Anchor Then Align: Anomaly-Aware CLIP Adaptation with Disentangled Text Anchors for Weakly-Supervised VAD

Chuanhao Liu
Lanzhou Jiaotong University
Lanzhou, China
liuchuanhao2004@outlook.com

***Abstract*—Weakly-supervised video anomaly detection (VAD) aims to localize anomalies using only video-level supervision, but it is often challenged by temporal ambiguity and semantic noise. Recent CLIP-based approaches benefit from vision–language priors, yet most of them keep CLIP frozen and rely on generic prompt matching. We observe that CLIP is not sufficiently anomaly-aware in its pretrained semantic space: it struggles to clearly distinguish normal from anomalous prompts, resulting in an insufficient similarity margin and unstable segment-level scoring. To address this limitation, we propose an anomaly-aware CLIP adaptation framework that equips CLIP with lightweight residual adapters and a text-space disentanglement objective, explicitly constructing more separable normal–anomaly semantic anchors. On top of the adapted representations, we further introduce a parameter-efficient temporal adapter and a dual-branch weak supervision design that unifies binary MIL classification with clip-to-text MIL-Align learning, optimized via a three-stage curriculum for stable training. Extensive experiments on UCF-Crime and XD-Violence demonstrate that our method achieves state-of-the-art performance under the weakly-supervised setting with lightweight trainable modules.**

***Index Terms*—video anomaly detection, weak supervision, vision-language model, CLIP adaptation, temporal modeling, transformer networks, representation & reasoning, neural network applications**

I. Introduction

Video anomaly detection (VAD) aims to identify abnormal events in long, untrimmed videos and is crucial for surveillance and public safety. In practice, anomalous events are rare and diverse, making dense temporal annotation expensive and often unavailable. Therefore, weakly-supervised VAD has become a widely adopted setting, where only video-level labels are provided during training and the model must infer segment-level anomaly scores under severe temporal ambiguity and label noise.

Recently, vision–language models (VLMs) have brought strong cross-modal priors and open-vocabulary generalization to visual recognition. CLIP [1]-style dual encoders enable matching visual content with natural language prompts, which has inspired a line of CLIP-based VAD methods that leverage prompts to guide anomaly scoring. Representative approaches such as VadCLIP [2] combine video-level discrimination with vision–language alignment, supporting both coarse-grained anomaly detection and finer semantic alignment. Subsequent studies further enrich CLIP-based VAD by incorporating additional cues into prompts or cross-modal reasoning, e.g., introducing spatio-temporal cues in STPrompt [3] and integrating audio cues for cross-modal anomaly retrieval in VarCMP [4].

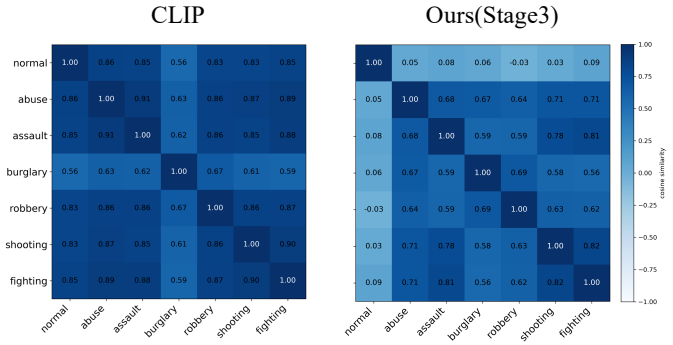


Fig. 1. Text-space cosine similarity among normal and anomaly class prompts. Left: frozen CLIP. Right: our Stage-3 adaptation.

Despite encouraging progress, a fundamental bottleneck remains when transferring CLIP to VAD: anomaly is typically defined as a deviation from normal patterns, rather than a standalone semantic category. As CLIP is pretrained to align holistic image content with text descriptions, it tends to emphasize coarse semantics and may fail to capture subtle context-dependent deviations. Consequently, generic normal/anomalous prompts can be insufficiently separated in CLIP's text space, leading to weak semantic margins and unstable segment scoring in long videos. Fig. 1 illustrates this issue: the frozen CLIP text embeddings exhibit high pairwise similarities across normal and anomaly prompts, while our Stage-3 adaptation significantly improves their separability.

To address CLIP's limited anomaly awareness in weakly-supervised VAD, we propose an anomaly-aware CLIP adaptation framework built upon disentangled text anchors. Specifically, we insert lightweight residual adapters into both the visual and textual branches and introduce a text-space disentanglement objective, which explicitly enlarges the semantic margin between normal and anomalous prompts and yields robust anomaly-aware anchors. On top of the adapted anchors, we adopt a dual-branch weak-supervision design: a classification branch learns stable anomaly confidence from video-

level binary labels, while an alignment branch (activated in later stages) strengthens fine-grained anomaly semantics via clip-to-text matching with MIL-Align pooling. In addition, we incorporate a lightweight local-to-global temporal adapter as an auxiliary module to refine clip representations, and optimize the whole framework with a three-stage curriculum to progressively stabilize anchor geometry and video representations for reliable anomaly localization and detection.

Our main contributions are:

- We identify a key limitation when transferring CLIP to weakly-supervised VAD: normal/anomalous prompts can be semantically entangled in the text space, resulting in weak margins and unstable segment scoring.
- We propose an anomaly-aware CLIP adaptation framework that learns disentangled normal/anomaly text anchors via lightweight residual adapters and an explicit disentanglement loss, and couples them with a dual-branch weak-supervision design under a three-stage curriculum for stable optimization.
- Extensive experiments on UCF-Crime and XD-Violence demonstrate state-of-the-art performance under the weakly-supervised setting, and ablations further validate the contribution of each component.

II. Related Work

A. Vision–Language Models and CLIP

Recent years have witnessed a rapid rise of vision–language foundation models, which learn transferable cross-modal representations from large-scale image–text corpora and enable prompt-driven inference across tasks [5]–[7]. Among them, CLIP [1] popularizes a simple dual-encoder paradigm trained with contrastive objectives, aligning an image with its paired text in a shared embedding space and supporting zero-shot recognition. Owing to its generalization ability, CLIP has been widely adopted as a backbone for downstream tasks [6], [7].

Despite its broad success, CLIP's pretraining objective is inherently biased toward global semantics: it matches an entire image to a typically coarse sentence (often category-like), without explicit supervision for localized regions, fine-grained attributes, or context-dependent deviations [1]. To improve fine-grained perception, follow-up works explore region-aware or grounding-oriented pretraining, e.g., coupling language with localized visual cues or detection-style supervision [8], [9]. However, these improvements largely target conventional categories/attributes, whereas abnormality is rarely a standalone concept: it is more naturally characterized as a deviation from normal patterns under a specific context, which requires sensitivity to subtle and context-dependent evidence.

B. Weakly-Supervised VAD and CLIP-based Methods

Weakly-supervised video anomaly detection (WS-VAD) learns from video-level labels without temporal boundaries, and typically infers segment-level anomaly scores under severe temporal ambiguity and label noise. Many WS-VAD methods adopt Multiple Instance Learning (MIL) to aggregate segment scores into video-level supervision [10]–[14].

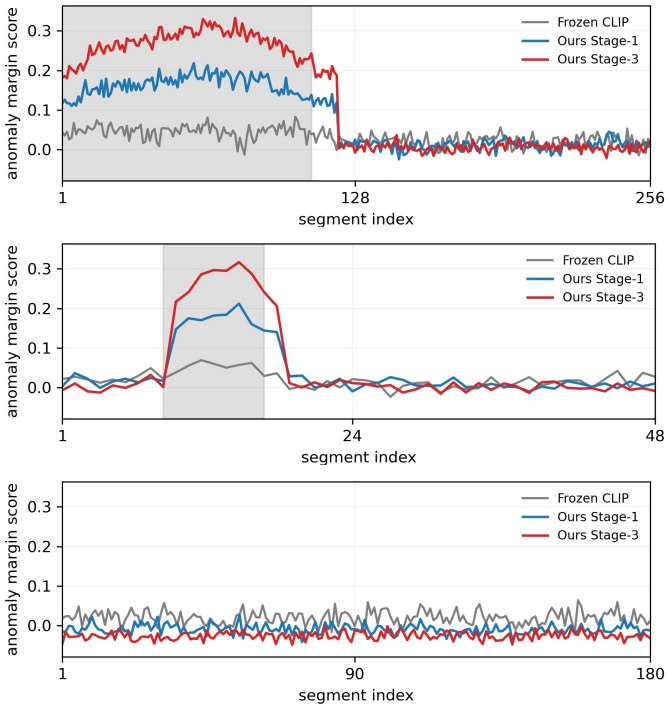


Fig. 2. Segment-level anomaly margin scores on representative videos.

With the rise of VLMs, recent WS-VAD studies increasingly leverage CLIP to inject semantic guidance through text prompts and vision–language similarity. VadCLIP [2] is an early attempt that combines coarse video-level discrimination with clip-to-text alignment in a dual-branch framework. Subsequent works enrich the guidance signal by introducing spatio-temporal prompting [3], LLM-generated explanations [15], or additional cross-modal cues such as audio [4]. Despite their promising progress, most CLIP-based methods still treat CLIP as a frozen feature extractor and focus on designing heads/prompts on top of fixed representations, while leaving CLIP's limited anomaly awareness unaddressed.

In contrast, our work directly enhances CLIP's anomaly sensitivity by learning anomaly-aware semantic anchors through lightweight adaptation and an explicit separation objective, and then integrates the improved representations into a stable weakly-supervised training pipeline for untrimmed videos.

III. Motivation and Analysis

A. CLIP is Weakly Separated for Normal vs. Anomaly Prompts

Although CLIP provides strong cross-modal priors, we observe that its text embedding space is not inherently anomaly-aware. In VAD, abnormality is defined as a context-dependent deviation from the norm rather than a standalone semantic category. Consequently, generic prompts describing "normal" and "anomalous" behaviors can be weakly separated in CLIP, yielding a small semantic margin for video–text matching. As shown in Fig. 1, the frozen CLIP text embeddings exhibit high pairwise similarities between normal and anomaly prompts, suggesting an entangled prompt space. Such entanglement

implies that the prompt space itself provides limited separability, which may directly translate into unreliable segment-level scoring when CLIP similarities are used for localization.

B. Entangled Semantics Undermine Segment Scoring in VAD

Weakly-supervised VAD requires assigning segment-level anomaly scores while training only with video-level labels. A common CLIP-based scoring strategy compares each segment feature with normal/anomaly text anchors. We define a simple margin-style score:

$$\Delta_t = s(\mathbf{v}_t, \mathbf{t}_{\text{anom}}) - s(\mathbf{v}_t, \mathbf{t}_{\text{norm}}), \tag{1}$$

where $\mathbf{v}_t$ is the feature of segment t, $\mathbf{t}_{\text{norm}}$ and $\mathbf{t}_{\text{anom}}$ denote the normal and anomaly text anchors, respectively, and $s(\cdot,\cdot)$ is cosine similarity. If $\mathbf{t}_{\text{norm}}$ and $\mathbf{t}_{\text{anom}}$ are close in the embedding space, the resulting margin Δ_t becomes weak and susceptible to background variations, making it difficult to reliably separate anomalous segments from normal dynamics. Fig. 2 illustrates this effect on representative abnormal and normal videos: compared with Frozen CLIP, our anomaly-aware adaptation (Stage-1) yields more distinguishable responses within anomalous intervals, while the jointly trained model (Stage-3) further enlarges the margin and suppresses spurious fluctuations on normal videos, leading to more reliable localization.

C. Takeaway

These observations motivate a simple principle: anomaly-aware semantic anchors should be constructed before performing weakly-supervised optimization. Accordingly, our method first adapts CLIP with lightweight modules and an explicit separation objective to enlarge the normal–anomaly margin in the embedding space, and then performs weakly-supervised learning on top of the improved representations.

IV. Method

A. Problem Setup and Overview

Given an untrimmed video, we uniformly split it into T clips and extract clip-level visual features $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, $\mathbf{x}_t \in \mathbb{R}^d$ using a *frozen* CLIP image encoder (or pre-extracted CLIP features). We consider a label set $\mathcal{C}$ with M text classes, where $c = 0$ denotes *Normal* and $c \in \{1, \ldots, M-1\}$ correspond to anomaly categories. Each training video is associated with a video-level multi-hot label $\mathbf{y} \in \{0,1\}^M$ and a binary label $y^{\text{bin}} \in \{0,1\}$ indicating whether the video contains anomalies.

As shown in Fig. 3, our framework follows a dual-branch design: (i) a coarse-grained classification branch (**C-branch**) trained with binary weak supervision to produce clip-level anomaly confidence, and (ii) a fine-grained alignment branch (**A-branch**) that constructs a clip-to-text alignment map and applies MIL-Align pooling for category supervision. To overcome CLIP's limited anomaly awareness, we adapt CLIP representations via lightweight residual adapters and explicitly disentangle normal vs. anomaly anchors in the text space. The whole model is optimized with a three-stage curriculum (Fig. 3b).

B. Temporal Adapter: Local Transformer + Global Graph

CLIP is pretrained on images and does not explicitly model temporal relations. We introduce a lightweight temporal adapter $g_{\text{temp}}(\cdot)$ to enhance clip features with both local continuity and global structure. First, we add learnable positional embeddings $\mathbf{P}$ and apply a windowed Transformer encoder to capture local temporal dependencies:

$$\tilde{\mathbf{X}} = \text{Trans}_w(\mathbf{X} + \mathbf{P}), \tag{2}$$

where w is the attention window size. Then, we build a global clip graph and apply a graph attention layer (GAT) to propagate long-range context:

$$\mathbf{H} = \text{GAT}(\tilde{\mathbf{X}}, \mathbf{A}), \tag{3}$$

where $\mathbf{A}$ is an affinity graph constructed from feature similarity (optionally combined with temporal-distance priors). Finally, we obtain temporally enhanced features:

$$\mathbf{X}^{\text{temp}} = g_{\text{temp}}(\mathbf{X}) = \text{Proj}([\tilde{\mathbf{X}}; \mathbf{H}]), \tag{4}$$

with $[\cdot;\cdot]$ denoting concatenation and $\text{Proj}(\cdot)$ a linear projection. This module is parameter-efficient and can be trained separately from CLIP adaptation to stabilize weakly-supervised optimization.

C. Anomaly-Aware CLIP Adaptation via Residual Adapters

We adapt CLIP representations using lightweight residual adapters while keeping CLIP encoders frozen. Given an embedding $\mathbf{z} \in \mathbb{R}^d$, a residual adapter is defined as:

$$g(\mathbf{z}) = \mathbf{z} + \alpha \cdot \mathbf{W}_2\, \phi(\mathbf{W}_1\, \text{LN}(\mathbf{z}))\,, \tag{5}$$

where $\text{LN}(\cdot)$ is LayerNorm, $\phi(\cdot)$ is a nonlinearity (e.g., GELU), and α is a scaling factor. We use: (i) a **visual adapter** $g_v(\cdot)$ applied on top of temporally enhanced features, and (ii) a **text adapter** $g_t(\cdot)$ applied on top of text embeddings. The adapted visual features are:

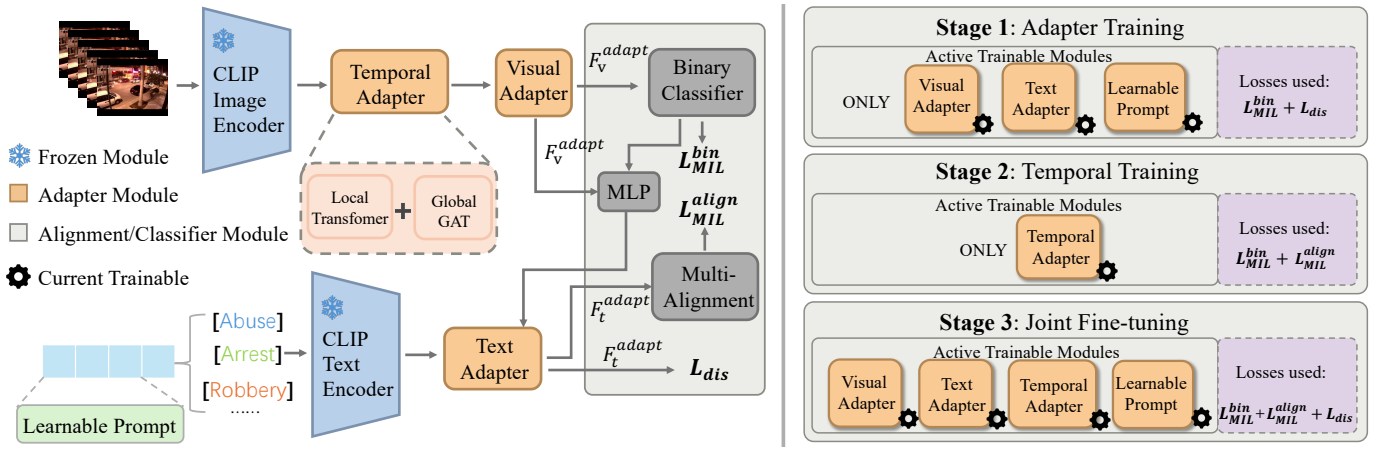
$$\mathbf{F}_v^{\text{adapt}} = \{\mathbf{f}_{v,t}^{\text{adapt}}\}_{t=1}^{T}, \quad \mathbf{f}_{v,t}^{\text{adapt}} = g_v(\mathbf{x}_t^{\text{temp}}). \tag{6}$$

D. Learnable Prompts and Text Anchors

For each class $c \in \mathcal{C}$, we construct a prompt by inserting L learnable context tokens around the class name. The prompt is encoded by the frozen CLIP text encoder to obtain $\mathbf{t}_c \in \mathbb{R}^d$, and then adapted by the text adapter:

$$\mathbf{f}_{t,c}^{\text{adapt}} = g_t(\mathbf{t}_c). \tag{7}$$

The set $\{\mathbf{f}_{t,c}^{adapt}\}_{c=0}^{M-1}$ serves as *text anchors* for normal/anomaly categories. We will explicitly enforce normal–anomaly separation among these anchors in Sec. IV-F.



a) Overall Model Structure b) Three-Stage Training Process Flow

Fig. 3. Overall framework of our anomaly-aware CLIP adaptation with dual-branch weak supervision and three-stage training.

E. Dual-Branch Weak Supervision

a) C-branch: binary MIL classification.: Given adapted clip features $\mathbf{f}_{v,t}^{adapt}$, we compute clip-level anomaly confidence via a binary classifier:

$$a_t = \sigma(\mathbf{w}^\top \mathrm{MLP}(\mathbf{f}_{v,t}^{adapt})), \tag{8}$$

where $\sigma(\cdot)$ is sigmoid and $\mathrm{MLP}(\cdot)$ is a shared projection head (trained in all stages). We adopt top-k MIL pooling to obtain a video-level prediction:

$$\bar{a} = \frac{1}{k} \sum_{t \in \mathrm{TopK}(\{a_t\})} a_t, \quad \mathcal{L}_{\mathrm{MIL}}^{\mathrm{bin}} = \mathrm{BCE}(\bar{a}, y^{\mathrm{bin}}). \tag{9}$$

b) A-branch: clip-to-text alignment map with MIL-Align.: We construct a fine-grained alignment map between clips and text anchors using normalized dot-product similarity:

$$\mathbf{M}_{t,c} = \frac{\langle \mathrm{norm}(\mathrm{MLP}(\mathbf{f}_{v,t}^{adapt})),\ \mathrm{norm}(\mathbf{f}_{t,c}^{adapt}) \rangle}{\tau}, \tag{10}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization and τ is a temperature. Importantly, Multi-Alignment has no learnable parameters: it is purely the similarity computation in Eq. (10) followed by MIL-Align pooling. Specifically, for each class c, we aggregate top-k clip responses:

$$s_c = \frac{1}{k} \sum_{t \in \mathrm{TopK}(\mathbf{M}_{:,c})} \mathbf{M}_{t,c}, \tag{11}$$

and optimize a video-level multi-class (or multi-label) objective:

$$\mathcal{L}_{\mathrm{MIL}}^{\mathrm{align}} = -\sum_{c=0}^{M-1} \tilde{y}_c \log \left(\mathrm{Softmax}(\mathbf{s})\right)_c, \quad \mathbf{s} = [s_0, \ldots, s_{M-1}], \tag{12}$$

where $\tilde{\mathbf{y}}$ is the normalized multi-hot label (one-hot in the single-label case).

F. Text-Space Disentanglement Loss

To explicitly improve anomaly awareness, we enforce a clearer normal–anomaly boundary in the adapted text space. Let $\mathbf{f}_{t,0}^{adapt}$ be the *Normal* anchor and $\mathbf{f}_{t,c}^{adapt}$ $(c \geq 1)$ be anomaly anchors. We penalize high similarity between normal and anomaly anchors with a margin:

$$\mathcal{L}_{\mathrm{dis}} = \frac{1}{M-1} \sum_{c=1}^{M-1} \max\Big(0,\ \cos(\mathbf{f}_{t,0}^{\mathrm{adapt}}, \mathbf{f}_{t,c}^{\mathrm{adapt}}) - m\Big), \tag{13}$$

where m is a margin hyperparameter and $\cos(\cdot,\cdot)$ is cosine similarity. This objective directly reshapes the prompt/text embedding geometry, making "normal" less entangled with anomaly categories.

G. Three-Stage Training Curriculum

Direct end-to-end optimization can be unstable under weak supervision and entangled semantics. We therefore use a three-stage curriculum (Fig. 3b), where the MLP/classifier heads are trained in all stages to stay compatible with the currently active feature space.

a) Stage 1: Adapter training (anchor shaping).: We freeze the temporal adapter and optimize visual adapter, text adapter, learnable prompts, and MLP/classifier heads. The goal is to build anomaly-aware anchors and binary separability without introducing additional temporal learning noise:

$$\mathcal{L}^{(1)} = \mathcal{L}_{\mathrm{MIL}}^{\mathrm{bin}} + \lambda_{\mathrm{dis}} \mathcal{L}_{\mathrm{dis}}. \tag{14}$$

b) Stage 2: Temporal training.: We freeze the visual/text adapters and prompts, and train only the temporal adapter (plus MLP/classifier heads). We introduce alignment supervision here to make temporal features better support clip-to-text matching:

$$\mathcal{L}^{(2)} = \mathcal{L}_{\mathrm{MIL}}^{\mathrm{bin}} + \mathcal{L}_{\mathrm{MIL}}^{\mathrm{align}}. \tag{15}$$

c) Stage 3: Joint fine-tuning.: We jointly fine-tune temporal adapter, visual adapter, text adapter, learnable prompts, and MLP/classifier heads, using the full objective:

$$\mathcal{L}^{(3)} = \mathcal{L}_{\text{MIL}}^{\text{bin}} + \mathcal{L}_{\text{MIL}}^{\text{align}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}. \tag{16}$$

V. Experiments

A. Datasets

We evaluate our approach on two widely used weakly-supervised video anomaly detection benchmarks: UCF-Crime and XD-Violence. Following standard protocols, we train with video-level labels and infer segment/frame-level anomaly scores.

B. Evaluation Metrics

UCF-Crime. We report frame-level AUC computed from ROC-AUC. In our implementation, clip-level anomaly scores are converted to frame-level scores by repeating each clip score with a fixed stride (e.g., 16 frames per clip) before computing ROC-AUC.

XD-Violence. We report Average Precision (AP) for coarse-grained anomaly recognition following prior works.

C. Implementation Details

Backbone and features. We use a frozen CLIP image encoder (ViT-B/16) as the visual backbone and keep it fixed during training. Each video is uniformly divided into $T = 256$ clips, producing a sequence of clip-level features.

Temporal adapter. The temporal adapter consists of a windowed Transformer for local modeling and a global graph attention module for long-range context. Unless otherwise specified, we use 5 Transformer layers, 1 attention head, and an attention window size of 8.

Prompt and adapters. We adopt learnable context tokens for prompts, with 10 prefix tokens and 10 postfix tokens around each class name. We enable lightweight residual adapters for both visual and text branches. In our three-stage UCF training setup, we set the adapter hidden dimension to 256 and the adapter scaling factor to 0.2.

MIL setting. For MIL pooling, we select the top-k clips for video-level aggregation, where k is proportional to video length (approximately one clip out of every 16 clips, i.e., $k = \lfloor \text{len}/16 \rfloor + 1$).

Optimization. We use AdamW with a MultiStep learning rate scheduler. For UCF-Crime in our three-stage script, we use a batch size of 64 and a learning rate of 2×10^{-5}. The default number of training epochs is 6. The disentanglement loss weight is set to 0.1 unless otherwise stated.

D. Comparison with State-of-the-Art

Analysis. Tables I and II summarize the performance of our method on XD-Violence and UCF-Crime, reported in terms of AP(%) and frame-level AUC(%), respectively. We compare against representative weakly-supervised baselines and recent CLIP-based approaches under the same weakly-supervised setting.

TABLE I
COMPARISON ON XD-VIOLENCE IN TERMS OF AP(%).

Category	Method	AP(%)
Semi	SVM baseline	50.80
Semi	OCSVM [16]	28.63
Semi	Hasan et al. [17]	31.25
Weak	Ju et al. [18]	76.57
Weak	Sultani et al. [10]	75.18
Weak	Wu et al. [19]	80.00
Weak	RTFM [14]	78.27
Weak	AVVD [20]	78.10
Weak	DMU [21]	82.41
Weak	CLIP-TSA [22]	82.17
Weak	VadCLIP [2]	84.51
Weak	STPrompt [3]	83.97
Weak	PEL4VAD [23]	85.59
Weak	Ex-VAD [15]	86.52
Weak	**Ours**	**86.84**

TABLE II
COMPARISON ON UCF-CRIME IN TERMS OF AUC(%)

Category	Method	AUC(%)
Semi	SVM baseline	50.10
Semi	OCSVM [16]	63.20
Semi	Hasan et al. [17]	51.20
Weak	Ju et al. [18]	84.72
Weak	Sultani et al. [10]	84.14
Weak	Wu et al. [19]	84.57
Weak	AVVD [20]	82.45
Weak	RTFM [14]	85.66
Weak	DMU [21]	86.75
Weak	UMIL [24]	86.75
Weak	CLIP-TSA [22]	87.58
Weak	OPVAD [25]	86.05
Weak	VadCLIP [2]	88.02
Weak	STPrompt [3]	88.08
Weak	PEL4VAD [23]	85.62
Weak	Ex-VAD [15]	88.29
Weak	**Ours**	**88.97**

Overall, CLIP-based dual-branch methods consistently outperform traditional MIL baselines, suggesting that vision–language priors are beneficial under weak supervision. However, most existing CLIP-based approaches keep CLIP frozen and rely on prompt matching, which may suffer from limited anomaly awareness. Our method explicitly disentangles normal versus anomaly anchors in the text space and stabilizes weakly-supervised optimization via a staged curriculum and a lightweight temporal adapter, leading to improved localization robustness. As a result, we achieve state-of-the-art performance under the weakly-supervised setting on both UCF-Crime and XD-Violence, supporting the effectiveness of our anomaly-aware adaptation design.

TABLE III
ABLATION STUDY ON UCF-CRIME.

Setting	AUC(%)
Frozen CLIP + MIL (baseline)	83.91
+ Visual/Text Adapter + LearnablePrompt (Stage-1)	86.77
+ Text disentanglement loss $\mathcal{L}_{\text{dis}}$ (Stage-1)	87.52
+ Temporal adapter (Stage-2)	87.64
+ MIL-Align loss $\mathcal{L}_{\text{MIL}}^{\text{align}}$ (Stage-2)	88.26
Stage-3 joint fine-tuning (full)	**88.97**

E. Ablation Study

Baseline: Frozen CLIP + MIL. The baseline relies on frozen CLIP features and a MIL-style binary classifier. Although CLIP provides strong generic semantics, its representation is not tailored for anomaly awareness, and the weak video-level supervision further amplifies ambiguity at the segment level. As a result, the baseline performance is limited.

Effect of anomaly-aware adaptation. Adding the visual/text adapters together with learnable prompts yields a clear gain. This improvement indicates that lightweight adaptation is effective for reshaping CLIP representations toward the anomaly detection objective, even without changing the frozen backbone. In particular, learnable prompts alleviate the mismatch between generic text templates and dataset-specific anomaly semantics, while adapters provide a parameter-efficient way to adjust both visual and textual embeddings for better discrimination under weak supervision.

Role of text-space disentanglement $\mathcal{L}_{\text{dis}}$. Further introducing $\mathcal{L}_{\text{dis}}$ brings an additional improvement. This supports our motivation that generic normal/anomaly prompts can be semantically entangled in CLIP text space, leading to weak matching margins. By explicitly pushing the normal anchor away from anomaly anchors, $\mathcal{L}_{\text{dis}}$ enlarges the normal–anomaly margin and makes similarity-based scoring more stable, which is critical in weakly-supervised settings where the model must localize anomalies without segment-level labels.

Contribution of temporal adapter. Adding the temporal adapter results in a modest yet consistent gain . This suggests that temporal modeling is helpful for refining clip representations by capturing local continuity and global context, but its benefit is secondary compared with establishing anomaly-aware semantic separation. This observation aligns with our design philosophy: temporal learning is most effective after the representation space becomes more anomaly-discriminative.

Benefit of MIL-Align supervision. Introducing the MIL-Align loss $\mathcal{L}_{\text{MIL}}^{\text{align}}$ further improves performance. This indicates that clip-to-text alignment provides complementary supervision beyond binary MIL classification, encouraging the model to associate clips with category-level semantics and improving the calibration of the alignment map under weak labels. Such alignment regularization is particularly valuable when anomalies are diverse, as it reduces reliance on a single binary decision boundary and promotes semantically meaningful matching.

Fig. 4. t-SNE visualization of clip embeddings. **Left:** our anomaly-aware adapted embeddings. **Right:** frozen CLIP embeddings.

Stage-3 joint fine-tuning. Finally, joint fine-tuning delivers the best performance. This shows that once anomaly-aware anchors and temporal/alignment objectives are established, jointly optimizing all trainable components allows cross-modal cues and temporal structure to be integrated more coherently, leading to a more stable and effective solution. Overall, the ablation trend validates our staged optimization strategy: we first construct anomaly-aware semantic anchors, then introduce temporal and alignment learning, and finally unify them via joint fine-tuning for the best performance.

F. Visualization and Qualitative Results

We provide two types of visualizations to better understand how our anomaly-aware adaptation improves weakly-supervised VAD: (1) t-SNE plots of learned embeddings to inspect the geometry of normal/anomaly categories; (2) qualitative localization examples to illustrate temporal scoring behaviors and key-clip evidence.

a) t-SNE embedding analysis.: Fig. 4 compares the embedding distributions produced by our method and the CLIP features. We sample clips from multiple categories and visualize their embeddings using t-SNE, where each point denotes a clip and colors indicate its category. With CLIP, embeddings from normal and different anomaly categories are highly mixed, suggesting that the pretrained space is not naturally structured for anomaly discrimination. In contrast, our method yields a more organized embedding space: normal clips form a more compact region and exhibit a clearer separation from abnormal categories, while different anomaly types become relatively more clustered. This observation is consistent with our motivation that explicitly enlarging the normal–anomaly margin helps produce more discriminative representations for downstream weakly-supervised localization.

b) Qualitative localization results.: Fig. 5 shows representative examples of coarse-grained WSVAD. For each video, we plot the predicted frame-level anomaly score curve, where the shaded region indicates the ground-truth anomalous interval. We also visualize several key clips: the highlighted thumbnails correspond to clips with high predicted anomaly scores (i.e., high-confidence evidence), while the remaining thumbnails provide surrounding context. For normal videos,

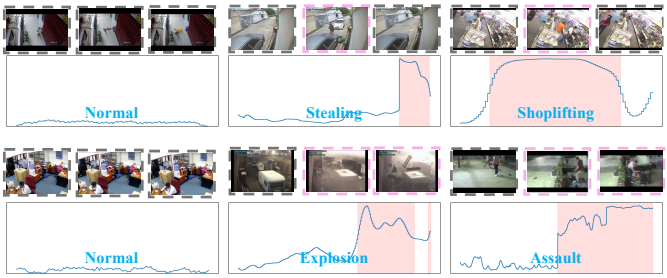


Fig. 5. Qualitative results of coarse-grained WSVAD.

our scores remain low and stable, indicating effective suppression of background fluctuations. For abnormal videos, the score curves rise prominently within the anomalous interval and drop outside it, demonstrating improved temporal consistency and localization reliability. Overall, these qualitative results suggest that our anomaly-aware adaptation not only strengthens the response on truly abnormal segments, but also reduces spurious peaks on normal content, which is critical under weak supervision.

VI. Conclusion

We present an anomaly-aware CLIP adaptation framework for weakly-supervised video anomaly detection. Our key idea is to construct more separable normal–anomaly semantic anchors by training lightweight visual/text adapters with a text-space disentanglement objective, which strengthens the margin for similarity-based localization. We combine a temporal adapter and a dual-branch weak supervision scheme (binary MIL classification and MIL-Align based clip-to-text alignment) within a three-stage curriculum for stable optimization. Experiments on UCF-Crime and XD-Violence show consistent improvements over strong CLIP-based baselines, and ablations validate each component.

Limitations and future work. We observe occasional false alarms or misses for small-scale anomalies and under low-light conditions. Future work will improve robustness in these challenging cases and extend our approach to richer open-vocabulary anomaly descriptions and more diverse prompt ensembles.

Acknowledgment

The authors used an AI language model (ChatGPT) to assist with drafting and language polishing in Sections I–V, and all such content was reviewed and edited by the authors for accuracy and clarity [26].

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[2] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: Adapting vision-language models for weakly supervised video anomaly detection," in *AAAI*, 2024, pp. 6074–6082.

[3] P. Wu, X. Zhou, G. Pang, Z. Yang, Q. Yan, P. Wang, and Y. Zhang, "Weakly supervised video anomaly detection and localization with spatio-temporal prompts," in *ACM Multimedia*, 2024, pp. 9301–9310.

[4] P. Wu, W. Su, X. He, P. Wang, and Y. Zhang, "Varcmp: Adapting cross-modal pre-training models for video anomaly retrieval," in *AAAI*, 2025, pp. 8423–8431.

[5] J. Li, R. R. Selvaraju, A. Gotmare, S. R. Joty, C. Xiong, and S. C. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021, pp. 9694–9705.

[6] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 12 888–12 900.

[7] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *ICCV*. IEEE, 2023, pp. 11 941–11 952.

[8] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, and J. Gao, "Regionclip: Region-based language-image pretraining," in *CVPR*. IEEE, 2022, pp. 16 772–16 782.

[9] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao, "Grounded language-image pre-training," in *CVPR*. IEEE, 2022, pp. 10 955–10 965.

[10] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *CVPR*, 2018, pp. 6479–6488.

[11] M. Z. Zaheer, A. Mahmood, M. Astrid, and S. Lee, "CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *ECCV*. Springer, 2020, pp. 358–376.

[12] M. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 12 742–12 752.

[13] J. Wu, W. Zhang, G. Li, W. Wu, X. Tan, Y. Li, E. Ding, and L. Lin, "Weakly-supervised spatio-temporal anomaly detection in surveillance video," in *IJCAI*, 2021, pp. 1172–1178.

[14] 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 *ICCV*, 2021, pp. 4955–4966.

[15] C. Huang, Y. Shi, J. Wen, W. Wang, Y. Xu, and X. Cao, "Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models," in *ICML*, 2025.

[16] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *NIPS*. The MIT Press, 1999, pp. 582–588.

[17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *CVPR*, 2016, pp. 733–742.

[18] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *ECCV (35)*, ser. Lecture Notes in Computer Science, vol. 13695. Springer, 2022, pp. 105–124.

[19] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *ECCV*, 2020, pp. 322–339.

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

[21] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *AAAI*. AAAI, 2023, pp. 3769–3777.

[22] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: clip-assisted temporal self-attention for weakly-supervised video anomaly detection," in *ICIP*. IEEE, 2023, pp. 3230–3234.

[23] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," *IEEE Trans. Image Process.*, vol. 33, pp. 4923–4936, 2024.

[24] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *CVPR*, 2023, pp. 8022–8031.

[25] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang, "Open-vocabulary video anomaly detection," *CoRR*, vol. abs/2311.07042, 2023.

[26] OpenAI, "Chatgpt," https://chat.openai.com/, 2026, accessed: 2026-01-25.