VAPrompt: Variational Prompting for Weakly Supervised Video Anomaly Detection

1st Wenbo Pang
Jiangnan University
Wuxi, China
6223112037@stu.jiangnan.edu.cn

2rd Tao Zhang*
Jiangnan Univ. & Central South Univ.
Wuxi & Changsha, China
taozhang@csu.edu.cn

3nd Gaoe Qin
Jiangsu Huaying Intelligent Technology Co
Wuxi, China
qingaoe@hwaintek.com

***Abstract*—Weakly supervised video anomaly detection (WSVAD) aims to identify abnormal events in long, untrimmed videos using only video-level labels, which is both practical and challenging. Existing methods often suffer from semantic ambiguity, caused by the diverse and context-dependent nature of anomalies, and attention dispersion, where models fail to focus on subtle anomalous cues. To address these issues, we propose VAPrompt, a semantic distribution-aware framework that explicitly models anomaly uncertainty and enhances spatiotemporal localization. The core component, Variational Text Prompt Refinement (VTPR), represents textual prompts as stochastic latent distributions rather than fixed embeddings, enabling the model to capture diverse anomaly semantics and improve generalization. In addition, a Prompt-Infused Spatiotemporal Saliency Aggregation (PISSA) module integrates cross-modal attention to highlight anomaly-relevant regions while suppressing background noise. Furthermore, a Multi-Anchor Semantic Alignment Objective (MASAO) is introduced to enforce fine-grained semantic separation under weak supervision. Extensive experiments on the UCF-Crime and XD-Violence benchmarks demonstrate that VAPrompt consistently outperforms state-of-the-art methods. Specifically, our approach achieves 89.01% AUC and 72.32% AnoAUC on UCF-Crime, and 86.16% AP on XD-Violence, along with substantial improvements in temporal localization accuracy. These results indicate that modeling semantic uncertainty and integrating prompt-guided saliency are effective for robust weakly supervised video anomaly detection, making VAPrompt a practical and extensible solution for complex real-world surveillance scenarios.**

***Index Terms*—video anomaly detection, prompt learning, variational inference**

I. Introduction

Video Anomaly Detection (VAD) [1]–[4] is a critical task in video surveillance that aims to identify abnormal events such as trespassing or violent altercations in long, untrimmed videos. The immense cost of annotating precise temporal boundaries makes fully supervised methods impractical. Consequently, the weakly supervised setting—relying on a single binary label per video—has become the dominant paradigm, typically addressed through Multiple Instance Learning (MIL) [5]–[7] formulations.

Despite recent progress [8]–[12], existing weakly supervised approaches face two significant hurdles: **semantic ambiguity** and **attention dispersion**. Semantic ambiguity arises because anomalies are diverse and context-dependent, making it difficult for models like RTFM and UMIL to generalize. Furthermore, video-level supervision often leads to attention dispersion, where models focus on irrelevant background motions rather than subtle anomalous cues, resulting in poor temporal localization.

Recent advancements in pre-trained Vision-Language Models (VLMs) [13]–[15] have opened new avenues for VAD by aligning visual content with rich semantic concepts, naturally complementing context-aware anomaly understanding [16]–[20]. However, prior prompt-based methods such as VadCLIP and CLIP-TSA rely on fixed or lightly perturbed textual embeddings, effectively treating prompts as single-point estimates. Such deterministic modeling overlooks the intrinsic distribution of anomaly semantics, restricting their capacity to capture fine-grained and diverse spatiotemporal patterns.

To overcome these limitations, we propose **VAPrompt**, a semantic distribution-aware framework for weakly supervised VAD. It integrates three modules that jointly enhance semantic diversity, spatiotemporal localization, and inter-class discrimination. First, the *Variational Text Prompt Refinement (VTPR)* module treats prompts as stochastic latent variables to capture the semantic uncertainty of anomalies. Second, the *Prompt-Infused Spatiotemporal Saliency Aggregation (PISSA)* module injects cross-modal attention and spatial clustering to focus on anomaly-relevant regions. Finally, the *Multi-Anchor Semantic Alignment Objective (MASAO)* employs multiple textual anchors to enforce fine-grained semantic separation. Given the growing interest in neural network–based multimodal learning and weak supervision, this work provides a practical and extensible solution for uncertainty-aware representation learning in complex video understanding tasks.

II. Related Work

A. Weakly Supervised Video Anomaly Detection

In weakly supervised VAD, models are trained with video-level labels but lack temporal annotations. Early approaches often relied on reconstruction-based objectives [21], [22] or self-supervised learning [23], [24] to model normal patterns. More recent works formulate the task as MIL, treating a video as a bag of clips and predicting whether any instance is anomalous. Methods such as RTFM [8] and UMIL [9] follow this paradigm, while others like DMU [25] employ

This research is partly supported by National Science and Technology Major Project (2025ZD1009908).

*Corresponding author.

memory modules to represent normal and abnormal events. Although effective, these methods often depend on limited feature representations and suffer from attention dispersion, where models spread focus across irrelevant regions instead of salient anomalies.

B. Vision-Language Models for Anomaly Detection

Leveraging pre-trained Vision-Language Models (VLMs) such as CLIP has emerged as a promising direction. By aligning visual and textual modalities, these models capture context-dependent semantics beyond traditional features. VadCLIP [16] adapts a frozen CLIP model with a dual-branch design, while CLIP-TSA [19] incorporates temporal self-attention to model dependencies. Such methods outperform classical approaches but remain constrained by static or simple prompts that inadequately capture anomaly diversity.

C. Prompt Engineering and Contrastive Learning

To overcome these limitations, recent works focus on prompt learning and advanced objectives. STPrompt [26] introduces spatio-temporal prompt embeddings to guide attention, and TPWNG [27] employs learnable prompts with normality guidance to improve pseudo-labels. While promising, these methods often treat prompts as fixed vectors or apply simple noise, failing to fully model the uncertainty and variability of real-world anomalies.

III. Our Method

A. Problem Definition

Video Anomaly Detection (VAD) aims to identify abnormal events—such as trespassing, traffic violations, or violent altercations—in long, untrimmed surveillance videos. Given a video $V = \{F_1, \ldots, F_T\}$ with T frames, the goal is to estimate an anomaly score $s_t \in [0, 1]$ for each frame F_t. In the weakly supervised setting, only a video-level label $y \in \{0, 1\}$ is available, without temporal annotations. This limited supervision introduces two key challenges: (1) *semantic ambiguity*, as anomalies are diverse and context-dependent; and (2) *attention dispersion*, where models may focus on irrelevant motion or background regions.

B. Framework Overview

We propose **VAPrompt**, a *Semantic Distribution-Aware Weakly Supervised VAD* framework that improves semantic diversity, spatiotemporal localization, and inter-class discrimination (Fig. 1). Its three modules form a unified design: VTPR models prompts as latent distributions to capture anomaly variability, PISSA injects cross-modal saliency to sharpen spatial focus, and MASAO applies multi-anchor contrastive learning to enforce fine-grained category separation.

C. Variational Text Prompt Refinement (VTPR)

Conventional prompt-based methods employ fixed textual embeddings, which inadequately capture the diverse manifestations of anomalies. To address this limitation, VTPR models each prompt embedding as a *distribution* rather than a single vector. Intuitively, instead of forcing a single deterministic prompt to represent all anomaly semantics, we allow the prompt to explore a local semantic neighborhood. For a given textual prompt t, we learn a mean vector μ and variance σ^2, and sample a latent embedding z via the reparameterization trick:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\odot$ denotes element-wise multiplication. This variational formulation encourages the model to explore semantically plausible variations of the prompt, thereby improving generalization to unseen anomaly types.

To maintain semantic consistency between the sampled textual embedding and the associated visual context, we introduce a prompt–visual alignment loss:

$$\mathcal{L}_{\text{PRI}} = 1 - \cos(z, v), \tag{2}$$

where v is the corresponding frame-level visual feature. This regularization constrains z to remain faithful to visual evidence while still modeling semantic uncertainty.

D. Prompt-Infused Spatiotemporal Saliency Aggregation (PISSA)

Weak supervision often causes the network to attend to irrelevant motion patterns. PISSA mitigates this issue by explicitly injecting semantic guidance from textual prompts into the saliency estimation process.

Let $\{t_k\}_{k=1}^{K}$ denote the set of class-specific textual embeddings produced by VTPR, and $\{v_t\}$ the frame-level visual features. Cross-modal attention scores are computed as

$$A_{t,k} = \text{softmax}\left(\frac{v_t t_k^\top}{\sqrt{d}}\right), \tag{3}$$

yielding prompt-conditioned saliency maps. These maps are temporally aggregated using saliency-weighted pooling, and a spatial clustering mechanism groups regions with coherent semantic responses. The aggregated representation is projected back to the feature space and fused with the original visual features, forming a feedback loop that highlights anomaly-relevant regions and suppresses background noise.

E. Multi-Anchor Semantic Alignment Objective (MASAO)

Binary classification losses are insufficient for distinguishing multiple anomaly categories under weak supervision. MASAO introduces a contrastive learning objective with multiple textual anchors to enforce finer-grained semantic separation.

Let $\bar{v}$ be the global video representation obtained by average pooling over frame features, and $\{t_k\}_{k=1}^{K}$ the set of textual anchors (e.g., "normal activity", "fighting", "intrusion"). The loss is defined as

$$\mathcal{L}_{\text{MASAO}} = -\frac{1}{K}\sum_{k=1}^{K} \log \frac{\exp(\cos(\bar{v}, t_k)/\tau)}{\sum_{j=1}^{K} \exp(\cos(\bar{v}, t_j)/\tau)}, \tag{4}$$

where τ is a learnable temperature. This objective pulls the video representation closer to its ground-truth anchor while

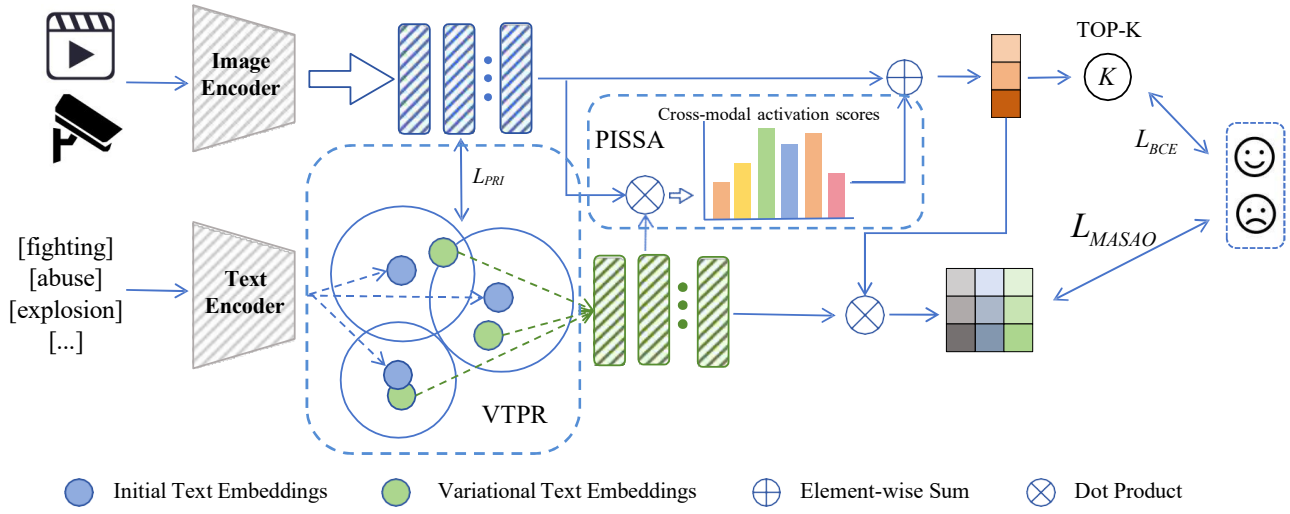

Fig. 1. Overall pipeline of the proposed **VAPrompt** framework. It integrates variational prompt learning, prompt-guided saliency aggregation, and multi-anchor semantic alignment for robust weakly supervised video anomaly detection. VTPR improves semantic diversity, PISSA enhances localization, and MASAO sharpens decision boundaries.

pushing it away from mismatched anchors, thereby refining decision boundaries and promoting category-aware anomaly detection.

F. Training Objective

The overall loss combines binary classification, semantic alignment, and prompt–visual consistency:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{MASAO}} + \lambda_3 \mathcal{L}_{\text{PRI}}. \tag{5}$$

Here $\mathcal{L}_{\text{BCE}}$ follows prior weakly supervised VAD work [16], [28], while $\lambda_{1,2,3}$ balance the terms.

IV. Experiments

A. Datasets and Evaluation Metrics

We evaluate our framework on two widely adopted weakly supervised video anomaly detection (WSVAD) benchmarks: **UCF-Crime** [29] and **XD-Violence** [28]. UCF-Crime contains 1,900 untrimmed surveillance videos across 13 anomaly categories, and XD-Violence comprises 4,754 multimodal videos spanning six anomaly types.

For coarse-grained anomaly detection, we report **frame-level AUC** and anomaly-specific AUC (**AnoAUC**) on UCF-Crime, and **Average Precision (AP)** on XD-Violence. For fine-grained temporal localization, we adopt mean Average Precision (**mAP**) at IoU thresholds from 0.1 to 0.5 and also report the averaged score (**AVG**). Following standard practice, mAP is computed only on abnormal test videos.

B. Implementation Details

Our model is implemented on top of the CLIP ViT-B/16 backbone with frozen vision and text encoders. The three proposed components—Variational Text Prompt Refinement, Prompt-Infused Spatiotemporal Saliency Aggregation, and Multi-Anchor Semantic Alignment Objective—are newly introduced and trained from scratch. We adopt AdamW optimizer with batch size 64 and a learning rate of 2×10^{-5} for XD-Violence and 1×10^{-5} for UCF-Crime. Models are trained for 20 and 10 epochs on XD-Violence and UCF-Crime, respectively, using a single NVIDIA RTX 3090 GPU.

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

We compare our approach against state-of-the-art weakly supervised and vision-language-based anomaly detection methods.

Coarse-grained results. Table I shows that our method achieves new state-of-the-art results on both benchmarks. On XD-Violence, our model reaches **86.16%** AP, while on UCF-Crime, we obtain **89.01%** AUC and **72.32%** AnoAUC. Compared with other method, our method yields consistent improvements, underscoring the gains from variational prompt refinement and saliency-guided aggregation.

Fine-grained results. Tables II and III summarize temporal localization performance. On XD-Violence, our model achieves **27.92%** AVG mAP, substantially exceeding all existing methods. On UCF-Crime, we obtain **9.86%** AVG mAP, consistently outperforming prior works. The improvements are especially pronounced at higher IoU thresholds, indicating enhanced boundary sensitivity and robustness to temporal ambiguity.

D. Ablation Study

We conduct ablation studies on UCF-Crime and XD-Violence to analyze the contribution of each component and the design choices of VTPR.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON XD-VIOLENCE AND UCF-CRIME (COARSE-GRAINED). THE BEST RESULTS ARE IN **BOLD** AND THE SECOND BEST ARE UNDERLINED.

Method	XD-Violence	UCF-Crime	
	AP	AUC	AnoAUC
RTFM (ICCV'21)	78.27	85.66	63.86
DMU (AAAI'23)	82.41	86.75	68.62
UMIL (CVPR'23)	81.90	86.75	68.68
CLIP-TSA (ICIP'23)	82.17	87.58	–
TPWNG (CVPR'24)	83.69	87.79	69.84
STPrompt (MM'24)	84.43	88.08	70.21
VadCLIP (AAAI'24)	84.51	88.02	70.23
Fed-WSVAD (AAAI'25)	84.39	88.32	70.21
Huang et al. (TMM'25)	85.61	87.98	70.14
Ours	**86.16**	**89.01**	**72.32**

TABLE II
FINE-GRAINED TEMPORAL LOCALIZATION RESULTS ON XD-VIOLENCE (MAP@IOU %).

Method	0.1	0.2	0.3	0.4	0.5	AVG
Baseline	1.82	0.92	0.48	0.23	0.09	0.71
Deep-MIL	22.72	15.57	9.98	6.20	3.78	11.65
HL-Net	30.51	25.75	20.18	14.83	9.79	20.21
VadCLIP	37.03	30.84	23.38	17.90	14.31	24.70
Ours	**39.43**	**33.54**	**27.59**	**20.91**	**18.16**	**27.92**

Effect of major components. Table IV summarizes the results of progressively adding our proposed modules. Starting from a **baseline trained with binary cross-entropy (BCE) only**, adding VTPR improves both XD-Violence AP and UCF-Crime AUC by enriching prompt diversity. Incorporating PISSA further enhances spatiotemporal localization, reflected in higher XD-Violence average mAP and UCF-Crime mAP. Finally, MASAO sharpens category boundaries, yielding consistent gains across all metrics. The full model achieves the best performance, demonstrating the complementary benefits of all three components.

Effect of VTPR design choices. Table V compares different VTPR variants. Using fixed textual prompts results in the weakest performance, as the model cannot capture semantic uncertainty. Gaussian noise injection improves robustness but remains inferior to the full stochastic formulation. Removing the consistency loss destabilizes prompt learning and reduces performance. The full VTPR consistently achieves the best results, confirming the importance of both variational modeling and visual-consistency regularization.

Qualitative analysis. Fig. 2 presents frame-level anomaly scores on several video examples from UCF-Crime and XD-Violence. The highlighted regions correspond to ground-truth anomalous segments. Compared with the baseline, our full model clearly assigns higher scores to true anomaly frames while suppressing background noise. These qualitative results corroborate the quantitative improvements observed in Tables IV and V, demonstrating the effectiveness of VTPR in enriching semantic prompts and PISSA in focusing attention on relevant spatiotemporal regions.

TABLE III
FINE-GRAINED TEMPORAL LOCALIZATION RESULTS ON UCF-CRIME (MAP@IOU %).

Method	0.1	0.2	0.3	0.4	0.5	AVG
Baseline	0.21	0.14	0.04	0.02	0.01	0.08
Deep-MIL	5.73	4.41	2.69	1.93	1.44	3.24
HL-Net	10.27	7.01	6.25	3.42	3.29	6.05
VadCLIP	11.72	7.83	6.40	4.53	2.93	6.68
Ours	**14.33**	**10.46**	**9.33**	**7.89**	**7.31**	**9.86**

TABLE IV
ABLATION ON THE MAJOR COMPONENTS OF OUR FRAMEWORK. "V" = VTPR, "P" = PISSA, "M" = MASAO.

Modules			XD-Violence		UCF-Crime	
V	P	M	AP (%)	AVG mAP (%)	AUC (%)	AVG mAP (%)
			82.13	0.71	86.47	0.02
✓			83.95	15.57	87.22	3.31
✓	✓		85.17	22.36	88.36	5.73
✓	✓	✓	**86.16**	**27.92**	**89.01**	**9.86**

E. Computational Complexity Analysis

To assess the practical efficiency of our method, we compare the computational cost with representative baselines in Table VI. We report the number of trainable parameters (in millions), floating-point operations per video (GFLOPs), and average inference time per video on a single NVIDIA RTX 3090 GPU. All CLIP-based methods use frozen ViT-B/16 encoders, with only task-specific modules counted as trainable parameters.

As shown in Table VI, VAPrompt introduces additional computational cost primarily from two components: the variational encoder in VTPR for modeling semantic distributions, the cross-modal attention module in PISSA for saliency aggregation. Despite having 8.62M trainable parameters and requiring 32.81 GFLOPs per video, our method maintains practical inference speed at 452ms per video. Notably, on XD-Violence, VAPrompt achieves 86.16% AP and 27.92% AVG mAP, outperforming VadCLIP by 1.65% and 3.22% respectively, demonstrating that modeling semantic uncertainty is both effective and efficient for real-world surveillance deployment.

V. CONCLUSION

We presented VAPrompt, a variational prompting framework addressing key challenges in weakly supervised video anomaly detection. By jointly modeling semantic uncertainty through VTPR, enhancing anomaly localization with PISSA, and enforcing fine-grained discrimination via MASAO, our method achieves consistent improvements over existing approaches. Experiments on UCF-Crime and XD-Violence show that VAPrompt not only delivers state-of-the-art performance in anomaly detection accuracy but also substantially improves

Fig. 2. Qualitative results of coarse-grained WSVAD.

TABLE V
ABLATION ON THE DESIGN CHOICES OF VTPR.

Variant	XD-Violence	UCF-Crime	
	AP (%)	AUC (%)	AVG mAP (%)
Fixed textual prompts	83.02	86.78	6.79
Gaussian noise prompts	83.61	87.05	7.12
w/o Consistency Loss	83.34	87.11	7.04
Full VTPR	**83.95**	**87.22**	**7.31**

TABLE VI
COMPUTATIONAL COMPLEXITY COMPARISON ON A 300-FRAME VIDEO. PARAMS REFER TO TRAINABLE PARAMETERS ONLY (CLIP ENCODERS ARE FROZEN). TIME IS MEASURED ON NVIDIA RTX 3090 GPU.

Method	Params (M)	FLOPs (G)	Time (ms)
RTFM	2.15	8.73	142
VadCLIP	7.42	28.65	401
CLIP-TSA	8.19	31.47	438
Ours	8.62	32.81	452

temporal localization precision, particularly under challenging temporal conditions. Beyond these empirical results, our study highlights the importance of treating prompts as semantic distributions and integrating cross-modal saliency for robust anomaly understanding. However, VAPrompt may struggle in extremely low-data regimes where variational sampling introduces noise, and when anomalies lack clear semantic descriptions that align with predefined textual anchors. Future work may explore extending this framework to open-set anomaly detection, incorporating multimodal inputs beyond RGB videos, and investigating adaptive prompt sampling strategies for real-time applications.

REFERENCES

[1] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, "Video anomaly detection in 10 years: A survey and outlook," *Neural Computing and Applications*, vol. 37, no. 32, pp. 26 321–26 364, 2025.

[2] Y. Liu, S. Liu, X. Zhu, H. Yang, J. Li, J. Guo, L. Teng, D. Yang, Y. Wang, and J. Liu, "Privacy-preserving video anomaly detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 37, no. 1, pp. 2–21, 2026.

[3] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 583–591.

[4] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 592–14 601.

[5] B. Ezard, L. Li, and S. An, "Multiple instance learning for visual grain quality analysis without instance-level annotation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5352–5359.

[6] B. Zhao, S. Kim, H. Chen, C. Zhou, Z.-h. Gao, G. Wang, and X. Li, "Ptcmil: Multiple instance learning via prompt token clustering for whole slide image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 502–512.

[7] D. Zhang, W. Fang, Y. Liu, Z. Lyu, C. Xiong, and Z. Wang, "Two-stage video anomaly detection based on dual-stream networks and multi-instance learning," *IET Image Processing*, vol. 18, no. 14, pp. 4843–4851, 2024.

[8] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4975–4986.

[9] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8022–8031.

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

[11] 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," *arXiv preprint arXiv:2108.03825*, 2021.

[12] B. Wang, C. Huang, J. Wen, W. Wang, Y. Liu, and Y. Xu, "Federated weakly supervised video anomaly detection with multimodal prompt," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 017–21 025.

[13] H. Cheng, H. Ye, X. Zhou, X. Liu, F. Chen, and M. Wang, "Vision-language pre-training via modal interaction," *Pattern Recognition*, vol. 156, pp. 110 809–110 810, 2024.

[14] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[15] K. Chen, D. Shen, H. Zhong, H. Zhong, K. Xia, D. Xu, W. Yuan, Y. Hu, B. Wen, T. Zhang *et al.*, "Evlm: An efficient vision-language model for visual understanding," *arXiv preprint arXiv:2407.14177*, 2024.

[16] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6074–6082.

[17] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing large language models for training-free video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 527–18 536.

[18] Y. Jiang and L. Mao, "Vision-language models assisted unsupervised video anomaly detection," *arXiv preprint arXiv:2409.14109*, 2024.

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

[20] C. Huang, W. Huang, Q. Jiang, W. Wang, J. Wen, and B. Zhang, "Multimodal evidential learning for open-world weakly-supervised video anomaly detection," *IEEE Transactions on Multimedia*, vol. 27, pp. 3132–3143, 2025.

[21] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "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*, 2021, pp. 13 588–13 597.

[22] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, pp. 107 706–107 707, 2021.

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

[24] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, and D. Zhang, "Self-supervised attentive generative adversarial networks for video anomaly detection," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 9389–9403, 2022.

[25] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3769–3777.

[26] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9301–9310.

[27] Z. Yang, J. Liu, and P. Wu, "Text prompt with normality guidance for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 18 899–18 908.

[28] 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 *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 322–339.

[29] 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.