Supplementary Material

Joint Learning of General and Diverse Patterns with Mixture of Memory Experts for Weakly-Supervised Video Anomaly Detection

Bo Sun[1,2]*, Junxi Chen[1]*, Zhe Wu[2,3]†, Feng Gao[4], Fan Yang[4], Li Su[1]†, Yaowei Wang[2,5],
[1]University of Chinese Academy of Sciences
[2]Institute of Perceptual Intelligence, Pengcheng Laboratory
[3]Pazhou Lab, [4]Peking University, [5]Harbin Institute of Technology, Shenzhen,
{sunbo24, chenjunxi22}@mails.ucas.ac.cn, wuzh02@pcl.ac.cn,
{gaof, fyang.eecs}@pku.edu.cn, suli@ucas.ac.cn, yaoweiwang@gmail.com

This supplementary material is organized as follows:

- Additional details for methods are provided in Section A.
- Implementation details are described in Section B.
- Datasets details are described in Section C.
- Additional analysis are presented in Section D.
- More visualization results are shown in Section E.

A. Additional Details for Methods

A.1. Temporal Modeling

As CLIP [1] is pretrained on static images, it lacks inherent mechanisms to capture temporal relationships in videos. To bridge this gap and enable effective modeling of video dynamics based on CLIP-extracted features, we adopt the Local-Global Temporal Adapter (LGT-Adapter) from VadCLIP [4]. This lightweight module is designed to learn temporal patterns, making it well-suited for our weakly supervised video anomaly detection task. VadCLIP's implementation is open-source and has demonstrated strong effectiveness in prior work; we refer readers to its repository for full code and ablation studies. Below, we detail the key components of LGT-Adapter for clarity and reproducibility.

The LGT-Adapter operates on frame-level features $\mathbf{X}_{\text{clip}} \in \mathbb{R}^{n\times d}$ extracted from the frozen image encoder of CLIP, where n is the number of frames and $d = 512$ is the feature dimension. It consists of two complementary modules: a *local module* for capturing short-range dependencies and a *global module* for modeling long-range interactions. Both modules employ residual connections to mitigate oversmoothing.

Local Module. To efficiently capture local temporal dependencies—often dominant since current events are highly correlated with adjacent frames—we employ a transformer encoder layer with restricted self-attention. Unlike standard global self-attention, our approach limits computation to overlapping local windows along the temporal dimension. Specifically, input features are divided into equal-length temporal windows, and self-attention is performed independently within each window. The resulting local representations, denoted as $\mathbf{X}_l$, retain fine-grained and context-aware temporal information.

Global Module. To complement the local module and capture holistic temporal relations, we introduce a lightweight Graph Convolutional Network (GCN) that models dependencies from both feature similarity and positional distance perspectives. Following established practices in video anomaly detection [3, 5], the GCN aggregates information across the entire sequence using two adjacency matrices: $\mathbf{H}_{\text{sim}}$ for similarity-based relations and $\mathbf{H}_{\text{dis}}$ for distance-based relations.

The similarity adjacency matrix $\mathbf{H}_{\text{sim}}$ is computed via frame-wise cosine similarity:

$$\mathbf{H}_{\text{sim}} = \frac{\mathbf{X}_l \mathbf{X}_l^\top}{\|\mathbf{X}_l\|_2 \cdot \|\mathbf{X}_l\|_2}, \tag{1}$$

followed by thresholding to filter weak connections and softmax normalization to ensure row-wise summation to 1.

The distance adjacency matrix $\mathbf{H}_{\text{dis}}$ encodes long-range positional dependencies:

$$\mathbf{H}_{\text{dis}}(i,j) = \frac{-|i-j|}{\sigma}, \tag{2}$$

where $\sigma = 1$ controls the decay rate, and softmax is applied for normalization. The global features $\mathbf{X}_g$ are then obtained as:

$$\mathbf{X}_g = \text{GELU}\left([\text{Softmax}(\mathbf{H}_{\text{sim}}); \text{Softmax}(\mathbf{H}_{\text{dis}})]\mathbf{X}_l\mathbf{W}\right), \tag{3}$$

with $\mathbf{W} \in \mathbb{R}^{2d\times d}$ as the sole learnable parameter, emphasizing the module's efficiency.

*Equal contribution.
†Corresponding authors.

A.2. Snippet-to-Expert Distribution Algorithm

Given the input temporal video features $X \in \mathbb{R}^{T\times d}$ and the routing score matrix $G \in \mathbb{R}^{T\times E}$, where T denotes the number of video snippets, d is the feature dimension, and E is the number of experts. For each snippet, we select the top k experts based on the routing scores:

$$I, R = \mathrm{TopK}(G, k), \tag{4}$$

where $I \in \mathbb{R}^{T\times k}$ stores the indices of the selected experts, and $R \in \mathbb{R}^{T\times k}$ contains the corresponding routing weights.

To explicitly represent the assignment between snippets and experts, we construct a binary routing matrix $P \in \mathbb{R}^{T\times E}$ by applying a one-hot expansion over I:

$$P_{t,e} = \begin{cases} 1, & \text{if } e \in I_t, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Accordingly, each expert E_e receives its corresponding subset of snippet features:

$$X_e = \{X_t \mid P_{t,e} = 1\}, \tag{6}$$

which defines the snippet-to-expert data partition for subsequent expert-specific processing.

A.3. Training and Inference

The anomaly detector employs a two-layer MLP architecture with a ReLU activation between the layers and a Sigmoid function at the output. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{m}} + \lambda_1\mathcal{L}_{\text{b}} + \lambda_2\mathcal{L}_{\text{reg}} + \lambda_3\mathcal{L}_{\text{div}} + \lambda_4\mathcal{L}_{\text{r}}, \tag{7}$$

where λ_1-λ_4 are weighting coefficients for the auxiliary losses.

During inference, test videos are processed by both MoNE and MoAE. The augmented features from the two modules are fused via memory-similarity weights, and the anomaly detector outputs snippet-level scores, which are then aggregated to frame-level predictions

B. Implementation Details

Data Pre-processing. To extract video features, we follow the standard pipeline used in recent works [4]. Each video is first segmented into snippets of 16 frames, and 512-dimensional visual features are extracted using the CLIP visual encoder (ViT-B/16) [1]. For fair comparison, we apply the same data augmentation strategy as [4]. Specifically, for the UCF-Crime [2] and XD-Violence [3] datasets, we adopt the 10-crop augmentation, which includes the center crop, four corner crops, and their horizontal flips. For text features, we encode the anomaly prototypes using the CLIP text encoder, obtaining 512-dimensional textual embeddings.

Hyperparameter settings. During training, the batch size is 128, consisting of 64 normal and 64 abnormal videos. The learning rate is initialized at 1×10^{-4} for UCF-Crime and decayed by 0.5 after the first epoch, while for XD-Violence it is set to 4×10^{-5} and decayed by 0.1 after the second epoch. The loss weights $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are set to $(1 \times 10^{-2}, 1, 1 \times 10^{-5}, 1 \times 10^{-3})$ for UCF-Crime and $(7\times10^{-5}, 4\times10^{-5}, 2\times10^{-5}, 7\times10^{-4})$ for XD-Violence. The temperature parameter τ is fixed at 3×10^{-3}. The memory sizes are $M_{\text{in}} = M_{\text{ext}} = 60$. The learning rate α for the external memory is fixed at 1×10^{-3}.

C. Datasets Details

UCF-Crime. UCF-Crime is a large-scale real-world surveillance dataset containing 1,900 videos with a total duration of over 128 hours. It covers 13 categories of anomalous events, such as fighting, robbery, and accidents, while also including extensive normal scenes to prevent category bias. Only video-level binary labels (normal/abnormal) are provided, making it a typical benchmark for weakly supervised video anomaly detection. Its diverse scenes, varying illumination, and complex camera motions pose significant challenges for anomaly modeling.

XD-Violence. XD-Violence is a multi-modal large-scale violence detection dataset composed of 4,754 video clips totaling more than 217 hours. It provides both RGB and audio modalities, with video-level violence/non-violence labels. The dataset contains a wide variety of violent activities such as riots, fighting, and weapon attacks across indoor and outdoor environments. Its cross-modal design enables richer supervision signals, making it suitable for multi-modal anomaly or violence detection research. The dataset statistics are summarized in Table 1.

D. Additional Analysis

D.1. Efficiency Study

In Table 2, we compare the efficiency and performance of UR-DMU, VadCLIP, and our proposed MoME framework. All results are obtained on a single 910C NPU, excluding the feature extraction time. Although MoME contains substantially more parameters due to its multi-expert architecture, the sparse MoE design activates only a small subset of experts for each input, resulting in a lower computational cost (6.3 GFLOPs) than VadCLIP, which has far fewer parameters. Notably, MoME can be easily scaled to larger datasets without increasing computational overhead, since the sparse MoE mechanism allows the number of experts to grow while keeping FLOPs nearly constant. While UR-DMU is a much smaller model, its AUC is considerably lower. Despite having a larger model capacity, MoME still achieves an inference time of 11.02s, demonstrating its potential for real-time deployment.

D.2. Hyper-parameter Study

The number of experts determines how many diverse patterns the model can capture, while the number of explicit anomaly experts controls the balance between explicit and

Dataset	Paper	Year	#Videos	Length	Label Type	Modalities	Anomaly Number
UCF-Crime	[2]	2018	1,900	128hrs	Video-level	RGB	13 anomaly categories
XD-Violence	[3]	2020	4,754	214hrs	Video-level	RGB + Audio	6 anomaly categories

Table 1. Statistics of the UCF-Crime and XD-Violence datasets used in our experiments.

	UR-DMU	VadCLIP	MoME (Ours)
GFLOPs	3.42	10.12	6.30
Parameters (M)	5.71	12.64	42.29
Inference Time (s)	2.62	7.45	11.02
AUC (%)	86.97	88.02	**88.32**

Table 2. Comparison of efficiency-performance trade-offs.

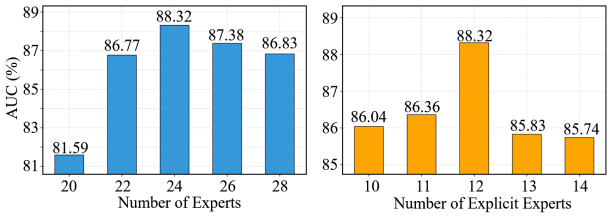


Figure 1. Hyper-parameter analysis on UCF-Crime. The effect of (left) the total number of experts and (right) the number of explicit anomaly experts on model performance.

implicit pattern learning. Therefore, we first conduct a hyperparameter analysis on these two factors. To encourage each expert to specialize in different patterns and to allow scaling to more experts without increasing computational cost, we fix the number of activated experts k to 1. In addition, we investigate the influence of the weights of the four loss terms to assess the sensitivity of our model to these parameters.

Number of Experts. The total number of experts determines the model's representation capacity and the diversity of learned patterns. As illustrated in Fig. 1(a), the performance consistently improves when the number of experts increases from 20 to 24, demonstrating that additional experts allow the model to capture richer and more diverse representations. However, as the number of experts continues to grow beyond 24, the performance begins to decline. We attribute this phenomenon to the limited availability of training samples relative to the growing number of parameters, which can lead to expert redundancy and unstable optimization. In particular, excessive experts tend to overlap in their functionality, causing conflicting gradients and reducing the effectiveness of expert specialization. This observation indicates that a moderate number of experts achieves a good balance between representation diversity and model generalization.

Number of Explicit Experts. Explicit experts are guided by semantic supervision and play a critical role in encouraging the model to learn interpretable and semantically grounded representations. To study the effect of explicit experts, we fix the total number of experts at 24 and vary the ratio of explicit ones. As shown in Fig. 1(b), the performance peaks when the proportion of explicit experts is around 50%. Increasing this ratio beyond 50% leads to a gradual decline in accuracy, suggesting that too much reliance on semantic supervision restricts the model's capacity to explore implicit and complementary patterns. Conversely, when too few explicit experts are used, the model lacks sufficient semantic guidance and may converge to suboptimal representations. These results highlight the importance of maintaining a balanced ratio, where explicit experts provide high-level semantic cues while implicit experts freely discover complex visual regularities.

Weights of Proxy Loss Functions. Our framework incorporates several proxy loss functions to guide training toward multiple objectives, such as classification, balance regularization, and diversity learning. To assess the sensitivity of the model to these hyperparameters, we vary the corresponding coefficients λ_1-λ_4 and summarize the results in Fig. 2. We observe that the model is relatively insensitive to λ_2-λ_4, indicating robust convergence and stable optimization behavior. In contrast, the model is more sensitive to the balance loss weight. When the weight of the balance loss is set too high, performance drops noticeably because the balance loss typically has a larger magnitude than the classification loss. An overly large coefficient forces the network to overemphasize inter-expert balance at the expense of discriminative learning, thereby weakening anomaly recognition capability. This finding suggests that while the balance loss is essential for avoiding expert collapse, its contribution should be carefully controlled to prevent it from dominating the overall objective.

D.3. Per-class analysis.

Figure 3 reports per-class results on UCF-Crime and XD-Violence. VadCLIP induces diversity primarily at the *class* level: its mechanism encourages representations to separate according to category labels, which is effective when anomalies exhibit consistent, class-specific visual signatures. However, this class-centric diversity does not explicitly enforce the learning of *general* (class-agnostic) patterns that are shared across multiple anomaly types. In contrast, MoME promotes *semantic* diversity by allocating experts according to semantically meaningful cues and, crucially, jointly optimizes for both *specialization* and *generality*. Concretely, some experts specialize in discriminat-

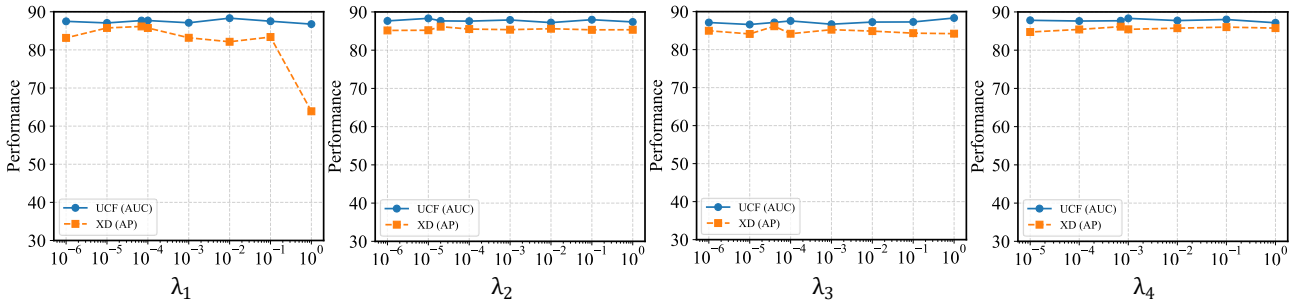
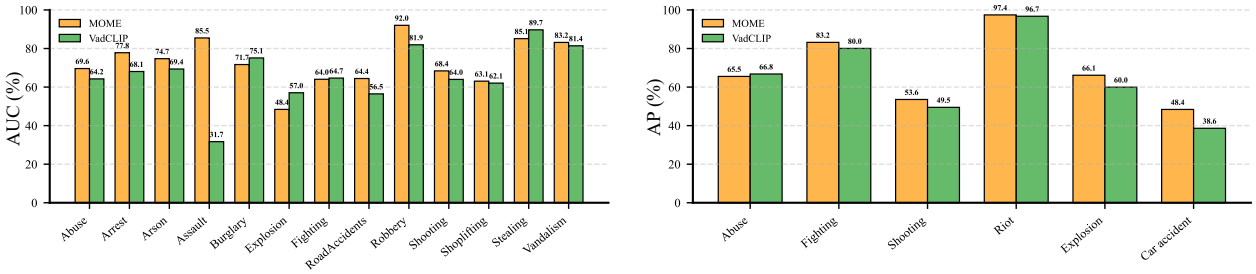


Figure 2. Sensitivity analysis of four proxy loss functions.



(a) Per-class Results on UCF-Crime

(b) Per-class Results on XD-Violence

Figure 3. Category-wise performance comparison with VadCLIP on UCF-Crime and XD-Violence.

ing fine-grained, class-specific cues (e.g., the detailed appearance or motion patterns unique to *Burglary* or *Explosion*), while other experts are encouraged to capture class-agnostic, broadly useful anomaly priors (e.g., abrupt motion bursts, sudden scene changes, or unusual object-scene interactions).

This joint learning of semantic specialization and generality explains the empirical gains observed for MoME. On **UCF-Crime**, MoME achieves substantial improvements over VadCLIP in categories characterized by complex dynamics. For example, the AUC rises from 31.7% to 85.5% in *Assault* (+53.8%) and from 81.9% to 92.0% in *Robbery* (+10.1%). These gains highlight that the combination of class-agnostic cues and multiple complementary semantic experts enables more reliable anomaly discrimination under highly dynamic scenes. In contrast, for categories with relatively homogeneous visual signatures, such as *Abuse* (69.6% vs. 64.2%) both methods perform comparably, suggesting that class-level diversity alone can sufficiently model visually consistent anomalies. On **XD-Violence**, MoME also exhibits stable advantages in most categories. Notably, the AP increases from 60.0% to 66.1% for *Explosion* (+6.1%) and from 38.6% to 48.4% for *Car accident* (+9.8%), demonstrating that the ability to learn shared anomaly priors significantly enhances robustness under scene variations and occlusions. For large-scale or highly distinctive events such as *Riot*, both methods reach near-saturation performance (97.4% vs. 96.7%), where the anomaly signal is already salient and easily separable.

D.4. Additional Ablation Study

Prototype Quality. To investigate whether fine-grained anomaly prototypes provide better representations than coarse-grained anomaly classes, we replace the anomaly prototypes in the Anomaly Prototype Router (APR) with class-level anomaly labels. The ablation results are reported in Fig. 6(a). Replacing the fine-grained prototypes with coarse-grained class labels leads to performance drops of -1.9% on UCF-Crime and -8.5% on XD-Violence. This observation indicates that fine-grained semantic prototypes capture richer anomaly characteristics and provide more discriminative routing signals than coarse-grained class-level representations.

Memory Size. We further analyze the influence of the memory size in the MoME module. The ablation results are shown in Fig. 6(b). When the memory size is small, the model performance degrades due to the limited capacity to store representative patterns. As the memory size increases, the performance gradually stabilizes, indicating that the model becomes less sensitive to larger memory capacities. Considering both performance and computational efficiency, we set the memory size to 60 in all experiments.

E. More Visualization Results

E.1. Extended Visualization of Expert-Prototype Allocation Matrix

In the main paper, we visualize only the Top-1 expert-prototype assignments for clarity. Here, we provide the full expert-prototype allocation matrix, which includes all as-

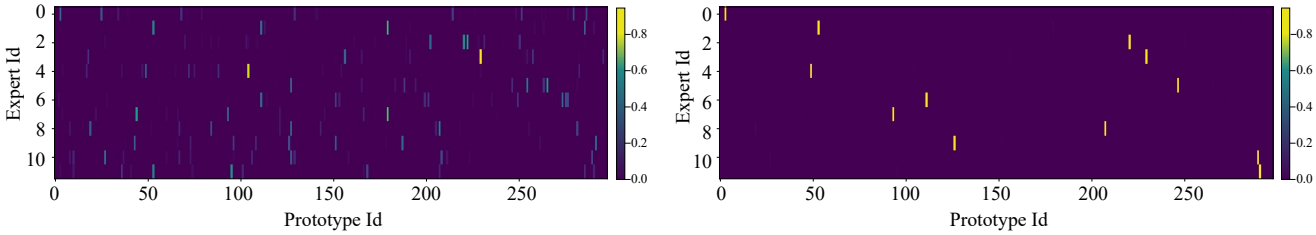


Figure 4. Visualization of the learned expert-prototype allocation matrix. The left heatmap shows results *without* regularization, where multiple experts focus on the same prototypes, leading to redundant specialization. The right heatmap corresponds to training *with* regularization, where experts are encouraged to specialize in distinct prototypes, yielding a more balanced and strong allocation.

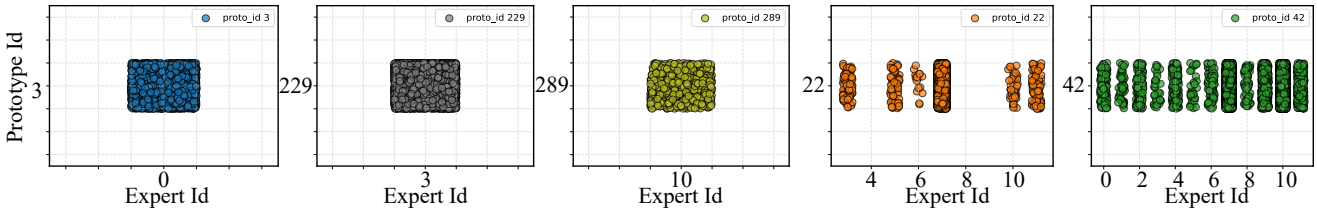
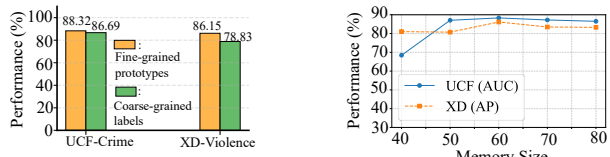


Figure 5. Visualization of data allocation in the Anomaly Prototype Router (APR). Each point denotes an anomaly sample routed to an expert by its semantic prototype. Highly correlated prototypes (e.g., 3, 229, 289) form compact expert-specific clusters, while ambiguous ones (e.g., 22, 42) spread across multiple experts, showing how APR promotes semantic specialization with shared coverage for transferable anomalies.



(a) Prototype quality ablation (b) Memory size ablation

Figure 6. Ablation of prototype quality and memory size

signment scores. This visualization offers a more comprehensive view of how experts attend to multiple prototypes during training.

E.2. Data Allocation.

To better illustrate the functionality of the APR module, we visualize the data allocation process in Fig. 5. For clearer visualization, a small random jitter is added to integer indices. We observe that when a prototype is highly matched with a specific expert (as shown in Fig. 4), the corresponding data are exclusively allocated to that expert. In contrast, ambiguous prototypes are distributed across multiple experts. Interestingly, the highly matched prototypes often represent more generalizable information. For example, prototype 3 corresponds to "Several people began to push each other, causing panic in the crowd." Such a description does not pertain to a specific anomaly class but directly reflects anomalous actions themselves, making it more transferable. On the other hand, ambiguous prototypes usually encode broader categories. For instance, prototype 42 corresponds to "explosive device detonation," which can cover various anomalous events such as "car collision" or "fire." It is therefore reasonable that such prototypes are allocated to multiple experts.

E.3. Anomaly Scores

More anomaly score results on XD-Violence are shown in Fig. 7 and Fig. 8. Overall, our method consistently produces higher and more discriminative anomaly scores than VadCLIP across various challenging videos in XD-Violence. In scenes involving shooting, fighting, explosions, or aggressive physical interactions, our method generates sharp and distinct peaks that precisely align with true abnormal moments while remaining low and stable during normal segments. In contrast, VadCLIP often yields smoother and less expressive curves, which makes it difficult to distinguish abnormal events from normal background motion, especially in videos with fast camera movement, dark environments, or cluttered backgrounds. These results verify the effectiveness of our proposed joint learning framework in capturing both general and semantically diverse patterns.

Another important observation is that our method demonstrates strong robustness to annotation noise and is able to reveal mislabeled abnormal segments. For example, in *God.Bless.America.2011*, the frames between 875 and 1250 clearly include threatening actions, yet they are annotated as normal; our model still assigns high anomaly scores to this region. Similarly, in *Desperado.1995*, shoot-

ing occurs between frames 1122 and 1309, but the annotation marks the segment as normal; our method detects the anomaly with clear high responses that better match the visual content.Across all these cases, our model highlights semantically abnormal events even when the annotations fail to do so, whereas VadCLIP tends to follow the annotation too closely and often produces nearly flat predictions in these mislabeled regions.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2

[2] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 3

[3] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 1, 2, 3

[4] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 1, 2

[5] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 1

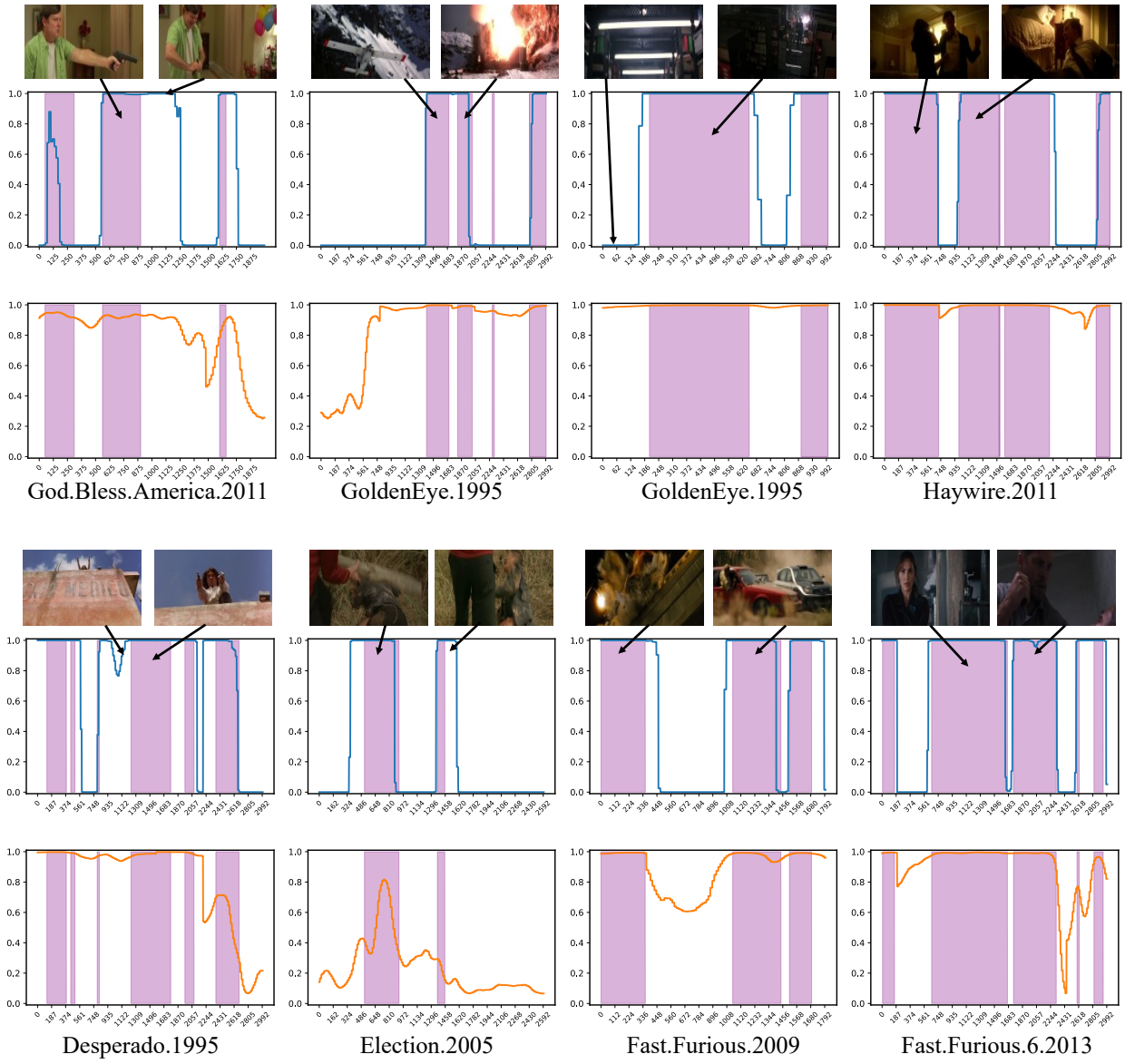


Figure 7. More qualitative results of our method and VadCLIP on XD-Violence test videos. The blue and orange lines denote our method and VadCLIP, respectively. The purple shaded regions highlight abnormal segments. The Y-axis shows anomaly scores, and the X-axis indicates video frame indices.

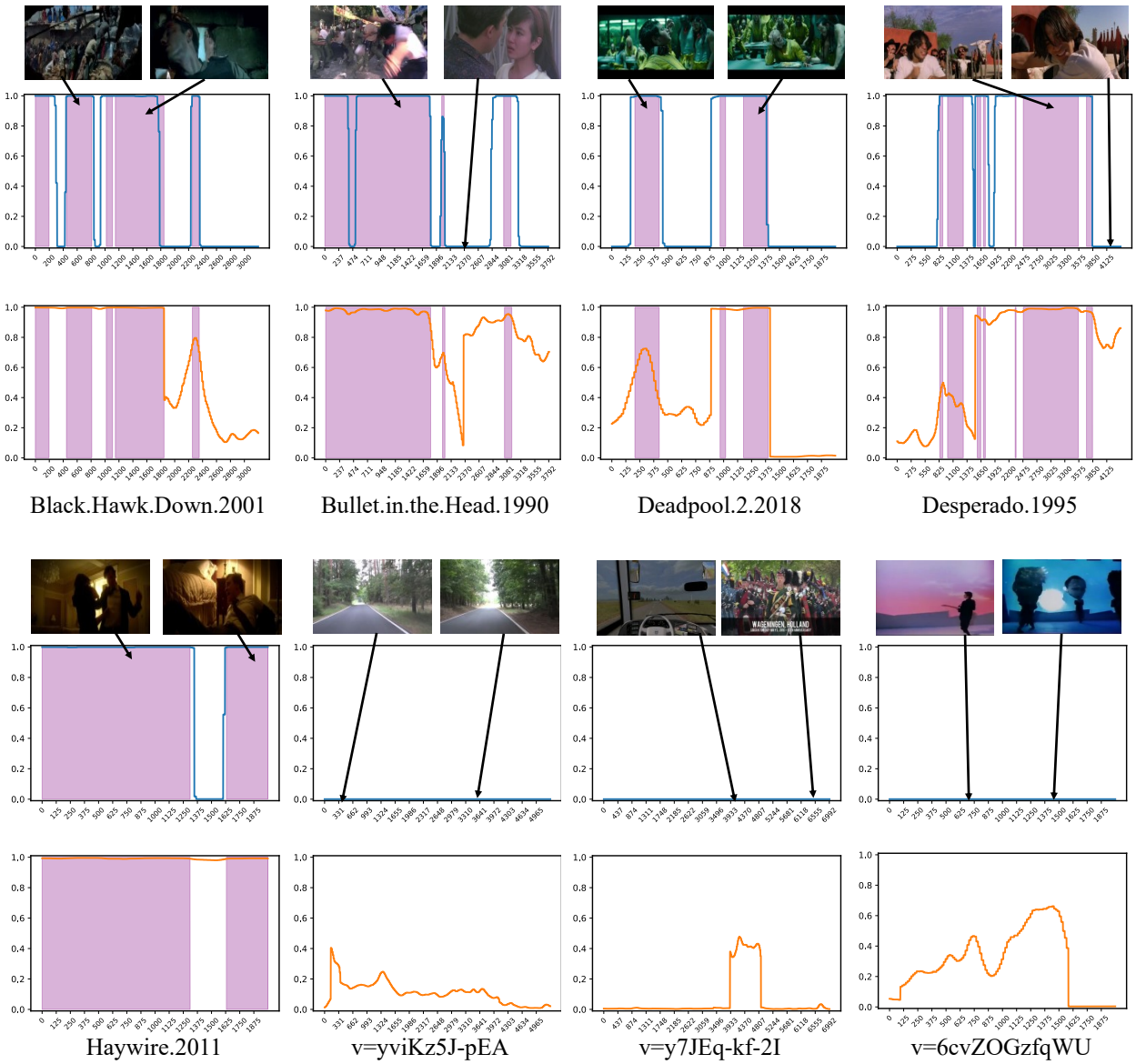


Figure 8. More qualitative results of our method and VadCLIP on XD-Violence test videos. The blue and orange lines denote our method and VadCLIP, respectively. The purple shaded regions highlight abnormal segments. The Y-axis shows anomaly scores, and the X-axis indicates video frame indices.