.93
Binary	Definition + Prior	71.66
Binary	Chain-of-Thought + Prior	73.03

Prompt Example 1: Binary Output without Anomaly Prior and Definition

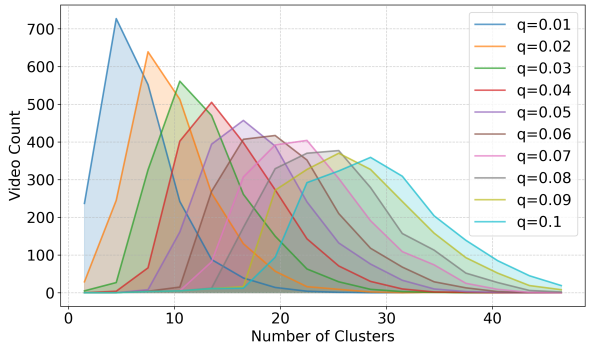
```
You are an anomaly detection assistant
    . Analyze the video clip carefully
    .
If the video clip contains any
    abnormal activity, reply with only
     "1".
If the video clip is normal, reply
    with only "0".
Do not include any explanation or
    extra text.
```

Prompt Example 2: Binary Output with Anomaly Prior

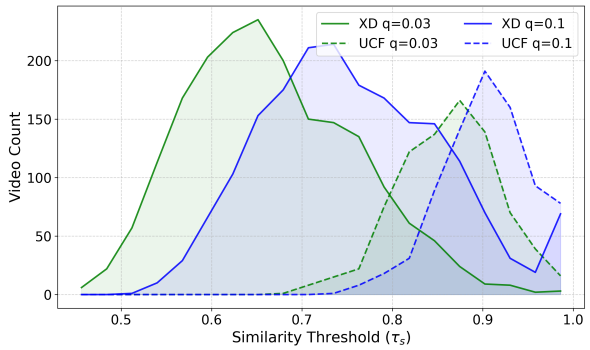
```
You are a video anomaly detection
    analyst. You are given a video
    clip and your task is to detect
    whether it contains any abnormal
    events like: [{anomaly}].
```
Definitions:
```
- Abnormal Event: Intentional, harmful
    , unlawful, or dangerous
    activities that threaten safety,
    break laws, or strongly deviate
    from normal daily routines.
    Examples include abuse, arrest,
    arson, assault, road accidents,
    explosions, fighting, shooting,
    vandalism, shoplifting, robbery,
    burglary, and similar threatening
    behaviors.
- Normal Event: Routine and harmless
    daily activities such as walking,
    talking, driving normally,
    shopping, exercising, or working.
```
Evaluation Criteria:
```
- Focus specifically on whether the
    given clip contains the abnormal
    event type [{anomaly}].
- Consider human actions, interactions
    , objects, and context.
```
Output Rule:
```
- If the clip shows [{anomaly}]
    Reply with: 1
- If the clip does not show [{anomaly
    }]     Reply with: 0
- Reply with only the digit (0 or 1),
    no explanations or extra text.
```

5.6. Effect of Temporal Clustering hyperparameters

As stated in the main paper, the hyperparameter q controls the number of clusters formed for each video, which is adaptive since it depends on the quantile value of the similarity scores within the video. Figure 1a shows how the distribution of clusters changes with different q values for the XD-Violence dataset. As q increases, the number of clusters also increases because the similarity threshold becomes stricter, resulting in more fine-grained clustering. Figure 1b shows the adaptive similarity threshold for each video across both datasets.



(a)



(b)

Figure 1. (a) Cluster distribution for varying q in XD-Violence. (b) Similarity threshold (τ_s) distribution for varying q.

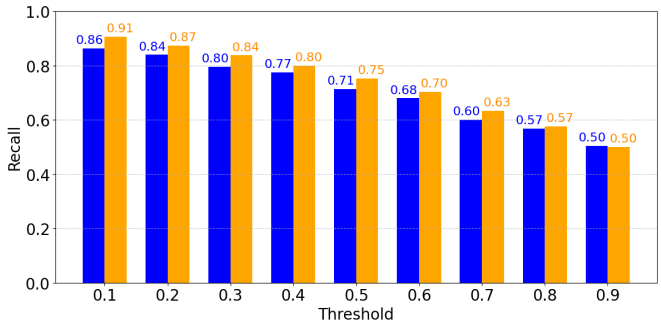
6. Additional Experiments

6.1. Performance comparison among different VLMs

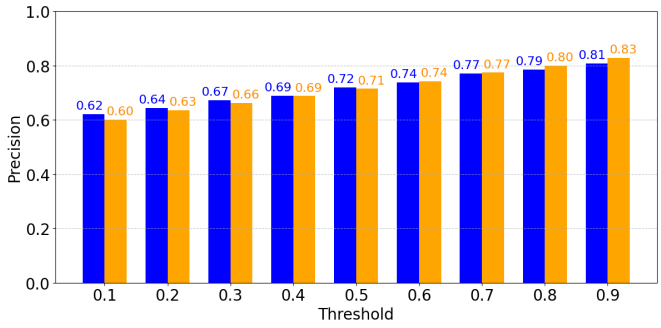
In the Table 10, we compared the effectiveness of a recent SOTA multimodal large language model trained on a video dataset using the "Detailed Likelihood" prompt with a temporal window of 48 segments and 8 frame selections per window. Constrained by inference resources, we select InternVL3-14B for our knowledge fusion module.

Table 10. Evaluation of different VLMs.

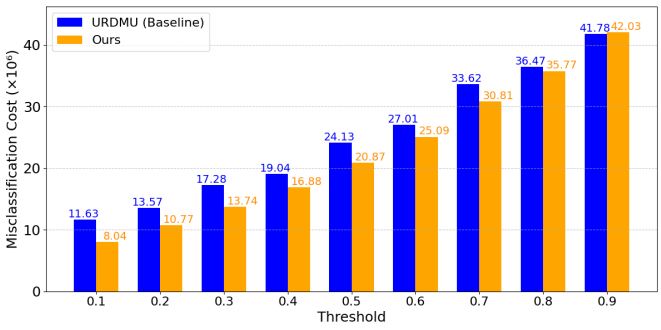
Model	Size	AUROC (Ab)	AUROC
InternVL3-9B [17]	9B	66.89	73.82
InternVL3-14B	14B	**69.60**	**80.28**
InternVL3-5-14B [11]	14B	69.59	77.97
LLaVA-NeXT-Video [15]	7B	66.11	74.9
Vera LLM Output [13]	–	62.46	75.26



(a) Recall



(b) Precision

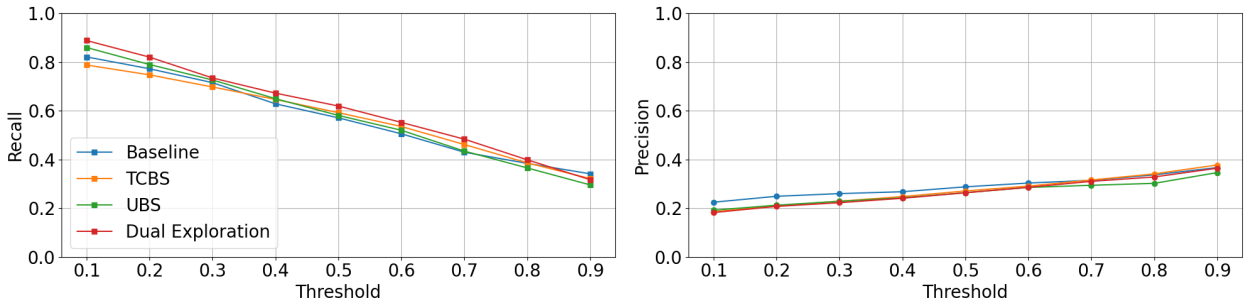


(c) Misclassification Cost

Figure 2. Comparison of recall, precision, and misclassification cost with the baseline (UR-DMU) across different thresholds on the XD-Violence dataset.

6.2. Result on XD-Violence

In Figure 2, we analyze results for the XD-Violence dataset. We observe that our method achieves higher recall with minimal compromise in precision, similar to the UCF-Crime dataset as described in the main paper. We also ob-



(a) Recall (b) Precision

Figure 3. Comparison of Recall and Precision across thresholds for different ablation settings on UCF-Crime Dataset.

serve that the misclassification cost for our method is lower than the UR-DMU baseline. The lower recall at 0.9 threshold is due to temporal clustering-based selection (TCBS).

6.3. Generalization to novel anomalies

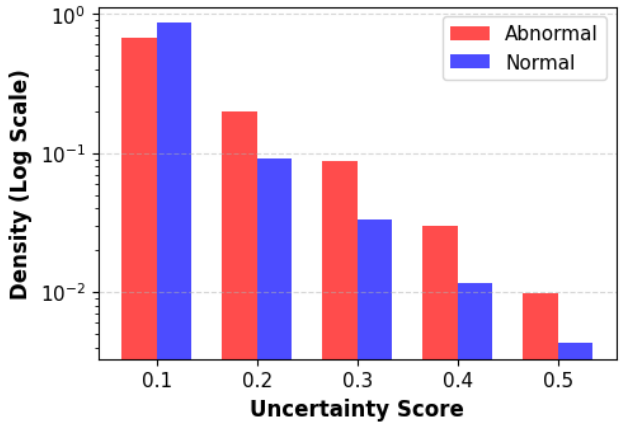


Figure 4. Uncertainty distribution on XD-Violence under cross-dataset evaluation (trained on UCF-Crime).

Since we estimate uncertainty via deep ensembles, a Bayesian approximation approach, unseen and rare events (anomalies) naturally induce higher uncertainty. Figure 4 shows the generalization capability of our method for unseen events, where a model trained on UCF-Crime exhibits significantly higher uncertainty for abnormal segments on XD-Violence than for normal segments.

6.4. Ablation on Dual Exploration

We study the effect of each exploration strategy of our method, i.e., TCBS and UBS, in Figure 3. We observe that both TCBS and UBS alone can improve recall at a reasonable threshold of 0.4-0.7. Moreover, combining both achieves the highest recall along the threshold range with minimal compromise in precision.

7. Qualitative Analysis

Figure 5 illustrates cases where our method detects abnormal events more effectively than the baseline. The first two examples, Figures 5a and 5b, show videos where both models identify them as abnormal; however, the baseline method fails to capture all abnormal events that occur at distinct times. In contrast, our dual exploration strategy successfully explores other segments in the video. Moreover, as shown in Figures 5c and 5d, our diverse exploration can detect abnormal videos entirely missed by the baseline. Figures 5e to 5h present normal video cases where the baseline misclassifies them as abnormal. Figure 6 shows an example from the training dataset where the model exhibits high uncertainty but a low anomaly score. This particular case is missed by temporal clustering but captured by uncertainty exploration.

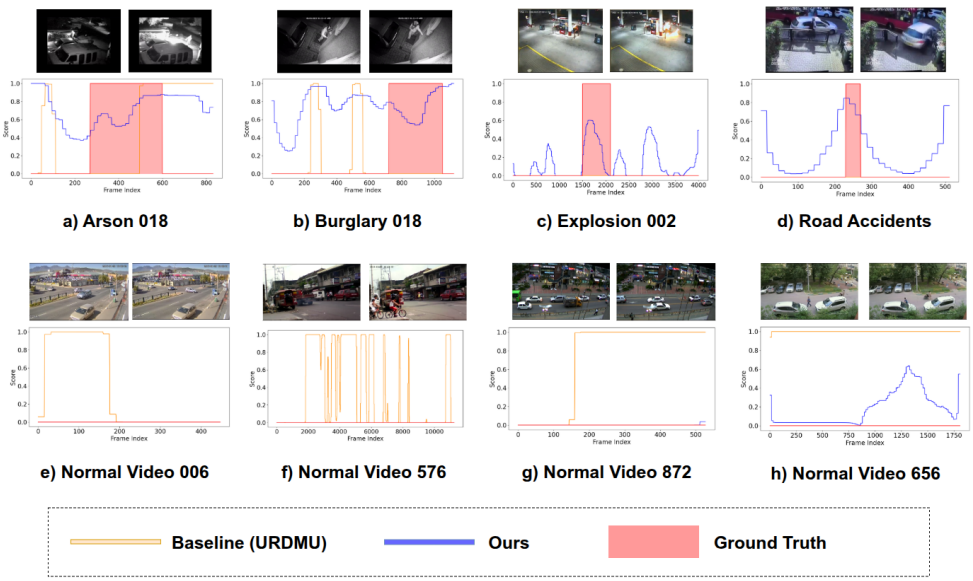


Figure 5. Qualitative analysis of baseline (UR-DMU) and our method on the UCF-Crime dataset.

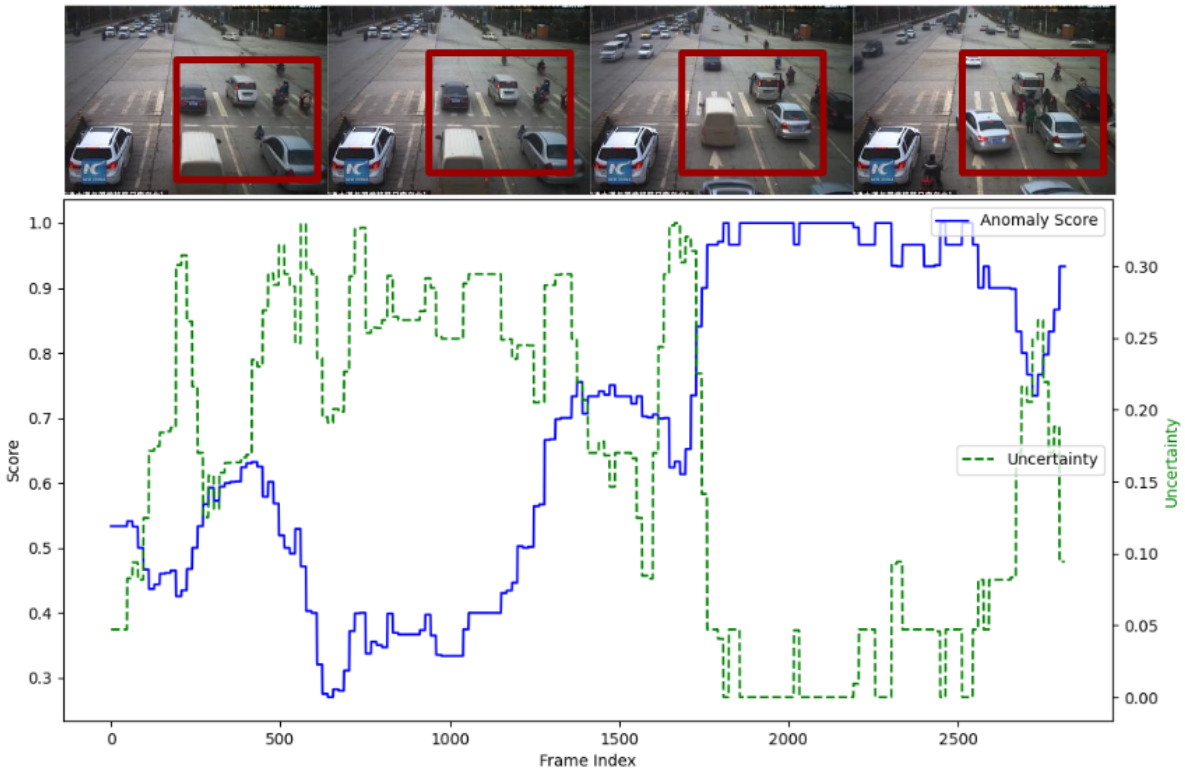


Figure 6. Top: Frames sampled from an abuse video where a person is repeatedly hit by a car, depicting subtle abnormal events. These events receive low prediction scores (blue line in the bottom graph) but exhibit high prediction uncertainty (green line). The model detects drastic motion changes, reflected by high anomaly scores with low uncertainty, after a group of people gathers near the scene, but it misses the actual abnormal event.

References

[1] Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond, and Egor Bondarev. Mixture of experts guided by gaus-

sian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10275–10285, 2025. 3

[2] Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18319–18329, 2024. 3

[3] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 387–395, 2023. 3

[4] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023. 3

[5] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023. 3

[6] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and François Brémond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24265–24274, 2025. 3

[7] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024. 3

[8] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3212–3221, 2022. 3

[9] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 3

[10] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 3

[11] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[12] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18899–18908, 2024. 3

[13] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688, 2025. 6

[14] Wenti Yin, Huaxin Zhang, Xiang Wang, Yuqing Lu, Yicheng Zhang, Bingquan Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12027–12035, 2026. 3

[15] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 6

[16] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3769–3777, 2023. 2, 3

[17] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3, 6