CAViR: A Synthetic Video-Text Benchmark for Real-World Anomaly Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Language-based anomaly clip retrieval searches pre-segmented short clips whose content matches a natural-language description of a normal or anomalous event. However, real-world collection, *e.g.*, the XD-Violence dataset, faces prohibitive cost and privacy constraints, and covers only a fraction of anomalous event types. We introduce **CAViR** (**C**ross-modal **A**nomaly **Vi**deo **R**etrieval), a synthetic video-text benchmark comprising 41,315 short video clips (1.36M frames) across 30 normal activity types and 68 anomalous event types. Each clip is paired with 5 captions, yielding 206,575 video-text pairs. To test whether the generated data is learnable and transferable, we further propose the **FLAME** (**F**low-guided **L**anguage-video **A**lignment with **M**ixture-of-**E**xperts) framework, which integrates Token-Aware Mixture-of-Experts (TAM) and a Cross-modal Flow Predictor (CFP). TAM adaptively routes visual and textual tokens to specialized experts for fine-grained event representation learning, while CFP uses auxiliary optical-flow prediction to align language with short-term visual dynamics. FLAME achieves the highest text-to-video/video-to-text Recall@1 among competitive methods: **71.6%/87.3%** on CAViR, **41.0%/39.7%** on UCFCrime-AR, and **33.4%/33.8%** on XD-Violence, respectively. Extensive experiments verify synthetic-to-real transfer. Without relying on any real anomaly clips, pre-training on CAViR alone improves text-to-video Recall@1 on UCFCrime-AR from **22.4%** to **33.8%**.

1 Introduction

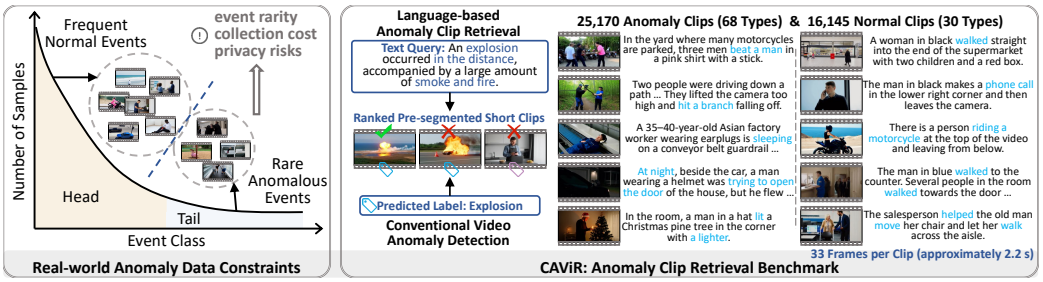


Figure 1: Due to the long-tail distribution of real-world anomalies and privacy constraints, large-scale data collection remains challenging (*left*). To address this limitation, we construct CAViR, a large-scale cross-modal anomaly clip retrieval benchmark (*right*), comprising $41,315$ pre-segmented short video clips (each with 5 matched captions). The benchmark covers 68 anomaly and 30 normal event types, and evaluates clip-level semantic retrieval.

Video Anomaly Retrieval (VAR) (Wu et al., 2024) aims to retrieve video content depicting specific anomalous events using cross-modal queries, such as natural-language descriptions. This capability can support incident investigation and workplace safety review by helping users find relevant footage of intrusions, falls, or unsafe actions in large video archives. While existing formulations commonly retrieve or localize relevant segments within long, untrimmed videos, we focus on *language-based anomaly clip retrieval*, a clip-level setting that ranks pre-segmented short video clips by their semantic relevance to a natural-language query. Each candidate clip contains one dominant normal or anomalous event. Unlike traditional Video Anomaly Detection (VAD) (Cao et al., 2024), which

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with existing datasets for anomaly detection, understanding, and retrieval. CAViR uses pre-segmented 2.2-second clips for retrieval and covers PAB's 1,600 fine-grained anomaly types.

Dataset	Modality	Visual Sample Unit	Annotation	#Visual Samples	#Texts	#Anomaly Types	Anomaly: Normal Samples	Visual Data Source
PAB (Yang et al., 2025b)	Image, Text	Image	Image-level Text	1.02M	1.02M	1,600	3:2	Collection & Synthesis
UBnormal (Acsintoae et al., 2022)	Video	Video	Frame-level Tag	543	–	22	2:3	Synthesis
ShanghaiTech (Luo et al., 2017)	Video	Video	Frame-level Tag	437	–	11	1:18	Collection
UCF-Crime (Sultani et al., 2018)	Video	Untrimmed Video	Video-level Tag	1,900	–	13	1:1	Collection
XD-Violence (Wu et al., 2020)	Video, Audio	Untrimmed Video	Frame-level Tag	4,754	–	6	1:1	Collection
UCA (Yuan et al., 2024)	Video, Text	Untrimmed Video	Temporal Event Text	1,854	23,542	13	1:1	Collection
UCFCrime-AR (Wu et al., 2024)	Video, Text	Untrimmed Video	Video-level Text	1,900	1,900	13	1:1	Collection
HIVAU-70k (Zhang et al., 2025)	Video, Text	Untrimmed Video	Multi-granular Text	5,443	70K	13	1:1	Collection
CAViR (Ours)	Video, Text	**Short Clip (2.2 s)**	**Clip-level Text**	41,315	206,575	68	3:2	Synthesis

focuses on detecting or categorizing anomalies, this setting supports fine-grained event queries, such as "multiple intruders breaking into a house at night" (Figure 1, right). By separating clip retrieval from temporal boundary prediction, this formulation provides a controlled setting for learning correspondences between language descriptions and short-term visual dynamics.

Existing anomaly video-text benchmarks (Yuan et al., 2024; Wu et al., 2024; Zhang et al., 2025) primarily augment UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020) with textual annotations (Table 1). Despite richer textual supervision, their visual diversity and event coverage remain constrained by the underlying source datasets. Expanding this coverage through real-world data collection is challenging: the long-tailed distribution of anomalous events makes it difficult to obtain sufficient examples of rare event types (Figure 1, left). Moreover, collecting and publicly releasing footage involving identifiable individuals or sensitive activities can raise privacy concerns, while manual annotation incurs substantial costs.

To address these limitations, we combine Large Language Models (LLMs) with text-to-video generation to produce synthetic video-text data at scale, reducing reliance on real-world footage and manual annotation. Making these data useful for real-world anomaly retrieval requires addressing two challenges. First, *synthetic-to-real domain discrepancies* can hinder the transfer of learned representations. We therefore guide video generation with annotations from real-world video datasets and LLM-expanded descriptions of event categories from existing benchmarks to encourage semantic and contextual alignment with real-world videos. Second, *diverse linguistic supervision* is needed to support varied retrieval queries. We supplement each generated clip's base description with four complementary captions that differ in linguistic style and descriptive focus, capturing event semantics, overall video dynamics, and keyframe details. The resulting benchmark, **CAViR** (**C**ross-modal **A**nomaly **Vi**deo **R**etrieval), comprises **41,315 short video clips** covering **30 normal activity types** and **68 anomalous event types** (Figure 1, right). With five captions per clip, CAViR provides 206,575 video-text pairs. Cross-dataset experiments show that training on CAViR improves retrieval performance across multiple models on real-world benchmarks, supporting the transferability of representations learned from synthetic video-text data.

To verify the value of the constructed dataset, we propose the **F**low-guided **L**anguage-video **A**lignment with **M**ixture-of-**E**xperts (**FLAME**) framework. FLAME addresses two complementary challenges. Event evidence varies in spatial granularity: some events involve conspicuous interactions, whereas others depend on local subject-object details. Token-Aware Mixture-of-experts (TAM) provides input-dependent routing of visual and textual tokens to represent this heterogeneity. A Cross-modal Flow Predictor (CFP) complements appearance features with auxiliary motion supervision, associating event descriptions with short-term visual dynamics. Both modules are trained jointly with video-text alignment objectives. Our main contributions are summarized as follows:

- To address limited event coverage and the difficulty of collecting real-world anomaly footage, we introduce **CAViR**, a large-scale synthetic benchmark for language-based anomaly clip retrieval. CAViR comprises 41K short clips and 206K video-text pairs, covering 30 normal activity types and 68 anomalous event types.
- Based on CAViR, we also propose **FLAME**, a flow-guided language-video alignment framework that integrates Token-Aware Mixture-of-experts (TAM) and Cross-modal Flow Predictor (CFP). TAM adaptively routes visual and textual tokens to specialized experts for fine-grained event representation, while CFP introduces auxiliary optical-flow prediction to encourage alignment between language descriptions and short-term visual dynamics.

- FLAME achieves the highest text-to-video Recall@1 among the compared methods on CAViR, UCFCrime-AR, and XD-Violence, reaching 71.6%, 41.0%, and 33.4%, respectively. Cross-dataset experiments further show that training on CAViR consistently improves retrieval on real-world datasets UCFCrime-AR, XD-Violence, and OOPS!, indicating the utility of synthetic supervision for real-world anomaly retrieval.

2 Related Work

Video Anomaly Detection and Retrieval. Video anomaly detection (VAD) identifies deviations from normal activity under different supervision settings (Micorek et al., 2024), including normal-only training (Park et al., 2020; Sabokrou et al., 2017; Liu et al., 2018; Wu et al., 2019), fully unsupervised learning (Zaheer et al., 2022), and supervised or weakly supervised learning (Wu et al., 2021; 2020; Sultani et al., 2018; Feng et al., 2021). As summarized in Table 1, benchmarks range from fixed surveillance scenes (Lu et al., 2013; Ramachandra & Jones, 2020; Adam et al., 2008; Li et al., 2013) to broader scene coverage in ShanghaiTech (Luo et al., 2017) and synthetic UBnormal (Acsintoae et al., 2022), and larger, more diverse untrimmed collections in UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020), the latter also providing audio. UCA (Yuan et al., 2024) and UCFCrime-AR (Wu et al., 2024) add temporally grounded and video-level descriptions, respectively, while HIVAU-70k (Zhang et al., 2025) enriches UCF-Crime and XD-Violence with LLM-generated annotations at clip, event, and video levels. PAB (Yang et al., 2025b) supports text-based person anomaly search using synthetic image-text training pairs. CAViR expands event coverage through synthetic videos across 68 anomalous and 30 normal categories, with five stylistically diverse captions per clip to support fine-grained retrieval and synthetic-to-real transfer.

Video-Text Retrieval. Video-text retrieval (VTR) learns cross-modal correspondence for bidirectional ranking, progressing from joint embeddings and fusion (Torabi et al., 2016; Kiros et al., 2014; Yu et al., 2018; Kaufman et al., 2017) to scalable end-to-end learning (Miech et al., 2020; Lei et al., 2021; Bain et al., 2021; Wang et al., 2022) and large-scale vision-language pretraining (Radford et al., 2021; Li et al., 2022; 2021; Zeng et al., 2022), which improve downstream transfer (Gan et al., 2023; Wu et al., 2023; Wang et al., 2023). CE, MMT, and T2VLAD explore expert aggregation, multimodal temporal modeling, and global-local alignment (Liu et al., 2019b; Gabeur et al., 2020; Wang et al., 2021), while CLIP4Clip and X-CLIP advance CLIP-based retrieval through temporal aggregation and multi-grained matching (Luo et al., 2022; Ma et al., 2022). XML (Lei et al., 2020) further studies moment retrieval by localizing query-relevant segments within a video corpus. For anomaly analysis, HL-Net (Wu et al., 2020) targets audio-visual anomaly detection, while ALAN (Wu et al., 2024) supports cross-modal retrieval of long, untrimmed videos. FLAME combines token-aware expert routing with auxiliary optical-flow prediction to capture fine-grained event semantics and short-term dynamics for anomaly clip retrieval.

3 Benchmark

Event Description Collection and Expansion. As shown in Figure 2, we first collect event descriptions from UCA (Yuan et al., 2024), UCFCrime-AR (Wu et al., 2024), and OOPS! (Epstein et al., 2020). Descriptions from UCA and UCFCrime-AR are retained without modification, while ambiguous OOPS! descriptions, *e.g.*, those containing "don't know", are removed. To broaden event coverage, we incorporate 23 additional anomaly categories from UBnormal (Acsintoae et al., 2022) and NTU RGB+D 120 (Liu et al., 2019a), including fighting, falling, and car crashes. For each added category, we use Qwen3 (Yang et al., 2025a), Seed-Thinking-v1.5 (Seed et al., 2025), and DeepSeek-R1 (DeepSeek-AI, 2025) to generate multiple candidate descriptions that vary the number and appearance of subjects, event context, environment, and camera viewpoint. We manually discard descriptions with ambiguous events, unclear subjects, or content unsuitable for a single short clip. After manual processing, these 23 types of anomalies retain 50 diverse descriptions for each category. The resulting set contains 41,315 base descriptions: 16,145 normal and 25,170 anomalous.

Efficient Short-Clip Generation. We synthesize one clip per base description using Wan2.1 (Wan et al., 2025), an open-source text-to-video model, with FP8 quantization and TeaCache acceleration for efficiency. Each clip contains 33 frames at 15 fps (approximately 2.2 s) with a resolution of 480×832 (height $\times$ width), yielding approximately 1.36M frames in total. The short duration

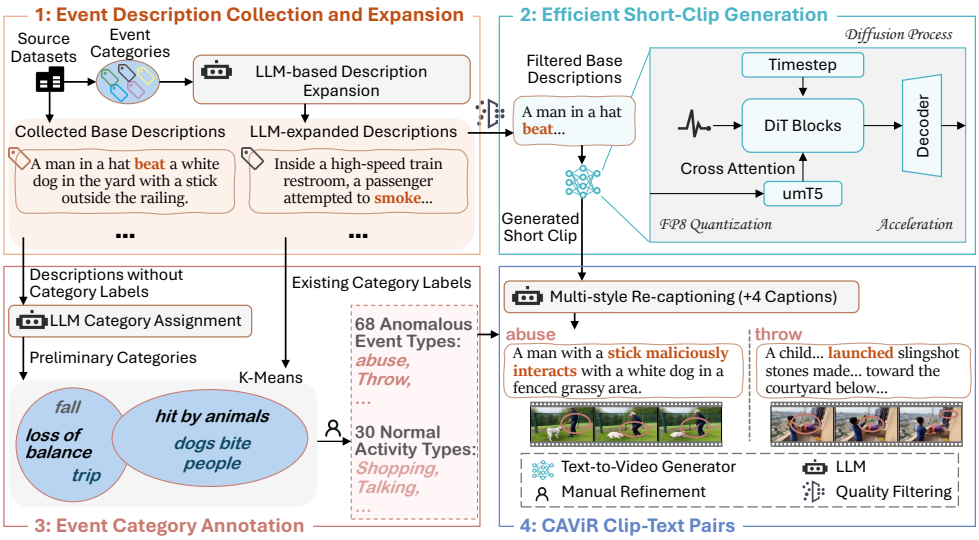
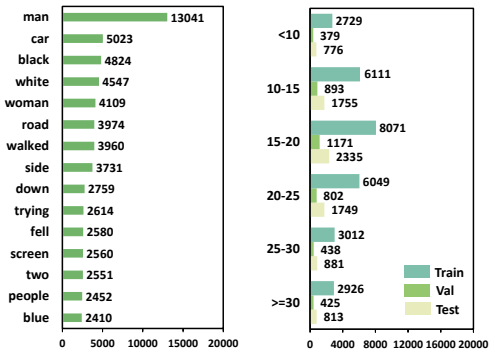


Figure 2: Construction pipeline of CAViR. Source event descriptions are expanded with LLMs and filtered for ambiguity before video synthesis. Each generated clip is paired with its base description and four additional captions in complementary styles. Event categories are assigned and consolidated through LLM annotation, K-means clustering, and manual verification.

targets a dominant event and its local dynamics; it does not represent the full temporal context of a long surveillance recording. The landscape format and resolution balance surveillance-style framing with generation cost. FLAME resizes sampled frames to 224×224 for encoding, so generation resolution and model input resolution are distinct.

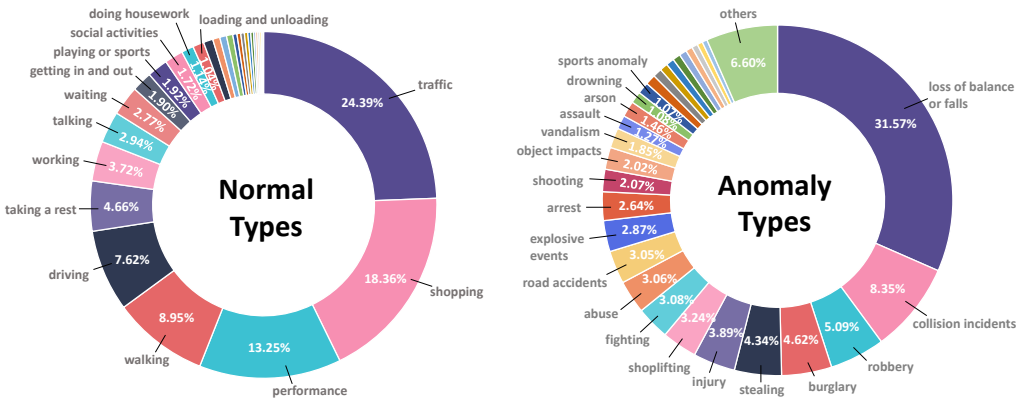
Multi-Style Re-captioning. To diversify linguistic supervision, we augment each generated clip with four captions in complementary styles: 1) UCFCrime-AR-style: a concise, factual event description; 2) OOPS!-style: a natural-language description emphasizing an unexpected moment; 3) Holistic: a Qwen-generated summary of the complete clip; and 4) Middle-frame: a fine-grained description of the middle frame. Qwen3-VL (Bai et al., 2025) generates the first two styles following the linguistic patterns of the corresponding datasets. General-purpose video and image prompts are used for the holistic and middle-frame captions, respectively. Together with the base description, these captions yield five descriptions per clip and 206,575 clip-text pairs.



(a) Word-frequency. (b) Sentence-length.

Figure 3: Caption statistics of CAViR: (a) word frequencies and (b) sentence-length distributions across the train/val/test splits.

Event Category Annotation. For base descriptions without category labels, Qwen3 (Yang et al., 2025a) assigns preliminary labels from the event text. A manual audit of 10% of these annotations yields over 97% agreement with verified labels. We consolidate semantically equivalent category names, *e.g.*, drone crash and drone falling, through K-means clustering followed by manual merging, renaming, and verification. The resulting taxonomy contains **30 normal activity categories** and **68 anomalous event categories** as shown in Figure 4. Normal categories include basic human motions, such as walking, and activities consistent with their context, such as playing soccer. Anomalous events fall into four broad groups: 1) Biomechanical deviations, such as falling; 2) Contextual inconsistencies, such as running in a laboratory; 3) Social norm violations, such as hitting another person; and 4) Environmental anomalies, such as natural disasters.



(a) Normal activity categories. (b) Anomalous event categories.

Figure 4: Category distributions of CAViR: (a) normal activities and (b) anomalous events.

Dataset Analysis. *1) Coverage and balance:* As summarized in Table 1, CAViR[1] combines broad event coverage with multiple captions per clip. The normal-to-anomalous sample ratio is approximately 2:3, while Figure 4 reports the fine-grained category distributions. *2) Scope and limitations:* CAViR provides pre-segmented clips with one dominant event for semantic retrieval, without temporal boundary prediction. It complements long-video benchmarks such as UCFCrime-AR (Wu et al., 2024) and UCA (Yuan et al., 2024), whose sample counts are not directly comparable to short-clip counts. Generation with a single model may not capture the full range of real-world surveillance conditions, including severe illumination changes, motion blur, compression artifacts, and complex camera motion. FID, FVD, and CLIPSIM provide partial quality evidence; they do not establish perceptual realism or the absence of a synthetic-to-real domain gap. *3) Privacy:* Synthesizing clips avoids recording identifiable participants in dangerous situations, reducing reliance on privacy-sensitive footage. Privacy, copyright, and bias concerns inherited from source descriptions and pretrained generators may nevertheless remain.

4 Method

4.1 Token-Aware Mixture-of-Experts (TAM)

Video Encoding. Without loss of generality, we employ Swin Transformer (Swin-B) (Liu et al., 2021) as the vision encoder. Swin-B consists of 4 stages, each composed of transformer blocks and a linear embedding layer. To enable the model to adaptively process diverse visual tokens, we insert a Token-Aware Mixture-of-experts layer (denoted as TAM_V) between the second and third stages of Swin-B. This design allows the video encoder to route different input tokens to specialized experts, enhancing its representational capacity. Formally, inspired by Switch Transformer (Fedus et al., 2022), TAM comprises N feed-forward network (FFN) experts $\{\mathrm{FFN}_i\}_{i=1}^{N}$ and a router G that activates the top-k experts for each token, as illustrated in Figure 5. Let $\mathcal{I}_k(x) \subseteq \{1, \ldots, N\}$ denote the indices of the k largest router logits. For an input token x, the TAM output y and routing weights $g_i(x)$ are computed as:

$$y = \mathrm{Norm}\left(x + \sum_{i=1}^{N} g_i(x)\,\mathrm{FFN}_i(x)\right), \qquad g_i(x) = \begin{cases} \dfrac{\exp\left([G(x)]_i\right)}{\sum_{j=1}^{N}\exp\left([G(x)]_j\right)}, & i \in \mathcal{I}_k(x), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Here, $G(x)$ is a learnable linear function. In our implementation, we set $N = 3$ and $k = 1$, *i.e.*, each token is routed to a single expert. For a given video, we sample N_f frames, and sequentially feed them into the vision encoder to obtain the corresponding frame embeddings $\{F_1, F_2, \ldots, F_{N_f}\}$. To further model the temporal dimension across all video frames, we add temporal information (a learnable weight matrix $W \in \mathbb{R}^{N_f \times D}$) to each frame embedding and compute the average along the

[1]Anonymous dataset demo: `https://cavir-dataset.github.io/`. For more benchmark details, please refer to the appendix.

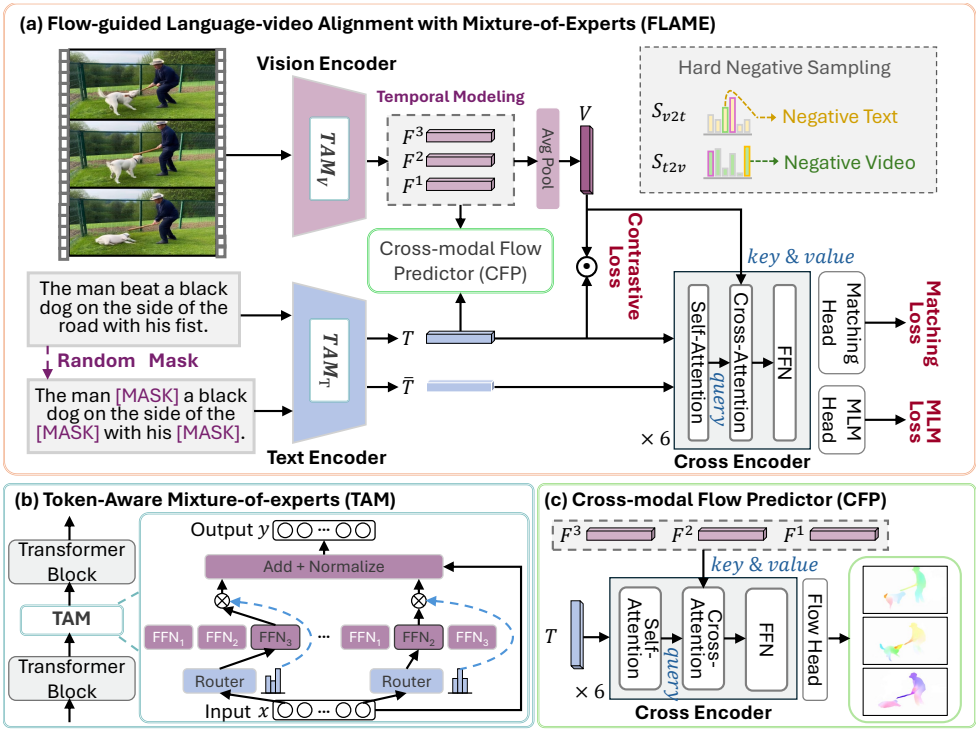


Figure 5: (a) Overview of our FLAME (Flow-guided Language-video Alignment with Mixture-of-experts) framework, which contains two core modules: (b) Token-Aware Mixture-of-experts (TAM) for fine-grained spatial representation via token-aware experts distinguishing normal and anomalous behaviors, and (c) Cross-modal Flow Predictor (CFP) linking motion text with visual dynamics through optical flow for temporal modeling. The framework is jointly optimized by optical flow prediction loss and video-text cross-modal alignment objectives.

temporal dimension. This average is denoted as $V = AVG([F_1, F_2, \ldots, F_{N_f}] + W)$, which we use as the video embedding.

Text Encoding. Following established practices (Yang et al., 2025b), we adopt the first 6 layers of BERT (Devlin et al., 2019) as the text encoder. Similar to the vision encoder, we insert a TAM layer (denoted as TAM_T) between the third and fourth layers of the text encoder. Given a video description text, we first tokenize it into N_t tokens and feed them into the text encoder to obtain high-dimensional text embeddings $T = \{t^{CLS}, t^1, t^2, ..., t^{N_t-1}\}$. Here, $t^n (n \in [1, N_t - 1])$ represents the embedding of the $(n+1)$-th token.

4.2 Cross-modal Flow Predictor (CFP)

For a matched text-video pair, we fuse the video frame and text representations via a cross-encoder and use the resulting representation to predict optical flow. Specifically, we employ the last six layers of BERT as the cross encoder and insert a cross-attention module in each layer of BERT. As shown in Figure 5, the text embedding T serves as the input to the cross-encoder and is fused with the frame embedding $F_m (m \in [1, N_f])$ in the cross-attention module. In this process, the text embedding acts as the query, while the frame embedding acts as the key and value. We transform the *[CLS]* embedding of the output feature from the cross encoder through a Flow Head (an MLP) to obtain the predicted optical flow value $\hat{p}^{\text{flow}}(F_m, T)$. The optical flow prediction loss is defined as the MSE loss between the predicted optical flow value and the ground-truth optical flow:

$$\mathcal{L}_{\text{flow}} = \frac{1}{N_f} \sum_{m=1}^{N_f} \|\hat{p}^{\text{flow}}(F_m, T) - p_m^{\text{flow}}\|^2. \tag{2}$$

Here, we use the optical flow frames p_m^{flow} obtained from the MEMFOF (Bargatin et al., 2025) model as pseudo-ground-truth optical flow.

Table 2: Multimodal text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall Rate (R@1, R@5, R@10), Median Rank (MdR), and Mean Rank (MnR) on CAViR.

Method	Train Set	T2V					V2T				
		R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip-tightTransf (Luo et al., 2022)	-	0.0	0.1	0.1	4320.0	4266.3	0.0	0.0	0.1	4320.0	4289.7
CLIP4Clip-seqTransf (Luo et al., 2022)	-	31.0	51.3	59.7	5.0	128.8	43.7	68.9	78.0	2.0	19.2
CLIP4Clip-seqLSTM (Luo et al., 2022)	-	34.4	56.1	64.8	4.0	108.7	48.0	73.3	82.2	2.0	14.3
CLIP4Clip-MeanP (Luo et al., 2022)	-	36.0	58.3	66.8	3.0	89.0	51.3	76.3	84.7	**1.0**	11.8
X-CLIP (Ma et al., 2022)	-	34.5	56.3	64.8	4.0	90.6	49.2	74.9	82.9	2.0	12.4
FLAME (Ours)	-	55.6	74.6	80.4	**1.0**	52.7	70.3	87.9	92.4	**1.0**	6.5
CLIP4Clip-tightTransf (Luo et al., 2022)	CAViR	54.8	79.6	86.6	**1.0**	10.9	69.9	91.2	95.5	**1.0**	3.0
CLIP4Clip-seqTransf (Luo et al., 2022)	CAViR	62.0	84.4	90.2	**1.0**	7.6	74.5	93.4	96.8	**1.0**	2.3
CLIP4Clip-seqLSTM (Luo et al., 2022)	CAViR	61.5	84.1	89.9	**1.0**	8.1	75.6	93.7	96.9	**1.0**	2.3
CLIP4Clip-MeanP (Luo et al., 2022)	CAViR	61.7	84.0	89.8	**1.0**	8.4	76.1	93.5	96.9	**1.0**	2.2
X-CLIP (Ma et al., 2022)	CAViR	62.7	84.8	90.4	**1.0**	7.4	75.6	93.8	97.1	**1.0**	2.2
FLAME (Ours)	CAViR	**71.6**	**88.6**	**92.8**	**1.0**	**6.6**	**87.3**	**97.0**	**98.6**	**1.0**	**1.6**

4.3 Video Text Alignment

To align video and text representations, we incorporate three cross-modal alignment objectives: video-text contrastive learning, video-text matching, and masked language modeling.

Video-Text Alignment. For each video-text pair in a batch of size N_v, we obtain video and text embeddings V^j and T^j ($j \in [1, N_v]$) from their respective encoders. Their *[CLS]* embeddings are projected via two MLPs to low-dimensional features z_v^j and z_t^j. The video-to-text similarity is computed as $S_{\text{v2t}} = \frac{\exp(s(z_v^j, z_t^j)/\tau)}{\sum_{l=1}^{N_v} \exp(s(z_v^j, z_t^l)/\tau)}$, where $s(\cdot,\cdot)$ is cosine similarity and τ a temperature parameter. The text-to-video similarity S_{t2v} is defined analogously. The contrastive loss is: $\mathcal{L}_{\text{cl}} = -\frac{1}{2}\mathbb{E}[\log S_{\text{v2t}} + \log S_{\text{t2v}}]$.

Video-Text Matching. To learn fine-grained alignment, for each video, we construct a hard negative pair by sampling a text example based on the similarity score S_{v2t}, and vice versa. Positive pairs together with hard negatives are fed into a cross encoder to obtain fused embeddings. A Matching Head (MLP) then predicts the matching probability $\hat{p}^{\text{match}}$. The loss is binary cross-entropy: $\mathcal{L}_{\text{match}} = -\mathbb{E}[p^{\text{match}} \log \hat{p}^{\text{match}} + (1 - p^{\text{match}}) \log(1 - \hat{p}^{\text{match}})]$, where p^{match} is the ground-truth label.

Masked Language Modeling. Following prior works (Yang et al., 2025b; Li et al., 2021; Zeng et al., 2022), for a given aligned video-text pair, we randomly mask 25% text tokens. The masked text is encoded as $\bar{T}$, fused with the corresponding video embedding V in the cross encoder, and an MLM Head (MLP with softmax) predicts the masked tokens via cross-entropy: $\mathcal{L}_{\text{mlm}} = -\mathbb{E}[p^{\text{mask}} \log \hat{p}^{\text{mask}}(V, \bar{T})]$, where, $\hat{p}^{\text{mask}}$ is the predicted token distribution, and p^{mask} is the ground-truth one-hot vector. The overall objective is: $\mathcal{L} = \beta\mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{mlm}}$.

5 Experiment

CAViR is randomly split within each action category into $28{,}898$ training, $4{,}108$ validation, and $8{,}309$ test videos (approximately $7:1:2$) with corresponding 5 captions. FLAME has 390.7M trainable parameters: 93.1M, 80.6M, and 59.1M in the vision, text, and cross encoders, respectively; 155.8M in the CFP branch (discarded at inference); and approximately 2.1M in the MLP projection modules for contrastive and matching losses. TAM_V and TAM_T contain 6.3M and 14.2M parameters, respectively. Initialized from X-VLM (Zeng et al., 2022), FLAME is trained for 5 epochs using AdamW with batch size 20 and a learning rate linearly decayed from $2.5e^{-5}$ to $2.5e^{-6}$. Inputs comprise three 224×224 frames per video and up to 56 tokens per text. We evaluate text-to-video (T2V) and video-to-text (V2T) retrieval on CAViR, UCFCrime-AR (Wu et al., 2024), and XD-Violence (Wu et al., 2020), and further examine cross-dataset transfer with OOPS! (Epstein et al., 2020) as an additional target.

5.1 Video Anomaly Retrieval Results

Quantitative evaluation on CAViR. Table 2 compares FLAME with four CLIP4Clip (Luo et al., 2022) variants and X-CLIP (Ma et al., 2022), before and after fine-tuning on CAViR. Without CAViR fine-tuning, FLAME achieves T2V and V2T R@1 scores of 55.6% and 70.3%, respectively, outperforming these baselines. Fine-tuning increases the scores to 71.6% and 87.3%, exceeding X-CLIP in T2V by 8.9 percentage points and CLIP4Clip-MeanP in V2T by 11.2 percentage points. FLAME

Table 3: Performance comparison on the real-world datasets UCFCrime-AR and XD-Violence for video anomaly retrieval. We adopt the video summary from HIVAU-70k (Zhang et al., 2025) as the matching text for each XD-Violence video. *The result is reported in (Wu et al., 2024)

Method	UCFCrime-AR						XD-Violence					
	T2V			V2T			T2V			V2T		
	R@1	R@10	MdR	R@1	R@10	MdR	R@1	R@10	MdR	R@1	R@10	MdR
CE* (Liu et al., 2019b)	6.6	32.4	23.5	5.5	32.4	21.0	-	-	-	-	-	-
MMT* (Gabeur et al., 2020)	8.3	39.3	16.0	-	-	-	-	-	-	-	-	-
T2VLAD* (Wang et al., 2021)	7.6	39.7	15.5	6.2	43.1	14.0	-	-	-	-	-	-
HL-Net* (Wu et al., 2020)	5.5	38.3	19.5	5.5	35.5	20.0	-	-	-	-	-	-
XML* (Lei et al., 2020)	6.9	42.4	14.0	6.6	43.4	13.0	-	-	-	-	-	-
ALAN* (Wu et al., 2024)	9.0	44.8	14.0	7.3	46.9	12.0	-	-	-	-	-	-
CLIP4Clip-tightTransf (Luo et al., 2022)	17.5	63.4	6.0	17.0	60.9	7.0	11.5	43.1	15.0	12.6	44.4	13.0
CLIP4Clip-seqTransf (Luo et al., 2022)	28.8	74.0	4.0	27.6	72.4	4.0	20.6	53.9	8.0	21.6	54.9	8.0
CLIP4Clip-seqLSTM (Luo et al., 2022)	32.2	75.7	4.0	31.0	72.1	**3.0**	22.1	53.6	8.0	20.1	55.3	8.0
CLIP4Clip-MeanP (Luo et al., 2022)	29.5	74.0	4.0	26.9	71.1	4.0	21.9	53.6	9.0	20.5	55.0	9.0
X-CLIP (Ma et al., 2022)	32.5	74.3	**3.0**	29.9	72.8	4.0	24.5	56.9	7.0	21.1	56.1	7.0
FLAME (Ours)	**41.0**	**77.9**	**3.0**	**39.7**	**76.6**	**3.0**	**33.4**	**67.9**	**4.0**	**33.8**	**69.0**	**3.0**

Table 4: Cross-dataset generalization performance on four target benchmarks, including the synthetic CAViR dataset and three real-world datasets, UCFCrime-AR, XD-Violence, and OOPS!. For brevity, we only report text-to-video results here.

Train Set	Method	CAViR		UCFCrime-AR		XD-Violence		OOPS!	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
-	CLIP4Clip-MeanP (Luo et al., 2022)	36.0	66.8	17.1	56.5	9.2	28.9	14.1	42.3
	X-CLIP (Ma et al., 2022)	34.5	64.8	17.1	59.9	11.6	33.3	13.9	42.1
	FLAME (Ours)	55.6	80.4	22.4	63.4	16.5	45.1	12.9	40.0
CAViR	CLIP4Clip-MeanP (Luo et al., 2022)	61.7	89.8	24.7	65.8	15.6	40.1	17.5	49.6
	X-CLIP (Ma et al., 2022)	62.7	90.4	26.7	66.8	15.0	41.6	**18.3**	**51.7**
	FLAME (Ours)	**71.6**	**92.8**	33.8	70.0	20.9	47.5	**18.3**	50.0
UCFCrime-AR	CLIP4Clip-MeanP (Luo et al., 2022)	40.5	72.5	29.5	74.0	13.0	36.2	14.5	44.1
	X-CLIP (Ma et al., 2022)	41.7	73.5	32.5	74.3	13.4	37.0	16.1	47.1
	FLAME (Ours)	59.4	84.2	**41.0**	**77.9**	19.8	49.3	16.1	45.8
XD-Violence	CLIP4Clip-MeanP (Luo et al., 2022)	39.7	70.8	20.5	62.0	21.9	53.6	14.5	44.2
	X-CLIP (Ma et al., 2022)	39.8	70.7	19.9	65.1	24.5	56.9	14.0	42.2
	FLAME (Ours)	58.6	83.8	27.9	68.6	**33.4**	**67.9**	15.3	44.8

also achieves the highest R@5 and R@10 and the lowest MnR in both directions among the methods in Table 2, with an MdR of 1.0.

Quantitative evaluation on real-world datasets. Table 3 compares FLAME with CLIP-based baselines and earlier retrieval methods on UCFCrime-AR and XD-Violence. After fine-tuning on UCFCrime-AR, FLAME achieves 41.0% T2V R@1 and 39.7% V2T R@1, improving over the strongest competing results by 8.5 and 8.7 percentage points, respectively. On XD-Violence, we adopt the video summary from HIVAU-70k (Zhang et al., 2025) as the paired caption for each video. FLAME reaches 33.4% T2V R@1 and 33.8% V2T R@1, with corresponding gains of 8.9 and 12.2 percentage points. It also obtains the highest R@10 and the best or tied-best MdR in both directions on both datasets. These results indicate FLAME's effectiveness on real-world anomaly video-text retrieval.

Cross-dataset transfer. Table 4 reports T2V retrieval across four benchmarks for models fine-tuned on CAViR, UCFCrime-AR, or XD-Violence. For cross-dataset evaluation, models are directly applied to the target benchmark without further fine-tuning. Relative to their pretrained counterparts, all three methods benefit from CAViR training, with consistent improvements in R@1 and R@10 on all three real-world targets. For FLAME, R@1 increases from 22.4% to 33.8% on UCFCrime-AR, an improvement of 11.4 percentage points. It also increases from 16.5% to 20.9% on XD-Violence and from 12.9% to 18.3% on OOPS!. The consistent gains across architectures support the effectiveness of CAViR as training data for retrieval on real-world videos. Among models trained on CAViR, FLAME achieves the highest R@1 and R@10 on UCFCrime-AR and XD-Violence. On UCFCrime-AR, it exceeds X-CLIP by 7.1 and 3.2 percentage points in R@1 and R@10, respectively. On OOPS!, FLAME ties X-CLIP at 18.3% R@1, while X-CLIP achieves a higher R@10 (51.7% versus 50.0%). Thus, CAViR training benefits all evaluated models, while their relative performance depends on the target dataset and retrieval metric.

Qualitative Results. In Figure 6, we show the ranking list of FLAME on our CAViR dataset. We observe that our dataset remains challenging for prevailing cross-modality models. Some fine-grained actions are not well captured by the existing approaches, and this difficulty provides a competitive test base for future exploration by the community.

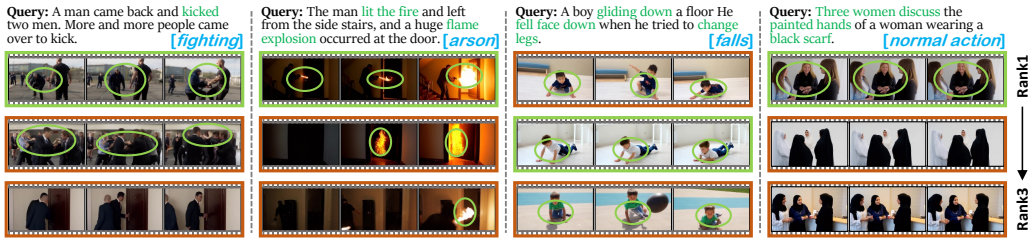
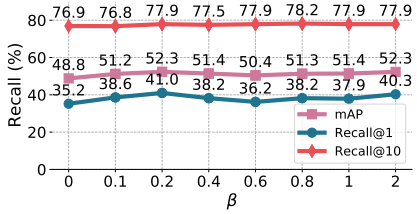


Figure 6: Text-to-video top-3 retrieved examples of FLAME on CAViR. (correct; incorrect).



(a) Ablation study on the value of β.

Method	CFP	TAM_V	TAM_T	R@1 ↑	R@5 ↑	R@10 ↑	mAP ↑
Baseline				33.45	63.10	**78.97**	47.56
M1	✓			37.59	64.14	76.21	50.31
M2		✓		34.83	62.76	75.86	48.05
M3			✓	36.21	65.87	77.24	49.34
M4	✓	✓		35.52	**68.97**	78.28	49.87
M5	✓		✓	37.24	65.17	77.93	50.17
M6		✓	✓	35.17	64.14	76.90	48.85
FLAME	✓	✓	✓	**41.03**	67.93	77.93	**52.30**

(b) Ablation study on the key component.

Figure 7: Ablation Study on β and the key component of the proposed method. In (a), we set $\beta = \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 2\}$ to study the impact of β. In (b), we present ablation results on the key component of FLAME, *i.e.*, Token-Aware Mixture-of-experts module at vision encoder (TAM_V) and text encoder (TAM_T), and Cross-modal Flow Predictor (CFP). The text-to-video anomaly retrieval recall rates on UCFCrime-AR (Wu et al., 2024) are reported.

5.2 Ablation Studies

Effect of the Key Components. We conduct ablation studies on the key components of the FLAME model. As shown in Table 7b, for fairness, all variants are trained under identical settings. Specifically, we train all the models on UCFCrime-AR (Wu et al., 2024) for 5 epochs and report the text-to-video retrieval recall rates. First, we evaluate the effectiveness of the CFP module by comparing the Baseline and M1 models. The comparison indicates that CFP is a critical component of FLAME. Furthermore, M2 and M3 incorporating TAM_V and TAM_T obtain 1.38% and 2.76% R@1 improvement compared to the Baseline, respectively. The results indicate that adding these modules improves feature discrimination. FLAME, equipped with all components, achieves the best performance, with 41.03% R@1 and 52.30% mAP.

Parameter Sensitivity of β. To explore the effectiveness of the value of β (the weight of optical flow prediction loss), we conduct further ablation experiments. We set β to $0, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 2$ separately and train the model on UCFCrime-AR (Wu et al., 2024). We report the text-to-video anomaly retrieval accuracy in Figure 7a. The model performs better when $\beta = 0.2$. Based on this observation, we set $\beta = 0.2$. ***For ablation studies on expert and frame-count, the position of TAM, and different text styles, please refer to the appendix.***

6 Conclusion

We introduce CAViR, a synthetic benchmark for language-based anomaly clip retrieval, comprising 41,315 short clips and 206,575 video-text pairs across 68 anomalous and 30 normal event categories. CAViR expands event and linguistic coverage while reducing the need to collect privacy-sensitive real-world footage. To effectively harness synthetic data, we propose FLAME, a flow-guided language-video alignment mixture-of-experts framework that integrates two novel modules: Token-Aware Mixture-of-experts (TAM) for fine-grained spatial anomaly modeling, and Cross-modal Flow Predictor (CFP) for enhanced temporal dynamics understanding. Experiments show that FLAME achieves text-to-video R@1 of 71.6%, 41.0%, and 33.4% on CAViR, UCFCrime-AR, and XD-Violence, respectively, under within-dataset training. Cross-dataset gains across several retrieval architectures support the value of synthetic clip-text supervision for real-world retrieval. These results underscore the practical value of CAViR as a benchmark and the effectiveness of FLAME in bridging synthetic and real-world anomaly retrieval for public safety analytics.

AI USE STATEMENT

In this work, we used generative AI tools for generating a synthetic dataset and assisting with translation. We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, pp. 20143–20153, 2022.

Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pp. 1728–1738, 2021.

Vladislav Bargatin, Egor Chistov, Alexander Yakovenko, and Dmitriy Vatolin. Memfof: High-resolution training for memory-efficient multi-frame optical flow estimation. In *ICCV*, 2025.

Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing*, 33: 1810–1825, 2024. doi: 10.1109/TIP.2024.3372466.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, pp. 919–929, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(1), January 2022. ISSN 1532-4435.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, pp. 14009–14018, 2021.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pp. 214–229, Cham, 2020. Springer International Publishing.

Tian Gan, Qing Wang, Xingning Dong, Xiangyuan Ren, Liqiang Nie, and Qingpei Guo. Cnvid-3.5 m: Build, filter, and pre-train the large-scale public chinese video-text dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14815–14824, 2023.

Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. Temporal tessellation: A unified approach for video analysis. In *ICCV*, pp. 94–104, 2017.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 447–463, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58589-1.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *PCVPR*, pp. 7331–7341, 2021.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900. PMLR, 2022.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019a.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. In *CVPR*, Jun 2018.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, September 2019b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pp. 2720–2727, 2013.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, pp. 341–349, 2017.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, pp. 638–647, 2022.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Koziński. MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection. In *CVPR*, pp. 18868–18877, June 2024.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pp. 9879–9889, 2020.

Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, Jun 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *WACV*, pp. 2569–2578, 2020.

Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, pp. 1992–2004, Apr 2017.

ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pp. 6479–6488, 2018.

Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xiao Wang, Yaoyu Li, Tian Gan, Zheng Zhang, Jingjing Lv, and Liqiang Nie. Rtq: Rethinking video-language understanding based on image-text model. In *ACM MM*, pp. 557–566, 2023.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *CVPR*, pp. 5075–5084, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00504. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00504`.

Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*, 25:6079–6089, 2022.

Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *IJCAI*, Aug 2021.

Mengxia Wu, Min Cao, Yang Bai, Ziyin Zeng, Chen Chen, Liqiang Nie, and Min Zhang. An empirical study of frame selection for text-to-video retrieval. *arXiv preprint arXiv:2311.00298*, 2023.

Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, Jan 2019.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, pp. 322–339. Springer, 2020.

Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Transactions on Image Processing*, 33:2213–2225, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. Beyond walking: A large-scale image-text benchmark for text-based person anomaly search. In *ICCV*, 2025b.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pp. 471–487, 2018.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *CVPR*, pp. 22052–22061, 2024.

M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, pp. 14744–14754, 2022.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ICML*, 2022.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *CVPR*, pp. 13843–13853, 2025.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Benchmark **3**

4 Method **5**
4.1 Token-Aware Mixture-of-experts (TAM) 5
4.2 Cross-modal Flow Predictor (CFP) 6
4.3 Video Text Alignment 7

5 Experiment **7**
5.1 Video Anomaly Retrieval Results 7
5.2 Ablation Studies 9

6 Conclusion **9**

A Experiment Details and Further Experiments **14**
A.1 Training Details 14
A.2 Inference Details 15
A.3 Computational Efficiency 15
A.4 Comparison with a Strong Multimodal Embedding Model 15
A.5 Expert and Frame-Count Ablations 16
A.6 Ablation Study on TAM Combination Position 16
A.7 Exploring the Effectiveness of TAM on Cross Encoder 17
A.8 The Impact of Different Style Text 18
A.9 More Qualitative Result Examples 18

B More Benchmark Details **19**
B.1 Dataset Visualization 19
B.2 Automatic Evaluation of Generated Data 24
B.3 Caption Obtaining Details 24
B.4 Attribute Annotation Details 24

APPENDIX

A EXPERIMENT DETAILS AND FURTHER EXPERIMENTS

A.1 TRAINING DETAILS

Implementation and Initialization. FLAME is implemented in PyTorch and initialized from X-VLM (Zeng et al., 2022), pretrained on 16M image-text pairs from datasets including MS-COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017).

Frame Sampling and Preprocessing. For each training video, we exclude the first and last frames, uniformly partition the remaining frames into three groups, and randomly sample one frame per group. Frames are resized to 224×224 with random horizontal flipping during training, and text sequences are limited to 56 tokens.

Optimization and Training Cost. We train for five epochs using AdamW with batch size 20, weight decay 0.01, and a learning rate linearly decayed from 2.5×10^{-5} to 2.5×10^{-6}. On CAViR, training uses four NVIDIA RTX 3090 GPUs, with the first 500 iterations devoted to warm-up, and takes approximately 63.5 minutes. On UCFCrime-AR (Wu et al., 2024), training uses one NVIDIA RTX 3090 GPU, with warm-up over the first 10% of iterations, and takes approximately 12.5 minutes.

Table 5: Computational measurements on UCFCrime-AR with one RTX 3090. The full FLAME configuration reranks 128 candidates per query. Parameters are the reported counts for each configuration and exclude FLAME's training-only flow prediction component. R@1 is in percent.

Method	Params (M)	FLOPs (G)	Latency (ms)	Memory (GB)	R@1
CLIP4Clip-tightTransf	136.5	32.4	11.0	1.4	17.5
CLIP4Clip-seqTransf	136.2	1.7	8.7	1.5	28.8
CLIP4Clip-seqLSTM	125.7	1.0	4.9	1.5	32.2
CLIP4Clip-MeanP	123.5	0.8	4.4	2.0	29.5
X-CLIP	136.7	1.5	4.9	1.9	32.5
FLAME (without reranking)	150.6	3.2	1.0	3.6	31.4
FLAME (top-128 reranking)	234.9	419.1	61.5	4.5	**41.0**

A.2 Inference Details

The CFP branch is discarded at inference. For each test video, we exclude the first and last frames, uniformly partition the remaining frames into three groups, and select the middle frame of each group. We extract embeddings for all test videos and texts and compute cross-modal similarities. For each query in either retrieval direction, all gallery samples are ranked by similarity, and the top 128 candidates are re-ranked using matching probabilities predicted by the cross encoder and matching head. Full test evaluation takes approximately 7.3 minutes on CAViR using four NVIDIA RTX 3090 GPUs and 38 seconds on UCFCrime-AR using one NVIDIA RTX 3090 GPU.

A.3 Computational Efficiency

We report text-to-video retrieval measurements on UCFCrime-AR using one NVIDIA RTX 3090. Table 5 distinguishes embedding retrieval from full FLAME with top-128 reranking. Within these measurements, reranking increases R@1 from 31.4% to 41.0% ($+9.6$ percentage points), while increasing FLOPs from 3.2G to 419.1G and query latency from 1.0 to 61.5 ms. Thus, the full model trades additional cross-modal interaction cost for retrieval accuracy.

The reported training total is 390.7M parameters, whereas removing the 155.8M flow prediction component gives the reported full inference count of 234.9M. The vision, text, and cross encoders are reported as 93.1M, 80.6M, and 59.1M, respectively, and the visual and textual TAM modules as 6.3M and 14.2M.

A.4 Comparison with a Strong Multimodal Embedding Model

Table 6 compares FLAME with VLM2Vec-V2 (Meng et al., 2025), a multimodal embedding model built on Qwen2-VL (Wang et al., 2024), for text-to-video retrieval. VLM2Vec-V2 has $2B$ parameters, whereas FLAME uses $0.39B$. VLM2Vec-V2 achieves higher zero-shot R@1 and R@10 on both datasets. Its zero-shot R@1 reaches 58.5% on CAViR and 27.9% on UCFCrime-AR, compared with 55.6% and 22.4% for FLAME, respectively. After dataset-specific fine-tuning, FLAME outperforms VLM2Vec-V2 on both benchmarks. On CAViR, FLAME achieves 71.6% R@1 and 92.8% R@10, exceeding VLM2Vec-V2 by 4.7 and 3.1 percentage points, respectively. On UCFCrime-AR, FLAME reaches 41.0% R@1 and 77.9% R@10, with corresponding gains of 15.8 and 11.0 percentage points. For VLM2Vec-V2, fine-tuning on UCFCrime-AR reduces R@1 from 27.9% to 25.2%, while increasing R@10 from 65.5% to 66.9%. In cross-domain evaluation, models fine-tuned on CAViR are directly evaluated on UCFCrime-AR. FLAME achieves 33.8% R@1 and 70.0% R@10, outperforming VLM2Vec-V2 by 5.2 and 4.4 percentage points, respectively. These results show that FLAME achieves stronger fine-tuned and cross-dataset retrieval performance with only 19.5% of the parameters of VLM2Vec-V2.

Table 6: Text-to-video retrieval comparison between FLAME and VLM2Vec-V2 (Meng et al., 2025) on CAViR and UCFCrime-AR. Fine-tuning uses the training split of each respective dataset. Cross-domain evaluation denotes training on CAViR and testing on UCFCrime-AR without further fine-tuning.

Method	#Parameter	CAViR				UCFCrime-AR					
		Zero-shot		Finetune		Zero-shot		Finetune		Cross-domain	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VLM2Vec-V2	2B	**58.5**	**82.3**	66.9	89.7	**27.9**	**65.5**	25.2	66.9	28.6	65.6
FLAME (Ours)	0.39B	55.6	80.4	**71.6**	**92.8**	22.4	63.4	**41.0**	**77.9**	**33.8**	**70.0**

Table 7: Effect of the number of TAM experts on UCFCrime-AR text-to-video retrieval. Metrics are percentages.

Number of experts	R@1	R@10	mAP
1	35.17	76.55	48.87
2	37.59	76.90	50.84
3 (default)	**41.03**	**77.93**	**52.30**
4	40.00	77.28	51.53
5	37.93	77.59	50.64
6	37.59	77.59	50.61

A.5 Expert and Frame-Count Ablations

The main paper studies the flow-loss weight β. Here, we report the additional expert-count and frame-count experiments on UCFCrime-AR. Table 7 shows that three experts achieve the highest R@1 and mAP among the evaluated expert counts. More experts do not consistently improve performance.

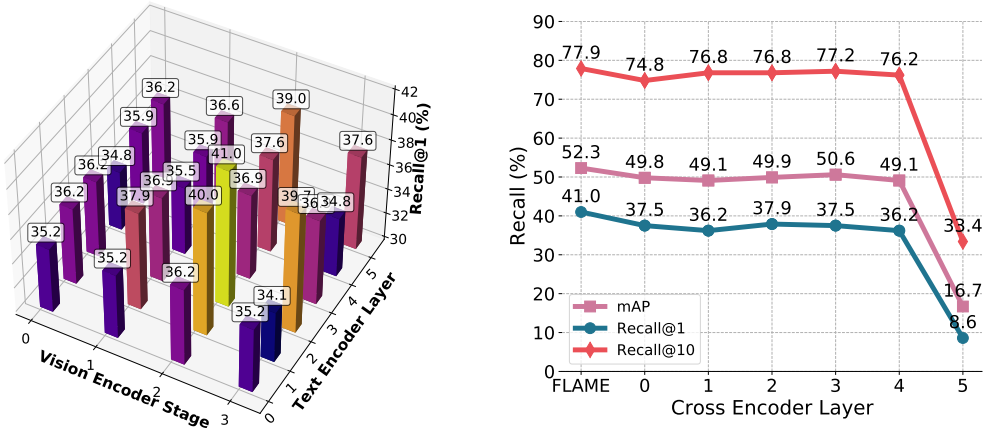
Among the configurations in Table 8, three frames achieve the highest R@1 and mAP, while six frames achieve the highest R@10. The three-frame setting also takes less training time than the four- to six-frame alternatives. Because the five- and six-frame settings use smaller batches, their accuracy differences cannot be attributed solely to frame count. The results support the selected accuracy–cost trade-off under the tested configurations.

A.6 Ablation Study on TAM Combination Position

To investigate the optimal placement of the proposed Token-aware Mixture-of-experts (TAM) module, we conduct ablation studies on both the vision and text encoders. Specifically, we simultaneously insert TAM before different stages of the vision encoder (Swin Transformer (Liu et al., 2021)) and after different layers of the text encoder (the first six layers of BERT (Devlin et al., 2019)), and report the video anomaly retrieval performance (text-to-video Recall@1) on UCFCrime-AR (Wu et al., 2024). As illustrated in Figure 8a, we observe that the insertion position significantly affects

Table 8: Effect of sampled frame count on UCFCrime-AR text-to-video retrieval. Retrieval metrics are percentages, and time is measured per training epoch. Batch size decreases for the five- and six-frame configurations, so these rows do not isolate frame count at a fixed batch size.

Frames	Batch size	Time/epoch (min)	R@1	R@10	mAP
1	20	2.2	35.86	73.10	48.45
2	20	2.4	34.48	77.93	48.11
3 (default)	20	2.5	**41.03**	77.93	**52.30**
4	20	4.0	39.66	76.21	51.55
5	16	4.6	34.48	77.59	48.45
6	12	5.0	36.21	**78.62**	49.07



(a) TAM combinations at different positions of vision and text encoders

(b) Exploring the addition of TAM at different cross-encoder layers.

Figure 8: Ablation study on the position of TAM on UCFCrime-AR. (a) We simultaneously insert TAM before different stages of the vision encoder (4 stages in total) and after different layers of the text encoder (6 layers in total). The best performance (text-to-video Recall@1: 41.0%) is achieved when TAM is inserted before stage 2 of the vision encoder and after layer 2 of the text encoder. (b) Comparison of TAM insertion at different cross-encoder layers. Adding TAM to the cross encoder consistently degrades performance across all metrics, with deeper layers experiencing larger drops. Our FLAME model (*w/o* TAM in cross encoder) is shown for reference.

Table 9: The effectiveness of different text styles. We report zero-shot transfer text-to-video retrieval results, in which models are fine-tuned on CAViR and directly tested on UCFCrime-AR (Wu et al., 2024).

Method	Style 0	Style 1	Style 2	Style 3	Style 4	R@1	R@10	mAP
FLAME	✓					22.8	62.0	36.6
	✓	✓				29.0	68.6	41.3
	✓	✓	✓			30.3	70.0	42.7
	✓	✓	✓	✓		30.3	69.6	43.6
	✓	✓	✓	✓	✓	**33.8**	**70.0**	**45.0**

the retrieval performance. The combination of TAM inserted before stage 2 of the vision encoder and after layer 2 of the text encoder achieves the highest Recall@1 of 41.0%, which supports our design choice. These results suggest that TAM is most effective when applied to mid-level features, where tokens begin to capture semantic information but are not yet fully abstracted.

A.7 Exploring the Effectiveness of TAM on Cross Encoder

We further explore the effect of inserting the Token-aware Mixture-of-experts (TAM) module at different layers of the cross encoder (the last six layers of BERT (Devlin et al., 2019)). As shown in Figure 8b, we compare the performance of our FLAME model (without TAM in the cross encoder) against variants where TAM is inserted at layers 0 to 5 of the cross encoder on UCFCrime-AR (Wu et al., 2024). The results show that adding TAM to the cross encoder consistently degrades text-to-video retrieval performance across all metrics (mAP, Recall@1, and Recall@10). Notably, the degradation becomes more severe as TAM is inserted at deeper layers, with the worst performance observed at layer 5 (16.7% mAP, 8.6% R@1, 33.4% R@10). This result suggests that token-aware routing is detrimental when applied to cross-modal interaction layers, likely due to interference with the fine-grained alignment learned by the cross encoder. Consequently, our final FLAME model does not incorporate TAM into the cross encoder.



Figure 9: Some video anomaly retrieval examples of FLAME on UCFCrime-AR. We visualize top 3 retrieved videos (green: correct; orange: incorrect).

A.8 The Impact of Different Style Text

To investigate the contribution of each text style in CAViR, we conduct ablation experiments by progressively incorporating different styles during fine-tuning and evaluating zero-shot transfer performance on the UCFCrime-AR (Wu et al., 2024) dataset. As shown in Table 9, using only Style 0 texts (directly collected or category-based LLM-generated descriptions) yields a model of 22.8% R@1 and 36.6% mAP. Adding Style 1 texts, which mimic the concise annotations of UCFCrime-AR (Wu et al., 2024), substantially boosts performance to 29.0% R@1 and 41.3% mAP, highlighting the importance of style alignment with the target domain. Further including Style 2 captions, *i.e.*, OOPS!-style descriptions, improves results to 30.3% R@1 and 42.7% mAP, suggesting that diverse narrative styles enhance model generalization. Incorporating Style 3 texts (Qwen-generated holistic captions) yields a marginal gain in mAP (43.6%), while further adding keyframe-focused Style 4 descriptions (including all five style texts) achieves the best performance across all metrics: 33.8% R@1, 70.0% R@10, and 45.0% mAP. These results indicate that the multi-style text annotations in CAViR provide complementary supervision, enabling the model to learn more robust video-text representations and achieve superior cross-dataset transferability.

A.9 More Qualitative Result Examples

We present qualitative results of FLAME for text-based video anomaly retrieval on the real-world dataset UCFCrime-AR (Wu et al., 2024) in Figure 9. For each query (anomaly or normal), we show the top-3 retrieved videos. Correct retrievals are marked with green boxes, while incorrect matches are marked with orange boxes. The action phrases in the text queries are highlighted in green. Notably, even the mismatched videos shown in the figure are somewhat relevant to the query sentence.

B More Benchmark Details

B.1 Dataset Visualization

In Figure 10, Figure 11, and Figure 12, we present additional video-text instances from the proposed dataset, CAViR. Figure 10 and Figure 11 showcase samples across nine anomaly categories. Figure 12 displays videos and corresponding descriptions for nine normal events. Each sample consists of a video and five aligned text descriptions of distinct styles:

- Style 0, directly collected or category-based LLM-generated video descriptions;
- Style 1, *i.e.*, *UCFCrime-AR style*, mimicking the concise, factual annotations of the UCFCrime-AR (Wu et al., 2024) dataset;
- Style 2, *i.e.*, *OOPS! style*, resembling the natural language descriptions of unexpected moments in the OOPS! (Epstein et al., 2020) dataset;
- Style 3, *i.e.*, *Qwen-generated video description*, a holistic caption summarizing the entire video content;
- Style 4, *i.e.*, *Qwen-generated keyframe description*, a fine-grained caption focusing on the middle frame of the video.

Additionally, each sample is annotated with specific anomaly or normal event categories to facilitate further fine-grained learning and evaluation. Human inspection reveals that the content of generated videos and captions is realistic and reasonable to some extent, though there is minor noise inherited from text-to-video generative models and LLMs. As shown in Table 11 (a) and (b), quantitative evaluations (FID, FVD, CLIPSIM) indicate the utility of CAViR in training and testing scenarios. Our experiments also show that such minor noise is tolerable for deep learning methods.



Figure 10: Examples of CAViR samples with nine anomaly categories (part 1). Each sample includes an anomaly video and five styles of aligned text descriptions: Style 0 (directly collected or category-based), Style 1 (*UCFCrime-AR style*), Style 2 (*OOPS! style*), Style 3 (*Qwen-generated video description*), and Style 4 (*Qwen-generated keyframe description*). Minor noise in videos or captions may arise from the limitations of video generative models and LLMs. (Best viewed on a computer screen with zoom.)



Figure 11: Examples of CAViR samples with nine anomaly categories (part 2). Each sample includes an anomaly video and five styles of aligned text descriptions: Style 0 (directly collected or category-based), Style 1 (*UCFCrime-AR style*), Style 2 (*OOPS! style*), Style 3 (*Qwen-generated video description*), and Style 4 (*Qwen-generated keyframe description*). Minor noise in videos or captions may arise from the limitations of video generative models and LLMs. (Best viewed on a computer screen with zoom.)



Figure 12: Examples of CAViR samples with nine normal categories. Each sample includes a normal video and five styles of aligned text descriptions: Style 0 (directly collected or category-based), Style 1 (*UCFCrime-AR style*), Style 2 (*OOPS! style*), Style 3 (*Qwen-generated video description*), and Style 4 (*Qwen-generated keyframe description*). Minor noise in videos or captions may arise from the limitations of video generative models and LLMs. (Best viewed on a computer screen with zoom.)

Table 10: Prompts for diverse video descriptions.

Diverse Video Descriptions Acquisition
Goal: Generating diverse video descriptions from an anomaly category name **Specific LLMs**: Qwen3 (Yang et al., 2025a), Seed-Thinking-v1.5 (Seed et al., 2025), and DeepSeek R1 (DeepSeek-AI, 2025) **Instruction**: "I want to generate a large-scale anomaly video dataset using Wan. I have several categories of anomalous behaviors. Can you help me design a sufficient number of video generation prompts? I will provide you with one anomalous behavior category at a time, and you will generate 50 corresponding prompts for it. Important Note: The prompts you provide should describe abnormal, unreasonable, or even dangerous behaviors in specific contexts. For example, a student running on a playground is normal and reasonable, not anomalous. I'd like you to refine the prompts to ensure the behavior is clearly out of place, illogical, or hazardous in the given scenario. The prompts should include details such as the person's age, gender, nationality, camera angle, and shot scale. First anomalous behavior category: {*anomaly category name*}."
Multi-style Re-captioning
Goal: Obtaining text style from unique dataset **Specific LLMs**: Qwen3-Max (Bai et al., 2025) **Instruction**: "Please summarize the language style of all sentences in this text file and output all the templates with different styles."
Goal: generating style-consistent descriptions for each video **Specific LLMs**: Qwen3-VL-8B-Instruct (Bai et al., 2025) **Instruction**: "Based on the video content, select a suitable template from the templates below. Using the selected template, output a sentence describing the video content. If no template matches the video content, select the template that best fits the video content, make slight modifications, and output a sentence describing the video content. Note that I expect your answer to only contain one sentence describing the video content, without showing the analysis process. Templates: {*UCFCrime-AR/OOPS! style templates*}"
Goal: Obtaining video/keyframe caption **Specific LLMs**: Qwen3-VL-8B-Instruct (Bai et al., 2025) **Instruction**: "Based on the video/frame content, output a sentence describing the video/frame content. Note that I expect your answer to only contain one sentence describing the video content, without showing the analysis process."

Table 11: Automatic evaluation of CAViR. (a) FID and FVD measure feature-distribution distances between datasets. (b) CLIPSIM measures text-clip similarity.

(a) Cross-dataset feature-distribution distances.

Dataset Pair	FID ↓	FVD ↓
UCFCrime-AR vs. OOPS!	156.00	2554.12
CAViR vs. UCFCrime-AR	146.60	2424.67
CAViR vs. OOPS!	45.61	1154.95

(b) CLIP-based text-clip similarity.

Dataset	CLIPSIM ↑
UCFCrime-AR	27.01
OOPS!	27.87
CAViR	28.29

B.2 Automatic Evaluation of Generated Data

Feature-Distribution Similarity. We evaluate feature-distribution similarity using Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD), with UCFCrime-AR (Wu et al., 2024) and OOPS! (Epstein et al., 2020) as real-world references. Under the same evaluation protocol, CAViR vs. OOPS! yields lower FID and FVD than UCFCrime-AR vs. OOPS! (Table 11a), indicating that CAViR is closer to OOPS! in the corresponding feature spaces.

Text-Clip Alignment. We assess text-clip alignment using CLIP-based similarity (CLIPSIM). Under the same protocol, CAViR achieves a score of 28.29, compared with 27.01 for UCFCrime-AR and 27.87 for OOPS! (Table 11b).

B.3 Caption Obtaining Details

Event Description Collection and Expansion Details. To further enhance anomaly diversity, we collect 23 unique anomaly categories from UBnormal (Acsintoae et al., 2022) and NTU RGB+D 120 (Liu et al., 2019a): throw, running, falling, fighting, sleeping, crawling, having a seizure, laying down, dancing, stealing, rotating 360 degrees, shuffling, walking injured, walking drunk, stumbling walk, people and car accident, car crash, running injured, fire, smoke, jaywalking, driving outside lane, and jumping. Adopting LLMs (Qwen3 (Yang et al., 2025a), Seed-Thinking-v1.5 (Seed et al., 2025), and DeepSeek R1 (DeepSeek-AI, 2025)), we expand these 23 anomaly categories into 23×50 diverse textual descriptions. The specific instruction is presented in Table 10. We manually filter unreasonable or low-quality samples, retaining 50 descriptions per category with variations in age, gender, cultural context, environment, and camera angles. Finally, we acquire 41,315 video descriptions, comprising 16,145 for normal and 25,170 for anomalous ones.

Multi-style Re-captioning Details. We enhance the diversity of text descriptions in CAViR and reduce the domain gap between synthetic data and real-world benchmarks by augmenting each video with multiple captions of varying styles. Initially, every video in CAViR is equipped with one matched textual description collected from existing datasets or generated by large language models (based on anomaly category tags). Building upon this, we introduce four additional captions per video, each reflecting a distinct style: (1) *UCFCrime-AR style*; (2) *OOPS! style*; (3) *Qwen-generated video description*; and (4) *Qwen-generated keyframe description*. For the first two styles, we first leverage an LLM (see Table 10) to analyze the textual patterns and templates prevalent in the respective source datasets. Based on the extracted stylistic templates (as shown in Table 12), we design tailored prompts (see Table 10) and employ Qwen3-VL-8B-Instruct (Bai et al., 2025) to generate style-consistent descriptions for each video. The latter two styles are directly produced by Qwen3-VL-8B-Instruct (Bai et al., 2025) with general-purpose prompts (see Table 10), capturing either the overall video dynamics or a static snapshot. This multi-style re-captioning strategy yields a rich set of textual annotations that better approximate the variability of real-world data.

B.4 Attribute Annotation Details

For attribute annotation, we employ the Qwen3 (Yang et al., 2025a) model to assign anomaly and normal categories to each CAViR sample without preexisting category labels. Leveraging the video descriptions, we design distinct prompts for anomalous and normal samples to query LLMs for attribute extraction, with detailed prompts provided in Table 13.

Table 12: Text style templates of UCFCrime-AR and OOPS!

UCFCrime-AR style templates
1. [Subject] [verb] [object/location]. 2. [Action 1], and then [action 2]. / [Action 1] and [action 2]. 3. In/At [location], [scene description]. 4. During the day/At night/[time phrase], [event]. 5. There [is/was] a constant flow/stream of [entities] [location]. 6. A [vehicle/object 1] collided/crashed with [vehicle/object 2] [location details]. 7. A [person description] [entered/approached] [location], [action with weapon/threat], and [stole/fled with items]. 8. There was an explosion/fire [location], [consequence description]. 9. [Event 1] causing/and [event 2]. 10. [Actor 1] [action 1], while/and [actor 2] [action 2]. 11. [Entities] [verb] on/in/at [location]. 12. [Perpetrator] attempted [crime] but [failed outcome].
OOPS! style templates
1. [Subject] was/were trying/attempting to [action]. [Failure outcome]. 2. To [infinitive phrase]. [What went wrong]. 3. [Subject] [action verb]. [Unexpected result]. 4. [Gerund phrase]. [Failure description]. 5. [Subject] is/are [verb-ing]. [Failure]. 6. Non-standard constructions, often missing subjects or verbs. 7. Implied goal with explicit failure (or vice versa).

Table 13: Prompts for attribute annotation on anomalous and normal samples.

Attribute Annotation
Goal: Annotating anomalous categories based on anomaly video descriptions **Specific LLMs**: Qwen3-VL-8B-Instruct (Bai et al., 2025) **Instruction**: "Help me categorize text content based on the type of exception or unexpected event. The following English text describes the abnormal or unexpected events that occurred in the video. Can you classify them by event type? Only answer the most likely type of abnormal event. If unable to determine, estimate the most likely type of abnormal or unexpected event. For example: falling, falling into water, slipping, being hit by an object, exploding, colliding, something breaking, jumping, being bitten, losing balance, getting injured, getting wet, being frightened, etc. English text: {*anomaly video description*}"
Goal: Annotating normal types based on normal video descriptions **Specific LLMs**: Qwen3-VL-8B-Instruct (Bai et al., 2025) **Instruction**: "Help me categorize the text content according to the type of event. The following English text describes the events that occurred in the video. Can you classify them by event type? Only answer one most likely event type; if unable to determine, estimate one most likely type. For example: jumping, skateboarding, running, walking, sitting, cycling, dancing, playing, playing hula hoops, grabbing, lifting weights, exercising, riding motorcycles, playing soccer, driving cars, crawling, kissing, etc. English text: {*normal video description*}"