TargetVAU: Multimodal Anomaly-Aware Reasoning for Target Behavior Understanding in Videos

Lingru Zhou[1], Peng Wu[1*] Manqing Zhang[2], Qingsheng Wang[1], Guansong Pang[3], Peng Wang[1]

[1]School of Computer Science, Northwestern Polytechnical University, China
[2]School of Software, Northwestern Polytechnical University, China
[3]School of Computing and Information Systems, Singapore Management University, Singapore
{lingruzhou, zmqgeek, wqshmzh}@mail.nwpu.edu.cn, xdwupeng@gmail.com, gspang@smu.edu.sg, peng.wang@nwpu.edu.cn

Abstract

Understanding anomalous human behaviors at a fine-grained level remains a major challenge in complex scenarios. Existing video anomaly understanding (VAU) methods often rely on coarse frame-level cues or overlook structured modeling of individual actions, limiting their capacity for reasoning about human interactions and accountability. To address these challenges, we propose TargetVAU, a multimodal anomaly-aware reasoning framework designed for individual-level anomaly recognition and explanation. TargetVAU first extracts both global-level and human-centric visual features using a frozen Vision Transformer (ViT) encoder. An Anomaly-focused Temporal Sampler is then employed to select behaviorally informative frames via a density-aware strategy guided by predicted anomaly scores. A Spatio-Temporal Interaction Graph is constructed to explicitly model interactions among individuals across time and space. These structured representations are fused with prompt embeddings via a frozen Q-Former to form a unified semantic representation. Finally, a large language model fine-tuned with low-rank adaptation (LoRA) performs instruction-guided reasoning to identify anomalous individuals and generate natural language explanations. Extensive experiments on UCCD and HIVAU-70K demonstrate that TargetVAU significantly outperforms existing methods in both accuracy and interpretability, advancing the state of individual-level anomaly understanding in surveillance videos.

Introduction

In traditional video anomaly detection (VAD) methods, anomalies are usually identified based on the predicted anomaly scores when only pre-defined video-level anomaly categories are provided. However, such methods can only roughly tell us whether an anomaly exists, but it is difficult to deeply reveal what happened and why it is considered an anomaly. Therefore, in order to understand anomalous events more thoroughly at the semantic level, it is necessary to break out of the limitations of simple detection. With the rapid development of multimodal technology, large models, and the continuous promotion of higher security standards, the single existence of anomalies is no longer sufficient to meet actual needs; instead, a more comprehensive insight into the anomalous events themselves is required. Based on this, video anomaly understanding (VAU) came into being and became an indispensable task in the fields of public safety, monitoring, and risk prevention. It not only focuses on whether there are anomalies in the video, but also focuses on the specific circumstances of the anomalous events, thereby providing favorable support for subsequent comprehensive analysis.

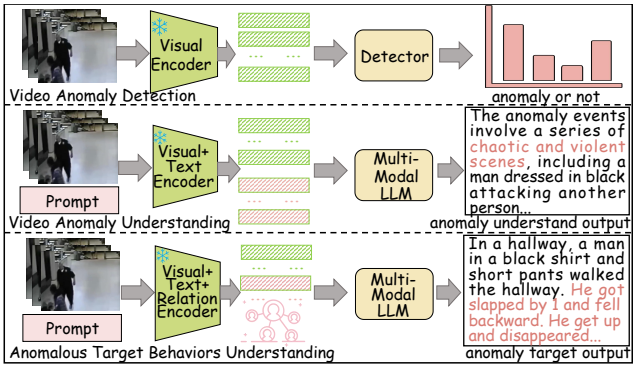

Figure 1: Comparison of three paradigms of video anomaly analysis.

Recently, researchers have explored more about VAU (Du et al. 2024a,b; Zhang et al. 2025), mainly by integrating multimodal information such as vision and text to capture richer semantics and provide interpretable anomaly descriptions. These approaches aim to shift anomaly recognition from rigid classification toward more flexible semantic understanding. However, most existing methods remain limited to video- or event-level analysis, offering only coarse-grained, global explanations. This is because they often treat the entire video as a unified whole or focus solely on high-level scene features, overlooking individual behaviors and interactions. However, in densely populated or public surveillance scenarios, anomalies frequently stem from complex interpersonal interactions. Without capturing the behaviors and relationships of individuals within the scene, it becomes difficult to comprehensively identify anomalous elements and accurately localize the key actors responsible for the

*Corresponding Author

anomalies. To address this limitation, we propose the task of Anomalous Target Behavior Understanding (ATBU). This task aims to enable fine-grained, individual-level anomaly recognition and interpretation by modeling the appearance attributes, behaviors, and interaction dynamics of each individual. It seeks to answer the critical questions of who is doing what, and why the behavior is considered anomalous. This new perspective not only facilitates a deeper understanding of anomalies in complex scenarios but also supports timely intervention, offering more effective decision-making strategies for surveillance and public safety management.

In order to solve the problems of difficulty in identifying individual behaviors, lack of interactive relationship modeling, and insufficient semantic interpretation capabilities faced in the ATBU task, we propose TargetVAU, a multimodal framework for the identification and behavior understanding of anomalous individuals in videos. This framework focuses on modeling and interpretation at the individual level, aiming to improve the ability of model to understand who is doing what and why in complex anomalous events. Structurally, TargetVAU uses pre-trained visual models to extract global features and human-centric features from videos, and combines text semantic information encoded by language models to construct individual interaction graphs across time and space to capture behavioral changes and interactive relationships between individuals. On this basis, its instruction-based multimodal large language model (MLLM, LLM) is inferred to generate individual-level anomaly identification and natural language interpretation, thereby achieving a more fine-grained, more readable and interpretable understanding of anomalous events. Figure 1 illustrates three different paradigms of video anomaly analysis, namely the detection-oriented paradigm based on VAD, the coarse-grained understanding paradigm based on VAU, and the proposed TargetVAU based on ATBU.

In summary, our main contributions are as follows:

- We introduce a novel task, Anomalous Target Behavior Understanding, which aims to detect anomalous individuals in videos and interpret their behaviors at a fine-grained level.
- We propose TargetVAU, a multimodal reasoning framework that extracts both global and individual-level features, integrates textual semantics and relational cues, and performs joint identification and explanation of anomalous behaviors.
- We conduct extensive experiments on UCCD and HIVAU-70K datasets, demonstrating the effectiveness of our approach in both accurately localizing anomalous individuals and generating interpretable explanations.

Related Work

Video Anomaly Detection

Video anomaly detection aims to temporally identify anomalous frames in untrimmed video sequences. Existing methods generally fall into three main categories: unsupervised, weakly-supervised, and fully-supervised approaches. Unsupervised methods (Gong et al. 2019; Xu et al. 2017; Yang et al. 2023; Liu et al. 2018, 2021) focus on modeling normal patterns by training exclusively on normal videos, employing self-supervised strategies such as reconstruction-based (Gong et al. 2019; Xu et al. 2017; Yang et al. 2023) or prediction-based techniques (Liu et al. 2018). Weakly-supervised methods (Feng, Hong, and Zheng 2021; Li, Liu, and Jiao 2022; Sultani, Chen, and Shah 2018; Tian et al. 2021; Wu et al. 2022, 2020; Zhou, Yu, and Yang 2023; Wu et al. 2025a, 2024c; Tian, Li, and Xu 2020; Ghadiya et al. 2024) use both normal and abnormal videos with only video-level labels, significantly reducing the annotation cost. Fully-supervised methods (Landi, Snoek, and Cucchiara 2019; Liu and Ma 2019) rely on precise frame-level annotations to achieve accurate localization, yet their practical deployment is limited due to high annotation expenses. However, traditional video anomaly detection methods predominantly depend on single-modal visual data, thus restricting their ability to interpret complex scenarios and contextual relationships. Multi-modal approaches address these limitations by integrating complementary visual and textual information, thereby enhancing the ability of model to understand and accurately characterize anomalies.

Multimodal Video Anomaly Understanding

Traditional video anomaly detection methods usually neglect deeper semantic understanding, while multimodal approaches offer richer context beneficial for anomaly retrieval and interpretation. Wu et al.(Wu et al. 2024a) introduced the Video Anomaly Retrieval task, using detailed textual descriptions or synchronized audio to identify anomalous segments within untrimmed videos, and proposed anomaly-guided sampling and masked phrase modeling to establish cross-modal associations. To further improve retrieval accuracy, Wu et al.(Wu et al. 2025b) developed VarCMP, leveraging cross-modal pre-trained models with unified hierarchical alignment at video, frame, and patch-levels, and an anomaly-biased weighting mechanism. Zhang et al.(Zhang et al. 2025) concurrently proposed Holmes-VAU for comprehensive anomaly understanding at clip, event, and video-levels, constructing the hierarchical HIVAU-70K dataset and integrating anomaly temporal sampling with multimodal large models. Zhou et al.(Zhou et al. 2024) developed the human-centric UCCD dataset with detailed person-level behavioral annotations, supporting fine-grained analysis of individual behaviors in surveillance scenarios. Unlike these methods, our research specifically targets the detection, identification, and explanation of individuals involved in anomalous events, emphasizing fine-grained human behaviors and interactions, representing a challenging yet practical advancement toward real-world anomaly interpretation. interpretation.

Methodology

Overview of TargetVAU

The overall pipeline of TargetVAU is illustrated in Figure 2. Given an input video, frames are sampled and processed

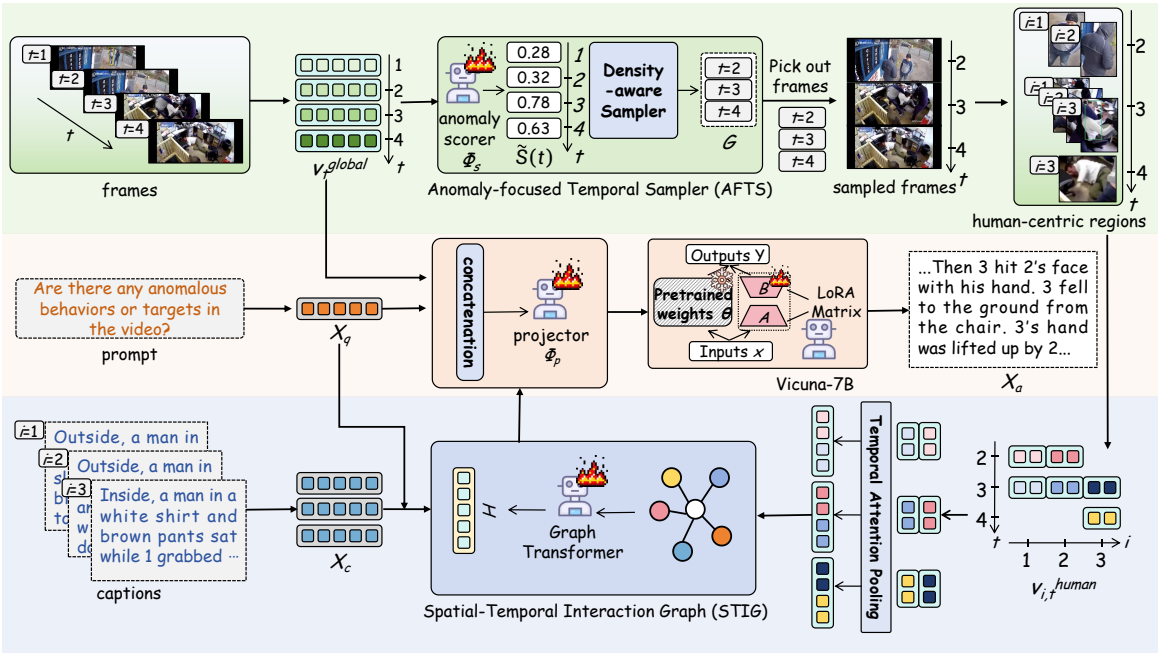

Figure 2: The pipeline of our proposed TargetVAU.

by ViT-L/14 (Dosovitskiy et al. 2021) to extract global and human-centric visual features. The Anomaly-focused Temporal Sampler (AFTS) selects keyframes with high anomaly potential. A Spatio-Temporal Interaction Graph (STIG) is then constructed to model individual behaviors and interactions, which are further encoded using a Graph Transformer (Dwivedi and Bresson 2020). Visual features are fused with textual prompts through Q-Former (Li et al. 2023), and the resulting representation is passed to Vicuna-7B (Zheng et al. 2023) to identify the anomalous individual and generate an explanation.

Video and Text Embedding

The frozen ViT-L/14 from CLIP is adopted as the visual backbone ϕ_v. For each input video, frames are densely sampled at a fixed interval of 16. From every sampled frame, two types of visual features are extracted: global-level features obtained by feeding the full frame into the encoder, and human-centric features derived from processing the cropped regions corresponding to detected person instances. The global-level features are computed as:

$$v_t^{global} = \phi_v(\mathcal{V}_t^{full}) \tag{1}$$

where $\mathcal{V}_t^{full}$ denotes the full image of frame t. In parallel, we extract human-centric features by cropping each detected the bounding box of person from the frame and processing these cropped regions individually through the same visual encoder:

$$v_{i,t}^{human} = \phi_v(\mathcal{V}_{i,t}^{crop}) \tag{2}$$

where $\mathcal{V}_{i,t}^{crop}$ represents the cropped region of person i in frame t. Hence, each frame yields both a global-level representation capturing the overall scene context, and human-centric embeddings that focus on individual appearances and actions.

The text encoder ϕ_t is initialized from Vicuna-7B and consists of a tokenizer and a textual embedding layer. Given a video caption and a prompt, the text encoder converts them into separate textual embedding sequences:

$$X_c = \phi_t(\text{Caption}), \quad X_q = \phi_t(\text{Prompt}) \tag{3}$$

These global-level and human-centric visual features, along with the textual embeddings, form the multimodal inputs for the subsequent modules in our anomaly understanding framework.

Anomaly-focused Temporal Sampler

Building upon the previously extracted global-level visual features, we introduce the AFTS to selectively identify keyframes that are most likely to contain anomalous behaviors. The AFTS module is composed of two components: a lightweight anomaly scorer and a density-aware sampler.

The anomaly scorer ϕ_s is implemented as a feature-based video anomaly detection network, following the efficient architecture DMU (Zhou, Yu, and Yang 2023). It takes the sequence of global class tokens $\{v_1^{global}, v_2^{global}, \ldots, v_T^{global}\}$ as input and produces anomaly scores for each frame:

$$s_t = \phi_s(v_t^{global}), \quad s_t \in [0, 1] \tag{4}$$

where s_t denotes the predicted anomaly score for the t-th frame. These scores collectively form a temporal sequence $S = [s_1, s_2, \ldots, s_T]$, where each s_j reflects the anomaly likelihood of the j-th video frame.

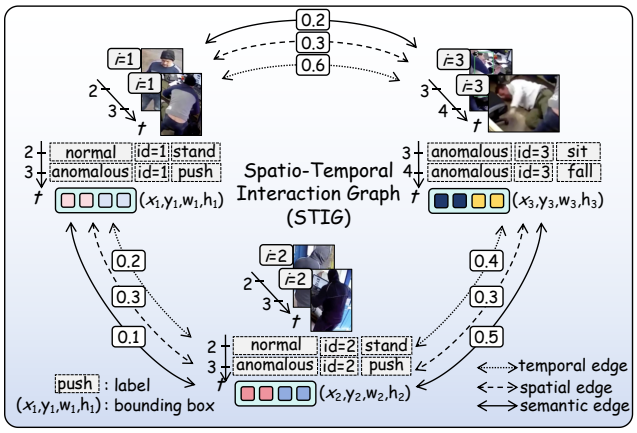

Figure 3: Illustration of node attributes and edge types in the STIG.

To concentrate computational efforts on informative regions, we adopt a density-aware non-uniform sampling strategy. Rather than directly summing scores as in traditional cumulative sampling, a piecewise normalized cumulative anomaly curve is computed:

$$\widetilde{S}(t) = \frac{\sum_{i=1}^{t}(s_i + \tau)}{\sum_{j=1}^{T}(s_j + \tau)} \tag{5}$$

Here, τ is a small smoothing constant that prevents zero gradients and controls sampling dispersion. During inference, the sampler applies $\tau = 0.1$ to smooth anomaly scores. The denominator aggregates the adjusted anomaly scores across all frames, ensuring $\widetilde{S}(t)$ is normalized to the range $[0, 1]$.

We then select N equally spaced anchor points $\{a_1, a_2, \ldots, a_N\}$ from the range $[0, 1]$, and for each anchor point a_j, identify the closest timestamp t_j such that $\widetilde{S}(t_j) \geq a_j$. The resulting set of sampled frame indices is denoted as:

$$G = \{t_1, t_2, \ldots, t_N\} \tag{6}$$

Denser sampling is encouraged in time segments with high anomaly likelihood, while minimal coverage is maintained elsewhere. The selected keyframes G then serve as the temporal backbone for downstream spatio-temporal interaction modeling and prompt-based reasoning.

Spatio-Temporal Interaction Graph

Based on the temporal sampling results from the AFTS module, we construct a STIG to model individual behaviors and their interactions over time. This graph serves as an intermediate representation that captures fine-grained human dynamics and supports multimodal reasoning.

Each node in STIG corresponds to a detected person in a sampled keyframe. For each individual i in frame t, we extract a human-centric feature $v_{i,t}^{\text{human}} = \phi_v(\mathcal{V}_{i,t}^{\text{crop}})$, where $\mathcal{V}_{i,t}^{\text{crop}}$ is the cropped region of the person. This is concatenated with bounding box coordinates (x, y, w, h) and category labels to form the node feature. To enhance temporal consistency, person features across sampled timestamps $G = \{t_1, t_2, \ldots, t_N\}$ are aggregated using temporal attention pooling.

Edges in STIG capture spatial, temporal, and semantic relationships. Spatial edges are constructed based on bounding box overlap and distance within a frame. Temporal edges link individuals across keyframes if their bounding boxes are similar, preserving motion continuity. This approach helps address person occlusion. When an individual is temporarily occluded and later reappears, temporal co-referencing edges enable correct association by linking nodes that represent the same person across different time steps. Semantic edges are derived from caption embeddings $X_c = \phi_t(\text{Caption})$, which are parsed into triplets and mapped to corresponding nodes to form directed interaction edges. Figure 3 provides a detailed illustration of node attributes and the three types of edges in STIG. A Graph Transformer encodes the sequence of structured graphs to model the temporal evolution of individuals and their relationships, enabling downstream multimodal understanding and reasoning.

Projector and Grounding-aware Reasoning

Building on the spatio-temporal interaction representations $\mathcal{H}$ from the Graph Transformer, the framework integrates three modalities to enable grounded reasoning about individual-level anomalies: global visual embeddings v_t^{global} from the AFTS module, human interaction features $\mathcal{H}$, and textual prompt embeddings X_q from the Vicuna-based text encoder ϕ_t. These modalities respectively capture the global scene context, structured human interactions, and instruction semantics.

To combine global and individual-level cues, a hierarchical fusion strategy is adopted. Human-centric features are first used to construct nodes in STIG and aggregated into $\mathcal{H}$, while global scene features are preserved in a separate branch. The three representations, v_t^{global}, $\mathcal{H}$, and X_q, are concatenated and projected into a shared semantic space via a frozen Q-Former-based module ϕ_p, enabling joint reasoning over detailed behaviors and the holistic scene.

To enhance semantic consistency between modalities, a grounding-aware contrastive learning objective aligns human-centric visual features with their corresponding textual descriptions.

The resulting unified embedding X_{ins} is then fed into a Vicuna-7B LLM, fine-tuned using LoRA for efficiency, which generates the final output $X_a = \{x_1, x_2, \ldots, x_L\}$ in an autoregressive manner, as formulated below:

$$p(X_a \mid X_{\text{ins}}) = \prod_{i=1}^{L} \text{LM}_\theta(x_i \mid \{x_1, \ldots, x_{i-1}\}, X_{\text{ins}}), \tag{7}$$

where θ denotes the LLM parameters. This process yields two outputs: the identity of the anomalous individual and a natural language explanation of the reason.

Loss Functions

TargetVAU is trained with a unified objective combining anomaly detection, grounding-aware contrastive, and cap-

tion generation losses, which guide anomaly localization, visual-text alignment, and language generation, respectively.

The anomaly detection loss guides the anomaly scorer ϕ_s in the AFTS module. Given predicted scores s_t and ground truth labels $\hat{y}_t$ from the UCCD dataset, the loss is defined as:

$$\mathcal{L}_{AS} = -\sum_{t=1}^{T}(s_t \log \hat{y}_t + (1 - s_t)\log(1 - \hat{y}_t)) \quad (8)$$

This encourages higher scores for abnormal frames and supports effective temporal sampling.

The contrastive loss promotes alignment between visual features v_i and corresponding textual descriptions t_j. For matched pairs $(v_i, t_j) \in \mathcal{P}$, the loss is:

$$\mathcal{L}_{contrast} = -\sum_{(v_i,t_j)\in\mathcal{P}} \log \frac{\exp(sim(v_i, t_j))}{\sum_{t_k} \exp(sim(v_i, t_k))} \quad (9)$$

Here, $sim(\cdot, \cdot)$ is the cosine similarity, and the denominator includes all negative samples t_k.

To supervise caption generation, we apply a token-level cross-entropy loss. Given the fused embedding X_{ins} and answer tokens $X_a = \{x_1, x_2, \ldots, x_L\}$, the loss is:

$$\mathcal{L}_{caption} = -\sum_{i=1}^{L} \log p_\theta(x_i | X_{ins,<i}, X_{a,<i}) \quad (10)$$

where θ are the trainable parameters of the Vicuna-7B model under LoRA fine-tuning.

The total training loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_0 \mathcal{L}_{AS} + \lambda_1 \mathcal{L}_{contrast} + \lambda_2 \mathcal{L}_{caption} \quad (11)$$

with λ_0, λ_1, and λ_2 controlling the weight of each component. We initially set them as 1, 1, and 0.5, allowing us to observe and adjust them subsequently.

Training and Testing

Training. We train TargetVAU on the UCCD dataset in two sequential stages. In the first stage, we optimize only the anomaly scorer ϕ_s using $\mathcal{L}_{AS}$, leveraging the ground truth frame-level labels to learn reliable anomaly scores. In the second stage, we freeze the parameters of ϕ_s and train the remaining components of the framework using the full set of caption and prompt annotations from UCCD. The Q-Former ϕ_p is kept frozen, while the Vicuna-7B language model is fine-tuned using LoRA to preserve its general capabilities while adapting to the anomaly reasoning task.

Testing. During inference, given an input surveillance video and a textual prompt, TargetVAU outputs the index of the anomalous individual, along with a natural language explanation describing the anomalous behavior. This output is conditioned on the fused global-level, human-centric, and textual information and strictly adheres to the instruction of users.

Experiment

Experiment Setup

Datasets. We evaluate VAU performance using two datasets: UCCD (Zhou et al. 2024) and HIVAU-70K (Zhang et al. 2025). UCCD contains 1,012 surveillance videos with 7,820 annotated individuals across 13 anomaly types, supporting fine-grained behavior and identity analysis. HIVAU-70K includes over 5,400 videos and 70,000 clips with hierarchical annotations at the clip, event, and video levels, enabling multi-level semantic information. We also assess VAD performance on UCF-Crime (Sultani, Chen, and Shah 2018) and XD-Violence (Wu et al. 2020) datasets.

Evaluation Metrics. We evaluate VAU performance on the UCCD dataset, which is repartitioned to ensure a balanced distribution of behavior categories, with 262 videos in the test set. Six metrics are used: BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), BERTScore (Zhang et al. 2019), and ParaScore (Shen et al. 2022). BLEU, CIDEr, METEOR, and ROUGE assess lexical and structural similarity with human-annotated captions, while BERTScore and ParaScore evaluate semantic consistency. The model is also tested on the HIVAU-70K dataset using the official train-test splits, focusing on video-level captioning for consistency. Additionally, we evaluate VAD performance on UCF-Crime and XD-Violence using AP and AUC as metrics. Qualitative analysis is further conducted to assess how well the model captures individual behaviors and interactions.

Implementation Details. For the constructed TargetVAU framework, we adopt Vicuna-7B as the backbone language model, given its superior performance in fine-grained anomaly detection and explanation. In comparison, InternVL2-2B demonstrates lower accuracy in semantic reasoning on UCCD and HIVAU-70K, justifying our choice. The Anomaly-focused Temporal Sampler is optimized independently using the Adam optimizer with a learning rate of 1e-4. For instruction tuning, we train the model with a batch size of 512 for 1 epoch using the AdamW optimizer, employing cosine learning rate decay and a warm-up phase. LoRA settings are defined as: $r = 64$, $\alpha = 128$, and learning rate = 4e-5. All experiments are conducted on 4 NVIDIA A100 GPUs.

Main Results

Anomaly Understanding Results. We evaluate the quality of anomaly-related text generated by TargetVAU in comparison with state-of-the-art general MLLMs (Zhang, Li, and Bing 2023; Lin et al. 2023; Maaz et al. 2023; Zhang et al. 2024; Wang et al. 2024; Chen et al. 2024)on the UCCD and HIVAU-70K datasets. As shown in Table 1, under the zero-shot setting, TargetVAU achieves slightly better performance than existing approaches on surface-level similarity metrics. In addition, Table 2 shows that TargetVAU performs competitively on semantic similarity metrics. Although it falls slightly behind Holmes-VAU on CIDEr and ParaScore, the model demonstrates overall robustness and effectiveness across both datasets. Furthermore, after fine-tuning on the UCCD training set, TargetVAU surpasses all baselines across all metrics reported in Table 3, highlighting its capabilities in anomalous target recognition and behavior understanding. These improvements are attributed to its fine-grained interaction modeling and instruction-guided

		UCCD				HIVAU-70K			
Method	**Params**	**BLEU**	**CIDEr**	**METEOR**	**ROUGE**	**BLEU**	**CIDEr**	**METEOR**	**ROUGE**
Video-LLaMA (2023)	7B	0.086	0.013	0.036	0.085	0.104	0.017	0.057	0.090
Video-LLaVA (2023)	7B	0.043	0.008	0.007	0.031	0.055	0.013	0.014	0.045
Video-ChatGPT (2023)	7B	0.062	0.008	0.023	0.076	0.066	0.013	0.044	0.079
LLaVA-Video (2024)	7B	0.093	0.026	0.075	0.093	0.120	0.031	0.096	0.106
QwenVL2 (2024)	7B	0.102	0.038	0.098	0.112	0.155	0.044	0.112	0.137
InternVL2 (2024)	8B	0.127	0.029	0.093	0.107	0.145	0.035	0.101	0.122
Holmes-VAU (2025)	2B	0.463	0.432	0.103	0.281	0.566	**1.437**	0.121	0.355
TargetVAU (Ours)	**7B**	**0.582**	**0.665**	**0.212**	**0.317**	**0.613**	1.424	**0.357**	**0.482**

Table 1: Comparison of understanding performance with state-of-the-art Multimodal Large Language Models on UCCD and HIVAU-70K datasets. BLEU denotes the cumulative score from BLEU-1 to BLEU-4.

	UCCD		HIVAU-70K	
Method	**BeScore**	**PaScore**	**BeScore**	**PaScore**
Video-LLaMA (2023)	0.708	0.752	0.752	0.793
Video-LLaVA (2023)	0.725	0.771	0.778	0.823
Video-ChatGPT (2023)	0.713	0.762	0.771	0.819
LLaVA-Video (2024)	0.752	0.798	0.806	0.851
QwenVL2 (2024)	0.768	0.813	0.814	0.859
InternVL2 (2024)	0.762	0.808	0.802	0.846
Holmes-VAU (2025)	0.795	0.832	0.834	**0.909**
TargetVAU (Ours)	**0.862**	**0.891**	**0.879**	0.901

Table 2: Comparison of semantic-level performance with state-of-the-art Multimodal Large Language Models. BeScore corresponds to BERTScore, and PaScore corresponds to ParaScore.

Method	**BLEU**	**CIDEr**	**METEOR**	**ROUGE**
Video-LLaMA (2023)	0.294	0.371	0.130	0.020
Video-LLaVA (2023)	0.317	0.309	0.124	0.171
Video-ChatGPT (2023)	0.336	0.301	0.128	0.194
LLaVA-Video (2024)	0.348	0.432	0.164	0.201
QwenVL2 (2024)	0.361	0.446	0.182	0.217
InternVL2 (2024)	0.382	0.439	0.175	0.231
Holmes-VAU (2025)	0.529	0.613	0.193	0.286
TargetVAU (Ours)	**0.582**	**0.665**	**0.212**	**0.317**

Table 3: Comparison of understanding performance of various MLLMs fine-tuned on UCCD dataset.

multimodal reasoning mechanisms.

Anomaly Detection Results. We compare our method with state-of-the-art VAD approaches, including non-explainable models like RTFM (Tian et al. 2021), MGFN (Chen et al. 2023), and VadCLIP (Wu et al. 2024d), as well as explainable multimodal models such as LLAVA-1.5 (Liu et al. 2024), LAVAD (Zanella et al. 2024), and Holmes-VAU (Zhang et al. 2025). Table 4 summarizes their performance on UCF-Crime and XD-Violence. Our method achieves an AP of 87.82% on XD-Violence and an AUC of 89.57% on UCF-Crime, slightly outperforming Holmes-VAU and all other baselines. While methods like VadCLIP and CLIP-TSA perform well in detection, they lack interpretability. In contrast, our model offers both accurate localization and fine-grained semantic understanding of anomalies. Compared to explainable models like LAVAD, which lack supervised anomaly learning, our method benefits from structured multimodal modeling and anomaly-aware representations, enabling more effective and interpretable VAD.

Methods	**Backbone**	XD-Violence	UCF-Crime
		AP%	AUC%
Non-explainable VAD			
GODS (2019)	I3D	N/A	70.46
Wu *et al.* (2020)	I3D	78.64	82.44
MIST (2021)	I3D	N/A	82.30
RTFM (2021)	I3D	77.81	84.30
S3R (2022)	I3D	80.26	85.99
MSL (2022)	I3D	78.28	85.30
GCL (2022)	ResN	N/A	71.04
DYANNET (2023)	I3D	N/A	84.50
MGFN (2023)	I3D	79.19	86.98
UR-DMU (2023)	I3D	81.66	86.97
CLIP-TSA (2023)	ViT	82.19	87.58
VadCLIP (2024d)	ViT	84.51	88.02
Yang *et al.* (2024)	ViT	83.68	87.79
Wu *et al.* (2024b)	ViT	66.53	86.40
Explainable Multi-modal VAD			
Zero-Shot CLIP (2021)	ViT	17.83	53.16
LLAVA-1.5 (2024)	ViT	50.26	72.84
LAVAD (2024)	ViT	62.01	80.28
Holmes-VAU (2025)	ViT	87.68	88.96
TargetVAU (Ours)	**ViT**	**87.82**	**89.57**

Table 4: We compare detection performance with state-of-the-art video anomaly detection methods, including both explainable and non-explainable approaches.

Ablation Study

Impact of Visual Feature Types. We evaluate the impact of using global and human-centric visual features in our framework. As shown in Table 5, human-centric features alone offer modest gains, emphasizing the value of individual appearance and motion. Combining both types of features significantly improves performance, showing that global context and detailed individual cues complement each other and enhance model representation.

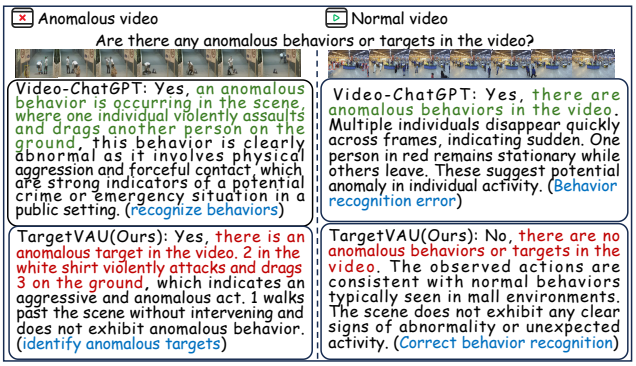

Figure 4: The figure shows a comparison of visualization results between TargetVAU and another method.

Visual Features Hu	Glo+Hu	AFTS	STIG	Q-Former	BLEU	CIDEr
✓		✓	✓	✓	0.448	0.526
	✓		✓	✓	0.509	0.591
	✓	✓		✓	0.472	0.553
	✓	✓	✓		0.521	0.603
	✓	✓	✓	✓	**0.582**	**0.665**

Table 5: Effectiveness of designed modules on UCCD. Glo and Hu denote global and human-centric features.

Impact of the Anomaly-focused Temporal Sampler. We evaluate the effectiveness of the Anomaly-focused Temporal Sampler by comparing the full model with a variant that replaces AFTS with uniform frame sampling. As shown in Table 5, removing AFTS leads to a CIDEr score drop from 0.665 to 0.591, along with a decrease in BLEU score, indicating that the absence of anomaly-guided keyframe selection reduces temporal focus and introduces redundancy. These results demonstrate that selecting keyframes based on predicted anomaly likelihood enables the model to focus on behaviorally informative moments and enhances overall performance.

Impact of Spatio-Temporal Interaction Graph. We further investigate the importance of explicitly modeling spatio-temporal interactions through the STIG module. Without STIG, the model relies solely on visual and textual information fusion, causing a noticeable performance drop. This decline demonstrates that explicitly modeling individual interactions and temporal continuity through structured graph representations significantly contributes to understanding complex human behaviors and identifying anomalous individuals.

Impact of Q-Former. We evaluate the effectiveness of the Q-Former-based multimodal fusion strategy. Removing the Q-Former and directly concatenating the embeddings of each modality using a simple linear projection reduces the performance to BLEU 0.521 and CIDEr 0.603. The significant performance gap compared to the full model demonstrates that the frozen Q-Former module can effectively align the features of each modality into a unified semantic space, thereby facilitating more efficient multimodal reasoning and more accurate anomaly explanation generation.

Loss ratio of λ_0:λ_1:λ_2	BLEU	CIDEr
1:1:1	0.572	0.658
2:1:1	0.556	0.641
1:2:1	0.564	0.649
1:1:2	0.582	0.665

Table 6: The impact of the weight of each loss ratio on UCCD.

Impact of Loss Function Weights. We conduct an ablation study to analyze the impact of different weight ratios among the three loss terms: anomaly detection, grounding-aware contrastive, and caption generation. As shown in Table 6, equal weighting yields balanced results with BLEU 0.572 and CIDEr 0.658. Increasing the anomaly detection or contrastive loss degrades performance, indicating that excessive local supervision weakens global reasoning. In contrast, emphasizing caption generation gives the best results, highlighting the importance of language-aligned optimization and balanced supervision across perception, alignment, and generation.

Qualitative Comparision

Identifying anomalous individuals in complex scenes requires fine-grained reasoning over visual cues and interactions. As shown in Fig. 4, TargetVAU outperforms baselines like Video-ChatGPT, which often misidentify individuals or miss the true anomaly. TargetVAU accurately localizes the responsible person, provides clear reasoning, and distinguishes bystanders. It also correctly detects normal scenes, demonstrating strong robustness. These results show that it combines multimodal cues and structured reasoning for accurate, interpretable anomaly identification.

Conclusion

We present TargetVAU, a multimodal reasoning framework for individual-level anomaly understanding in videos. It extracts global and human-centric features via a frozen ViT encoder, selects keyframes using an anomaly-focused temporal sampler, and builds a spatio-temporal graph to model human behaviors and interactions. Structured visual features are fused with textual prompts through a frozen Q-Former to form a unified representation. A Vicuna-7B language model, fine-tuned with LoRA, performs instruction-guided reasoning to identify anomalous individuals and provide explanations. Experiments on two datasets confirm the effectiveness of the framework in both accuracy and interpretability.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306240), China Postdoctoral Science Foundation (No. 2023TQ0272) and Natural Science Basic Research Program of Shaanxi Province (2024JC-DXWT-07).

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; and Wu, Y.-C. 2023. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 387–395.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, H.; Nan, G.; Qian, J.; Wu, W.; Deng, W.; Mu, H.; Chen, Z.; Mao, P.; Tao, X.; and Liu, J. 2024a. Exploring What Why and How: A Multifaceted Benchmark for Causation Understanding of Video Anomaly. *arXiv preprint arXiv:2412.07183*.

Du, H.; Zhang, S.; Xie, B.; Nan, G.; Zhang, J.; Xu, J.; Liu, H.; Leng, S.; Liu, J.; Fan, H.; et al. 2024b. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18793–18803.

Dwivedi, V. P.; and Bresson, X. 2020. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14009–14018.

Ghadiya, A.; Kar, P.; Chudasama, V.; and Wasnik, P. 2024. Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1965–1974.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1705–1714.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.

Landi, F.; Snoek, C. G.; and Cucchiara, R. 2019. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1395–1403.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, K.; and Ma, H. 2019. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1490–1499.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545.

Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13588–13597.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shen, L.; Liu, L.; Jiang, H.; and Shi, S. 2022. On the evaluation metrics for paraphrase generation. *arXiv preprint arXiv:2202.08479*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Thakare, K. V.; Raghuwanshi, Y.; Dogra, D. P.; Choi, H.; and Kim, I.-J. 2023. Dyannet: A scene dynamicity guided

self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, 5541–5550.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 436–454. Springer.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, J.; and Cherian, A. 2019. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8201–8211.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, J.-C.; Hsieh, H.-Y.; Chen, D.-J.; Fuh, C.-S.; and Liu, T.-L. 2022. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, 729–745. Springer.

Wu, P.; Liu, J.; He, X.; Peng, Y.; Wang, P.; and Zhang, Y. 2024a. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Transactions on Image Processing*, 33: 2213–2225.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 322–339. Springer.

Wu, P.; Su, W.; Pang, G.; Sun, Y.; Yan, Q.; Wang, P.; and Zhang, Y. 2025a. AVadCLIP: Audio-Visual Collaboration for Robust Video Anomaly Detection. *arXiv preprint arXiv:2504.04495*.

Wu, P.; Su, W.; Xiangteng, H.; Wang, P.; and Zhang, Y. 2025b. VarCMP: Adapting Cross-Modal Pre-Training Models for Video Anomaly Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 42, 6074–6082.

Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; and Zhang, Y. 2024b. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18297–18307.

Wu, P.; Zhou, X.; Pang, G.; Yang, Z.; Yan, Q.; Wang, P.; and Zhang, Y. 2024c. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9301–9310.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2024d. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6074–6082.

Xu, D.; Yan, Y.; Ricci, E.; and Sebe, N. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156: 117–127.

Yang, Z.; Liu, J.; and Wu, P. 2024. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18899–18908.

Yang, Z.; Liu, J.; Wu, Z.; Wu, P.; and Liu, X. 2023. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14592–14601.

Zaheer, M. Z.; Mahmood, A.; Khan, M. H.; Segu, M.; Yu, F.; and Lee, S.-I. 2022. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14744–14754.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; and Sang, N. 2025. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13843–13853.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3769–3777.

Zhou, L.; Gao, Y.; Zhang, M.; Wu, P.; Wang, P.; and Zhang, Y. 2024. Human-centric Behavior Description in Videos: New Benchmark and Model. *IEEE Transactions on Multimedia*.