# PA-VAD: Diffusion-Based Pseudo-Only Video Anomaly Detection via Domain-Aligned Memory Updates

Satoshi Hashimoto*, Yanan Wang, Hitoshi Nishimura, and Mori Kurokawa

KDDI Research, Inc., Fujimino, Saitama, Japan
{st-hashimoto,wa-yanan,ht-nishimura,mo-kurokawa}@kddi.com

**Abstract.** Deploying video anomaly detection (VAD) in the real world is often constrained by the scarcity, privacy, and cost of collecting real abnormal footage. We propose PA-VAD, a novel pseudo-only framework that trains an anomaly detector without using any real abnormal videos, by pairing real normal videos with diffusion-synthesized pseudo-abnormal videos generated from a small set of real normal images. Beyond proposing a generation-driven training pipeline, we make a key empirical discovery: pseudo anomalies exhibit a characteristic spatiotemporal magnitude bias in feature space, which can dominate Multiple Instance Learning and degrade generalization if left unaddressed. To counter this pseudo-induced bias, we introduce the Domain-Aligned Regularized Module (DARM), which combines domain alignment with usage-aware memory updates to balance prototype coverage and stabilize optimization under biased pseudo supervision. Extensive experiments demonstrate that PA-VAD achieves 98.2% AUC on ShanghaiTech, 82.5% on UCF-Crime, and 95.1% on XD-Violence, and further improves generalization to unseen anomaly classes in open-set evaluations. Notably, PA-VAD surpasses the best real-abnormal WVAD baselines on ShanghaiTech and XD-Violence by +0.6% and +0.9% , respectively, and improves over the UVAD state of the art on UCF-Crime by +1.9% —showing that high-accuracy VAD is attainable without collecting real abnormal videos.

**Keywords:** Anomaly Detection· Diffusion Model· Vision Language Model

## 1 Introduction

Ensuring public safety and order has driven extensive research on video anomaly detection (VAD) in computer vision. VAD aims to identify abnormal events such as assaults, traffic accidents, and theft in surveillance videos. Prior works span two main families: (i) unsupervised VAD (UVAD), which trains only on normal data [7, 11, 12, 17, 19, 20, 29, 30, 34]; (ii) weakly supervised VAD (WVAD), which relies on normal and abnormal videos with video-level labels [6,8,13,18,22,24,26,

* Corresponding author

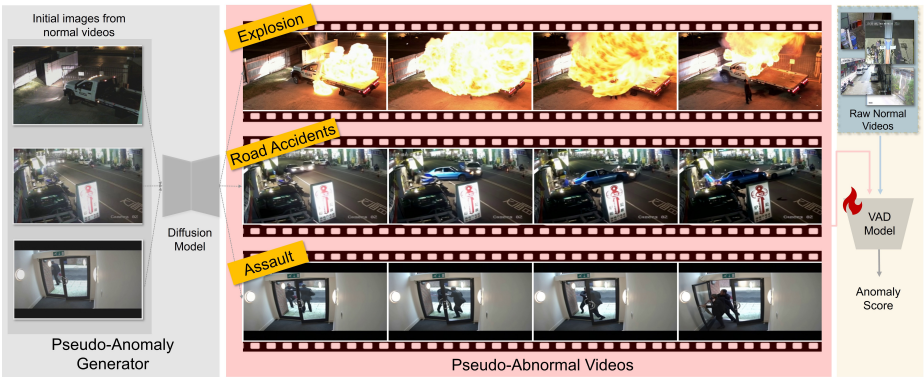



**Fig. 1:** Our PA-VAD framework generates class-aware pseudo-abnormal videos (e.g., Explosion, Road Accidents, Assault) from a small set of normal images. These controllable pseudo anomalies provide scalable, diverse supervision, removing the need for real abnormal data and broadening the range of learnable abnormal patterns.

28,31,32,35,36]. Despite rapid progress, deployment remains challenging. UVAD methods can be brittle under distribution shift and label noise, whereas WVAD approaches typically assume access to substantial real abnormal footage, which is costly and often constrained by safety, privacy, or rarity. A complementary direction synthesizes pseudo anomalies to ease data scarcity [3,16]. Yet, in most cases the synthesized clips supplement rather than replace real anomalies, and the synthesis is driven by rule-based edits or narrow generative priors, so the reliance on real abnormal data is reduced but not removed.

**Our approach.** We introduce **PA-VAD**, a generation-driven framework that learns an anomaly detector from pseudo-anomalous videos synthesized from a small set of real normal images. Training uses no real abnormal videos, while evaluation follows the standard benchmark splits and protocols. Unlike prior pseudo-anomaly works that mainly supplement real anomalies and rely on heuristic edits or narrow generators, our synthesis pipeline is designed to replace real anomalies: we select class-relevant seed images with CLIP and perform VLM-based prompt rewriting before driving a video diffusion model, which improves fidelity and scene consistency without manual prompt search.

When training solely on diffusion-synthesized pseudo anomalies without any real abnormal videos, a naïve detector can become biased; we discover that synthesis artifacts (e.g., exaggerated motions, physical implausibility, and temporal inconsistency) often yield abnormally large feature norms—a spatiotemporal magnitude bias—that skews learning toward magnitude cues. To counter this pseudo-induced bias, we introduce the Domain-Aligned Regularized Module (DARM), combining domain alignment between real Normal and pseudo-Normal streams with usage-aware memory updates to suppress magnitude-driven bias and stabilize learning for stronger anomaly discrimination.

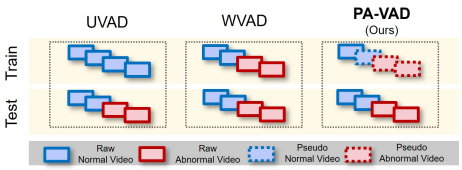



**Fig. 2:** Overview of our framework. PA-VAD trains on real *Normal* and diffusion-generated pseudo videos (no real *Abnormal*) and is evaluated on standard splits with real *Normal*/*Abnormal*.

We conduct extensive experiments on multiple benchmarks, showing that our approach outperforms state-of-the-art UVAD baselines and can surpass strong methods that rely on real abnormal videos, while substantially reducing data collection costs and improving open-set generalization to unseen anomalies.

- We propose a generation-driven VAD framework that trains without real abnormal videos while evaluating under standard weakly supervised splits, substantially reducing data collection costs.
- We propose the Class-Aware Pseudo-Anomaly Generator (CA-PAG), a video diffusion–based pseudo-anomaly generator that attains high-fidelity abnormal videos through CLIP-guided initial image selection and VLM-driven prompt refinement.
- We propose the Domain-Aligned Regularized Module (DARM), an adaptive memory module that mitigates the pseudo-induced large-magnitude bias via domain alignment and usage-aware updates, enabling stable MIL training with synthesized anomalies.
- Through comprehensive experiments on multiple datasets, we demonstrate state-of-the-art results against UVAD methods and show that our approach can surpass strong methods that depend on real abnormal videos.

## 2 Related Work

### 2.1 Video Anomaly Detection

We group prior work into two families—unsupervised (UVAD) and weakly supervised (WVAD) based methods—and summarize their techniques and limitations. In UVAD, classical approaches include dictionary learning on normal-only data [17] and prediction/reconstruction paradigms [11, 12, 19, 20, 29, 30, 34]. Recent methods that integrate prediction with reconstruction have proven effective; to explicitly address small-scale anomalies and background interference, MGSTRL learns multi-grained spatio-temporal representations via three proxy tasks—continuity judgment, discontinuity localization, and contrastive missing-frame estimation in feature space—achieving state-of-the-art results, especially on small-scale anomalies [34]. A persistent deployment issue for UVAD is the assumption that training covers all normal variations, which rarely holds in the wild. When previously unseen normal patterns arise due to camera/view changes or noise, UVAD methods tend to suffer from false positives.

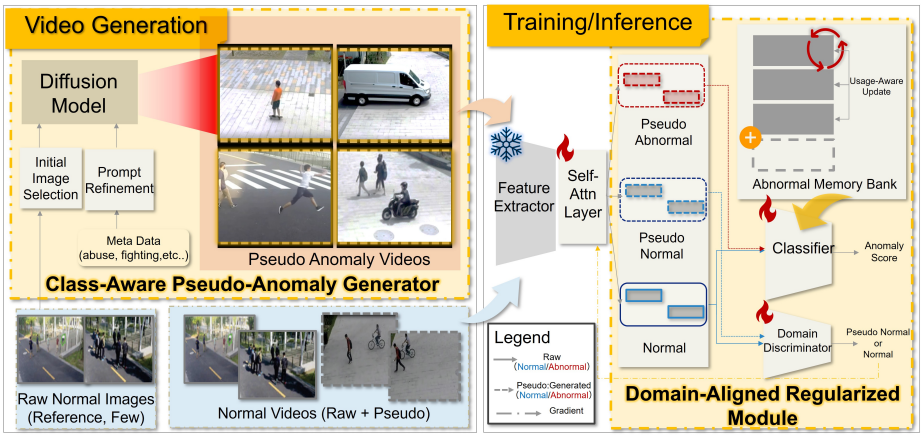



**Fig. 3:** Overview of our framework. Starting from a small set of real normal images and class texts, we synthesize pseudo-abnormal videos via an image-to-video diffusion process, then train a classifier on real normal and synthesized pseudo-abnormal videos. A Domain-Aligned Regularized Module—designed to account for the characteristic large spatiotemporal magnitude of synthesized anomalies—mitigates Multiple Instance Learning (MIL) bias and enables accurate detection.

In contrast, WVAD has become popular as a practical discriminative framework trainable with weak video-level labels. A large body of work explores architectures using video-level supervision to infer frame-level anomalies [5, 6, 13, 18, 22, 25, 26, 31, 35, 36]. Sultani *et al.* [18] introduced a MIL-based training scheme under weak labels. Each video is split into segments and represented as a bag of instances: a normal bag $\tilde{B}_n = \{\mathbf{f}_n^i\}_{i=1}^m$ and an abnormal bag $\tilde{B}_a = \{\mathbf{f}_a^i\}_{i=1}^l$, where $\mathbf{f} \in \mathbb{R}^K$ denotes a $K$-dimensional feature vector and $m, l$ are the numbers of segments. A detector $\mathcal{D}$ outputs an anomaly score $s = \mathcal{D}(\mathbf{f})$ for each segment. Under the MIL assumption, the detector is optimized so that the maximum score in abnormal bags exceeds that in normal bags, using ranking loss:

$$\mathcal{L}_{\text{MIL}} = \max\left(0,\ 1 - \max_{i \in \tilde{B}_a} s_a^i + \max_{i \in \tilde{B}_n} s_n^i\right). \tag{1}$$

Unlike UVAD, such weak supervision learns discriminative representations and tends to be more robust to multi-scene settings and complex backgrounds. WVAD continues to advance, with studies such as Chen *et al.* [5] extending it by leveraging single-frame supervision. Recently, WVAD methods have adopted vision–language models (VLMs) for better interpretability. Du *et al.* [8] target causal understanding of abnormal events using hard/soft prompts. Zhang *et al.* [32] fine-tune VLM on large captioned datasets and report leading performance. However, these WVAD approaches typically presuppose large-scale collection of real abnormal videos, which can be costly and constrained by safety and privacy, limiting practical deployment.

### 2.2 Pseudo Video Generation

Several works have explored generating pseudo data to boost VAD. Cai *et al.* [3] integrate video generation into weakly supervised training by synthesizing text-conditioned pseudo anomalies mixed with real abnormal videos. This alleviates but does not eliminate reliance on real anomalies, as full real sets remain used during training. Rai *et al.* [16] generate pseudo anomalies for UVAD by inpainting frames with a pretrained diffusion model and perturbing optical flow, but their design still limits the diversity and realism of synthesized anomalies.

Recent advances in diffusion models have produced powerful video generators. Two major conditions are common: text-to-video (T2V), which directly synthesizes videos from text prompts, and image-to-video (I2V), which extends a static image over time while preserving spatial structure under a prompt. Wan 2.2 is a diffusion-Transformer framework with noise-stage mixture of experts and highly compressed latent representations [21]. Stable Video Diffusion establishes a three-stage curriculum spanning text–image pretraining, video pretraining, and high-quality video finetuning [2]. While these models are strong general-purpose generators, they are not tailored for anomaly synthesis; prompting to obtain rare abnormal behaviors is challenging and typically requires careful search and conditioning.

## 3 Proposed Method

We propose **PA-VAD**, a pseudo-only framework that trains a video anomaly detector *without any real abnormal videos*. As illustrated in Fig. 2, training uses real *Normal* videos together with diffusion-generated pseudo-abnormal videos, while evaluation follows the standard splits with real *Normal*/*Abnormal*. An overview is provided in Fig. 3. PA-VAD comprises two components: (i) *Class-Aware Pseudo-Anomaly Generator* and (ii) *Domain-Aligned Regularized Module (DARM)*. We detail each module below.

### 3.1 Class-Aware Pseudo-Anomaly Generator

The goal is to synthesize pseudo-abnormal videos from only a small set of real normal images and class metadata (abnormal class names) by driving a video diffusion model. We adopt an image-to-video (I2V) inference process conditioned on (a) an initial image and (b) a textual prompt. A central challenge is the covariate shift between synthesized anomalies and real data: prior synthesis for VAD either mixes generated clips with real abnormal videos [3] or perturbs real images/videos [16], which keeps the domain gap relatively small. In contrast, because our aim is to eliminate reliance on real abnormal footage altogether, *reducing this gap while preserving quality* is crucial. We therefore introduce the Class-Aware Pseudo-Anomaly Generator (CA-PAG), whose synthesis is controlled by two levers: (1) **initial image selection** to anchor the generation near the target class domain, and (2) **prompt refinement** to tailor textual conditioning to the chosen initial image via zero-shot vision–language reasoning.

**Initial Image Selection.** Naively sampling initial images at random from the normal set risks large domain gaps for certain classes (e.g., selecting an indoor home scene for the target class "Road Accident" in UCF-Crime [18]). Instead, we propose to select *class-relevant* inits with CLIP [15] using a Top-$K$ strategy in the joint vision–text space (Fig. 4 (a)).

Concretely, given a class label $x_c$, we augment it with positive phrases $g_{\text{pos}}$ to enforce surveillance-style attributes such as camera footage and CCTV, and introduce a set of negative phrases $g_{\text{neg}}$ to suppress nuisances such as black screens or channel logos. Let $\Phi$ and $\psi$ denote the CLIP image and text encoders. Please refer to the supplementary material for the detailed description of each phase. Define unit-normalized embeddings $\hat{\mathbf{v}}(\mathbf{I}) = \Phi(\mathbf{I})/|\Phi(\mathbf{I})|_2$ and $\hat{\mathbf{t}}(u) = \psi(u)/|\psi(u)|_2$. We form a single positive text by concatenating the class name with positives,$t_c = \text{concat}(x_c,\, g_{\text{pos}})$ and compute the positive similarity:

$$s_{\text{pos}}(\mathbf{I}, c) = \big\langle \hat{\mathbf{v}}(\mathbf{I}),\, \hat{\mathbf{t}}(t_c) \big\rangle. \tag{2}$$

We compute each negative similarity:

$$s_{\text{neg}}(\mathbf{I}) = \big\langle \hat{\mathbf{v}}(\mathbf{I}),\, \hat{\mathbf{t}}(g_{\text{neg}}) \big\rangle. \tag{3}$$

The final score balances class affinity and nuisance suppression,

$$\tilde{s}(\mathbf{I}, c) = s_{\text{pos}}(\mathbf{I}, c) - \lambda\, s_{\text{neg}}(\mathbf{I}), \qquad \lambda \in [0, 1], \tag{4}$$

and select the initial set $\text{TopK}_c$ for class $c$ using a scene-balanced ranking scheme. We apply scene balancing to prevent seed selection from being dominated by populous camera/location scenes. Specifically, with $s$ denoting a camera/location scene, we subsample the normal-image pool with per-scene budgets proportional to $\text{count}(s)^{\alpha}$ and prioritize the highest-scoring candidate from each scene before filling the remaining Top-$K$ slots by the global ranking $\tilde{s}(\mathbf{I}, c)$.

**Prompt Refinement for Pseudo Anomalies** Using a raw class name as a prompt is simple but often ambiguous, which harms generation consistency and text–video alignment [10,14]. Exhaustive manual search, however, is impractical. We therefore refine prompts with a vision–language model (VLM) by extracting objects and scene cues from each initial image and turning them into concise, class-consistent abnormal descriptions (Fig. 4 (b)).

Given initial images $\{\mathbf{I}_c^{(k)}\}_{k=1}^{K}$ for class $c$, a class text $x_c$, and a refinement instruction $g_{\text{vlm}}$, the VLM produces short candidate phrases:

$$\hat{s}_c^{(k)} \;=\; \mathcal{G}_{\varphi}\big(\mathbf{I}_c^{(k)},\, g_{\text{vlm}},\, x_c\big). \tag{5}$$

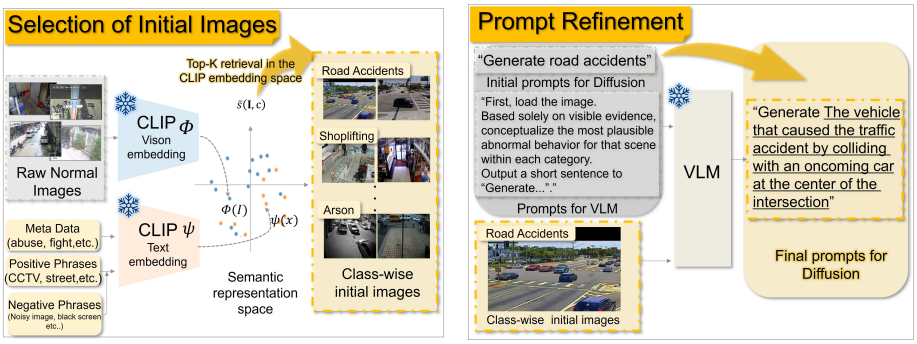
To standardize camera and rendering attributes and suppress artifacts, each phrase is concatenated with a set of template prompts $g_{\text{tem}}$ (e.g., *natural movement, fixed camera*):

$$p_c^{(k)} \;=\; \text{concat}\big(\hat{s}_c^{(k)},\, g_{\text{tem}}\big). \tag{6}$$

Finally, the video diffusion model $\mathcal{G}_\theta$ is conditioned on the refined prompt and the corresponding initial image to synthesize a pseudo-abnormal clip:

$$\tilde{\mathbf{V}}_c^{(k)} \;=\; \mathcal{G}_{\theta}\big(p_c^{(k)},\, \mathbf{I}_c^{(k)}\big). \tag{7}$$

Please see the supplementary material for details of the refinement instruction.




**(a)** Initial image selection in the vision–text space. We score normal images using a class text with positive phrases and subtract a negative similarity, taking the Top-$K$ per class.

**(b)** Prompt refinement. A VLM converts initial-image cues into concise, class-consistent abnormal descriptions, concatenating a template before driving the diffusion model.

**Fig. 4:** Overview of our class-aware pseudo-anomaly generation pipeline.

### 3.2 Domain-Aligned Regularized Module

This module trains and infers an anomaly detector using synthesized pseudo-abnormal videos $\tilde{\mathbf{V}}$ and real normal videos $\mathbf{V}$. When we simply plug synthesized videos into existing weakly supervised pipelines, we observe a characteristic failure mode: generated clips tend to exhibit excessive motion and viewpoint changes, inflating feature norms and biasing MIL Top-$k$ optimization toward a few high-magnitude pseudo instances. This bias induces a covariate shift at test time against real anomalies and degrades sensitivity. Table 1 reports the mean I3D feature norms on UCF-Crime for real vs. pseudo anomalies.

To counter this and exploit the diversity of synthesized data, we introduce the Domain-Aligned Regularized Module (DARM) with two components: (i) **Domain alignment:** a DANN-based objective [9] that aligns real and pseudo normal distributions to reduce covariate shift; (ii) **Usage-aware memory update:** an update rule that pulls under-utilized slots toward their responsibility-weighted centers to rebalance coverage across abnormal modes. The domain alignment term is applied to mean-pooled features of real normal and pseudo-normal streams. By encouraging domain-invariant representations via gradient reversal, it reduces magnitude shifts between real and pseudo data, preventing pseudo anomalies from dominating MIL Top-$k$ due to inflated norms. The usage-aware memory update balances the abnormal prototypes. It computes per-slot usage from soft assignments and updates under-utilized slots toward their responsibility-weighted centers, mitigating the tendency of a few high-norm slots to monopolize MIL Top-$k$ selection. We follow the UR-DMU [36] classifier, which provides a GL-MHSA–based encoder, dual (normal/abnormal) memory banks for prototype separation, and a MIL scoring head with uncertainty control. Adding the DARM objectives on top of this backbone effectively suppresses the MIL bias caused by excessive spatiotemporal magnitude in synthesized anomalies while preserving representation diversity across abnormal prototypes.

**Table 1:** Mean $\ell_2$ norms of frozen backbone features extracted from real and pseudo anomalies on UCF-Crime. Feature spaces differ across backbones.

| | I3D [4] | Qwen [1] | C3D [23] |
|---|---|---|---|
| Real | 20.52 | 24.22 | 19.14 |
| Pseudo | 23.03 | 25.55 | 20.41 |

**Formulation.** Given a video, a feature extractor $\Phi_{\text{vid}}$ produces segment features $\mathbf{f} = \Phi_{\text{vid}}(\mathbf{V}) \in \mathbb{R}^{T\times D}$. We embed them with temporal convolution and self-attention to obtain $\tilde{\mathbf{f}} \in \mathbb{R}^{B\times T\times d}$:

$$\tilde{\mathbf{f}} = \text{SelfAttn}\big(\text{Temporal}(\mathbf{f})\big). \tag{8}$$

Let $\mathbf{M}_A, \mathbf{M}_N \in \mathbb{R}^{K\times d}$ denote abnormal and normal memory banks. Their memory guided augmentations are $\mathbf{h}_A(\tilde{\mathbf{f}})$ and $\mathbf{h}_N(\tilde{\mathbf{f}})$. We concatenate these with $\tilde{\mathbf{f}}$ and predict framewise anomaly scores $\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} = \sigma\Big(\mathbf{W}\,\big[\,\tilde{\mathbf{f}}\,;\,\mathbf{h}_A(\tilde{\mathbf{f}}) + \mathbf{h}_N(\tilde{\mathbf{f}})\,\big] + \mathbf{b}\Big). \tag{9}$$

Let $\mathcal{S}_{\text{abn}}$ be indices of abnormal embeddings and $\tilde{\mathbf{f}}_{\text{abn}} = \tilde{\mathbf{f}}[\mathcal{S}_{\text{abn}}]$. Unfolding over time yields $\mathbf{Z} \in \mathbb{R}^{(B_a\,T)\times d}$. Define row-wise $\ell_2$ normalization by $\mathbf{D}_Z = \text{diag}(\|\mathbf{Z}_1\|_2, \ldots, \|\mathbf{Z}_{B_a T}\|_2)$ and $\bar{\mathbf{Z}} = \mathbf{D}_Z^{-1}\mathbf{Z}$. Similarly, define

$$\mathbf{D}_M = \text{diag}\big(\|\mathbf{m}_1\|_2, \ldots, \|\mathbf{m}_K\|_2\big), \tag{10}$$

and let $\bar{\mathbf{M}}_A = \mathbf{D}_M^{-1}\mathbf{M}_A$.

With temperature $\tau > 0$, we define the assignment matrix and slot usage as:

$$\mathbf{Q} = \text{softmax}\left(\frac{\bar{\mathbf{Z}}\bar{\mathbf{M}}_A^\top}{\tau}\right) \in \mathbb{R}^{(B_a\,T)\times K}. \tag{11}$$

$$\mathbf{u} = \frac{1}{B_a\,T}\,\mathbf{1}^\top\mathbf{Q} \in \mathbb{R}^K. \tag{12}$$

*(i) Domain alignment* We align real Normal and pseudo-Normal distributions through adversarial training. Let $\bar{\mathbf{f}}_N$ and $\bar{\mathbf{f}}_{\tilde{N}}$ denote the mean features of Normal and pseudo-Normal streams, and $D(\cdot)$ the domain discriminator with gradient reversal layer $G_{\lambda_{\text{da}}}(\cdot)$. The domain alignment loss is

$$\begin{aligned}\mathcal{L}_{\text{DA}} = {} & \text{BCE}\big(D(G_{\lambda_{\text{da}}}(\bar{\mathbf{f}}_N)),\, y_N\big) \\ & + \text{BCE}\big(D(G_{\lambda_{\text{da}}}(\bar{\mathbf{f}}_{\tilde{N}})),\, y_{\tilde{N}}\big) \\ & + \lambda_{\text{dist}}\big\|\bar{\mathbf{f}}_N - \bar{\mathbf{f}}_{\tilde{N}}\big\|_2^2,\end{aligned} \tag{13}$$

where $\lambda_{\text{da}} \geq 0$ controls the gradient reversal strength and $\lambda_{\text{dist}} \geq 0$ weights the explicit Normal–pseudo-Normal feature alignment term.

*(ii) Usage-aware update.* While domain alignment adjusts global statistics, some slots may remain underutilized. Let $\boldsymbol{\mu}_k = \big(\sum_n Q_{nk}\mathbf{Z}_n\big)/\big(\sum_n Q_{nk}\big)$ be

the responsibility-weighted center of slot $k$, $\mathbf{m}_k$ the $k$-th row of $\mathbf{M}_A$, and $\bar{u} = \frac{1}{K}\sum_k u_k$ the mean slot usage. We encourage balanced coverage by pulling underused slots more strongly toward their centers:

$$\mathcal{L}_{\text{upd}} = \frac{1}{K}\sum_{k=1}^{K}\Big(\frac{\bar{u}}{u_k+\varepsilon}\Big)^{\beta}\big\|\,\mathbf{m}_k - \boldsymbol{\mu}_k\,\big\|_2^2. \tag{14}$$

**Objective.** Our final loss augments the UR-DMU discrimination term with the two regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{UR-DMU}} + \lambda_1\mathcal{L}_{\text{DA}} + \lambda_2\mathcal{L}_{\text{upd}}. \tag{15}$$

## 4 Experiments

### 4.1 Experimental Setup

**Setup.** Unless otherwise specified, we follow an *unsupervised training* regime: the training data include only real normal videos and the names of abnormal classes that might appear at inference time; *no real abnormal videos* are used for training. At evaluation, both real normal and real abnormal videos are included, and we report frame-level metrics.

**Datasets.** We evaluate on three widely used benchmarks with distinct characteristics: ShanghaiTech (SHT) [11], UCF-Crime (Crime) [18] and XD-Violence (XD) [24]. SHT consists of fixed-view campus surveillance scenes with diverse illumination and viewpoints, spanning 13 locations and 437 video clips, and covering 14 abnormal categories (e.g., bicycle riding, skateboard intrusion, fighting), with 63 abnormal clips available in the training split. Following common practice for weakly supervised evaluation [6, 13, 22], we use the revised split of [33] at test time, while keeping training *abnormal-free*. For SHT, we synthesize 10 pseudo-abnormal clips per class (total 140) and additionally generate 50 pseudo-normal clips for domain alignment. Crime is a large-scale collection of real-world surveillance videos (about 1,900 videos, ∼128 hours) covering 13 abnormal classes (e.g., explosion, abuse), with 810 abnormal clips in the training split. For Crime, we synthesize 20 pseudo clips per class (total 260) and additionally generate 80 pseudo-normal clips. Following prior setups [36], we apply 10-crop augmentation. XD is a large-scale multi-scene violence benchmark with 4,754 untrimmed videos and 217 hours, spanning 6 violence categories. We follow the official weakly supervised split with 3,954 training videos and 800 testing videos. For XD, we synthesize 10 pseudo-abnormal clips for each of 6 classes and 50 pseudo-normal clips for domain alignment; details of pseudo-generation are provided in the supplementary material. Pseudo-video synthesis is performed once offline; generating 60 clips takes ∼10 hours on two RTX 6000 Ada GPUs (96GB).

**Metrics.** To assess the *generation quality* of the synthesized videos, we report standard video generation metrics. For image-level quality, we compute Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) between frame

**Table 2:** Examples of prompt refinement by the VLM.

| | |
|---|---|
| w/o refine | Generate Shoplifting |
| w/ refine | Generate Store theft involves an individual removing merchandise from shelves, concealing it in a backpack, and evading surveillance cameras. |


**Fig. 5:** Initial image examples produced by our CLIP-based selection for SHT and UCF-Crime. Images are class-relevant and scene-consistent, serving as anchors for I2V synthesis.

distributions of generated and real videos. For spatiotemporal fidelity, we report Fréchet Video Distance (FVD) and Kernel Video Distance (KVD) based on I3D features, capturing motion and temporal coherence. For *anomaly detection*, we follow the VAD literature and report frame-level Area Under the ROC Curve (AUC). In all quantitative comparisons, scores of prior methods are directly quoted from their original papers unless otherwise stated.

**Implementation Details.** We use the Wan 2.2 image-to-video diffusion model (I2V-A14B) [21] with a resolution of $832 \times 480$ and a frame length of 81. For prompt refinement, we employ the Qwen3 30B-A3B vision–language model [27]. As an auxiliary video feature extractor (for analysis/metrics), we use the final projector embedding (3,584-D) of Qwen2.5-VL 7B-Instruct [1]. CLIP ViT-B/32 [15] is used for initial image selection. Our implementation is based on PyTorch and HuggingFace Transformers.

**Initial Images and Prompt Design.** Using the init selection process in Sec. 3.1, we obtain class-relevant init images for SHT and Crime (Fig. 5). They exhibit minimal semantic mismatch to target classes. Representative examples of VLM-based prompt refinement are shown in Table 2, where the refined text better reflects the scene context of the initial image. These inits and refined prompts are then fed to the I2V generator to construct the pseudo-anomaly datasets. Please refer to the supplementary material for the detail.

### 4.2 Qualitative Evaluation

**Class-Aware Pseudo-Anomaly Generator (CA-PAG)** Fig. 6 shows examples of generated videos. Without prompt refinement (*w/o Refine*, i.e., using the raw class label), ambiguous instructions degrade consistency and text–video alignment, leading in SHT to artifacts such as the intrusion of large fighting crowds, object merging/disappearance across frames, and excessive motion that ignores physical plausibility. On Crime, we likewise observe multiple vehicles appearing/disappearing. With prompt refinement (*w/ Refine*), the VLM-guided prompts produce scenes that reflect the initial image context: on SHT we observe

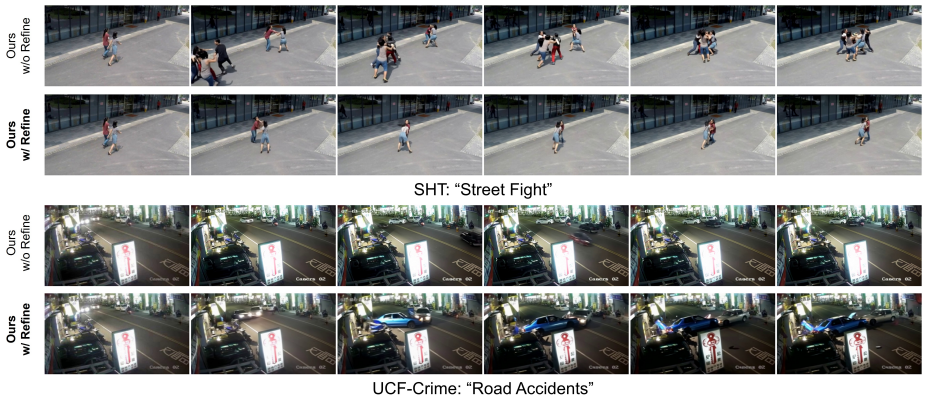



**Fig. 6:** Qualitative examples of synthesized pseudo anomalies.

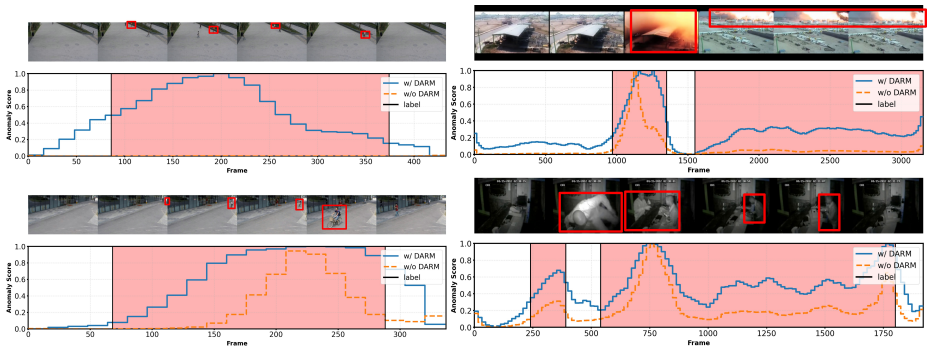



**Fig. 7:** Qualitative detection results. Left: SHT (top:skateboarder, bottom:biker). Right: UCF-Crime (top:explosion,bottom:burglary). The adaptive memory module mitigates MIL bias and improves sensitivity to subtle abnormal segments.

person fighting present in the initial image with more natural motion; on Crime we observe an inter-vehicle collision without disappearance/merging artifacts.

**Domain-Aligned Regularized Module(DARM).** Fig. 7 shows score traces (vertical: score; horizontal: time). Left: in "skateboard intrusion," *w/ DARM* responds sharply at abnormal segments, while *w/o DARM* barely rises. In the "biker" case, *w/ DARM* suppresses false positives and concentrates its peak within the annotated interval. Right: in "explosion," *w/ DARM* sustains higher sensitivity during the burning phase. In "burglary," it produces temporally tighter peaks aligned with abnormal actions. These examples demonstrate that DARM not only enhances responsiveness to true abnormal events but also suppresses magnitude-driven false negatives, yielding temporally sharper and more semantically aligned detection peaks.

### 4.3 Quantitative Evaluation

**Pseudo-anomaly videos quality.** Table 3 reports SHT generation quality. As prompt refinement is applied, all metrics improve. In particular, we observe a reduction of FVD by 97 points and KVD by 22.7 points, indicating better motion and temporal coherence. These gains are consistent with the qualitative trends in Fig. 6, suggesting that refined prompts enhance physical plausibility and scene consistency. In contrast, KID remains unchanged. As it reflects per-frame appearance rather than temporal dynamics, both settings share identical image statistics, while CA-PAG improves motion coherence captured by FVD/KVD.

**DARM for detection.** Table 4 reports frame-level AUC against prior UVAD and WVAD methods under the Real/Pseudo setting (i.e., training without any real abnormal videos). On SHT, PA-VAD with DARM achieves 98.2% AUC, surpassing the best UVAD method [34] and even outperforming the strongest WVAD competitor CMRL [6] (97.6%) by $+\mathbf{0.6}$ points. On XD-Violence, it attains 95.1% AUC and exceeds the strongest Real/Real baseline UR-DMU [36] (94.2%) by $+\mathbf{0.9}$ points. These results indicate that DARM mitigates pseudo-induced bias and improves transfer across diverse anomaly domains. On Crime, our Real/Pseudo setting is naturally disadvantaged against Real/Real baselines, yet PA-VAD with DARM remains competitive, and still surpasses the UVAD state-of-the-art [34] (80.6%) by $+\mathbf{1.9}$ points. UVAD methods are the closest comparison in abnormal-data access, though the level of supervision is not strictly equal, as PA-VAD additionally relies on large pretrained generative and vision-language priors; this trades hard-to-collect real abnormal footage for readily available pretrained knowledge—a practical advantage for real-world deployment. Within our PA-VAD framework, replacing the plain UR-DMU [36] with DARM consistently improves performance under the same setting, from $96.0\% \rightarrow 98.2\%$ on SHT, $80.2\% \rightarrow 82.5\%$ on Crime, and $92.5\% \rightarrow 95.1\%$ on XD; a consistent gain also holds for the Sultani et al. [18] classifier, indicating that domain alignment and the usage-aware update are important to fully exploiting pseudo anomalies without overfitting to their magnitude bias.

**Ablation of CA-PAG and DARM.** Table 5a disentangles the individual contributions of the CA-PAG and DARM components on SHT, allowing us to quantify the effect of each design choice. Starting from a random-init baseline (86.7%), CLIP-based initial image selection raises AUC to 94.9%, a substantial gain of +8.2 points, confirming the benefit of anchoring generation to class-relevant normal seeds. Enabling prompt refinement on top of init image selection already improves AUC from 94.9% to 96.0%, indicating that class-aware textual conditioning is important even without any regularization on the detector side. Within DARM, the usage-aware memory update is the primary driver of gains: adding the update term on top of CA-PAG boosts AUC to 97.6–97.7%, whereas domain alignment alone provides only modest improvement. The best result (98.2%) is obtained when domain alignment and usage-aware update are combined, supporting our design that DARM suppresses the magnitude bias by reducing the real/pseudo discrepancy and preventing high-magnitude prototypes from dominating MIL optimization.

**Table 3:** Quality of pseudo-anomaly videos.

| Data | FVD↓ | FID↓ | KVD↓ | KID↓ |
|---|---|---|---|---|
| Ours w/o Refine | 701 | 83.9 | 57.4 | **0.03** |
| Ours w/ Refine | **604** | **78.0** | **34.7** | **0.03** |

**Table 4:** Frame-level AUC (%) comparison on ShanghaiTech (SHT), UCF-Crime (Crime), and XD-Violence (XD). "Feature" denotes the frozen clip-level backbone used by the detector; "–" indicates methods without such features. "Data type (Nor/Abn)" denotes the source of normal/abnormal training videos (real or pseudo).

| Protocol | Methods | Reference | Feature | Data type (Nor/Abn) | SHT | Crime | XD |
|---|---|---|---|---|---|---|---|
| UVAD | LERF [19] | AAAI23 | – | Real/- | 78.6 | – | – |
| | HSC [20] | CVPR23 | – | Real/- | 83.4 | – | – |
| | MGSTRL [34] | CVPR24 | – | Real/- | 87.5 | 80.6 | – |
| | SeeKer [7] | ICCV25 | – | Real/- | 85.5 | – | – |
| WVAD | Sultani et al. [18] | CVPR18 | C3D | Real/Real | – | 75.4 | – |
| | CMRL [6] | CVPR23 | I3D | Real/Real | 97.6 | 86.1 | – |
| | UR-DMU [36] | AAAI23 | I3D | Real/Real | – | 87.0 | 94.2 |
| | VERA [28] | CVPR25 | InternVL2 | Real/Real | – | 86.6 | 88.3 |
| PA-VAD (Ours) | Ours w/Sultani [18] | | Qwen2.5-VL | Real/Pseudo | 93.7 | 77.8 | 81.8 |
| | Ours w/UR-DMU [36] | | Qwen2.5-VL | Real/Pseudo | 96.0 | 80.2 | 92.5 |
| | Ours w/DARM | | Qwen2.5-VL | Real/Pseudo | **98.2** | **82.5** | **95.1** |

**Table 5:** Ablation studies of PA-VAD.

**(a)** Effect of CA-PAG and DARM components on SHT.

| CA-PAG | | DARM | | AUC(%) |
|---|---|---|---|---|
| Initial | Refine | DA | Update | |
| | | | | 86.7 |
| ✓ | | | | 94.9 |
| ✓ | ✓ | | | 96.0 |
| ✓ | | | ✓ | 97.6 |
| ✓ | ✓ | | ✓ | 97.7 |
| ✓ | | ✓ | | 96.0 |
| ✓ | ✓ | ✓ | | 96.8 |
| ✓ | ✓ | ✓ | ✓ | **98.2** |

**(b)** Mean $\ell_2$ norm of real/pseudo abnormal features in the detector space.

| Feature norm | SHT | Crime | XD |
|---|---|---|---|
| Real Abn. | 9.13 | 11.06 | 39.08 |
| Pseudo Abn. w/o DARM | 199.48 | 52.94 | 420.10 |
| Pseudo Abn. w/ DA | 183.50 | 45.87 | 262.40 |
| Pseudo Abn. w/ DARM | **8.39** | **12.80** | **33.63** |

**Table 6:** Effect of the number of synthesized clips on SHT. We vary the number of generated pseudo-abnormal clips (total = 140) and report the frame-level AUC (%).

| Clips | 14 | 35 | 70 | 105 | 140 | 175 |
|---|---|---|---|---|---|---|
| SHT (%) | 85.4 | 94.6 | 96.9 | 97.3 | **98.2** | 97.7 |

**Table 7:** Comparison with pseudo-data generation methods and SF-VAD.

| Protocol | Method | Feature | Data type | | SHT | Crime |
|---|---|---|---|---|---|---|
| | | | Nor. | Abn. | | |
| WVAD | Cai et al. [3] w/RTFM | CLIP | Real+Pseudo | Real+Pseudo | – | 87.3 |
| | Cai et al. [3] w/LAP | | | | | 89.3 |
| SF-VAD | SF-VAD/FPL [5] | I3D RGB | Real | Real | 98.4 | 89.9 |
| PA-VAD | PA-VAD (ours) | Qwen2.5-VL | Real | Pseudo | **98.2** | **82.5** |

**Magnitude bias analysis.** To verify that DARM mitigates bias on the *anomaly* side, we measure the mean $\ell_2$ norm of real and pseudo abnormal features in the detector space (Table 5b). The mild gap in the frozen backbone (cf. Table 1) is strongly amplified where MIL Top-$k$ operates: without DARM, the pseudo-abnormal norm inflates to over 20× the real-abnormal norm on SHT (199.48 vs. 9.13), forming a magnitude shortcut. Comparing the last two rows of Table 5b isolates the usage-aware update: domain alignment alone (w/ DA) only partially narrows the gap, as it aligns the normal streams rather than the abnormal prototypes, whereas full DARM restores the pseudo-abnormal norms to the real-abnormal range across all datasets. This confirms the usage-aware update as the key component for correcting the bias on the abnormal side.

**Effect of the number of pseudo clips.** Table 6 further analyzes robustness to the amount of synthesized data on SHT. Increasing the number of pseudo-abnormal clips from 14 to 140 steadily improves performance from 85.4% to 98.2% AUC, and we already reach 96.9% with only 70 clips. Even when the number of synthesized clips is reduced to around the scale of the 63 real abnormal clips in the standard WVAD split, PA-VAD with DARM remains highly competitive. When we further increase the number of synthesized clips to 175, the AUC slightly drops to 97.7%, suggesting a mild saturation effect where overly many pseudo clips introduce redundancy or noise that does not further benefit, and can slightly harm, generalization. Overall, the results indicate that pseudo-only training can match or exceed Real/Real pipelines, offering a practical option when abnormal data is unavailable.

**Further Analysis with Related Methods.** We compare PA-VAD with pseudo-data generation methods and the single-frame-supervised SF-VAD/FPL. Table 7 reports two comparisons: prior pseudo-data results from [3] and the reported SF-VAD/FPL results. Cai *et al.* [3] mix generated pseudo anomalies with real abnormal videos; in contrast, PA-VAD uses no real abnormal videos. PA-VAD remains competitive with SF-VAD/FPL on SHT (98.2 vs. 98.4) despite not relying on the single precise abnormal-frame annotations that SF-VAD uses. Note that SF-VAD differs from PA-VAD in its baseline, as it employs single frame-level supervision together with real abnormal videos.

**Generalization to Unseen Anomalies.** A key concern in pseudo-only training is overfitting to anomaly categories used for pseudo generation, potentially harming robustness to unseen anomalies. Following OpenVAD's open-set protocol [37] on Crime and XD, we generate pseudo-abnormal videos from a restricted

**Table 8:** Open-set on Crime.

| **Seen anom.** | 1 | 3 | 6 | 9 |
|---|---|---|---|---|
| RTFM [22] | 75.91 | 76.98 | 77.68 | 79.55 |
| OpenVAD [37] | 76.73 | 77.78 | 78.82 | 80.14 |
| Ours | **77.65**$_{\pm 3.29}$ | **80.86**$_{\pm 2.03}$ | **82.18**$_{\pm 0.25}$ | **82.30**$_{\pm 0.14}$ |

**Table 9:** Open-set on XD.

| **Seen anom.** | 1 | 4 |
|---|---|---|
| OpenVAD [37] | 72.50 | 88.25 |
| Ours | **89.08**$_{\pm 4.52}$ | **94.99**$_{\pm 0.13}$ |

subset of anomaly classes (*seen*) and evaluate on the full test set including held-out classes (*unseen*). We repeat three random class splits and report mean AUC with standard deviation for PA-VAD; for prior methods, we report the values from their original papers.

Tables 8 and 9 show that PA-VAD consistently outperforms OpenVAD across all settings, even with only one seen class (XD: 72.50 → 89.08, +16.6 points), indicating transfer beyond seen-class artifacts. We attribute this open-set robustness primarily to DARM. As discussed in the paper, DARM suppresses pseudo-induced shortcut cues (e.g., feature-magnitude bias) that can dominate MIL-based training and lead to brittle decision rules tied to synthesized artifacts. More importantly for open-set transfer, DARM's *usage-aware memory update* acts as an anti-overfitting mechanism against seen-class specialization: it reweights memory updates by slot usage so that slots repeatedly activated by the same seen-class–specific modes are updated less, while under-used slots are promoted. This prevents memory from being monopolized by a few seen-driven prototypes and encourages the detector to rely on more transferable, mode-agnostic abnormal evidence, consistent with the improved sensitivity to unseen classes under open-set splits.

**Limitation.** On UCF-Crime, our method does not surpass the SoTA that trains with real abnormal videos under the WVAD setup. A likely factor is the difficulty of synthesizing minute-long, context-heavy anomalies (e.g., extended temporal dependencies), which remain challenging for current video generators; such anomalies are more prevalent in UCF-Crime than in SHT, making the dataset inherently harder under a pseudo-only setting.

## 5 Conclusion

We presented PA-VAD, a pseudo-only video anomaly detection framework with two components: CA-PAG for synthesizing class-consistent pseudo anomalies and DARM for mitigating pseudo-induced magnitude bias and stabilizing MIL-based learning. Without using any real abnormal footage, PA-VAD achieves competitive or superior performance to strong baselines on SHT, Crime, and XD, demonstrating that high-quality pseudo anomalies with regularized training provide a practical alternative to Real/Real pipelines. Future work includes improving motion priors for pseudo generation and extending evaluation to broader video generative models and settings.

## References


1. Bai, S., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Cai, S., Peng, X., Wang, C., Cai, X., Qian, J.: GV-VAD: Exploring Video Generation for Weakly-Supervised Video Anomaly Detection. arXiv preprint arXiv:2508.00312 (2025)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Chen, J., Li, L., Tu, Y., Su, L., Xue, Z., Huang, Q.: Generalizing single-frame supervision to event-level understanding for video anomaly detection. In: NeurIPS. vol. 38, pp. 6142–6173 (2025)
6. Cho, M., Kim, M., Hwang, S., Park, C., Lee, K., Lee, S.: Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In: CVPR. pp. 12137–12146 (2023)
7. Delić, A., Grcic, M., Šegvić, S.: Sequential keypoint density estimator: An overlooked baseline of skeleton-based video anomaly detection. In: ICCV. pp. 11579–11589 (2025)
8. Du, H., Zhang, S., Xie, B., Nan, G., Zhang, J., Xu, J., Liu, H., Leng, S., Liu, J., Fan, H., Huang, D., Feng, J., Chen, L., Zhang, C., Li, X., Zhang, H., Chen, J., Cui, Q., Tao, X.: Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In: CVPR. pp. 18793–18803 (2024)
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016)
10. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: ICCV. pp. 20406–20417 (2023)
11. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection – a new baseline. In: CVPR. pp. 6536–6545 (2018)
12. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV. pp. 13588–13597 (2021)
13. Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., Zhang, H.: Unbiased multiple instance learning for weakly supervised video anomaly detection. In: CVPR. pp. 8022–8031 (2023)
14. Mehrabi, N., Goyal, P., Verma, A., Dhamala, J., Kumar, V., Hu, Q., Chang, K.W., Zemel, R., Galstyan, A., Gupta, R.: Resolving ambiguities in text-to-image generative models. In: ACL. pp. 14367–14388 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. vol. 139, pp. 8748–8763 (2021)
16. Rai, A.K., Krishna, T., Hu, F., Drimbarean, A., McGuinness, K., Smeaton, A.F., O'Connor, N.E.: Video anomaly detection via spatio-temporal pseudo-anomaly generation: A unified approach. In: CVPRW. pp. 3887–3899 (2024)

17. Ren, H., Liu, W., Olsen, S.I., Escalera, S., Moeslund, T.B.: Unsupervised behavior-specific dictionary learning for abnormal event detection. In: BMVC. pp. 28.1–28.13 (2015)
18. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
19. Sun, C., Shi, C., Jia, Y., Wu, Y.: Learning event-relevant factors for video anomaly detection. In: AAAI. vol. 37, pp. 2384–2392 (2023)
20. Sun, S., Gong, X.: Hierarchical semantic contrast for scene-aware video anomaly detection. In: CVPR. pp. 22846–22856 (2023)
21. Team Wan, et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
22. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4975–4986 (2021)
23. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV. pp. 4489–4497 (2015)
24. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV. vol. 12375, pp. 322–339 (2020)
25. Wu, P., Zhou, X., Pang, G., Sun, Y., Liu, J., Wang, P., Zhang, Y.: Open-vocabulary video anomaly detection. In: CVPR. pp. 18297–18307 (2024)
26. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In: AAAI. vol. 38, pp. 6074–6082 (2024)
27. Yang, A., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
28. Ye, M., Liu, W., He, P.: VERA: Explainable video anomaly detection via verbalized learning of vision-language models. In: CVPR. pp. 8679–8688 (2025)
29. Yu, G., Wang, S., Cai, Z., Liu, X., Xu, C., Wu, C.: Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In: CVPR. pp. 13987–13998 (2022)
30. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR. pp. 14744–14754 (2022)
31. Zhang, C., Li, G., Qi, Y., Wang, S., Qing, L., Huang, Q., Yang, M.H.: Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In: CVPR. pp. 16271–16280 (2023)
32. Zhang, H., Xu, X., Wang, X., Zuo, J., Huang, X., Gao, C., Zhang, S., Yu, L., Sang, N.: Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In: CVPR. pp. 13843–13853 (2025)
33. Zhang, J., Qing, L., Miao, J.: Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In: ICIP. pp. 4030–4034 (2019)
34. Zhang, M., Wang, J., Qi, Q., Sun, H., Zhuang, Z., Ren, P., Ma, R., Liao, J.: Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In: CVPR. pp. 17385–17394 (2024)
35. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a Plug-And-Play action classifier for anomaly detection. In: CVPR. pp. 1237–1246 (2019)
36. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: AAAI. vol. 37, pp. 3769–3777 (2023)

37. Zhu, Y., Bao, W., Yu, Q.: Towards open set video anomaly detection. In: ECCV. vol. 13694, pp. 395–412 (2022)