

Short-Window Sliding Learning for Real-Time Violence Detection via LLM-based Auto-Labeling

Seoik Jung*, Taekyung Song**, Yangro Lee**, Sungjun Lee†

PIA-SPACE Inc.

si.jung@pia.space*, tg.song@pia.space**, yr.lee@pia.space**, sj.lee@pia.space†

Abstract—This paper proposes a Short-Window Sliding Learning framework for real-time violence detection in CCTV footages. Unlike conventional long-video training approaches, the proposed method divides videos into 1-2 second clips and applies Large Language Model (LLM)-based auto-caption labeling to construct fine-grained datasets. Each short clip fully utilizes all frames to preserve temporal continuity, enabling precise recognition of rapid violent events. Experiments demonstrate that the proposed method achieves 95.25% accuracy on RWF-2000 and significantly improves performance on long videos (UCF-Crime: 83.25%), confirming its strong generalization and real-time applicability in intelligent surveillance systems.

I. INTRODUCTION

Recently, video-based violence and abnormal behavior detection has been gaining attention as an essential core technology in fields such as public safety, smart cities, and intelligent surveillance [1]. Especially in real-time CCTV environments, it is crucial to quickly and accurately recognize rapidly occurring violent situations or abnormal behaviors within a short time [1], [15].

However, existing research has mainly focused on training models using long-duration video data, and this approach has limitations that do not fit the temporal characteristics of actual surveillance environments [7]. Previous studies rely on frame sampling from long video units, failing to sufficiently reflect temporal continuity and contextual information, and are also unsuitable for real-time CCTV environments in terms of computational efficiency [7].

Therefore, this paper proposes a short-window sliding learning technique. The proposed method involves dividing existing long videos into short clips of 1-2 seconds and auto-labeling each clip using a Large Language Model (LLM: e.g., Gemini). The generated labels are then reviewed by humans to construct a high-quality, fine-grained action dataset. This allows for more detailed learning of events that occur over a short period, such as violence. Furthermore, this study minimizes information loss due to sampling by utilizing all possible frames within each short clip for learning. This addresses the problem of relying on traditional frame sampling and allows for a richer reflection of temporal features within short intervals.

†Corresponding author.

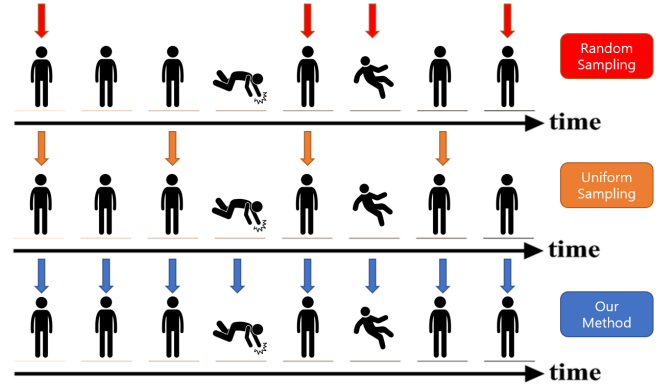


Fig. 1. Comparison of frame sampling methods.

This study moves beyond the existing learning paradigm centered on long videos and presents the possibility of short-window-centric learning suitable for real-time surveillance environments. It is expected that this can be applied to various video-based application fields in the future, such as real-time abnormal behavior monitoring, violence detection, and emergency response systems.

II. RELATED WORK

This section summarizes the trends in video anomaly detection and violence recognition research and discusses the key differentiators of the proposed method from existing studies. Related work is examined in three categories: (1) video abnormal behavior detection, (2) violence video detection research, and (3) clip-level learning and auto-labeling approaches.

A. Video Abnormal Behavior Detection

Video abnormal behavior detection, the problem of distinguishing between normal and abnormal actions, has been studied using supervised, semi-supervised, and weakly supervised learning. Since Sultani et al.'s [1] MIL (Multiple Instance Learning) framework, unsupervised learning using reconstruction-based and prediction-based approaches, as well as Transformer [16] models, has been actively pursued [2], [3]. Recently, research integrating temporal features through combination with vision-language models is expanding [4].

These studies mainly focus on detecting general abnormal behaviors such as theft, accidents, and congestion, and have limitations in precisely detecting behaviors that occur within a short time, like violence or attacks.

B. Violence Video Detection

Violence detection is a subfield of video anomaly detection, aiming to recognize aggressive acts such as physical conflict, striking, and chasing between people. Kaur et al. (2024) [5] analyzed various violence recognition models and pointed out that most existing approaches are based on long-video-centric learning, limiting their ability to capture rapidly changing violent acts in a short time. Generally, violence detection models use long clips or sample a number of frames at regular intervals as input, making it difficult to sufficiently reflect temporal granularity [6]. Recently, multimodal approaches integrating visual and text information using vision-language models have been attempted, but most methods still rely on clip-level inputs and a frame-sampling-centric structure [6].

C. Clip-level Learning and Auto-Labeling

Recently, attempts have been made to increase learning efficiency by dividing existing long video units into short clips or combining them with auto-labeling [7]. Traditional sampling-based learning extracts frames at regular intervals from a long video, which can cause temporal information loss.

Consequently, most existing studies are based on long video inputs, sampling-centric processing, and clip-level labeling, lacking a structural approach for the detailed detection of violent situations occurring in brief moments. To solve these limitations, this study proposes a new learning framework that combines sub-second clip division, auto-caption labeling using LLMs, and the utilization of all frames within the clip. This method simultaneously improves temporal precision and real-time applicability, which were overlooked by existing methods.

III. METHODOLOGY

This study proposes a short-window sliding learning framework to precisely detect violent situations that occur in brief moments in real-time CCTV environments. The proposed method consists of two main stages: (1) data construction and (2) learning. Figure 2 illustrates the structure of the data construction process.

A. Data Construction

Sliding window-based video segmentation. Existing CCTV violence video data consists of clips 5-30 seconds long, with the actual violent act occurring in only a portion (about 1-2 seconds) of that time. Therefore, this study segments the original video using a sliding window approach with a set length of L seconds, window size W , and stride S .

$$N = \frac{L - W}{S} + 1$$

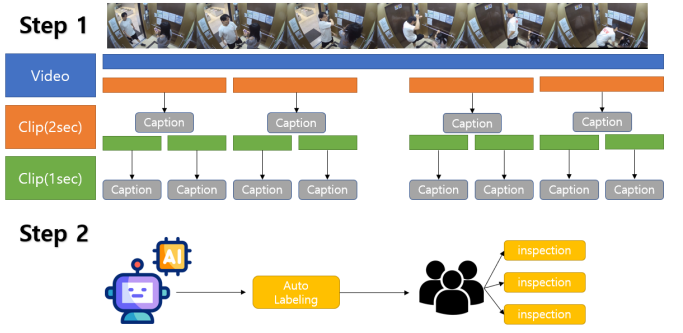


Fig. 2. Proposed data construction process. Step 1 involves dividing the original video into 1-2 second clips and generating automatic captions. Step 2 builds a refined short video dataset through LLM-based auto-labeling and human review.

Here, N represents the total number of clips generated. Each clip maintains the temporal order of the original video, allowing the video's entire temporal information to be subdivided without loss.

Auto-labeling. Captions were generated for each clip using Google's LLM service, Gemini. The model analyzes the main scenes of each clip and outputs sentence-form action descriptions such as "a scene of people pushing each other" or "an action of raising a fist". The generated captions are organized into a coarse-to-fine label structure. Subsequently, 3 reviewers manually checked the auto-generated captions to correct errors. The review work took about 40 hours per person, totaling 120 hours, and finally, a highly reliable short-video-based action dataset was constructed.

- **Coarse Level:** Violence / Non-violence
- **Fine Level:** Punching, Kicking, Pushing, Chasing, etc.

B. Learning and Experimental Setup

This study used the InternVL3 [17] architecture. This model is a VLM that processes video and text information integrally. It extracts features from each through a vision encoder and a language encoder, then aligns them in an intermediate representation space to learn semantic consistency.

By using all frames of a short interval (1-2 seconds) as input, temporal information loss that occurs in existing sampling-based learning was minimized. Caption information generated via LLM was provided as labels, enabling the model to learn not only visual patterns but also the meaning of the actions. In the inference phase, clips of 15-frame intervals were input in a sliding window fashion to predict violence in real-time. This structure allows for more accurate capturing of violent situations that occur in brief moments.

IV. EXPERIMENTS

In this section, experiments were conducted by setting up various datasets and learning scenarios to verify the effectiveness of the proposed short-window sliding learning technique. All experiments were evaluated based on the RWF-2000 dataset [15], which consists of 2000 short videos (5-10 seconds long) including violent and non-violent situations

TABLE I
EVALUATION RESULTS ON RWF-2000 AFTER TRAINING ON DIFFERENT DOMAIN DATA.

Experiment	AI hub / Short	RWF-2000 / Short	SCVD / Short	UCF-Crime / long	UCF-Crime / Short	Acc.
Exp.0						78.50%
Exp.1	✓					82.50%
Exp.2		✓				91.78%
Exp.3			✓			88.75%
Exp.4				✓		55.75%
Exp.5					✓	83.25%
Exp.6	✓	✓	✓			95.25%

filmed in real CCTV environments. The evaluation goals of this study are (1) to verify the effectiveness of the proposed short-window learning technique, and (2) to analyze the generalization performance when video datasets from different domains are combined.

A. Model and Training Environment

The base model is InternVL3 [17] with a multimodal Transformer structure, and all experiments were conducted with the same hyperparameters (input 12-15 frames, AdamW).

B. Dataset Configuration

A total of five datasets were used. For comparative experiments, each dataset was either configured in short clip units (proposed method) or kept as existing long videos.

- **RWF-2000 [15]:** Violence/non-violence video dataset collected in real CCTV environments (short videos).
- **AI Hub CCTV Dataset:** Short video clip dataset based on indoor/outdoor surveillance situations.
- **SCVD Dataset:** CCTV short video dataset including pedestrians, crowds, and abnormal situations.
- **UCF-Crime (long):** General abnormal behavior detection dataset composed of long videos several minutes long.
- **UCF-Crime (short):** The version segmented into 1-2 second units and auto-labeled using this study's method.

TABLE II
COMPARISON WITH EXISTING VIOLENCE DETECTION MODELS.

Model / Method	Year	Accuracy
MSTFDet [8]	2025	95.20%
CUE-Net [9]	2024	94.00%
Violence 4D [10]	2023	94.67%
Structured Keypoint Pooling [11]	2023	93.40%
Skeleton+Change Detection [12]	2023	90.25%
Semi-Supervised Hard Attention (SSHA) [13]	2022	90.40%
TwoStreamSepConv-LSTM [14]	2021	89.75%
Flow Gated Network [15]	2021	87.25%
Our Method	2025	95.25%

C. Quantitative Comparison

In Table 2, the proposed method achieved 95.25% accuracy on RWF-2000, slightly surpassing MSTFDet (95.2%) [8]. Achieving SOTA-level performance with only short clip learning, without skeleton or multi-stream structures, confirmed

that the proposed strategy is effective for capturing temporal features.

As shown in Table 1, the model trained on long videos (UCF-Crime/Long, Exp.4) showed a low accuracy of 55.75%, whereas the Short version, segmented into 1-2 second units (Exp.5), improved to 83.25%. The model that combined short clip data such as AI Hub, SCVD, and RWF-2000 (Exp.6) recorded the highest performance at 95.25%, confirming the generalization effect of short-window learning.

V. CONCLUSION

In this study, we proposed a short-window sliding learning technique for precisely detecting violent situations that occur over a short period in real-time surveillance environments. By dividing videos into 1-2 second clips and performing auto-labeling and caption-based data augmentation using LLM, we simultaneously improved temporal precision and data efficiency compared to existing methods.

Experimental results showed that the proposed method achieved 95.25% accuracy on the RWF-2000 dataset, outperforming previous state-of-the-art research. Furthermore, performance was significantly improved when long video data was converted to short clip units for training, and generalization performance was maximized when short video data from various domains was combined. These results suggest that the method proposed in this study can be effectively applied to the problem of real-time violence detection in actual CCTV environments.

Future research plans to expand in the direction of integrating multiple modalities (voice, subtitles, action features) to develop an intelligent real-time surveillance model that can cover more complex abnormal situations (fear, theft, suicide attempts, etc.).

ACKNOWLEDGMENT

This paper was conducted as part of the "Advancement and Overseas Expansion of VLM-based Automatic Anomaly Detection Real-time Video Analysis AI Solution" (Project No: PJT-25-031547), a research task of the "2025 Regional Digital Basic Fitness Support (Leading Enterprise Commercialization Support Project)," supervised by the National IT Industry Promotion Agency (NIPA) with support from the Ministry of Science and ICT.

REFERENCES

- [1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479-6488.
- [2] J. Liu, et al., "Networking systems for video anomaly detection: A tutorial and survey," *ACM Computing Surveys*, 2025, 57.10: 1-37.
- [3] M. Abdalla, et al., "Video anomaly detection in 10 years: A survey and outlook," *Neural Computing and Applications*, 2025, 1-44.
- [4] Y. Liu, et al., "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, 2024, 56.7: 1-38.
- [5] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," *Neurocomputing*, 2024, 597: 128113.
- [6] S. Jung, et al., "DUAL-VAD: Dual Benchmarks and Anomaly-Focused Sampling for Video Anomaly Detection," *arXiv preprint arXiv:2509.11605*, 2025.
- [7] H. Zhang, et al., "Holmes-vau: Towards long-term video anomaly understanding at any granularity," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13843-13853.
- [8] B. Qi, B. Wu, and B. Sun, "Automated violence monitoring system for real-time fistfight detection using deep learning-based temporal action localization," *Scientific Reports*, 2025, 15.1: 29497.
- [9] D. C. Senadeera, et al., "Cue-net: violence detection video analytics with spatial cropping enhanced uniformerv2 and modified efficient additive attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4888-4897.
- [10] M. Magdy, M. W. Fakhri, and F. A. Maghraby, "Violence 4D: Violence detection in surveillance using 4D convolutional neural networks," *IET Computer Vision*, 2023, 17.3: 282-294.
- [11] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22962-22971.
- [12] G. Garcia-Cobo and J. C. Sanmiguel, "Human skeletons and change detection for efficient violence detection in surveillance videos," *Computer Vision and Image Understanding*, 2023, 233: 103739.
- [13] H. Mohammadi and E. Nazerfard, "Video violence recognition and localization using a semi-supervised hard attention model," *Expert Systems with Applications*, 2023, 212: 118791.
- [14] Z. Islam, et al., "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1-8.
- [15] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *2020 25th International conference on pattern recognition (ICPR)*, IEEE, 2021, pp. 4183-4190.
- [16] A. Vaswani, et al., "Attention is all you need," *Advances in neural information processing systems*, 2017, 30.
- [17] J. Zhu, et al., "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.