Detection, Retrieval, and Explanation Unified: A Violence Detection System Based on Knowledge Graphs and GAT

Wen-Dong Jiang, *Graduate Student Member, IEEE*, Yu-Ting Chin, Yu-Ting Yang, Chih-Yung Chang, *Member, IEEE*, and Diptendu Sinha Roy, *Senior Member, IEEE*

***Abstract*—Recently, violence detection systems developed using unified multimodal models have achieved significant success and attracted widespread attention. However, most of them typically suffer from two major issues: lack of interpretability as opaque models and limited functionality, providing only classification or retrieval capabilities. This article proposes a novel interpretable violence detection system called the three-in-one (TIO) system to address these challenges. The proposed TIO system integrates knowledge graphs (KGs) and graph attention networks (GATs) to achieve three core functionalities: detection, retrieval, and explanation. Specifically, the system processes each video frame along with text descriptions generated by a large language model (LLM) for a given video containing potential violent behavior. It employs ImageBind to create high-dimensional embeddings for constructing a KG. Subsequently, the system uses GAT for reasoning, applies light time series modules to extract video embedding features, and finally connects a classifier and retriever for multifunctional output. Leveraging the interpretability of KG, the system verifies the rationale behind each output during the reasoning process. In addition, this article proposes a kernel function-adjusted distance-based computation method to reduce the computational complexity of the TIO system. Extensive experiments conducted on the XD-Violence, UCF-Crime, and UCFCrime-AR datasets demonstrate the effectiveness of the proposed TIO system.**

***Index Terms*—Artificial life, explanation, intelligent monitoring, multimedia system, smart city, systems management.**

I. Introduction

THE primary goal of violence detection is to monitor abnormal events in the real world to prevent violent behaviors and maintain social order [1]. With the rapid advancement of artificial intelligence, researchers are exploring the application of deep learning technologies in surveillance systems [2], [3] to replace the inefficiencies and high costs of traditional manual monitoring. In recent years, weakly supervised violence detection (WSVD), also known as weakly supervised video anomaly detection, and violence retrieval (VAR), also referred to as weakly supervised video anomaly retrieval, have become key research directions.

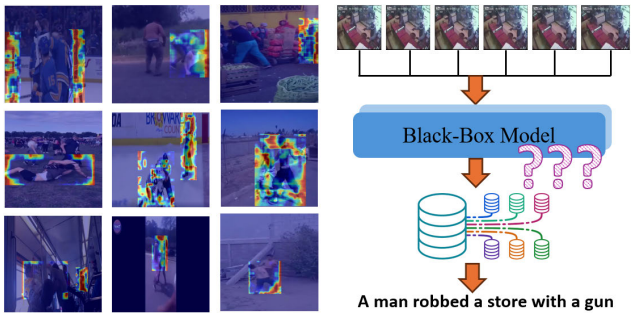


Fig. 1. Interpretable issue in (a) WSVD and (b) VAR tasks.

Unlike traditional methods that required frame-level annotations for videos and supervised learning for frame-level training, WSVD achieved frame-level anomaly detection using video-level annotations [4], [5]. Meanwhile, VAR focused on retrieving detailed textual descriptions corresponding to a given video [6], assisting law enforcement in post-event analysis and summarization.

However, current research on WSVD and VAR tasks [6], [7], [8], [9], [10], [11], [12] faced two major issues: first, the lack of interpretability in opaque models; and second, the need to design separate models for each task. Fig. 1 illustrates the interpretability challenges in WSVD and VAR tasks. In (a), although the WSVD model successfully identified violent scenes, interpretability methods like heatmaps reveal that the system focused on irrelevant areas rather than actual violent scenes. Similarly, in (b), although the VAR model retrieved the correct textual descriptions, users could not understand the rationale behind the retrieval results due to the opaque nature of the system.

Moreover, existing research typically designed independent models for WSVD and VAR tasks rather than adopting a

Received 6 July 2025; accepted 10 January 2026. Date of publication 21 January 2026; date of current version 19 March 2026. This work was supported by the National Science and Technology Council of Taiwan, Republic of China, under Grant NSTC 112-2622-E-032-005. This article was recommended by Associate Editor X. Le. *(Corresponding author: Chih-Yung Chang.)*

Wen-Dong Jiang and Chih-Yung Chang are with the Department of Computer Science and Information Engineering, Tamkang University, New Taipei City 25137, Taiwan (e-mail: wendongjiang@ieee.org; cychang@mail.tku.edu.tw).

Yu-Ting Chin is with the Department of Artificial Intelligence, Tamkang University, New Taipei City 25137, Taiwan (e-mail: kroyxyork@gmail.com).

Yu-Ting Yang is with the Holistic Education Center, Fu Jen Catholic University, New Taipei City 25137, Taiwan (e-mail: 142720@mail.fju.edu.tw).

Diptendu Sinha Roy is with the Department of Computer Science and Engineering, National Institute of Technology, Shillong 793003, India (e-mail: diptendu.sr@nitm.ac.in).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2026.3653874.

Digital Object Identifier 10.1109/TSMC.2026.3653874

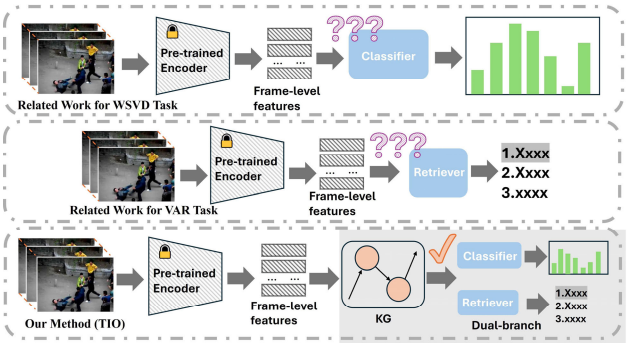


Fig. 2. Comparisons of different paradigms for WSVD and VAR.

unified end-to-end approach, leading to the following problems: 1) redundant design and training processes increased development and computational costs; 2) independent models limited information sharing and collaboration, hindering semantic reasoning and performance optimization; and 3) users must separately operate and maintain systems for each task, reducing overall usability and efficiency.

To address these issues, this article proposes a novel interpretable violence detection system called the three-in-one (TIO) system. As shown in Fig. 2, unlike previous systems for WSVD and VAR tasks [6], [7], [8], [9], [10], [11], [12] that suffer from limited functionality and poor interpretability, the proposed TIO system achieves interpretability through a knowledge graph (KG) module and supports end-to-end multitasking with a dual-branch design. Specifically, given video data with violent behaviors and corresponding labels, the system first uses a large language model (LLM) to generate keywords related to the labels. These, along with video data, are passed through a frozen ImageBind encoder [13] to produce unified multimodal embeddings. Next, a KG is constructed and refined through a graph attention network (GAT) [14]. Temporal sequence analysis is then performed using a lightweight temporal module, followed by a classifier and retriever for end-to-end multitask functionality. During usage, users not only receive system decisions and retrieval results but also gain explanations through KG-based triplet relationships (entity1-relation-entity2).

In practice, triplet construction is defined as follows: entity1 represents high-dimensional video embeddings, relations are LLM-generated keyword embeddings, and entity2 comprises object information extracted using the pretrained Yolo-World [15], supplemented with semantic information via Concept Net [20]. In the proposed system, the lightweight design is implemented by applying the GAT and temporal modules. For the GAT module, the original inner product operations are replaced with kernel-based Euclidean distance operations, which significantly simplifies the system complexity. For the temporal module, distance-based computation has been used to replace the traditional similarity calculations, improving computational efficiency and model performance. Detailed implementation is described in Section IV.

Experiments on benchmark datasets XD-Violence [18], UCF-Crime [19], and UCFCrime-AR [6] demonstrate the feasibility of the proposed TIO system. In summary, the main contributions of this article are as follows.

1) *Incorporation of a KG Module and Dual-Branch Design:* The proposed system achieves end-to-end processing for WSVD and VAR tasks, significantly enhancing the interpretability of the model, making system decisions and retrieval results more transparent.
2) *Innovative Use of Multimodal Unified Embeddings and Euclidean Distance for Model Optimization:* Unlike previous methods, the proposed TIO system employs a frozen ImageBind encoder to generate unified multimodal embeddings and uses kernel-based Euclidean distance instead of inner product operations in the GAT of the KG. This approach substantially reduces system complexity. Additionally, the temporal module leverages distance-based metrics rather than traditional similarity calculations, improving computational efficiency and model performance.
3) *Unified Multitask Model Design for Enhanced System Efficiency and User Experience:* The proposed TIO system replaces the traditional approach of designing separate models for WSVD and VAR tasks with a unified multitask model. This reduces redundancy in model design and training, facilitates information sharing and collaborative use, minimizes user operations and maintenance efforts, and achieves dual improvements in semantic reasoning and performance optimization.

The remainder of this article is organized as follows. Section II discusses and compares previous relevant studies. Section III describes the assumptions and problem formulation in detail. Section IV details the proposed TIO system in this article. Section V provides experiments and performance evaluation. The limitation and conclusion are discussed in Sections VI and VII, respectively.

II. Related Work

This section reviews the related literature, divided into two parts: WSVD and VAR.

A. Weakly Supervised Violence Detection

Sultani et al. [19] and Speer et al. [20] were the first to introduce the multi-instance learning (MIL) model into supervised violence monitoring (SVM). Their approach packaged surveillance videos and fed them into I3D and C3D networks for binary classification of violent and nonviolent events. Ji and Lee [21] proposed the one-class support vector machine (OCSVM) model for violence detection. However, due to the limitations of 3-D network-based designs in terms of real-time performance and hardware requirements in practical applications, research gradually shifted toward combining video frame extraction and temporal sequence modules for violence detection. Subsequent studies focused on utilizing self-attention mechanisms, Transformer, or Graph ConvNet (GCN) to capture temporal and contextual relationships within video content.

Zhong et al. [22] proposed a GCN-based method to calculate feature similarity and temporal consistency between video segments for monitoring violent behavior. Tian et al. [23] proposed robust temporal feature magnitude (RTFM), which utilized a self-attention network to capture the global temporal

context of violent behaviors in violent videos. Wu et al. [24] introduced S3R, which models anomalies at the feature level by exploring the synergy between dictionary-based representation and self-supervised learning. Li et al. [25] proposed spatial–temporal relation learning (STRL), which jointly utilizes spatial and temporal evolution patterns for feature learning.

It is noteworthy that in certain practical scenarios, only single-modal cameras are typically available, and traditional multimodal designs are often inapplicable. Thanks to pretrained models such as CLIP [26], which aligns images and text to accomplish multimodal tasks, and Imagebind [13], which unified data from images and seven different heterogeneous modalities, it became possible to achieve multimodal tasks even in single-modality scenarios. Some studies attempted to deploy these pretrained models for violence detection. For instance, Wang et al. [27] proposed ActionClip, which enhances video representation with richer semantic language supervision and supports zero-shot action recognition without additional labeled data or parameter requirements. Zanella et al. [28] combined large language and vision (LLV) models (e.g., CLIP) with MIL for joint violence detection. Wu et al. [7] introduced spatiotemporal prompting (STPrompt), a CLIP-based three-branch architecture to address classification and localization in violence detection. Wu et al. [8] proposed video anomaly CLIP (VadCLIP), a simple yet powerful baseline that efficiently adapted pretrained image-based visual-language models for robust general video understanding. Yun et al. [30] proposed MissionGNN, which integrated Imagebind with LLM to construct a multilevel graph neural network for anomaly detection.

Although these methods have made significant progress in WSVD, the models designed in these studies are essentially black boxes, raising serious concerns about interpretability. Furthermore, some methods, such as GCN, GNN, or Transformer-based networks for temporal sequence analysis, inherently demand significant computational resources.

B. Violence Retrieval

Events in videos typically reflected the interactions between actions and entities evolving, and people tended to use retrieval methods to obtain comprehensive descriptions. Benefiting from the success of cross-modal pretrained models, some studies focused on audio–text and audio–video retrieval. Liu et al. [29] proposed collaborative experts (CEs), a CE model that compressed multimodal information into compact representations. It utilized pretrained semantic embeddings along with specific cues such as automatic speech recognition (ASR) and optical character recognition (OCR) to query videos through text. Gabeur et al. [31] proposed the multimodal Transformer (MMT), a multimodal Transformer model designed to jointly encode different modalities in videos, enabling intermodal attention.

Wang et al. [32] introduced the text-to-video vector of locally aggregated descriptors (T2VLAD), which implemented an efficient global–local alignment method. It adaptively aggregated multimodal video sequences and textual features through a shared set of semantic centers and computed local cross-modal similarities between video and textual features within the same centers. Ma et al. [33] proposed a cross-modal clip (XClip), which calculated the correlation between coarse-grained features and each fine-grained feature through coarse-to-fine contrast. It filtered out unnecessary fine-grained features guided by coarse-grained features during similarity computation, thereby improving the accuracy of video-to-text retrieval. Lei et al. [34] introduced cross-modal late fusion (XML), which adopted a late fusion design and a novel convolutional start–end detector (ConvSE) to achieve efficient retrieval. Wu et al. [6] proposed the anomaly-led attention network (ALAN), which developed an anomaly-led sampling method to enhance the fine-grained semantic association between violent videos and text, further matching cross-modal content using two complementary alignment strategies.

Although these methods achieved significant success in retrieval tasks, these models still lacked interpretability. During inference, it was difficult for users to understand the reasons behind successful matches. Additionally, the models designed for these tasks were typically limited in functionality, focusing solely on retrieval while being unable to classify violent content on time.

III. Assumptions and Problem Formulation

This section presents the assumptions and problem statements of this study, divided into two parts: detection and retrieval.

A. Detection

Given a monitoring video sequence $\mathcal{V}$ of duration $\tau > 0$, partitioned into N equal-length temporal segments $\{\Phi_l\}_{l=1}^{N}$, the principal aim of this study is to detect anomalies (violent events). Let $\mathcal{V} = \{\Phi_t\}_{t=1}^{N}$, where each $\Phi_l \subseteq \mathcal{V}$ may consist of nonviolent ($\hat{\mathcal{V}}_l$) and violent ($\mathcal{V}_l$) portions, and therefore, $\Phi_l = \hat{\mathcal{V}}_l \cup \mathcal{V}_l$.

Let $\mathcal{C} = \{c_k\}_{k=1}^{\mathrm{K}} \cup \{\hat{c}\}$ denote all classes, where c_k denotes the kth violence class and $\hat{c}$ represents the nonviolence class, and let $\mathcal{L} = \{\ell_k\}_{k=1}^{\mathrm{K}} \cup \{\hat{\ell}\}$ be the corresponding labels. Assume that each Φ_l containing violence belongs to some c_k, where $k \in \{1, \ldots, \mathrm{K}\}$.

Consider a detection mechanism $M \in \mathfrak{M}$, which outputs for each Φ_l a score $\mathcal{P}_l \in [0, 1]$. Let $\delta_l^M(\cdot)$ be a decision function, which maps the output of model M associated with the input Φ_l to a decision 1 or 0. Let $\theta \in [0, 1]$ denote a decision threshold inducing $\delta_l^M(\theta) = 1[\mathcal{P}_l \geq \theta]$. Let $\delta_l = 1[\mathcal{V}_l \neq \emptyset]$ be the ground truth.

Let $\mathrm{TP}_\iota, \mathrm{TN}_\iota, \mathrm{FP}_\iota$, and FN_ι, represent true positive, true negative, false negative, and false positive, respectively, of the prediction result of input $\hat{\mathcal{V}}_\iota$ by applying mechanism M. Let $\delta_l^M(\theta)$ denote the probability that mechanism M, under decision threshold θ, predicts the input $\hat{\mathcal{V}}_l$ to be positive. Formally, $\delta_l^M(\theta)$ is defined as $\delta_l^M(\theta) = \mathbb{P}(M(\hat{\mathcal{V}}_l) = 1|\theta, l)$. Then $\mathrm{TP}_\iota = \delta_l \times \delta_l^M(\theta)$, $\mathrm{TN}_\iota = (1 - \delta_l) \times (1 - \delta_l^M(\theta))$, $\mathrm{FP}_\iota = (1 - \delta_l) \times \delta_l^M(\theta)$ and $\mathrm{FN}_\iota = \delta_l \times (1 - \delta_l^M(\theta))$, respectively.

Let TP, TN, FP, and FN represent the cumulative true positive, true negative, false positive, and false negative,

respectively. Summations over $l = 1, \ldots, N$ yield $\mathrm{TP} = \sum \mathrm{TP}_\iota$, $\mathrm{TN} = \sum \mathrm{TN}_\iota$, $\mathrm{FP} = \sum \mathrm{FP}_\iota$, and $\mathrm{FN} = \sum \mathrm{FN}_\iota$.

Let $\wp^{\mathcal{M}}$ and $\mathcal{R}^{\mathcal{M}}$ denote the precision and recall of the predictions by applying mechanism M to predict a given video $\mathcal{V}$. The values of $\wp^{\mathcal{M}}$ and $\mathcal{R}^{\mathcal{M}}$ can be calculated by $\wp^{\mathcal{M}} = (\mathrm{TP}/(\mathrm{TP} + \mathrm{FP}))$ and $\mathcal{R}^{\mathcal{M}} = (\mathrm{TP}/(\mathrm{TP} + \mathrm{FN}))$, respectively.

Let $\{\theta_j\}_{j=1}^{J} \subset [0, 1]$ denote a finite collection of thresholds. Let $\mathcal{AP}^{\mathcal{M}}$ denote the average precision of mechanism M. $\mathcal{AP}^{\mathcal{M}}$ can be calculated by the following equation:

$$\mathcal{AP}^{\mathcal{M}} = \sum_{j=1}^{J-1} \left(\mathcal{R}^{\mathcal{M}}\left(\theta_{j+1}\right) - \mathcal{R}^{\mathcal{M}}(\theta_j)\right) \times \wp^{\mathcal{M}}\left(\theta_{j+1}\right).$$

Similar to [1], [2], [3], [4], [5], [7], [13], [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28], the first objective of violence detection in this article is to develop mechanism M^{best} that satisfies $M^{\text{best}} = \arg \mathrm{Max}_{\mathcal{M} \epsilon \mathfrak{M}}(\mathcal{AP}^{\mathcal{M}})$.

Let $\mathrm{TPR}(\theta) = (\mathrm{TP}(\theta)/(\mathrm{TP}(\theta) + \mathrm{TN}(\theta)))$ and $\mathrm{FPR}(\theta) = (\mathrm{FP}(\theta)/(\mathrm{TP}(\theta) + \mathrm{TN}(\theta)))$ denote the true-positive and false-positive rates as functions of θ, respectively. Let $\mathcal{A}^{\mathcal{M}}$ denote the AUC of mechanism $\mathcal{M}$. The value of $\mathcal{A}^{\mathcal{M}}$ can be calculated by the following equation:

$$\begin{aligned}\mathcal{A}^{\mathcal{M}} &= \int_0^1 \mathrm{TPR}\,(\theta) d\,(\mathrm{FPR}\,(\theta)) \\ &= \int_0^1 \frac{\mathrm{TP}(\theta)}{\mathrm{TP}\,(\theta) + \mathrm{FN}(\theta)} d\left(\frac{\mathrm{FP}(\theta)}{\mathrm{FP}\,(\theta) + \mathrm{TN}(\theta)}\right).\end{aligned}$$

Similar to [1], [2], [3], [4], [5], [7], [13], [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28], the second objective of violence detection in this article is to develop mechanism $\mathcal{M}^{\text{best}}$ that satisfies $\mathcal{M}^{\text{best}} = \arg \mathrm{Max}_{\mathcal{M} \epsilon \mathfrak{M}}(\mathcal{A}^{\mathcal{M}})$.

B. Retrieval

Given a scenario involving both time-series data and multiple triplet relationships, let $\{\alpha_t\}_{t=1}^{T}$ represent a sequence of observations recorded at each time t, and let $\{(h_n, r_n, t_n)\}_{n=1}^{N}$ denote a collection of triplets, where h_n, r_n, and t_n are the head entity, relation, and tail entity, respectively.

Additionally, let $\{z_n\}_{n=1}^{N} \subseteq \mathbb{R}^d$ denote a set of embeddings, where each embedding z_n corresponds to a triplet (h_n, r_n, t_n) as well as its associated temporal information $\{\alpha_t\}$.

Consider a query vector $q \in \mathbb{R}^d$. Let $s(z_n, q) \in \mathbb{R}$ denote the similarity function, which measures the relevance of z_n to q.

Based on the similarity scores, all embeddings z_n are ranked in descending order of $s(z_n, q)$. Denote the corresponding ranking function as $\pi\colon \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$ such that $s(z_{\pi(1)}, q) \geq s(z_{\pi(2)}, q) \geq \cdots \geq s(z_{\pi(N)}, q)$.

Let $\delta_n \in \{0, 1\}$ denote an indicator function that specifies whether z_n is ground truth relevant to the query q. Let $\sum_{n=1}^{N} \delta_n$ represent the total number of relevant embeddings. Define $R@k$ as the fraction of relevant items retrieved within the top-k positions out of all relevant items. For any $k \in \mathbb{N}$, it is defined as shown in the following equation:

$$R@k = \sum_{r=1}^{k} \delta_{n(r)} \Big/ \sum_{r=1}^{k} \delta_n$$

when $k \in \{1, 5, 10\}$, this corresponds to $R@1$, $R@5$, and $R@10$, respectively. Similar to previous works [6], [29], [31], [32], [33], [34], this study aims to maximize retrieval performance. Within the retrieval strategy space $\mathfrak{R}$, the objective is to identify the optimal retrieval strategy $\mathcal{R}^{\text{best}}$ such that $\mathfrak{R}^{\text{best}} = \arg \mathrm{Max}_{\mathcal{R} \epsilon \mathfrak{R}}\{R@1, R@5, R@10\}$.

The above content constitutes the assumptions and problem formulation of this article. Section IV will introduce the proposed TIO system in detail.

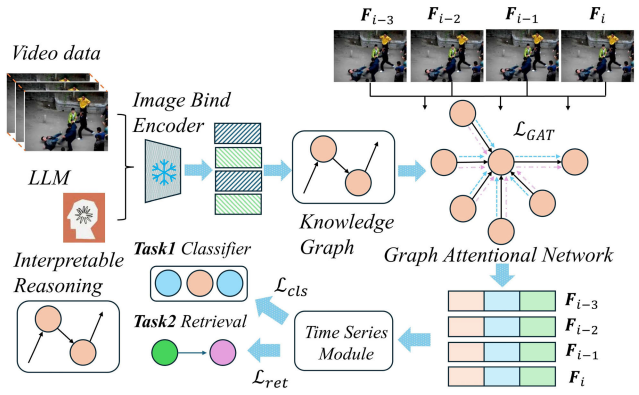


Fig. 3. Proposed TIO system.

IV. Proposed TIO System

This section provides a detailed introduction to the proposed TIO system. Unlike previous systems designed specifically for violence detection [7], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] or VAR [6], [29], [31], [32], [33], [34], the proposed TIO system in this article offers two significant advantages. First, TIO is a multifunctional system capable of both violence detection and retrieval tasks. Second, the design of TIO incorporates interpretability.

Fig. 3 illustrates the operational framework of the proposed TIO system. The system is divided into three main components: feature extraction, lightweight temporal sequence analysis and classification, and retrieval task deployment.

Specifically, in terms of feature extraction, the TIO system leverages LLM and ImageBind pretrained modules to process the information in each frame of training data, constructing a KG and optimizing it using a GAT module. After the feature extraction, the compressed representation of each video frame is further processed by a lightweight temporal sequence module to generate a high-dimensional embedding for the entire video. Finally, the system connects to both a classifier and a retriever to achieve its multitasking objectives. During inference, through the output of the triplet relationships in the KG, users can clearly understand the basis of the model's decisions, thereby achieving interpretability.

The specific designs of these three components are detailed in this section. Section IV-A introduces the construction of the KG and GAT optimization, Section IV-B describes the design of the lightweight temporal sequence module, and Section IV-C explains how to integrate the classifier and retriever to complete the system training.

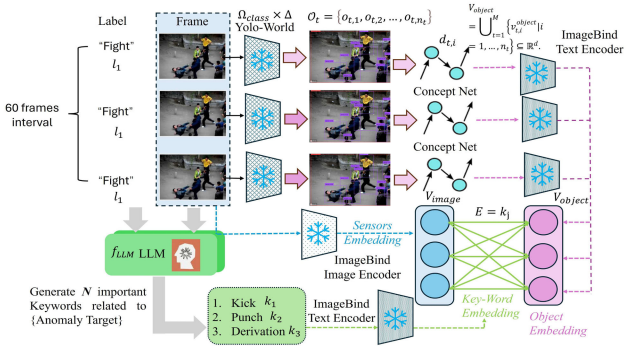


Fig. 4. Details of the acquisition of triplet relationships.

A. Feature Extraction

This section provides a detailed explanation of the feature extraction functionality of the TIO system. Notably, unlike previous works [6], [7], [19], [28], [29], [31], [32], [33], [34], which directly employed pretrained modules for feature extraction, this study adopts a novel approach by constructing a KG and leveraging GAT-based reasoning for frame-level feature extraction. This method not only captures the semantic relationships between multimodal data but also enhances the interpretability and discriminative power of features through structured reasoning, thereby improving the accuracy of recognition and retrieval.

The feature extraction process has three steps: acquisition of triplet relationships, constructing KG, and adjusting the GAT module.

The acquisition of triplet relationships is used to extract entities and their relationships from raw data. The construction of the KG is used to structure the extracted entities and relationships into graph data. The adjustment of the GAT module is used to optimize the graph data and extract high-level features. The specific details are described as follows.

Fig. 4 illustrates the overall framework for generating the triplet relationships. The triplet relationships consist of three components: Entity 1, Entity 2, and the relationship. Entity 1 corresponds to each video frame, while Entity 2 represents the descriptions of objects within the frame. The relationship is a synonym description generated by an LLM based on the current frame and its associated labels.

Through this construction, semantic connections between the image content and the objects are effectively captured, establishing richer contextual relationships. This approach not only enhances the diversity and accuracy of the triplet relationships but also provides high-quality data for the subsequent KG construction.

Specifically, let T denote the total number of frames in the original video. In practice, this article only samples one frame out of every 60 frames to reduce computational overhead. Hence, the total number of sampled frames is $m = T/60$, assuming that T is a multiple of 60 for simplicity. Denote the set of sample frames as $\mathfrak{I} = \{I_t\}_{t=1}^{m}$, where each I_t corresponds to the $(60 \times (t-1)+1)$th original frame.

For each frame I_t, assume that it lies in the space $\Omega_{\text{image}} \subseteq \mathbb{R}^p$. Using a pretrained Imagebind [15] encoder $E_I : \Omega_{\text{image}} \to \mathbb{R}^d$ to map each frame I_t to its corresponding d-dimensional embedding $v_t^{\text{image}} = E_I(I_t), v_t^{\text{image}} \in \mathbb{R}^d$. Accordingly, the set of frame-level entities is defined as $V_{\text{image}} = \{v_1^{\text{image}}, v_2^{\text{image}}, \ldots, v_T^{\text{image}}\} \subseteq \mathbb{R}^d$.

Next, let $L = \{l_1, l_2, \ldots, l_n\}$ denote a predefined set of labels, where each l_i corresponds to a particular video or semantic descriptor. Let $f_{\text{LLM}}(\cdot)$: $\mathcal{L} \times \Omega_{\text{image}} \to \mathcal{R}$ denote the LLM. Using it to generate a set of keyword-based relation types: $\mathrm{K} = \{k_1, k_2, \ldots, k_m\} \subseteq \mathcal{R}$. Each k_j arises by combining the label set $\mathcal{L}$ and frame samples $\mathfrak{I}$ via the LLM. That is, $k_j = f_{\text{LLM}}(I_t, l_i)$.

Let $\mathcal{O}_t = \{o_{t,1}, o_{t,2}, \ldots, o_{t,n_t}\}$ be a set of objects extracted from I_t by applying the YOLO world. For each frame I_t, the pretrained YOLO-World [15] obtains a set of detected objects, where each $o_{t,i}$ is represented as $(\text{class}_{t,i}, \text{bbox}_{t,i}) \in \Omega_{\text{class}} \times \Delta$, with Ω_{class} denoting the possible object classes and $\Delta \subseteq \mathbb{R}^4$ denoting the space of bounding box coordinates.

To enrich the semantic features of each detected object, the textual descriptions $d_{t,i}$ from ConceptNet values [20] and then feed $d_{t,i}$ into a pretrained text encoder $E_T : \Omega_{\text{text}} \to \mathbb{R}^d$, where Ω_{text} is the textual input space to obtain the semantic embedding $v_{t,i}^{\text{object}} = E_T(d_{t,i}), v_{t,i}^{\text{object}} \in \mathbb{R}^d$.

Collecting these embeddings for all objects across the sampled frames yields the object-level entity set $V_{\text{object}} = \bigcup_{t=1}^{M} \{v_{t,i}^{\text{object}} | i = 1, \ldots, n_t\} \subseteq \mathbb{R}^d$. Define a multimodal triple $\tau \in V_{\text{image}} \times \mathrm{K} \times V_{\text{object}}$ of the form $\tau = (v_t^{\text{image}}, k_j, v_{t,i}^{\text{object}})$ to represent the semantic relation k_j between the frame-level embedding v_t^{image} and the object-level embedding v_t^{object}.

The above describes the acquisition of triplet relationships, followed by the construction of the KG. The purpose of constructing the KG is to structure the extracted triplet relationships and represent the semantic connections between entities and relationships in the form of a graph. Through the KG, the associations within the data can be more intuitively represented, providing interpretability for subsequent analysis and reasoning. Additionally, the KG helps integrate data from different sources, achieving semantic-level information fusion, thereby enhancing the system's ability to handle complex relationships and further improving the effectiveness of knowledge inference.

Specifically, aggregating all such possible triples across every $t \in \{1, \ldots, T\}$, every $i \in \{1, \ldots, n_t\}$, and every $j \in \{1, \ldots, m\}$, the edge set E represents in:

$E = \bigcup_{t=1}^{T} \bigcup_{i=1}^{n_t} \bigcup_{j=1}^{m} \{(v_t^{\text{image}}, k_j, v_{t,i}^{\text{object}})\} \subseteq V_{\text{image}} \times \mathrm{K} \times V_{\text{object}}$.

The node set $V = V_{\text{image}} \cup V_{\text{object}}$, and together with the edge set E, they form a directed KG $G = (V, E)$. Finally, the relationship structure of G is encoded in a high-order adjacency tensor $A \in \{0,1\}^{|V| \times |V| \times |\mathrm{K}|}$, where $A_{(v,v,j)} = 1$ if and only if $(u, k_j, v) \in E$, with $u, v \in V$ and $k_j \in \mathrm{K}$. Otherwise, $A_{(v,v,j)} = 0$. This adjacency tensor representation captures not only the links between entities but also the specific type of relation k_j assigned to each edge, enabling a richer and more structured representation of the KG.

After constructing the KG, the GAT module was applied. Unlike previous studies [22], [23] that primarily utilized GNN architectures for time-series analysis, this study innovatively applying GAT to refine the KG.

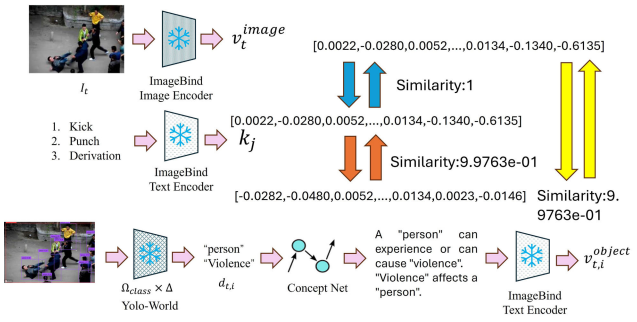


Fig. 5. Example of imagebind encoder.

The main advantage of this approach lies in the attention mechanism of GAT, which effectively assigns weights to the nodes and edges in the KG. This allows the refinement process to more precisely capture key semantic relationships within the graph, as illustrated in the diagram. For example, in Fig. 5, the violent events such as "kick," "punch," and "derivation" are processed through the ImageBind encoders to generate high-dimensional feature vectors (v_t^{image}, k_j, and $v_{t,i}^{\text{object}}$).

During experimentation, a key issue became apparent, as visualized in the diagram. The existing unified cross-modal pretrained model, ImageBind [13], has a high dimensionality (reflected in the feature vectors), leading to significant computational costs for similarity comparisons and dot product operations. Additionally, for violent events of the same category (e.g., "kick" and "punch") and their corresponding textual descriptions, the feature vectors extracted by ImageBind exhibit high similarity.

To address these issues, this study proposes a Euclidean distance-based correction method using an exponential kernel function as an alternative to the computationally expensive similarity comparisons and dot product operations.

Specifically, after constructing the task-related KG $G = (V, E)$ with each node $v \in V$ associated with a feature vector $\mathrm{h}_v \in \mathbb{R}^d$. Let $d_{u,v}$ denote the pairwise Euclidean distance. First, calculate $d_{u,v}$ for any edge $(u, k_j, v) \in E$ as shown in the following equation:

$$d_{u,v} = \sum_{k=1}^{d} \left| h_{u,k} - h_{v,k} \right|^2 = \|h_u - h_v\|_2^2.$$

Then, it finds $d_{\max} = \max_{(u,v)\in E} d_{u,v}$ and $d_{\min} = \min_{(u,v)\in E} d_{u,v}$, which applies interval normalization in the following equation:

$$d'_{uv} = \left(d_{u,v} - d_{\min}\right) \Big/ \left(d_{\max} - d_{\min}\right)$$

and uses a Gaussian kernel to obtain the edge weight $w_{u,v}$ in the following equation:

$$w_{u,v} = \exp\left(-\left(d'_{uv}\right)^2 / \sigma^2\right).$$

Rather than introducing any feature-based similarity or inner-product term, one can define the attention score purely $e_{u,v} = w_{u,v}$. The attention score applies to a softmax over neighbors $N(u)$ to obtain the following equation:

$$\alpha_{u,v} = \exp\left(e_{u,v}\right) / \Sigma_{k\in N(u)} \exp\left(e_{u,k}\right)$$

and finally update the node feature via $h'_u = \sigma\left(\sum_{v\in N(u)} \alpha_{u,v} h_v\right)$ so that attention depends solely on the geometrical separation of nodes as captured by the distance-based kernel.

Lower Complexity Computation Theorem: The proposed kernel-adjusted distance function is more lightweight compared to the attention calculation method.

Proof: Let $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$ denote a set of n high-dimensional vectors. This article compares two methods for computing pairwise similarity/attention scores on all $n(n-1)/2 \approx n^2/2$ pairs (x_i, x_j), where $i, j = 1, \ldots, n$. For multihead dot product method, each vector x_i is projected to query and key vectors for each head $\{1, \ldots, H\}$: $q_i^{(h)} = W_Q^{(h)} x_i$ and $k_i^{(h)} = W_{\mathrm{K}}^{(h)} x_i$, where $W_Q^{(h)}, W_{\mathrm{K}}^{(h)} \in \mathbb{R}^{d\times D}$. The total projection cost across n vectors and H heads is $T_{\text{proj}} = O(HnDd)$. Each pair (x_i, x_j) then requires $\langle q_i^{(h)}, k_i^{(h)} \rangle$ for $h = 1, \ldots, H$, at a cost of $O(d)$ per dot product. Hence, the attention score computation over all pairs scales as $T_{\text{atten}} = O(Hn^2 d)$. Summing both costs leads to total complexity $T_{\text{multi-head}} = T_{\text{proj}} + T_{\text{atten}} = O(HnDd) + O(Hn^2 d)$.

For the proposed distance kernel method, each pair (x_i, x_j), compute the squared Euclidean distance in (4), which takes $O(D)$ time per pair. Over $\approx n^2/2$ pairs, the cost is $T_{\text{dist-calc}} = O(n^2 D)$. A scalar kernel operation in (5), which does not change the overall order; thus, $T_{\text{kernel}} = O(n^2 D)$.

Suppose that there exist constants $C_1, C_2 > 0$ such that $n^2 D < C_1(HnDd)$ and $n^2 D < C_2(Hn^2 d)$, and then, we have $n^2 D < HnDd + Hn^2 d \Longrightarrow T_{\text{kernel}} < T_{\text{multi-head}}$.

B. Lightweight Temporal Sequence Module

Until now, after completing frame-level feature extraction, the extracted features capture instantaneous information but lack the global temporal context critical to video understanding. In this section, a lightweight temporal sequence module is proposed, analogous to the standard receptance weighted key value (RWKV) [35] encoder, incorporating self-attention, layer normalization (LN), and a feed-forward network (FFN). Since the positions of frames in sequential data are already fixed, positional encoding is unnecessary when training the temporal sequence module.

Compared to prior works [22], [23], [24], [25], [26], [27], the primary distinction lies in the computation method of self-attention. This method is designed based on relative distances rather than feature similarity, making the model lightweight. By focusing solely on the relative distances between frames and avoiding the computation of high-dimensional feature similarity, this approach effectively reduces the complexity of attention computation, thereby reducing computational costs and improving training and inference efficiency.

Specifically, let $V = \{v_1, v_2, \ldots, v_n\}$ denote the set of frames in the video V. Using the GAT module, the embeddings $h_{v_i} \in \mathbb{R}^d$ are obtained. All frame embeddings are integrated into a matrix $H = [h_{v_1}, h_{v_2}, \ldots, h_{v_n}] \in \mathbb{R}^{n\times d}$. Define $A_t \in \mathbb{R}^{n\times n}$, where n represents the total number of frames in V. For any $1 \le i \le n$, the elements of A_t are defined in the following equation:

$$(A_t)_{i,j} = \exp\left(-\left|h_{v_{i+1}} - h_{v_i}\right| / \sigma\right)$$

where $\sigma > 0$ is a hyperparameter that controls the degree of temporal distance decay. This design ensures that A_t depends

solely on the relative distances between frame indices, reducing reliance on high-dimensional feature similarity.

To enhance numerical stability, a symmetric normalization strategy is adopted. Define the diagonal degree matrix $D \in \mathbb{R}^{n\times n}$, with diagonal elements given by the following equation:

$$(D)_{i,i+1} = \sum_{i=1}^{n} (A_t)_{i,i+1}$$

and construct the symmetrically normalized adjacency matrix $\tilde{A}_t = D^{-1/2}A_tD^{-1/2}$. This symmetric normalization ensures the stability of gradient propagation and numerical operations during feature fusion.

In practice, the proposed temporal sequence module consists of two stages to integrate and enhance temporal context in the embedding matrix H. The first stage is temporal context fusion. The normalized temporal adjacent matrix $\tilde{A}_t$ is used to weight the frame embedding matrix H, as shown in the following equation:

$$H' = \text{softmax}\left(\tilde{A}_t\right) H \in \mathbb{R}^{n\times d}$$

where softmax($\tilde{A}_t$) performs softmax normalization row-wise on $\tilde{A}_t$ ensuring a reasonable distribution of weighting coefficients. Subsequently, LN is applied to H' to stabilize feature distributions, as shown in the following equation:

$$H'' = \text{LN}(H').$$

The next stage involves a nonlinear transformation and a residual connection. The fused features H'' are subjected to an FFN for nonlinear transformation, with residual connections preserving the original information. The operation is as follows:

$$\begin{aligned} H''' &= \text{LN}\left(\text{FFN}\left(H''\right) + H''\right) \\ &= \text{LN}\left(\text{ReLU}\left(xW_1 + b_1\right)W_2 + b_2\right) \end{aligned}$$

where W_1 and W_2 are the parameter matrices for linear transformations, b_1 and b_2 are bias vectors, and the ReLU function provides nonlinear mapping capabilities, enhancing the representational power of the model.

C. Classifier and Retriever

After completing the time-series computation, the classifier and retriever are integrated. The classifier is used to determine which predefined category the current image frame (or frame-level feature) belongs to. The retriever is used to locate the corresponding textual description in the database, enabling cross-modal retrieval.

Specifically, for the classifier, the loss function is defined as $\mathcal{L}_{\text{cls}}$, as shown in the following equation:

$$\underbrace{\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\delta_{y_i,c}\log p_{i,c}}_{\text{For Classcification}}$$

where N is the number of samples or video frames in a batch, C is the total number of categories, and $\delta_{y_i,c}$ is the indicator function defined as:

$$\delta_{y_i,c} = \begin{cases} 1, & \text{if } y_i = c \\ 0, & \text{if } y_i \neq c \end{cases}$$

$p_{i,c} = P(y_i = c|z_i)$ represents the predicted probability by the classifier that the ith sample belongs to category c and z_i is the feature representation of the frame or sample.

For the retriever, the loss function is defined as L_{ret}, as shown in the following equation:

$$\underbrace{\mathcal{L}_{\text{ret}} = \sum_{(i,j)\epsilon\Omega}\max\left\{0, \alpha + D\left(q_i, t_j^+\right) - \left(q_i, t_j^-\right)\right\}}_{\text{For Retrivel}}$$

where Ω is the set of indices for all text–video pairs, and each pair (i, j) corresponds to a specific match between the ith image (or video) and the jth text. Here, q_i represents the embedding of the ith image (or video) in the shared embedding space, while t_j^+ and t_j^- represent the text vectors matching q_i (positive sample) and not matching q_i (negative sample), respectively; $D(q_i, t_j^+)$ is the distance function; and $\alpha > 0$ is the margin hyperparameter. This loss encourages the distance of correctly matched image–text pairs to be smaller than that of mismatched pairs, enabling effective retrieval judgments.

Through backpropagation with these two loss functions, parameters in the GAT module are also updated. A self-regularization loss is designed for the GAT module, expressed as $\mathcal{L}_{\text{gat}}$ as shown in the following equation:

$$\underbrace{\mathcal{L}_{\text{GAT}} = \lambda\sum_{(i,j)\epsilon E^+} + \lambda\sum_{(i,j)\epsilon E^-} -\log(1-\alpha_{i,j})}_{\text{For GAT Moudle}}$$

where E^+ represents the set of node pairs that should be connected based on prior knowledge or inferred relationships, E^- represents the set of pairs that "should not be connected," $\alpha_{i,j}$ is the weight calculated by GAT, and λ is a balancing hyperparameter. This design allows GAT to automatically adjust weights using prior or detected relationships during the learning of frame-level attention allocation while suppressing interference from noisy edges.

Combining the above multiple objectives, the total loss function $\mathcal{L}_{\text{total}}$ can be expressed in the following equation:

$$\mathcal{L}_{\text{total}} = \alpha_{\text{cls}}\mathcal{L}_{\text{cls}} + \alpha_{\text{ret}}\mathcal{L}_{\text{ret}} + \alpha_{\text{GAT}}\mathcal{L}_{\text{GAT}}$$

where α_{cls}, α_{ret}, and α_{GAT} are weighting coefficients for the three loss components, used to adjust the relative importance of different task objectives during training. By performing end-to-end backpropagation with this total loss function, this method not only learns the discriminative capabilities required by the classifier and retriever for cross-modal matching but also refines and enhances the attention weights in the GAT module, ultimately achieving more comprehensive spatiotemporal relationship modeling and semantic understanding.

V. System Performance

In this section, relevant experimental performance and analysis are presented.

A. Dataset and Evaluation Metrics

For the detection task, the experiment employed two widely recognized datasets: UCF-Crime [18] and XD-Violence [19].

TABLE I
FINE-GRAINED RESULT OF THE SOTA METHODS AND THE PROPOSED TIO ON XD-VIOLENCE ($\mathcal{AP}^{\mathcal{M}}$)

Method	Class					
	Abuse	CarAccident	Explosion	Fighting	Riot	Shooting
CLIP [26]	0.32	12.21	22.26	25.25	66.60	1.26
RTFM [23]	9.25	25.36	53.53	61.73	90.38	18.01
S3R [24]	2.63	23.82	45.29	49.88	90.41	4.34
ActionCLIP [27]	2.73	25.15	55.28	58.09	89.31	12.87
AnomalyCLIP [28]	6.10	**31.31**	68.75	71.44	92.74	26.13
MISSIONGNN [30]	10.98	8.91	42.49	34.06	81.64	15.03
TIO (Ours)	**17.84**	28.89	**73.54**	**79.98**	**82.64**	**32.65**

XD-Violence, the largest publicly available dataset for violence monitoring, includes 4754 videos totaling 217 h and covers six categories of violent incidents: verbal abuse, car accidents, explosions, fights, riots, and shootings. The dataset is divided into a training set of 3954 videos and a test set of 800 videos, with the latter comprising 500 violent and 300 nonviolent videos. The UCF-Crime dataset consists of 1900 real-world surveillance videos, with 1610 allocated for training and 290 for testing.

Regarding evaluation metrics, the experiment adheres to the framework described in Section III-A. For XD-Violence, the metric $\mathcal{AP}^{\mathcal{M}}$ from (1) is used, while for UCF-Crime, the metric $\mathcal{A}^{\mathcal{M}}$ from (2) is applied. This approach aligns with prior research [1], [5], [7], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28].

For the retrieval task, the experiment utilized the UCFCrime-AR [6] dataset, which contains 1900 real-world surveillance videos accompanied by corresponding bilingual text descriptions. Of these, 1610 videos are used for training and 290 for testing. Evaluation follows the setup outlined in Section III-B. Specifically, the retrieval metric $R@k$ from (3) is applied to the XD-Violence dataset, consistent with methodologies from earlier studies [6], [29], [31], [32], [33], [34].

B. Experimental Settings

In the configuration of some pretraining modules, the multimodal model used is ImageBind [13], with the version imagebind_huge. For KG construction, the object detection model for Entity 2 is YOLO-World [15], specifically the YOLO-Worldv2-X. The software utilized for KG storage is ArangoDB. In the configuration of the LLM, the Glaude 3.5 Sonnet API is invoked.

For hyperparameter settings, in (6) and (8), σ is set to 0.25 and 0.3, respectively. In (14), α is set to 0.9; in (15), λ is set to 1; and in (16), $\alpha_{\text{cls}}, \alpha_{\text{ret}}$, and α_{GAT} are set to 1.4, 1.3, and 1, respectively.

The training hardware consists of NVIDIA Tesla A100 40 GB, and the system employs the Adam optimizer during training. The initial learning rate is set to 5×10^{-5}, with a multiplicative decay factor of 0.95 per epoch.

C. Comparison With State-of-the-Art Methods

Table I compares the proposed TIO system with prior studies on fine-grained violence detection using the XD-Violence dataset. The results show that the TIO system outperforms previous methods, especially in dynamic and subtle anomaly detection.

In the Abuse category, the TIO system achieves an $\mathcal{AP}^{\mathcal{M}}$ of 17.84, a 6.86% improvement over the best baseline, MISSIONGNN (10.98), highlighting its strength in low-dynamic scenarios. In the car accident category, it reaches an $\mathcal{AP}^{\mathcal{M}}$ of 28.89, surpassing RTFM (25.36) and S3R (23.82), indicating its ability to detect subtle anomalies in car accidents. For the more dynamic explosion and fighting categories, the TIO system achieves $\mathcal{AP}^{\mathcal{M}}$ of 73.54 and 79.98, respectively. In the shooting category, the TIO system achieves an $\mathcal{AP}^{\mathcal{M}}$ of 32.65, significantly surpassing AnomalyCLIP (26.13) by 6.52%, demonstrating its adaptability to high-speed, dynamic scenarios.

These results highlight the TIO system's effectiveness by combining KG and GAT modules to enhance anomaly detection. KG models semantic relationships between anomalies, while GAT strengthens attention on anomalous features. Additionally, the dual-branch architecture allows for better multimodal fusion, providing higher performance than single-branch systems.

Similarly, as shown in Table II, the proposed TIO system demonstrates outstanding the proposed TIO system excels in the UCF-Crime dataset, outperforming most competitors across various categories. The proposed TIO achieves the highest $\mathcal{A}^{\mathcal{M}}$ in key categories such as Abuse (96.32), Assault (97.02), RoadAcc (97.66), Stealing (97.81), and Vandalism (91.62). Compared to MISSIONGNN, the proposed TIO demonstrates notable improvements, especially in fine-grained anomaly detection, with significantly higher $\mathcal{A}^{\mathcal{M}}$ in RoadAcc and Stealing.

In high-dynamic categories like Explosion (94.88), Fighting (93.23), and Shooting (92.56), TIO outperforms AnomalyCLIP and other methods, showcasing strong audiovisual feature modeling in high-energy scenarios. Notably, TIO achieves 92.56 in Shooting, surpassing AnomalyCLIP's 87.45, demonstrating superior motion and sound capture. In low-dynamic categories, TIO's dual-branch design optimizes multimodal fusion for precise anomaly detection.

These results highlight TIO's architectural strengths, especially its dual-branch design, which models visual and auditory features separately but integrates them effectively. This allows TIO to outperform single-branch systems like RTFM, S3R, and STRL. Furthermore, TIO's use of KG and GAT modules enhances semantic modeling and selective feature attention, driving its exceptional performance across all categories.

TABLE II

FINE-GRAINED RESULT OF THE SOTA METHODS AND THE PROPOSED TIO ON UCF-CRIME ($\mathcal{A}^{\mathcal{M}}$)

Method	Class												
	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	RoadAcc	Robbery	Shooting	Shoplifting	Stealing	Vandalism
CLIP [26]	57.37	80.65	91.72	80.83	74.34	90.31	83.54	87.46	70.22	63.99	71.21	45.49	66.45
RTFM [23]	79.99	62.57	90.53	82.27	85.53	92.76	85.21	90.31	81.17	82.82	92.56	90.23	87.20
S3R [24]	86.38	68.45	92.19	93.55	86.91	93.55	81.69	85.03	82.07	85.32	91.64	94.59	83.82
STRL [25]	95.33	79.26	93.27	91.74	89.06	92.25	87.36	80.24	87.75	84.5	92.31	94.22	88.19
ActionCLIP [27]	91.88	90.47	89.21	86.87	81.31	94.08	83.23	94.34	82.82	70.53	91.60	94.06	89.89
AnomalyCLIP [28]	75.03	**94.56**	**96.66**	94.80	90.08	94.79	88.76	93.30	86.85	87.45	89.47	97.00	89.78
MISSIONGNN [30]	93.68	92.95	94.15	96.89	92.31	94.79	92.38	97.16	81.04	90.57	92.94	97.74	84.03
TIO (Ours)	**96.32**	93.68	96.08	**97.02**	**93.55**	**94.88**	**93.23**	**97.66**	**87.78**	**92.56**	**93.32**	**97.81**	**91.62**

TABLE III

COURSE-GRAINED RESULT OF THE SOTA METHODS AND THE PROPOSED TIO ON XD-VIOLENCE ($\mathcal{AP}^{\mathcal{M}}$) AND UCF-CRIME ($\mathcal{A}^{\mathcal{M}}$)

Method	XD-Violence $\mathcal{AP}^{\mathcal{M}}$	UCF-Crime $\mathcal{A}^{\mathcal{M}}$
CLIP [26]	27.21	58.63
RTFM [23]	77.81	84.03
S3R [24]	80.26	85.99
STRL [25]	(~)	87.43
ActionCLIP [27]	61.01	82.30
AnomalyCLIP [28]	78.51	86.36
MISSIONGNN [30]	98.42	84.48
TIO (Ours)	99.25	88.67

TABLE IV

COMPARISONS WITH THE SOTA METHODS ON UCFCRIME-AR (R@K)

Method	Video→Text		
	R@1	R@5	R@10
Random Baseline	0.3	1.0	3.1
CE [29]	5.5	19.7	32.4
MMT [31]	7.2	23.1	39.0
T2VLAD [32]	6.2	27.9	43.1
X-CLIP [33]	6.9	25.8	40.3
HL-Net [11]	5.5	22.8	35.5
XML [34]	6.6	25.9	43.4
ALAN [6]	7.3	24.8	46.9
TIO(Ours)	**9.8**	**32.1**	**68.8**

TABLE V

ABLATION STUDY IN XD-VIOLENCE ($\mathcal{AP}^{\mathcal{M}}$) AND UCF-CRIME ($\mathcal{A}^{\mathcal{M}}$)

KG	GAT	Time Temporal Sequence Module	$\mathcal{AP}^{\mathcal{M}}$	$\mathcal{A}^{\mathcal{M}}$
✔			78.12	66.54
	✔		81.45	70.32
		✔	76.78	64.89
✔	✔		87.34	76.12
	✔	✔	84.56	72.89
✔		✔	85.21	74.23
✔	✔	✔	99.25	88.67

Table III presents the coarse-grained results of the proposed TIO system on the XD-Violence and UCF-Crime datasets, showing superior performance. On XD-Violence, TIO achieves an $\mathcal{AP}^{\mathcal{M}}$ of 99.25, significantly outperforming MISSIONGNN (98.42). On UCF-Crime, TIO sets a new benchmark with an $\mathcal{A}^{\mathcal{M}}$ of 88.67, surpassing STRL (87.43), AnomalyCLIP (86.36), and S3R (85.99), demonstrating excellent stability and fine-grained anomaly detection. TIO excels in balancing multimodal features, providing stable and superior performance in both high- and low-dynamic scenarios.

In contrast, traditional methods like CLIP (58.63) and ActionCLIP (82.30) perform weaker, highlighting limitations in multimodal fusion and anomaly detection. Although MISSIONGNN nearly matches TIO on XD-Violence, its $\mathcal{A}^{\mathcal{M}}$ score on UCF-Crime (84.48) is notably lower, showing less adaptability to varied scenarios. STRL, while stable on UCF-Crime, lacks results on XD-Violence, limiting its comprehensiveness.

The TIO system employs a dual-branch architecture that independently models visual and auditory features, enhanced by KG to explore semantic relationships between anomalies. This design improves attention to critical features, enabling precise anomaly detection. Experimental results demonstrate that TIO excels in high-dynamic anomaly tasks by combining its dual-branch approach, multimodal fusion, and the integration of GAT with KG, achieving significant performance gains on both the XD-Violence and UCF-Crime datasets.

Table IV shows the ranking results (R@K) of the proposed TIO system and other state-of-the-art (SoTA) methods on the UCFCrime-AR dataset for the Video → Text task. TIO outperforms all methods across R@1, R@5, and R@10, with an R@1 score of 9.8, surpassing ALAN (7.3) and traditional methods like CE and HL-Net (5.5). For R@5 and R@10, TIO achieves 32.1 and 68.8, respectively, outperforming ALAN by 7.3% and 21.9%. TIO also shows a 25% improvement over methods like T2VLAD and XML in R@10, demonstrating strong adaptability in large-scale retrieval scenarios.

The TIO consistently beats the random baseline across all metrics, with the random baseline scoring only 3.1% for R@10. While ALAN performs well in small retrieval ranges, its performance drops for larger ranges, unlike TIO, which maintains stable, superior performance. Traditional methods like CE and HL-Net lag behind, with CE scoring only 32.4% in R@10, less than half of TIO's score.

In conclusion, the TIO system, with its innovative feature extraction using KG and GAT and its multibranch design, excels in both anomaly detection and cross-modal retrieval tasks, demonstrating exceptional adaptability and stability.

D. Ablation Study

Table V presents ablation study results on XD-Violence and UCF-Crime, showing the impact of KG, GAT, and the time temporal sequence module. Enabling a single module improves performance but has a limited effect. For example, KG alone achieves 78.12 $\mathcal{AP}^{\mathcal{M}}$ on XD-Violence and 66.54 $\mathcal{A}^{\mathcal{M}}$ on UCF-Crime. GAT improves these to 81.45 and 70.32, while the time temporal sequence module yields 76.78 and 64.89.

TABLE VI
ABLATION STUDY IN UCFCRIME-AR (R@K)

KG	GAT	Time Temporal Sequence Module	R@1	R@5	R@10
✔			5.4	21.8	52.7
	✔		6.2	23.4	56.1
		✔	4.9	20.1	49.8
✔	✔		8.1	28.3	62.5
	✔	✔	7.4	26.8	60.2
✔		✔	7.9	27.5	61.3
✔	✔	✔	**9.8**	**32.14**	**68.8**

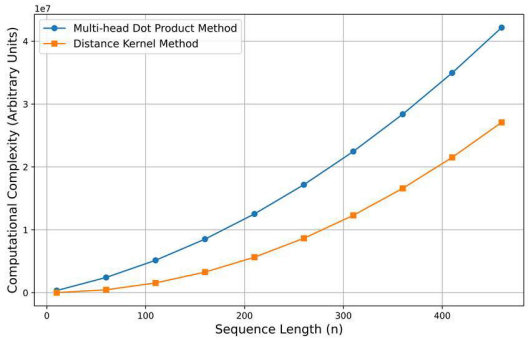


Fig. 6. Computational complexity for proof 1.

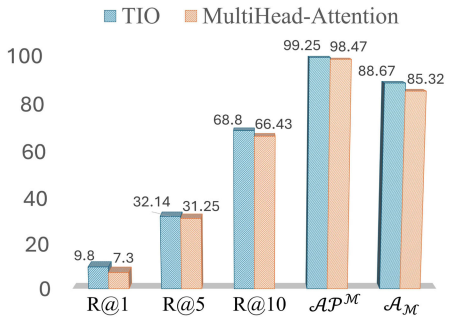


Fig. 7. Compare with multihead attention. (a) demonstrates the performance curve for XD-Violence dataset, (b) presents similar trends for the UCF-Crime dataset.

Combining two modules significantly boosts the performance. KG and GAT together achieve 87.34 $\mathcal{AP}^{\mathcal{M}}$ and 76.12 $\mathcal{A}^{\mathcal{M}}$. GAT with the time temporal sequence module results in 84.56 and 72.89, while KG and the time temporal sequence module reach 85.21 and 74.23. Enabling all three modules leads to optimal performance, with 99.25 $\mathcal{AP}^{\mathcal{M}}$ and 88.67 $\mathcal{A}^{\mathcal{M}}$. This shows that combining KG, GAT, and the time temporal sequence module maximizes performance by capturing semantic context, enhancing feature attention, and modeling temporal dynamics, significantly improving TIO system adaptability in diverse scenarios.

Table VI presents the ablation study on UCFCrime-AR, highlighting the critical roles of KG and GAT in retrieval tasks. Enabling KG significantly improves retrieval performance (R@K) by structuring semantic relationships, while GAT enhances feature attention by capturing key node interactions.

Combining KG and GAT leads to further performance gains, especially in R@1 and R@5, demonstrating their synergy in balancing semantic modeling and feature attention. Adding the time temporal sequence module incorporates temporal relationships, improving R@10 by providing contextual information.

When fully integrated, the TIO system achieves optimal performance (R@1 = 9.8, R@5 = 32.14, and R@10 = 68.8), showcasing the effectiveness of combining all modules.

Fig. 6 shows the computational complexity of the multihead dot product method and the distance kernel method as sequence length n increases. While both methods exhibit quadratic growth, their rates differ significantly. The multihead dot product method has a steeper curve due to the quadratic complexity of attention computation $T_{\text{atten}} = O(Hn^2d)$, which dominates as n grows.The projection cost $T_{\text{proj}} = O(HnDd)$ contributes minimally, resulting in sharply increasing costs for longer sequences.

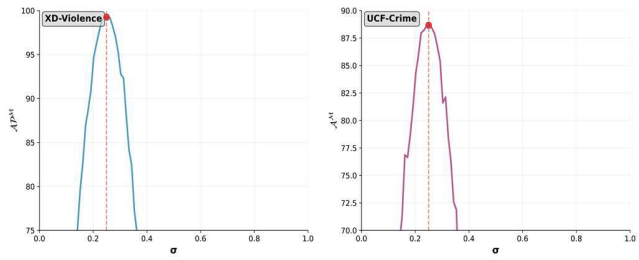
In contrast, the distance kernel method's growth is dominated by the cost of pairwise Euclidean distance calculations $T_{\text{dist-calc}} = O(n^2D)$ but grows more slowly, as it is unaffected by parameters like H or d. For short sequences, both methods have similar costs, but as n increases, the distance kernel method becomes significantly more efficient, while the multihead dot product method is better suited for short sequences where its attention flexibility adds value.

In conclusion, for large-scale data, the distance kernel method is computationally more efficient, aligning with theoretical analysis and validating its suitability for long-sequence scenarios.

Fig. 7 presents the ablation study, showing that the proposed TIO system outperforms the multihead attention method across all retrieval metrics (R@1, R@5, and R@10) and anomaly detection metrics ($\mathcal{AP}^{\mathcal{M}}$ and $\mathcal{A}^{\mathcal{M}}$). TIO achieves 9.8 for R@1, higher than multihead attention's 7.3, and scores 32.14 (R@5) and 68.8 (R@10), surpassing multihead attention's 31.25 and 66.43. In anomaly detection, TIO achieves $\mathcal{AP}^{\mathcal{M}} = 99.25$ on XD-Violence and $\mathcal{A}^{\mathcal{M}} = 88.67$ on UCF-Crime, outperforming multihead attention's 98.47 and 85.32.

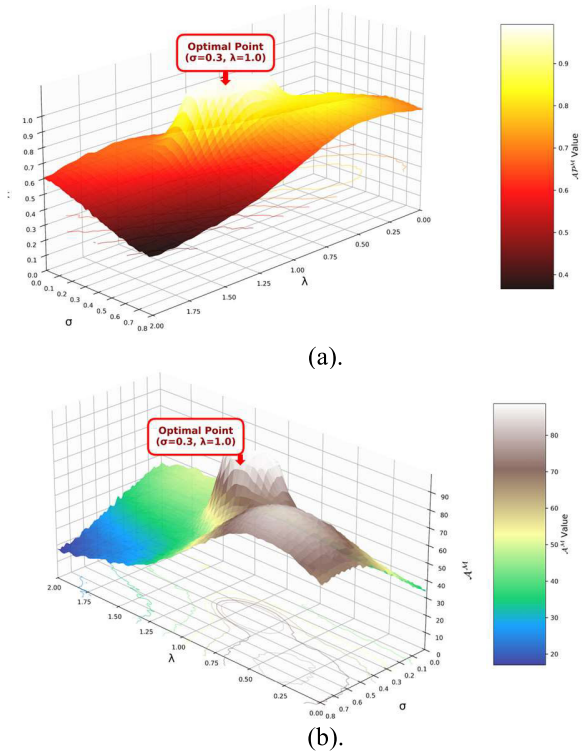
The proposed TIO's advantage lies in its use of the distance kernel method for self-attention, relying on relative distances rather than feature similarity, and incorporating a temporal sequence module. This approach reduces computational complexity by focusing on frame-to-frame relative distances, avoiding high-dimensional similarity calculations. The result is lower computational cost, improved efficiency, and superior performance in retrieval and anomaly detection tasks, as demonstrated by the ablation study.

Fig. 8 illustrates the system performance sensitivity to the σ parameter variations in (6) across both datasets. Fig. 8(a) demonstrates the performance curve for the XD-Violence dataset, where $\mathcal{AP}^{\mathcal{M}}$ exhibits a pronounced peak at $\sigma = 0.25$, achieving the maximum performance of 99.25%. The curve displays a sharp, symmetric bell-shaped distribution with rapid performance degradation on both sides of the optimal value. Performance drops significantly below $\sigma = 0.2$ and above $\sigma = 0.4$, indicating that the XD-Violence dataset requires precise σ tuning for optimal results.



(a). (b).

Fig. 8. System performance under different σ value settings in (6).



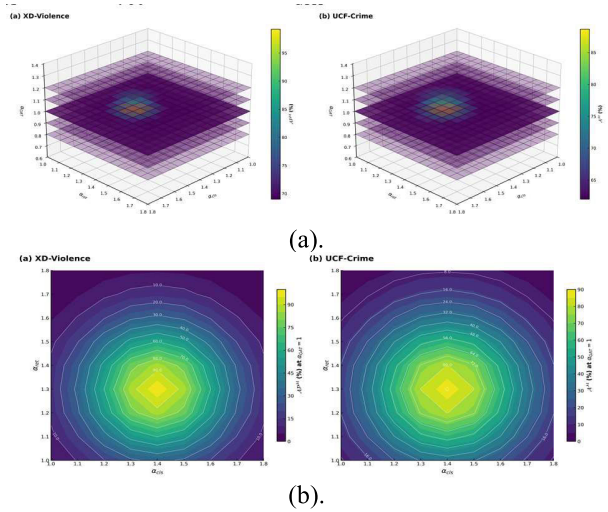
(a).

(b).

Fig. 9. System performance under different σ value settings in (8) and λ. (a) displays the performance landscape for the XDViolence dataset. (b) Hows the corresponding analysis for UCF-Crime dataset.

Fig. 8(b) presents similar trends for the UCF-Crime dataset, with $\mathcal{AP}^{\mathcal{M}}$ reaching its peak performance of 88.47 at the same optimal $\sigma = 0.25$. The performance curve exhibits comparable sensitivity patterns, though with slightly more gradual transitions compared to XD-Violence. The consistent optimal σ value across both datasets suggests that this parameter setting captures fundamental characteristics of violence detection tasks rather than dataset-specific artifacts. The narrow optimal range around $\sigma = 0.25$ emphasizes the critical importance of proper parameter selection, as deviations beyond ± 0.1 result in substantial performance losses exceeding 10% on both datasets. These findings validate the robustness of $\sigma = 0.25$ as the optimal configuration and demonstrate the method's consistent behavior across different violence detection scenarios.

Fig. 9 presents a comprehensive 3-D analysis of system performance sensitivity to both σ and λ parameters in (8). Fig. 9(a) displays the performance landscape for the XD-Violence dataset, revealing a smooth, mountain-like surface with a distinct peak at $\sigma = 0.3$ and $\lambda = 1.0$. The surface exhibits gradual performance transitions across the parameter space, with the highest performance regions concentrated in a relatively narrow valley around the optimal configuration. The color gradient from dark red (low performance) to bright yellow (high performance) clearly delineates the optimal parameter region, demonstrating that performance degrades progressively as parameters deviate from the optimal values.



(a).

(b).

Fig. 10. System performance under different settings in $\alpha_{\text{cls}}, \alpha_{\text{ret}}$, and α_{GAT}. (a) provides a three-dimensional visualization of the complete parameter space. (b) Presents detailed 2D contour maps with α_{GAT} fixed at 1.0.

Fig. 9(b) shows the corresponding analysis for the UCF-Crime dataset, presenting a more complex performance topology with the same optimal point at $\sigma = 0.3$ and $\lambda = 1.0$. The surface displays steeper gradients and more pronounced performance variations compared to XD-Violence, suggesting higher parameter sensitivity. The multicolored surface transitions from blue (lowest performance) through green and yellow to white (highest performance), indicating a broader performance range. Both visualizations consistently identify the same optimal parameter combination, validating the robustness of $\sigma = 0.3$ and $\lambda = 1.0$ across different datasets. The 3-D perspective reveals that while both parameters influence performance, σ exhibits more critical sensitivity, as evidenced by the steeper performance drops along the σ-axis compared to the λ-axis, emphasizing the importance of precise σ tuning for optimal violence detection performance.

Fig. 10 presents a comprehensive analysis of system performance sensitivity across different weight configurations for α_{cls}, α_{ret}, and α_{GAT}. Fig. 10(a) provides a three-dimensional (3-D) visualization of the complete parameter space, where semi-transparent layers represent different α_{GAT} values. The 3-D perspective reveals complex interaction patterns among all three parameters, with color intensity indicating performance levels. Red contour lines delineate high-performance regions exceeding 95% of maximum scores, demonstrating that optimal performance concentrates around $\alpha_{\text{cls}} = 1.4$, $\alpha_{\text{ret}} = 1.3$, and $\alpha_{\text{GAT}} = 1.0$.

Fig. 10(b) presents detailed two-dimensional (2-D) contour maps with α_{GAT} fixed at 1.0, enabling precise examination of α_{cls} and α_{ret} interactions. The concentric contour patterns exhibit smooth performance transitions with consistent optimal regions at $\alpha_{\text{cls}} = 1.4$ and $\alpha_{\text{ret}} = 1.3$, achieving $\mathcal{AP}^{\mathcal{M}} = 99.25$ on XD-Violence and $\mathcal{A}^{\mathcal{M}} = 88.67$ on UCF-Crime. The circu-

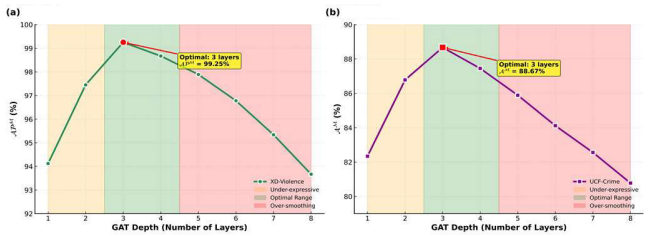


Fig. 11. System performance under different layers of GAT.

lar contour geometry suggests balanced sensitivity between classification and retrieval loss components, while tight contour spacing around the optimum emphasizes the importance of precise hyperparameter selection. These findings validate the robustness of the proposed weight configuration across different datasets and demonstrate the effectiveness of the multiobjective optimization strategy for enhanced violence detection performance.

Fig. 11 presents a comprehensive analysis of GAT depth sensitivity across both datasets, revealing critical insights into the network's architectural requirements. (a) demonstrates the performance trajectory for the XD-Violence dataset, where $\mathcal{AP}^{\mathcal{M}}$ exhibits a clear inverted-U pattern with optimal performance at three layers, achieving 99.25%. The curve shows three distinct phases: an underexpressive region (1–2 layers) highlighted in orange where shallow networks fail to capture complex node relationships, an optimal range (2–4 layers) marked in green representing the sweet spot for performance, and an oversmoothing region (>4 layers) indicated in red where excessive depth degrades performance due to gradient vanishing and node feature homogenization.

Fig. 11(b) reveals similar patterns for the UCF-Crime dataset, with $\mathcal{A}^{\mathcal{M}}$ peaking at 88.67% for three-layer GAT configuration. However, UCF-Crime exhibits higher sensitivity to depth variations, showing steeper performance degradation beyond the optimal point compared to XD-Violence. The performance drop from three to eight layers is more pronounced (7.89% versus 5.58% for XD-Violence), suggesting that UCF-Crime's graph structure is more susceptible to oversmoothing effects. Both datasets consistently identify three layers as the optimal GAT depth, validating the robustness of this architectural choice. The colored regions effectively illustrate the tradeoff between network expressiveness and oversmoothing, emphasizing that while deeper networks can theoretically model more complex relationships, they suffer from diminishing returns and eventual performance degradation, making careful depth selection crucial for optimal violence detection performance.

E. Visualization

As shown in Fig. 12, we further illustrate the alignment between the model-generated heatmaps and the KG triplets, highlighting the model's capability to explain violent events. In each example, the heatmap identifies the regions of visual attention, while the corresponding structured triplet provides semantic interpretation, forming a complete pathway from attention localization to causal reasoning.

The top-left example depicts a grabbing scenario, where the heatmap focuses on the upper bodies of two individuals,

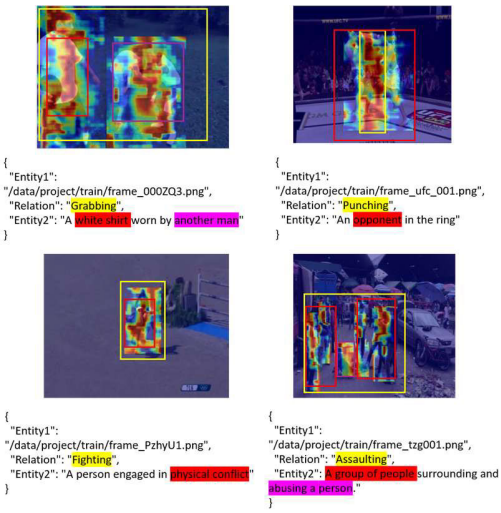


Fig. 12. Heatmap and KG visualization.

especially the area of physical contact. Based on this, the model generates the triplet ("Grabbing," "white shirt," and "another man"), accurately capturing the causal relationship between the action initiator, the type of interaction, and the target. In the top-right example, a fighting scene is shown where the model detects one participant attacking the other, resulting in the triplet ("Punching" and "opponent"), which aligns well with the attention region and clearly reflects the interaction between subject and object.

The bottom-left example presents a more abstract description of a multiperson physical conflict with the triplet ("Fighting" and "physical conflict"). In contrast, the bottom-right image illustrates a more complex humiliation scenario, with the heatmap concentrating on a group of individuals surrounding a person on the ground. The generated triplet ("Assaulting," "A group of people," and "abusing a person") not only distinguishes between perpetrators and the victim but also semantically encodes the violent behavior in a structured format.

Additionally, the use of color-coded labels and bounding boxes enhances both the readability and visual clarity of the information. Each generated triplet corresponds closely to the high-attention areas in the heatmap, demonstrating that the proposed explanation module can effectively extract causal relationships from visual cues and present them in a structured, human-understandable form.

Fig. 13 provides an example of traffic accident detection (RoadAcc), using the SAM2 [36] model to focus on objects in the scene. The heatmap highlights two vehicles, enabling the system to identify the accident. The KG structurally represents semantic relationships, such as distances and interactions between vehicles, while GAT enhances attention to critical features, allowing effective detection of anomalies.

The detection curve shows the decision score rising steadily as the scene becomes abnormal, exceeding the threshold in the detection region and successfully identifying the accident. This integration of multimodal information and contextual modeling is crucial for detecting traffic accidents and other fine-grained anomalies.

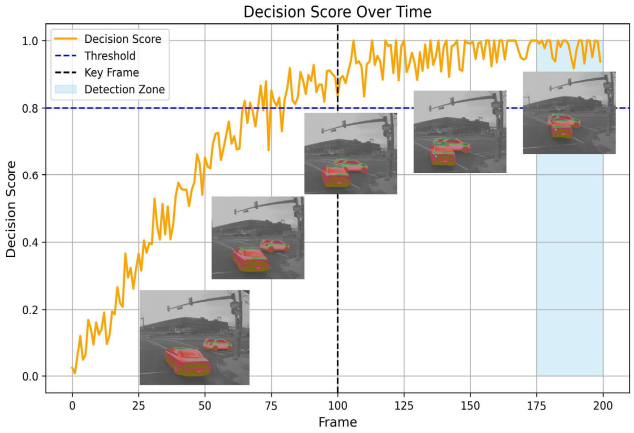


Fig. 13. Simple example of roadacc.

Fig. 14. Simple example of VAR.

As shown in Fig. 14, we present the interpretable outputs of the proposed TIO system in the task of violent event retrieval. Each example consists of a query video, an automatically generated natural language description (Output), and a set of structured triplet-based explanations (Interpretable Output). To further verify the semantic consistency between the model's explanations and the original textual descriptions, we highlight key entities and actions in the output text and compare them with the generated triplets, visually illustrating their semantic alignment.

In the case of "Burglary005," the model successfully captures multiple elements highly relevant to a burglary scenario, such as "a white shirt worn by the man," "kicking the door open," and "his companion," which are reflected across several triplets. For "Shooting004," the system identifies critical actions like "assault a police officer" and "was shot by a policeman with a gun" and expresses them through triplets such as ("Assault" and "A man attempting to attack a police officer") and ("Defense" and "A police officer defending himself with a gun"), effectively reconstructing the causal chain of the event.

Similarly, in "Robbery048," the model extracts fine-grained semantic components such as "a head-covered man holding a gun" and "a man in a suit being pulled forcefully," which are successfully mapped to corresponding visual segments. In "Stealing019," the system captures key details like "a man on the phone" and "white car" and further identifies important actions in the sequence, such as "leaving the scene," resulting in a complete and coherent semantic explanation.

Overall, the structured triplets generated by the proposed TIO system demonstrate strong semantic consistency with the natural language descriptions. The use of color-coded alignment helps validate the accuracy of the model's outputs in reflecting core actions, entities, and roles, indicating that the system is not only capable of precise retrieval but also provides a high level of interpretability.

VI. Limitation

Despite the promising performance of the proposed TIO system, several limitations warrant acknowledgment, particularly in four challenging scenarios.

1) *Occlusion-Related Limitations:* The TIO system struggles in heavily occluded scenarios where violent behaviors are obscured by objects or individuals. When critical visual information is unavailable, the YOLO-World module fails to accurately identify key entities, resulting in incomplete KG construction. This leads to suboptimal GAT attention distributions and compromised interpretability, as generated triplet relationships may not reflect actual violent behaviors in occluded regions.
2) *Low-Light Constraints:* Performance degrades significantly under poor lighting conditions. The ImageBind encoder produces less discriminative embeddings in low-light scenarios, undermining the effectiveness of the Euclidean distance-based kernel function in the GAT module. Additionally, reduced YOLO-World accuracy in object detection leads to incomplete KGs and subsequent reasoning failures.
3) *Rare Violence Type Recognition:* The system demonstrates limitations with rare or atypical forms of violence underrepresented in training datasets. LLM-generated keywords may not adequately capture the semantic nuances of uncommon violent scenarios, while ConceptNet-enhanced information may lack comprehensive coverage for rare violence types. This results in impoverished graph representations that fail to model unique characteristics of atypical violent behaviors.
4) *The Quality of the Prompts and Selected Keywords:* This research approach heavily depends on the quality of the prompts used for keyword generation, as well as

the accuracy and relevance of the selected keywords. Poorly designed prompts or low-quality keywords can lead to incomplete or imprecise representations in the KG, thereby limiting its semantic expressiveness and relational inference capabilities. As a result, the overall performance of downstream tasks may be significantly compromised.

These limitations highlight important directions for future research, including robust feature extraction for challenging visual conditions, expanded datasets covering diverse violence types, and computational optimization for practical deployment.

VII. Conclusion

This study introduces an innovative TIO system designed to address WSVD and VAR tasks. The system significantly enhances interpretability, performance, and user experience. It achieves end-to-end multitask processing through the integration of a knowledge graph-based (KG-based) approach, a kernel-adjusted GAT module, and a dual-branch design. By leveraging these components, the system overcomes interpretability challenges posed by the opaque nature of existing methods. It also reduces complexity using kernel function-based Euclidean distance within unified multimodal embeddings and GATs. Additionally, the lightweight temporal module enhances computational efficiency and model performance.

Unlike traditional independent model designs, the TIO system employs a unified multitask framework. This reduces redundancy in design and training, promotes information sharing, and optimizes semantic reasoning. As a result, it achieves improved system efficiency and interpretability. Experimental results on XD-Violence, UCF-Crime, and UCFCrime-AR demonstrate the proposed TIO system's superior performance and interpretability in violence detection and retrieval.

References

[1] M. Wang, D. Zhou, and M. Chen, "Anomaly monitoring of nonstationary processes with continuous and two-valued variables," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 1, pp. 49–58, Jan. 2023.

[2] Z. Chen, K. Huang, D. Wu, C. Yang, and W. Gui, "Global information-based lifelong dictionary learning for multimode process monitoring," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 12, pp. 1–13, Dec. 2024, doi: 10.1109/TSMC.2024.3442168.

[3] L. Ji, X. Liu, C. Zhang, S. Yang, and H. Li, "Fully distributed dynamic event-triggered pinning cluster consensus control for heterogeneous multiagent systems with cooperative–competitive interactions," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 1, pp. 394–404, Jan. 2023.

[4] W. Lin, S. Wang, W. Wu, D. Li, and A. Y. Zomaya, "HybridAD: A hybrid model-driven anomaly detection approach for multivariate time series," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 866–878, Feb. 2024.

[5] Q. Peng, Y. Liu, Y. Jin, X.-G. Yang, R. Wang, and K. Liu, "Coating feature analysis and capacity prediction for digitalization of battery manufacturing: An interpretable AI solution," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 55, no. 1, pp. 284–294, Jan. 2025, doi: 10.1109/TSMC.2024.3468010.

[6] P. Wu, J. Liu, X. He, Y. Peng, P. Wang, and Y. Zhang, "Toward video anomaly retrieval from video anomaly detection: New benchmarks and model," *IEEE Trans. Image Process.*, vol. 33, pp. 2213–2225, 2024.

[7] P. Wu et al., "Weakly supervised video anomaly detection and localization with spatio-temporal prompts," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 9301–9310.

[8] P. Wu et al., "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 6074–6082.

[9] P. Wu, X. He, M. Tang, Y. Lv, and J. Liu, "HANet: Hierarchical alignment networks for video-text retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3518–3527.

[10] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14592–14601.

[11] P. Wu, X. Liu, and J. Liu, "Weakly supervised audio-visual violence detection," *IEEE Trans. Multimedia*, vol. 25, pp. 1674–1685, 2023.

[12] Y. Zhang, B. Sun, J. He, L. Yu, and X. Zhao, "Multi-level neural prompt for zero-shot weakly supervised group activity recognition," *Neurocomputing*, vol. 571, Feb. 2024, Art. no. 127135.

[13] R. Girdhar et al., "ImageBind one embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15180–15190.

[14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–7.

[15] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-world: Real-time open-vocabulary object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16901–16911.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[17] B. Heo, S. Park, D. Han, and S. Yun, "Rotary position embedding for vision transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 289–305.

[18] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 322–339.

[19] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[20] R. E. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4444–4451.

[21] Y. Ji and H. Lee, "Event-based anomaly detection using a one-class SVM for a hybrid electric vehicle," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6032–6043, Jun. 2022.

[22] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[23] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4955–4966.

[24] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 729–745.

[25] G. Li, G. Cai, X. Zeng, and R. Zhao, "Scale-aware spatio-temporal relation learning for video anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 333–350.

[26] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[27] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," 2021, *arXiv:2109.08472*.

[28] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, "Delving into CLIP latent space for video anomaly recognition," *Comput. Vis. Image Understand.*, vol. 249, Dec. 2024, Art. no. 104163.

[29] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," 2019, *arXiv:1907.13487*.

[30] S. Yun, R. Masukawa, M. Na, and M. Imani, "MissionGNN: Hierarchical multimodal GNN-based weakly supervised video anomaly recognition with mission-specific knowledge graph generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 4736–4745.

[31] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 214–229.

[32] X. Wang, L. Zhu, and Y. Yang, "T2VLAD: Global–local sequence alignment for text-video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5075–5084.

[33] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 638–647.
[34] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A large-scale dataset for video-subtitle moment retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 447–463.
[35] B. Peng et al., "RWKV: Reinventing RNNs for the transformer era," 2023, *arXiv:2305.13048*.
[36] N. Ravi et al., "SAM 2: Segment anything in images and video," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 1–8.

Wen-Dong Jiang (Graduate Student Member, IEEE) received the B.S. degree in multimedia and game science development from the Lunghwa University of Science and Technology, Taoyuan, Taiwan, in 2021, and the M.S. degree in computer science and information engineering from Ming Chuan University, Taoyuan, in 2023, under the supervision of Prof. Yue-Shi Lee and Prof. Show-Jane Yen. He is currently pursuing the Ph.D. degree in computer science and information engineering with Tamkang University, New Taipei City, Taiwan, under the supervision of Distinguished Prof. Chih-Yung Chang.

He has published several research articles in leading journals, including IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL, IPM, and ESWA. His research interests include interpretable machine learning and its applications in smart cities.

Dr. Jiang was a recipient of the Best Paper Award at WASN 2024 and DLT 2024 and the Best Presentation Award at ICCE-TW 2025 Special Session on Artificial Intelligence Applications and Technologies in the Internet of Things. He serves as a reviewer for multiple international journals.

Yu-Ting Chin received the Ph.D. degree in computer science and information engineering from Tamkang University, New Taipei City, Taiwan, in 2017.

He is currently an Assistant Professor at the Department of Artificial Intelligence, Tamkang University. His current research interests include the Internet of Things, wireless sensor networks, ad hoc wireless networks, software-defined networking, and low-power wide-area network technologies.

Yu-Ting Yang received the Ph.D. degree in computer science and information engineering from Tamkang University, New Taipei City, Taiwan, in 2024.

She is currently an Assistant Professor with the Center for General Education, Fu Jen Catholic University, New Taipei City. Her research interests focus on adaptive education.

Chih-Yung Chang (Member, IEEE) received the Ph.D. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 1995.

He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering and the Department of Artificial Intelligence, Tamkang University, New Taipei City, Taiwan. His current research interests include artificial intelligence, deep learning and machine learning, the Internet of Things, and wireless sensor networks.

Dr. Chang is the Co-Chair of ACM SIGMOBILE, Taiwan. He has been serving as an Associate Guest Editor for several SCI-indexed journals, including *International Journal of Ad Hoc and Ubiquitous Computing* since 2011 and *Journal of Applied Science and Engineering* since 2018. He was an Associate Guest Editor of *International Journal of Distributed Sensor Networks* from 2012 to 2014, *IET Communications* in 2011, *Telecommunication Systems* in 2010, *Journal of Information Science and Engineering* in 2008, and *Journal of Internet Technology* from 2004 to 2008.

Diptendu Sinha Roy (Senior Member, IEEE) received the Ph.D. degree in engineering from the Birla Institute of Technology, Ranchi, India, in 2010.

In 2016, he joined the Department of Computer Science and Engineering, National Institute of Technology (NIT) Meghalaya, Shillong, India, as an Associate Professor, where he has been working as the Chair of the Department of Computer Science and Engineering since January 2017. Prior to his stint at NIT Meghalaya, he worked with the Department of Computer Science and Engineering, National Institute of Science and Technology, Berhampur, India. His current research interests include software reliability, distributed and cloud computing, and the Internet of Things (IoT), specifically applications of artificial intelligence/machine learning for smart integrated systems.

Dr. Roy is a member of the IEEE Computer Society.