Anomize: Better Open Vocabulary Video Anomaly Detection

Supplementary Material

1. Overview

This supplementary material is organized as follows:

- Data Division
- Impact of Data Addition
- Impact of Using Other Lightweight Temporal Encoder
- t-SNE Visualization of Video Features
- Visualization of Per-Class Results
- Prompt Design
- Limitation

2. Data Division

In the OVVAD task, we categorize anomaly labels into base and novel categories, utilizing only base category samples during training. Following standard open vocabulary learning practices, frequent labels are designated as base categories, while rare labels are classified as novel categories. Specifically, for UCF-CRIME dataset, the base categories include Abuse, Assault, Burglary, Road Accident, Robbery, and Stealing, with all other labels treated as novel. On XD-VIOLENCE dataset, the base categories are Fighting, Shooting, and Car Accident.

3. Impact of Data Addition

Tab. 6 indicates that the performance drop after adding new labels is acceptable. These new labels consist of both same-group and cross-group categories. Same-group labels can lead to closer encodings, potentially interfering with original predictions and slightly affecting performance. For XD-VIOLENCE and UCF-CRIME, the added overlapping categories include drug trafficking, harassment, stalking, loitering, and public intoxication. Additionally, XD-VIOLENCE incorporates shoplifting, arson, robbery, and arrest, while UCF-CRIME includes riot.

We provide a detailed analysis of the results presented in Tab. 6. After adding new labels and data, the model correctly categorizes 20 of 21 shoplifting samples, 3 of 5 arrest samples, 7 of 10 arson samples, 3 of 3 assault samples, and 91 of 99 riot samples. Among these, only shoplifting is not grouped with the base labels. For anomalous data within the same group of labels, the similarity in visual encodings suggests a higher risk of overfitting, potentially leading to miscategorization as base cases within the same group. Nonetheless, our method achieves excellent results, highlighting the effectiveness of the guided text encodings in reducing categorization confusion.

4. Impact of Using Other Lightweight Temporal Encoder

We implement an LSTM as a lightweight temporal encoder to mitigate overfitting to base classes and reduce label confusion in novel cases. This design choice is motivated by our observation that highly parameterized encoders are prone to overfitting, likely because the additional parameters tend to capture features that closely resemble those in the training data when handling novel data. As a result, when these visual features align with the label textual features, they are found to be closer to the labels of the base data, thereby leading to misclassification.

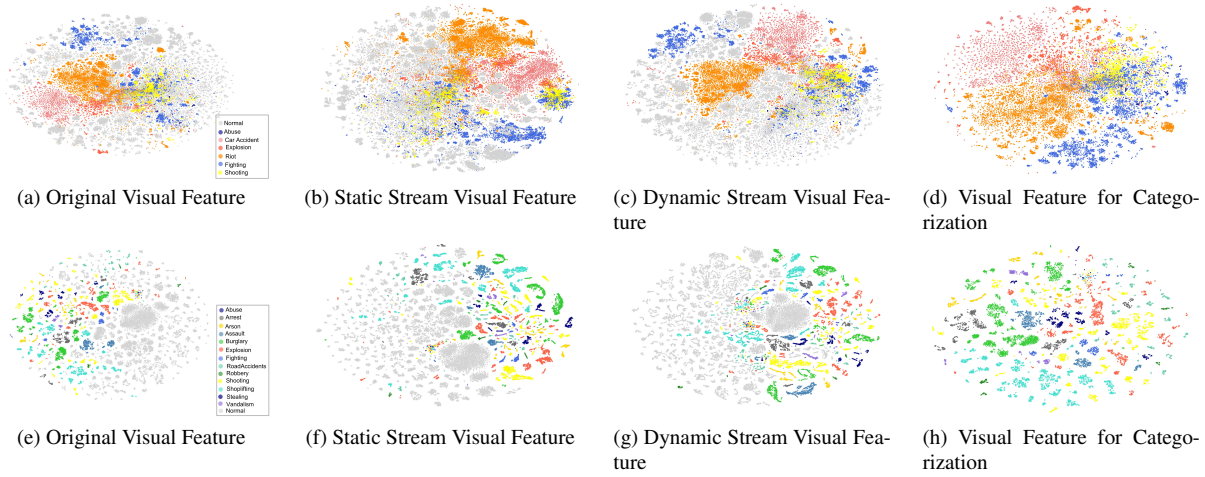
In the supplementary material, we also evaluate a transformer to validate our method, with results presented in Tab. 7. However, due to having more parameters, the transformer underperformed compared to the LSTM on both datasets. On UCF-CRIME, the transformer required a lower learning rate of 5×10^{-8} to prevent overfitting. While this adjustment improves performance for novel classes, it significantly reduces performance for base classes, resulting in poorer overall results. Regardless of the temporal encoder employed, our method consistently outperforms existing SOTA models, indirectly confirming the effectiveness of our group-guided text encoding mechanism.

Dataset	ACC	ACC_b	ACC_n
XD-VIOLENCE	83.89	95.04	68.59
UCF-CRIME	46.43	52.94	42.70

Table 7. **Top-1 Accuracy (%) with a Transformer-Based Lightweight Temporal Encoder.**

5. t-SNE Visualization of Video Features

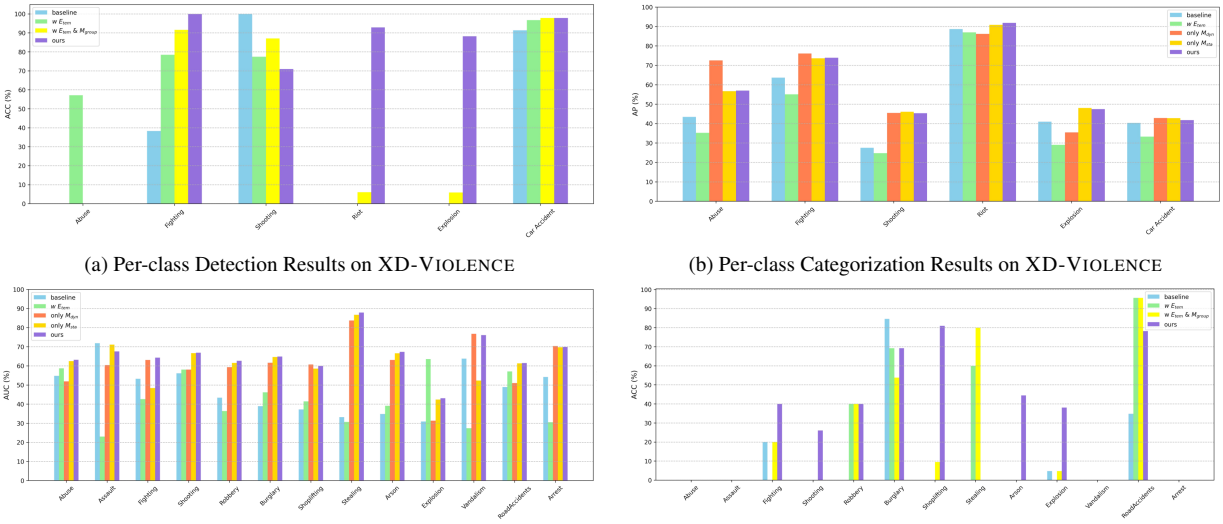
As shown in Fig. 6, the original CLIP encodings exhibit irregular patterns, making them unsuitable for anomaly detection and categorization. In contrast, Fig. 6(b), (c), (f), and (g) demonstrate that the text-augmented static and dynamic streams provide sufficient information, resulting in clearer boundaries between normal and anomalous data. This facilitates the detector in better distinguishing between them. After temporal modeling and feature fusion, as shown in Fig. 6(d) and (h), more distinct clusters emerge, with data points sharing the same label becoming closer in high-dimensional space, thereby enhancing categorization. Notably, the visualization reflects the frame-level features used for detection and categorization, and the presence of normal frames within anomalous videos contributes to minor overlaps.



(a) Original Visual Feature (b) Static Stream Visual Feature (c) Dynamic Stream Visual Feature (d) Visual Feature for Categorization

(e) Original Visual Feature (f) Static Stream Visual Feature (g) Dynamic Stream Visual Feature (h) Visual Feature for Categorization

Figure 6. **Scatter Plots of Features Reduced to 2D Using t-SNE.** (a–d) correspond to XD-VIOLENCE, while (e–h) represent UCF-CRIME. "Original Visual Feature" refers to scatter plots generated from CLIP frame encodings. "Static Stream Visual Feature" and "Dynamic Stream Visual Feature" denote features refined through the static and dynamic streams for detection, respectively. "Visual Feature for Categorization" represents the fused visual features used for classification.



(a) Per-class Detection Results on XD-VIOLENCE (b) Per-class Categorization Results on XD-VIOLENCE

(c) Per-class Detection Results on UCF-CRIME (d) Per-class Categorization Results on UCF-CRIME

Figure 7. **Per-Class Contribution of Each Design Element to Open Vocabulary Anomaly Detection on XD-VIOLENCE and UCF-CRIME.** The baseline uses original CLIP encodings, while "w E_{tem}" incorporates temporal encoding. "Only M_{dyn}" and "Only M_{sta}" denote the dynamic and static streams, respectively, with text augmentation and loss weight w_i. "w E_{tem} & M_{group}" further integrates the guided encoding mechanism.

6. Visualization of Per-Class Results

Fig. 7 illustrates the impact of key design elements on performance across different categories. In both categorization and detection branches, our method achieves the best or near-best results across most categories, demonstrating that our approach not only ensures strong performance on base anomalies but also effectively addresses novel anomalies in OVVAD.

7. Prompt Design

We provide the prompt designs described in the paper to offer deeper insights into our method and explain why the group-guided text encoding mechanism and *ConceptLib* designs lead to significant performance improvements. After generating group data using $\text{prompt}_{\text{group}}$, we manually refine the results to eliminate noise. For $\text{prompt}_{\text{desc}}$, we input a group of labels with length constraints to align with the CLIP text encoder requirements. The value of L in $\text{prompt}_{\text{conc}}$ is determined by the number of labels: XD-

VIOLENCE, with six anomaly labels, uses $L = 200$, while UCF-CRIME, with thirteen labels, uses $L = 500$.

$\mathrm{prompt}_{\mathrm{group}}$. Group the following anomaly labels: {labels}. Organize them based on similarities in visual characteristics, where behaviors with comparable actions, activities, or scene contexts during the anomaly are placed in the same group.

$\mathrm{prompt}_{\mathrm{desc}}$. Describe the anomaly of {labels}. Begin each description with one or two sentences that emphasize the common traits shared among all behaviors, followed by one or two sentences detailing the unique characteristics specific to each behavior. Ensure the descriptions clearly capture significant details of the anomaly, such as actions, movements, and scene context. Maintain a consistent sentence structure for each description, with word counts between 50 and 70 words.

$\mathrm{prompt}_{\mathrm{conc}}$. Given the anomaly labels: {labels}, generate $\{L\}$ noun phrases that accurately capture the key scene characteristics associated with each label. These phrases should be well-suited for CLIP model encoding and designed to complement visual features, enhancing the effectiveness of model in anomaly detection.

8. Limitation

In this paper, we distinguish dynamic and static elements at the semantic level, considering dynamic elements as label descriptions that appear in the form of sentences, and static elements as nouns that describe key features of the anomaly. There may be slight redundancy between these two elements, but it does not affect our ability to achieve outstanding performance. However, exploring feature-level disentanglement for them may further enhance performance.