ReactVAU: A Slow-Fast Decoupled Framework for Streaming Video Anomaly Understanding

Chia-Hui Chen[1], Shih-Ying Yeh[1], Fu-En Yang[2], Min-Hung Chen[2], and Shang-Hong Lai[1]

[1] National Tsing Hua University, Hsinchu, Taiwan
[2] NVIDIA, Taipei, Taiwan

Abstract. In this paper, we propose **ReactVAU**, a Slow-Fast Decoupled Framework for real-time streaming Video Anomaly Understanding (VAU). Existing VAU methods rely on offline inference with global temporal sampling, which violates causality and prevents deployment in live surveillance streams. Conversely, general streaming video models satisfy causal access but dilute rare transient anomalies during memory compression and often invoke heavyweight MLLMs uniformly over long normal intervals. ReactVAU addresses this gap with three synergistic components: a lightweight Fast Detection Module based on Spatial Grid Folding (SGF) for continuous anomaly filtering; an Anomaly-Aware Persistent Memory (AAPM) that protects critical visual cues from temporal decay; and a heavyweight Slow Reasoning Module that remains dormant during normal streams and is awakened only by suspicious events for semantic verification and causal description. Extensive experiments on multiple benchmarks demonstrate that ReactVAU operates under strict streaming constraints while simultaneously achieving competitive performance in both anomaly detection and causal reasoning, alongside significantly enhanced computational efficiency by minimizing heavyweight MLLM invocations. Project page is available at https://huiyuiui.github.io/ReactVAU/

Keywords: Video Anomaly Understanding · Streaming Video Understanding · Multimodal LLM · LLM Memory

1 Introduction

Video anomaly detection (VAD) serves as a foundational technology for intelligent surveillance, traffic monitoring, and public safety. Early VAD methodologies inherently operated at the frame level to identify deviations from normal patterns [15, 16, 32, 35, 42, 43, 46]. While effective for localization, they functioned largely as black boxes, providing only a scalar anomaly probability without semantic context regarding what occurred or why it was deemed anomalous. To meet the real-world demand for semantic explainability, recent advancements introduced Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) [1, 9, 47, 56, 61], driving a significant paradigm shift from traditional VAD to comprehensive Video Anomaly Understanding (VAU). However, current top-tier VAU models are inherently constrained by offline inference

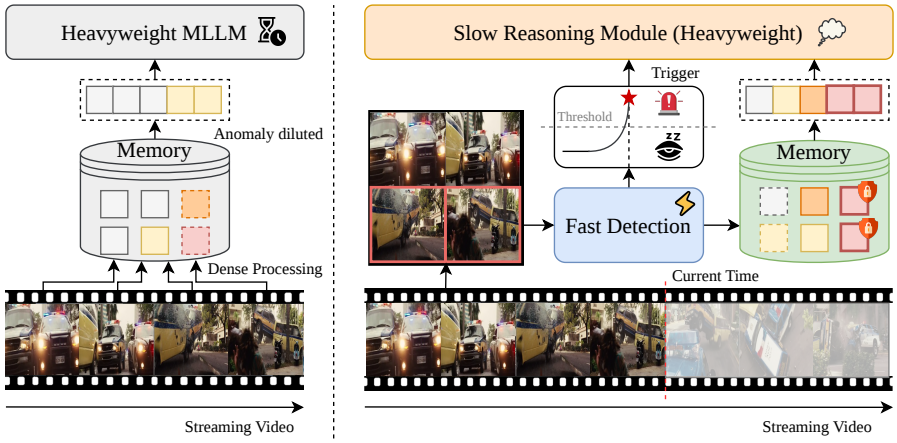

Fig. 1: Conceptual comparison of ReactVAU against general online models. (a) General Online Model: Continuously evaluating dense video streams with heavyweight MLLMs incurs prohibitive computational overhead, and anomaly features are easily diluted during memory compression. (b) ReactVAU: A lightweight Fast Module continuously monitors streams and reactively triggers the Slow Module only upon detection. Concurrently, the Anomaly-Aware Persistent Memory (AAPM) proactively safeguards critical visual cues from temporal decay.

paradigms [14,18,29,55,62,65]. These architectures rely on observing the entire video sequence to perform global temporal sampling or calculate holistic anomaly density distributions prior to language generation. This absolute dependence on future frames breaks causality, rendering them fundamentally incompatible with real-world streaming surveillance environments.

Concurrently, the broader field has seen the emergence of general streaming video understanding frameworks designed to process infinite video streams [8, 20, 49, 64]. By dynamically managing historical context—such as through token merging and persistent memory [31, 39, 57, 60] or KV-cache compression [24, 52]—these models overcome the sequence length limits of offline architectures, enabling continuous long-term video interaction. However, applying these frameworks directly to the anomaly domain exposes two critical flaws. First, the transient nature of real-world anomalies [14, 42] causes their sparse visual features to be easily diluted during continuous context updates [9,57], impairing downstream threat perception. Second, evaluating every streaming frame with a heavyweight MLLM incurs prohibitive computational overhead, hindering real-time responsiveness—a bottleneck that recent decoupled and event-gated architectures have only just begun to address [7,12,38].

To resolve these intertwined challenges, we propose **ReactVAU**, a Slow-Fast Decoupled Framework designed around a reactive event-gated mechanism for Streaming Video Anomaly Understanding. As illustrated in Figure 1, the

core insight is that a single high-capacity MLLM need not process every frame uniformly: normal surveillance footage dominates real streams, while anomalous intervals are rare and short. ReactVAU therefore separates high-frequency visual anomaly filtering from low-frequency semantic reasoning. A lightweight Fast Detection Module continuously monitors the stream, an Anomaly-Aware Persistent Memory (AAPM) preserves transient suspicious evidence while the reasoning model sleeps, and a heavyweight Slow Reasoning Module is awakened only when the Fast module detects a potential anomaly. This design keeps the pipeline causal by construction, reduces redundant MLLM computation on normal content, and still preserves the visual evidence required for fine-grained causal explanation when a reaction is needed.

The main contributions of our work are summarized as follows:

- We propose ReactVAU, a Slow-Fast Decoupled Framework equipped with a reactive event-gated mechanism that triggers heavyweight reasoning only upon detected anomalies, eliminating the reliance on future frames and enabling real-time streaming VAU with reduced computational overhead.
- We introduce Spatial Grid Folding (SGF), transforming short-term temporal anomaly detection into a highly efficient 2D spatial reasoning task that bypasses heavy temporal modeling.
- We design the Anomaly-Aware Persistent Memory (AAPM) mechanism, utilizing an Anomaly Priority Score and a threat-based Anomaly Pool to protect transient abnormal features during continuous memory compression.
- Extensive experiments across multiple benchmarks demonstrate that ReactVAU achieves competitive performance against state-of-the-art offline models while delivering significantly enhanced computational efficiency under a strict causal streaming inference protocol.

2 Related Work

Video Anomaly Detection. Traditional Video Anomaly Detection (VAD) prioritizes event localization via reconstruction-based one-class classification [15, 16, 35, 40] or 3D feature-based Multiple Instance Learning (MIL) [5, 6, 19, 42, 43, 46, 53]. While foundational, most methods require offline feature buffering. Pushing towards real-time applications, recent advancements like REWARD [23] propose an end-to-end architecture that learns directly from raw video segments, bypassing offline feature extraction to achieve highly efficient online detection. Although such low-latency models suit streaming inputs, they operate strictly as semantic black boxes. Outputting only scalar probabilities, they fundamentally lack the causal interpretability essential for comprehensive security analysis.

Video Anomaly Understanding. To bridge this gap, Video Anomaly Understanding (VAU) leverages Multimodal Large Language Models (MLLMs) for text generation. Initial zero-shot methods introduced open-world detection [1, 9, 21, 26, 34, 41, 61] by adapting frozen CLIP models (e.g., VadCLIP [47]), extracting

frame captions for LLMs (LAVAD [56]), or exploiting anomaly-sensitive attention heads (HeadHunt-VAD [4]). For deeper reasoning, state-of-the-art models employ supervised instruction-tuning [14,18,65]. For example, Holmes-VAU [62] utilizes an Anomaly-focused Temporal Sampler to process untrimmed videos, while VADER [11] employs context-aware sampling and relational encoders to explicitly model causal object interactions. Additionally, works like VERA [55] integrate verbalized learning, and agentic paradigms such as PANDA [54] and Unified [29] deploy automated workflows for holistic causal explanations. Despite their descriptive prowess, a critical flaw unites these top-tier methods: strict reliance on offline inference. The necessity to observe entire sequences for global temporal sampling breaks causality, rendering them fundamentally inapplicable to real-world streaming environments where future frames remain unobservable.

Streaming Video Understanding. To process infinite video streams, recent works developed streaming frameworks [7, 8, 20, 22, 31, 48–50, 64]. They manage unbounded context via KV-cache compression [24, 52] or persistent memory [39, 60]. For instance, StreamForest [57] organizes visual histories into a memory forest. To mitigate latency, Dispider [38] and StreamMind [12] introduce disentangled perception-reaction and event-gated cognition. Concurrently, MoniTor [51] deploys MLLMs directly for online anomaly detection. While effective, these models struggle with VAU's practical demands. First, memory compression algorithms heavily penalize older events to maintain fixed bounds, rapidly diluting sparse anomalous features amidst overwhelming normal backgrounds. Second, despite decoupled designs, continuously querying heavyweight MLLMs for frame-level evaluation creates severe computational bottlenecks, hindering true real-time responsiveness.

3 Method

To bridge the conflicting demands of real-time efficiency and deep causal explainability in streaming environments, we present **ReactVAU**, a Slow-Fast Decoupled Framework. As illustrated in Figure 2, ReactVAU strictly operates under a causal streaming protocol, accessing only current and historical visual information, and is structured around three components: **(1) Fast Detection Module** (Sec. 3.2), which continuously filters streaming frames for anomalies without heavy temporal modeling; **(2) Anomaly-Aware Persistent Memory (AAPM)** (Sec. 3.3), which accumulates visual context while shielding transient threat features from memory compression; and **(3) Slow Reasoning Module** (Sec. 3.4), a heavyweight MLLM that remains dormant during normal conditions and awakens only upon anomaly verification to perform fine-grained causal reasoning on the protected memory.

3.1 Preliminary: Continuous Memory Formulation

To process infinite video streams without catastrophic memory overflow, our framework explicitly leverages the StreamForest-7B architecture [57] as the back-

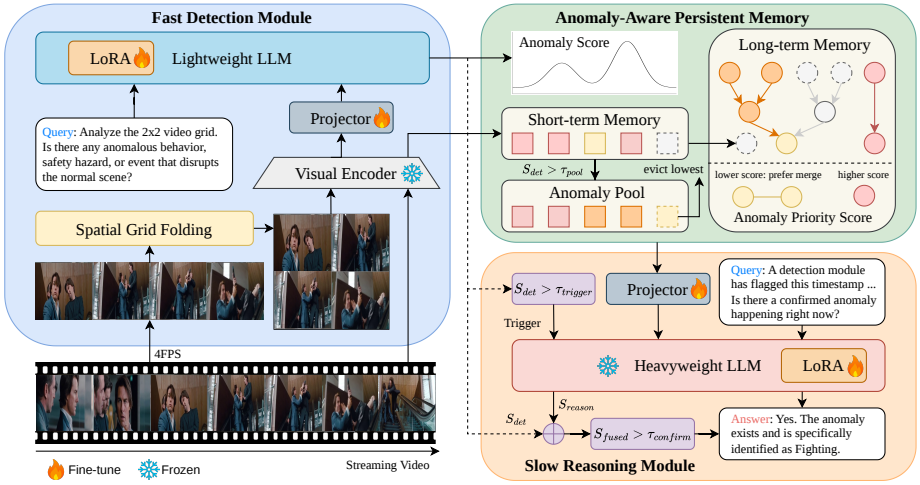

Fig. 2: Architecture of ReactVAU. The proposed Slow-Fast Decoupled Framework comprises three core components: a lightweight Fast Detection module, an Anomaly-Aware Persistent Memory (AAPM), and a heavyweight Slow Reasoning Module. The Fast Module continuously processes video streams via Spatial Grid Folding (SGF) to yield a real-time anomaly score (S_{det}). This score directs the AAPM to protect critical anomalous features from compression decay. Once an anomaly is detected ($S_{det} > \tau_{trigger}$), the dormant Slow Module is awakened to conduct fine-grained causal reasoning on the safeguarded memory sequence, generating a secondary verification score (S_{reason}) and a semantic description. The final detection outcome is dictated by the fused score (S_{fused}).

bone for our heavyweight Slow Reasoning Module. Designed exclusively for streaming comprehension, we formalize its native memory management mechanism into the following continuous tiers:

Real-Time Perception (RTP). RTP preserves a complete set of feature tokens from the current frame ($N_{RTP} = 729$), providing fine-grained spatial information near the current timestamp.

Fine-grained Spatiotemporal Window (FSTW). FSTW stores compressed visual features of recent frames ($N_{ST} = 128$ tokens per frame) within a fixed 12-frame sliding window. When the window capacity is exceeded, overflowing frames are grouped into meta-event nodes and evicted to long-term storage.

Persistent Event Memory Forest (PEMF). PEMF [57] organizes long-term visual history as compressed event nodes ($N_{PEMF} = 64$ tokens per node) under a token quota $C_{max} = 2048$. When this quota is exceeded, PEMF computes a composite penalty $\mathcal{P}_{total}$ from similarity, temporal distance, and merge frequency, and iteratively merges the adjacent node pair with the lowest penalty. This preserves recent and distinctive events while heavily compressing redundant or distant history.

3.2 Fast Detection Module and Spatial Grid Folding (SGF)

Existing VAD models often rely on computationally expensive 3D CNNs [6,43] to extract temporal features. While recent Video LLMs [10,27,36,59] capture comprehensive temporal sequences, their sequence-level processing in a streaming context imposes heavy overhead on computational resources. Inspired by recent findings demonstrating that Vision-Language Models can perform temporal reasoning spatially when frames are stitched together [13,25], we introduce **Spatial Grid Folding (SGF)**. This mapping of temporal dynamics into 2D spatial layouts allows the lightweight VLM to leverage its inherent spatial attention mechanisms, thereby exploiting the powerful priors acquired during large-scale pre-training without the computational burden of heavy temporal modeling.

We utilize PaliGemma2-3B [3] as the Fast Detection backbone. Its visual encoder (SigLIP-So400m [58]) $\mathcal{E}_{vis}$ matches the one used in the heavyweight Slow Reasoning Module, providing consistent visual extraction priors across the framework. To equip the module with anomaly perception capabilities tailored for SGF, the model is fine-tuned for 1 epoch on a specifically constructed **Grid Image Dataset** derived from UCF-Crime [42] and XD-Violence [46], updating only the projector and LoRA [17] weights attached to the LLM. The detailed construction and balancing strategies of this dataset are provided in the Supplementary Material.

During streaming inference at 4 FPS, a sliding window of 1 second (frames F_1 to F_4) is spatially folded into a 2×2 Grid Image, denoted as I_{grid}. This grid is processed by $\mathcal{E}_{vis}$ and mapped to the language layer via a projector Ψ_{det}. Combined with the task prompt $\mathcal{Q}_{det}$, the lightweight language model (LLM_{det}) predicts the binary anomaly target logits $z_{\text{"Yes"}}$ and $z_{\text{"No"}}$. The continuous detection anomaly score S_{det} is derived using a localized Softmax over these targets:

$$S_{det} = \frac{\exp(z_{\text{"Yes"}})}{\exp(z_{\text{"Yes"}}) + \exp(z_{\text{"No"}})} \tag{1}$$

This is mathematically equivalent to a sigmoid over $z_{\text{"Yes"}} - z_{\text{"No"}}$, but follows the native next-token prediction interface of the language model and focuses calibration on the decision boundary between the two target tokens. For a binary grid label $b \in \{0, 1\}$, the same localized distribution is trained with

$$\mathcal{L}_{det} = -b \log S_{det} - (1 - b) \log(1 - S_{det}). \tag{2}$$

The Slow Reasoning Module is awakened only when $S_{det} > \tau_{trigger}$, where $\tau_{trigger}$ is calibrated from the training split as described in the Supplementary Material.

3.3 Anomaly-Aware Persistent Memory (AAPM)

Because the Slow Reasoning Module remains dormant during normal streams, the visual history must be updated independently of heavyweight reasoning. Yet generic streaming memory treats dominant normal backgrounds and rare transient anomalies uniformly, allowing brief anomalous evidence to be compressed

into background events before later retrieval. We introduce **Anomaly-Aware Persistent Memory (AAPM)**, which retains the FSTW/PEMF hierarchy while using S_{det} at three complementary horizons: long-term event protection, high-density evidence retention, and fine-grained current perception.

Anomaly Priority Score. The native PEMF consolidates adjacent event nodes according to similarity (P_s), temporal distance (P_t), and merge frequency (P_m). Although effective for generic streaming content, these criteria cannot distinguish a rare anomaly from ordinary background, so an old but critical event may be repeatedly merged. We introduce an **Anomaly Priority Score (P_a)** that assigns high-threat node pairs an exponential protective weight:

$$P_a(i,j) = \exp\left(\kappa \cdot \frac{S_{det}^{(i)} + S_{det}^{(j)}}{2}\right) \tag{3}$$

where κ is a scaling constant and $S_{det}^{(i)}$ is the Fast Detection score associated with node i when it enters PEMF. This score is added to the native total penalty:

$$\mathcal{P}'_{total}(i,j) = w_s P_s(i,j) + w_t P_t(i,j) + w_m P_m(i,j) + w_a P_a(i,j) \tag{4}$$

where w_s, w_t, w_m inherit the native PEMF weights and w_a controls anomaly protection. Because PEMF merges the pair with the lowest penalty, adding P_a makes anomaly-rich pairs less likely to be consolidated. Normal segments retain the original PEMF behavior, while long-term anomalous events remain semantically distinguishable despite temporal aging.

Anomaly Pool. The Anomaly Priority Score protects event semantics, but PEMF nodes still undergo spatial compression and hierarchical merging. Fine visual details needed for causal verification may therefore be lost even when an event node survives. To preserve such evidence, we introduce an isolated **Anomaly Pool ($\mathcal{M}_{pool}$)** that stores higher-density features from suspicious frames as evidence anchors for the Slow Module. The pool holds eight frames at 128 tokens per frame and lies outside the $C_{max} = 2048$ PEMF quota, preventing anomaly evidence from displacing the normal background required for comparative reasoning. To avoid contamination by minor score fluctuations, only frames satisfying $S_{det} > \tau_{pool}$ enter $\mathcal{M}_{pool}$. Once full, the pool retains the strongest evidence by evicting the frame with the lowest anomaly score:

$$e_{evict} = \arg\min_{k \in \mathcal{M}_{pool}} S_{det}^{(k)} \tag{5}$$

Dynamic Dense Sampling and Real-Time Anomaly Perception. Long-term protection alone cannot recover rapid motion that was never encoded at sufficient temporal resolution. AAPM therefore adapts current perception to the Fast score. During normal intervals ($S_{det} \leq \tau_{trigger}$), it updates memory sparsely

at 1 FPS using the last frame of each one-second grid, avoiding redundant processing of static background. Once $S_{det} > \tau_{trigger}$, it switches to dense perception and independently encodes all four frames as $\mathcal{M}_{RTP}$, providing the Slow Module with local micro-dynamics of the suspected event alongside anomaly-protected historical context. The complete capacity allocation of the assembled memory is provided in Supplementary Material.

3.4 Slow Reasoning Module

When the trigger threshold ($\tau_{trigger}$) is breached, the dormant 7B MLLM is awakened and retrieves the fully constructed visual memory sequence $\mathcal{M}_{seq}$ from the AAPM. This sequence is structured chronologically and hierarchically:

$$\mathcal{M}_{seq} = \mathcal{M}_{PEMF} \oplus \mathcal{M}_{ST} \oplus \mathcal{M}_{pool} \oplus \mathcal{M}_{RTP} \tag{6}$$

where $\mathcal{M}_{ST}$ denotes the short-term component of FSTW and $\mathcal{M}_{RTP}$ denotes the dense RTP tokens produced by AAPM. The raw visual tokens comprising $\mathcal{M}_{seq}$ are generated by the shared vision encoder $\mathcal{E}_{vis}$.

To perform fine-grained secondary verification, the visual memory sequence $\mathcal{M}_{seq}$ is mapped into the language space via the reasoning projector Ψ_{reason} and combined with the reasoning prompt $\mathcal{Q}_{reason}$. The resulting sequence is processed by LLM_{reason}. Following the Fast Module's scoring formulation (Eq. (1)), a secondary anomaly score S_{reason} is computed from the "Yes" and "No" logits.

The final system anomaly score S_{fused} is a weighted fusion of the Fast Detection score (S_{det}) and the Slow Reasoning score (S_{reason}):

$$S_{fused} = \mu \cdot S_{det} + (1 - \mu) \cdot S_{reason} \tag{7}$$

where μ is set to 0.4 based on empirical validation. The slightly larger weight on S_{reason} reflects its deep semantic reasoning over both the current observation and accumulated history, while S_{det} preserves sensitivity to abrupt local changes.

To determine the final system action, we establish a confirmation threshold $\tau_{confirm}$. Only if $S_{fused} > \tau_{confirm}$ is the anomaly verified. Upon confirmation, the Reasoning Module generates the fine-grained anomaly description $\mathcal{D} = \{w_1, w_2, \ldots, w_N\}$ by maximizing the autoregressive likelihood conditioned on the accumulated memory sequence and prompt:

$$P(\mathcal{D} \mid \mathcal{M}_{seq}, \mathcal{Q}_{reason}) = \prod_{n=1}^{N} P(w_n \mid w_{<n}, \mathcal{M}_{seq}, \mathcal{Q}_{reason}) \tag{8}$$

During instruction tuning, the Slow Reasoning Module uses standard full-vocabulary teacher-forced next-token cross-entropy:

$$\mathcal{L}_{reason} = -\sum_{n=1}^{N} \log P(w_n \mid w_{<n}, \mathcal{M}_{seq}, \mathcal{Q}_{reason}). \tag{9}$$

This formulation ensures that the generated narrative is rigorously grounded in the verified visual evidence comprehensively retrieved from the anomaly-protected memory structures.

4 Experiments

4.1 Benchmark Datasets and Evaluation Metrics

To rigorously evaluate the performance of ReactVAU, we conduct experiments across three dataset benchmarks that encompass VAD and VAU tasks:
UCF-Crime [42]: Comprises 1,900 untrimmed surveillance videos covering 13 anomaly categories. We report the frame-level Area Under the Receiver Operating Characteristic Curve (AUC).
XD-Violence [46]: Contains 4,754 untrimmed videos focusing exclusively on violent events. Due to extreme class imbalance, we report the Average Precision (AP) alongside the AUC.
HIVAU-70K [62]: A comprehensive VAU benchmark providing over 70,000 annotations at Clip, Event, and Video granularities. To evaluate semantic explanations, we utilize standard generation metrics: BLEU [37], METEOR [2], ROUGE-L [28], and primarily CIDEr [44], which effectively prioritizes critical rare anomaly descriptors over generic text.

Detailed evaluation protocols, including causal online smoothing and frame-level score broadcasting mechanisms designed to adapt VAD/VAU tasks into streaming constraints, are provided in the Supplementary Material.

4.2 Experimental Results

Video Anomaly Detection (VAD) Performance. We compare ReactVAU against state-of-the-art VAD methods, categorizing them by their operational settings: weakly supervised, offline fine-tuned, offline training-free, and online architectures.

As presented in Tab. 1, traditional offline models benefit immensely from observing the entire video sequence, which inherently inflates their performance by leveraging future context. Notably, when existing robust offline models are forced into online settings (e.g., Online-LAVAD [56]), their performance degrades severely, dropping to 76.06% AUC on UCF-Crime and 52.63% AP on XD-Violence.

In stark contrast, operating under strict streaming constraints without any future information, ReactVAU reaches 88.44% AUC on UCF-Crime and 88.50% AP with 95.25% AUC on XD-Violence. It outperforms existing online methods, such as MoniTor [51] and the closest fine-tuned StreamForest† [57] baseline. Furthermore, our streaming framework remains competitive with the best offline fine-tuned methods (e.g., Holmes-VAD [61]) that access future frames. These results show that SGF and Slow-Fast verification can preserve strong anomaly perception without breaking causality.

Video Anomaly Understanding (VAU) Performance. To evaluate semantic explanation capabilities, we benchmark ReactVAU on the multi-granular HIVAU-70K dataset against generic MLLMs and specialized VAU models. Tab. 2 details the natural language generation metrics across Clip (C), Event (E), and

Table 1: Video Anomaly Detection Performance. Comparison on UCF-Crime and XD-Violence datasets, categorized by inference paradigms and training strategies. StreamForest† is fine-tuned on the HIVAU instruction dataset.

Setting	Method	UCF-Crime	XD-Violence	
		AUC (%)	AP (%)	AUC (%)
Offline Weakly Sup.	CLIP-TSA [21]	87.58	82.19	-
	VadCLIP [47]	**88.02**	**84.51**	-
Offline Fine-Tuned	Holmes-VAD [61]	**89.51**	**90.67**	-
	Holmes-VAU [62]	88.96	87.68	-
Offline Training-Free	LLaVA-1.5 [30]	72.84	50.26	79.62
	LAVAD [56]	80.28	62.01	85.36
	Unified [29]	84.28	68.07	**91.34**
	VERA [55]	86.55	70.54	88.26
	HeadHunt-VAD [4]	**87.03**	**82.63**	-
Online Training-Free	Online-LAVAD [56]	76.06	52.63	76.01
	MoniTor [51]	**82.57**	**55.01**	**79.11**
Online Fine-Tuned	REWARD [23]	86.94	77.71	-
	StreamForest† [57]	85.26	75.92	92.82
	ReactVAU (Ours)	**88.44**	**88.50**	**95.25**

Video (V) levels. We specifically emphasize the CIDEr metric in this analysis because it rigorously assigns higher weights to rare, informative tokens (e.g., specific anomaly behaviors) rather than frequently occurring background words, making it the most critical indicator of exact anomaly understanding.

As the data shows, generic multi-modal LLMs (e.g., Video-LLAMA [59], QwenVL2 [45], InternVL2 [10]) degrade sharply at the Event and Video levels (e.g., InternVL2 scores 0.022 on CIDEr-E), reflecting the difficulty of retaining transient anomaly evidence over long contexts. Conversely, at the Clip level, offline VADER performs best because its global sampler first observes the complete input and concentrates a fixed frame budget on anomaly-dense regions. ReactVAU instead ingests causally with one AAPM update per second, leaving short clips with less accumulated evidence. As the Event and Video contexts grow, AAPM preserves sufficient anomaly evidence for ReactVAU to achieve the best scores across all reported long-range metrics.

Crucially, the comparison between the native StreamForest and our ReactVAU framework under the same instruction-tuning setting highlights the impact of temporal feature dilution in general streaming architectures. ReactVAU consistently improves Event- and Video-level VAU. These gains support AAPM's role in preserving critical anomaly evidence from aggressive memory compression. This preservation ensures that the Slow Reasoning Module has access to anomaly-focused evidence and long-range context, enabling it to generate highly

Table 2: Video Anomaly Understanding Performance on HIVAU-70K. Methods are evaluated across three temporal granularities: clip (C), event (E), and video (V). Models denoted with † are fine-tuned on the HIVAU instruction dataset.

Method	BLEU			CIDEr			METEOR			ROUGE		
	C	E	V	C	E	V	C	E	V	C	E	V
Generic MLLMs (Zero-shot)												
Video-ChatGPT [36]	0.152	0.068	0.066	0.033	0.011	0.013	0.102	0.069	0.044	0.153	0.048	0.079
Video-LLAMA [59]	0.151	0.079	0.104	0.024	0.014	0.017	0.112	0.076	0.057	0.156	0.067	0.090
Video-LLAVA [27]	0.164	0.046	0.055	0.032	0.009	0.013	0.097	0.022	0.014	0.132	0.023	0.045
LLAVA-Next-Video [63]	0.435	0.091	0.120	0.102	0.015	0.031	0.117	0.085	0.096	0.198	0.080	0.106
QwenVL2 [45]	0.312	0.082	0.155	0.044	0.020	0.044	0.133	0.092	0.112	0.163	0.081	0.137
InternVL2 [10]	0.331	0.101	0.145	0.052	0.022	0.035	0.141	0.095	0.101	0.182	0.102	0.122
NVILA [33]	0.610	0.340	0.283	0.261	0.154	0.098	0.157	0.096	0.072	0.273	0.218	0.198
Offline VAU Models												
Holmes-VAU† [62]	0.913	0.804	0.566	0.467	1.519	1.437	0.190	0.165	0.121	0.329	0.370	0.355
VADER† [11]	**1.266**	1.246	1.268	**1.040**	1.763	1.812	**0.247**	0.216	0.164	**0.429**	0.463	0.446
Streaming VAU Models												
StreamForest [57] (Zero-shot)	0.422	0.252	0.271	0.138	0.076	0.070	0.128	0.097	0.094	0.233	0.175	0.180
StreamForest† [57]	1.245	1.334	1.320	0.947	1.999	1.931	0.243	0.221	0.178	0.427	0.483	0.456
ReactVAU† (Ours)	1.150	**1.366**	**1.337**	0.920	**2.032**	**2.016**	0.222	**0.222**	**0.179**	0.401	**0.488**	**0.465**

accurate, causally grounded narratives regardless of the video's total streaming duration.

Effectiveness of Slow-Fast Decoupling (Efficiency Analysis). A critical bottleneck in streaming Video Anomaly Understanding is the prohibitive computational cost of evaluating every frame with a heavyweight MLLM. Tab. 3 reports profiling on a single NVIDIA H100-80GB GPU, demonstrating how our Slow-Fast decoupling fundamentally resolves this issue. Evaluated on the UCF-Crime test set—where anomalies naturally occur in approximately 43% of the segments, ReactVAU successfully leverages the lightweight Fast Module to filter redundant nominal backgrounds. This reduces heavyweight 7B-parameter LLM inference queries from 34,670 to 15,955, achieving a 54.0% reduction in computational workload.

More importantly, this architectural decoupling directly translates to real-time responsiveness. While the baseline StreamForest (which operates entirely on a heavy 7B MLLM backbone) suffers from a constant 216.1 ms/query delay, ReactVAU efficiently restricts the heavyweight LLM to event verification, operating at a mere 22.1 ms/query during normal states. Consequently, the empirical weighted average streaming latency on UCF-Crime drops to 98.3 ms/query, cutting the total delay by more than half. Supplementary Material further profiles normal and triggered paths on an RTX 4090-24GB under a one-second streaming budget to better demonstrate the feasibility of real-world deployment.

Real-world Scalability Projection: It is crucial to note that the UCF-Crime test set is highly condensed with anomalous events. In actual real-world surveillance deployment, true anomalies are exceedingly rare and typically account for less than 5% to 10% of the continuous video stream [42]. As projected in the bottom section of Tab. 3, under a realistic 5% anomaly rate, the trigger fre-

Table 3: Efficiency and Scalability Analysis. Evaluated on the UCF-Crime test set using a single NVIDIA H100-80GB GPU. By employing the lightweight Fast Module as an active gatekeeper, ReactVAU significantly minimizes the redundant invocations of the heavyweight Slow Module. The weighted average latency is calculated as $L_{avg} = L_{normal} \times (1 - \text{Rate}_{anomaly}) + L_{anomaly} \times \text{Rate}_{anomaly}$, illustrating theoretical performance under realistic sparse anomaly occurrence rates [42].

Method / Scenario	7B LLM Overhead		Latency (ms/query)		
	Queries ↓	Reduction	Normal	Anomaly	Weighted Avg. ↓
Empirical Evaluation (UCF-Crime Test Set, ∼43% Anomaly Rate)					
StreamForest [57] (Slow-only)	34,670	-	216.1	216.1	216.1
ReactVAU (Ours)	15,955	54.0%	22.1	269.5	98.3
Theoretical Projection (Real-World Surveillance Deployment)					
ReactVAU (10% Anomaly Rate)	∼3,467	90.0%	22.1	269.5	39.9
ReactVAU (5% Anomaly Rate)	∼1,734	**95.0%**	22.1	269.5	**31.0**

quency of our Slow Module plummets. This theoretical projection indicates that ReactVAU could reduce 7B LLM queries by up to 95.0%, driving the average operational latency down to 31.0 ms/query. This confirms that ReactVAU is not only empirically effective but highly scalable for infinite real-world streams.

4.3 Qualitative Results

Evaluating comprehensive anomaly events requires understanding at multiple semantic granularities, from localized clip-level actions to holistic video-level summaries. Conventional offline VAU models typically adapt global temporal sampling to each specific question. This necessitates separate full-video processing passes for every query, which incurs massive redundant computation and fundamentally breaks causality.

In contrast, ReactVAU performs VAU within a causal streaming pipeline. As illustrated in Fig. 3, ReactVAU processes the stream sequentially without observing future frames, while the AAPM protects crucial anomalous evidence (e.g., escaping rioters and police formations) from being progressively diluted within normal streaming context. The resulting continuously updated memory can be queried at the corresponding timestamp, such as a clip-level query at $t = 20$s or a video-level summary at $t = 200$s, allowing ReactVAU to capture both transient localized actions and the global event narrative without resampling or re-encoding the full video.

4.4 Ablation Study

To rigorously validate the effectiveness of our proposed modules, we conduct an incremental ablation study (Tab. 4). Given the highly imbalanced nature of real-world anomalies, we primarily focus on the Average Precision (AP) metric on XD-Violence, which strictly evaluates a model's ability to suppress false positives—a critical capability enhanced by our framework.

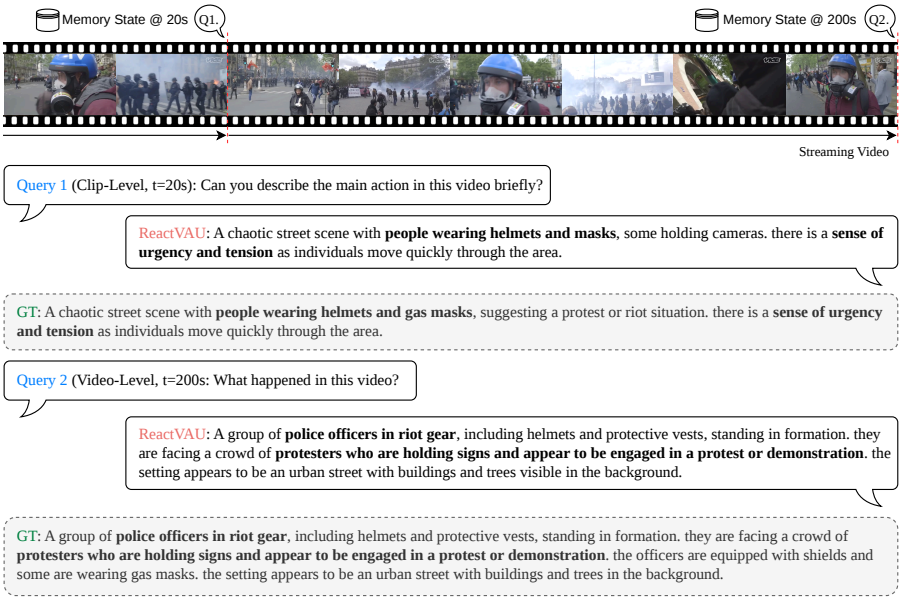

Fig. 3: Qualitative streaming VAU result. ReactVAU preserves anomaly evidence in AAPM while processing the stream causally. Each query is answered using the memory state available at its corresponding timestamp, e.g., 0–20s memory for a clip-level query at $t = 20$s and 0–200s memory for a video-level summary at $t = 200$s, without question-specific global resampling.

Effectiveness of Spatial Grid Folding (SGF): The upper block isolates the Fast Module. Transitioning PaliGemma2-3B from single-frame inputs to zero-shot grid images disrupts its pre-trained spatial priors, causing the AP to drop (53.71%→43.42%). However, upon fine-tuning on our Grid Image Dataset, the SGF mechanism achieves a massive AP surge to 79.55%. This validates that properly adapted Vision-Language Models can perform robust temporal anomaly perception within a folded 2D layout.

Evolution from Streaming Baseline to ReactVAU: The lower block contrasts the continuous memory baseline against our architectural evolution. The zero-shot StreamForest struggles severely with anomaly semantics (49.24% AP), though domain fine-tuning elevates it to 75.92% AP. Building upon this, introducing our Slow-Fast Decoupled Framework yields immediate gains. By leveraging the Fast Module (with SGF) as a rapid filter and the Slow Module as a rigorous semantic gatekeeper, this dual-verification mechanism effectively eradicates false alarms, driving the AP up to 86.90% when both modules are tuned.

Crucially, integrating the AAPM allows the lightweight Fast Module to directly govern evidence retention, effectively safeguarding transient abnormal features. This integration pushes the streaming performance to a highly competitive 88.50% AP (95.25% AUC) on XD-Violence and 88.44% AUC on UCF-

Table 4: Ablation Study. Incremental validation of key framework components. **PaliGemma2**: PaliGemma2-3B, **StreamForest**: StreamForest-7B, **SGF**: Spatial Grid Folding. The strictly causal online smoothing mechanism is uniformly applied across all streaming evaluations.

Method / Incremental Component	UCF-Crime	XD-Violence	
	AUC	AP	AUC
Fast Module Analysis (PaliGemma2 Base)			
Single Frame (Zero-shot)	74.42	53.71	79.71
SGF (Zero-shot)	69.45	43.42	75.44
SGF (Fine-tuned)	85.37	79.55	91.18
ReactVAU Architecture Evolution (StreamForest Base)			
Baseline (Zero-shot)	78.09	49.24	81.94
Baseline (Fine-tuned)	85.26	75.92	92.82
+ Slow-Fast Decoupling (Fast Tuned)	85.66	80.59	91.34
+ Slow-Fast Decoupling (Both Tuned)	87.17	86.90	94.16
+ AAPM (Ours)	**88.44**	**88.50**	**95.25**

Crime. These consistent improvements across anomaly benchmarks underscore the architectural synergy between decoupled verification and anomaly-protected memory, highlighting that both are structurally important for infinite stream understanding.

Threshold Trade-off: Figure 4 illustrates the Pareto front between computational cost and accuracy by modulating the trigger threshold ($\tau_{trigger}$). While lower thresholds increase sensitivity at the cost of higher overhead, higher thresholds reduce latency but risk missing subtle anomalies. A well-calibrated threshold successfully strikes an optimal balance, effectively filtering out the vast majority of redundant nominal frames while ensuring that critical threat deviations are not overlooked. This optimization significantly enhances the overall system utility and demonstrates the extreme flexibility of ReactVAU in adapting to varying hardware constraints. Cross-dataset threshold transfer is reported in Supplementary Material.

Score Fusion Strategy: Table 5 evaluates the fusion of the Fast (S_{det}) and Slow (S_{reason}) anomaly scores. Our empirical results reveal that purely overriding the Fast score with the reasoning MLLM's prediction (Replace strategy) yields suboptimal results, dropping the AUC to 87.28%. Although the Slow Module receives complete dense frames upon activation, its objective heavily prioritizes long-form semantic reasoning over the entire historical sequence. This broad contextual focus inadvertently smooths out instantaneous high-frequency threat spikes. Conversely, the Fast Module is exclusively calibrated for short-term spatial perception, maintaining extreme sensitivity to sudden visual disruptions. Consequently, an adaptive fusion approach falls short of optimal synergy, whereas a complementary weighted fusion ($S_{fused} = 0.4S_{det} + 0.6S_{reason}$) optimally combines localized reactive

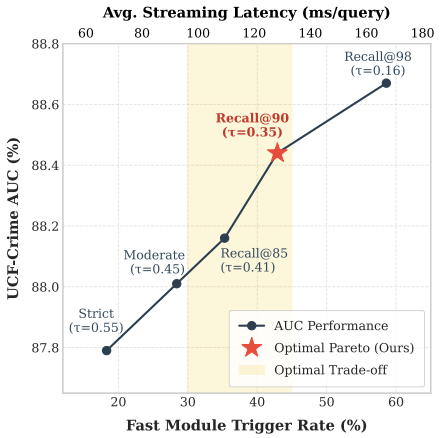

Fig. 4: Efficiency vs. Accuracy. Pareto trade-off analysis on UCF-Crime demonstrating the optimization of the trigger threshold.

Table 5: Score Fusion Strategies. Ablation on late fusion. For compact formulas, D, R, and F denote the full paper notation S_{det}, S_{reason}, and S_{fused}, respectively. Purely replacing the fused score with the Slow score (Replace) fails to capture temporal dynamics, while Weighted fusion combines the spatial sensitivity of Fast Detection and semantic verification of Slow Reasoning.

Strategy (Formulation)	**AUC (%)**
Replace ($F = R$)	87.28
Adaptive ($F = (1 - R)D + R^2$)	88.08
Weighted ($F = 0.3D + 0.7R$)	88.17
Weighted ($\mathbf{F = 0.4D + 0.6R}$)	**88.44**
Weighted ($F = 0.5D + 0.5R$)	88.26

5 Conclusion

In this paper, we introduced ReactVAU, a Slow-Fast Decoupled Framework for streaming Video Anomaly Understanding that reconciles strict causal access, low-latency operation, and high-capacity semantic reasoning. ReactVAU continuously monitors incoming streams with a lightweight Fast Detection Module built on Spatial Grid Folding, preserves transient threat evidence through Anomaly-Aware Persistent Memory, and reactively awakens a heavyweight Slow Reasoning Module only when semantic verification is needed. Extensive evaluations on UCF-Crime, XD-Violence, and HIVAU-70K demonstrate that ReactVAU remains competitive with offline methods that can observe future frames while substantially reducing redundant heavyweight MLLM invocations. Beyond improving accuracy alone, ReactVAU improves the accuracy–efficiency Pareto frontier for causal streaming VAU, offering a practical path toward deploying semantic anomaly reasoning in continuous real-world monitoring systems.

Acknowledgements

This work is supported by the NVIDIA Taiwan AI Research & Development Center (TRDC).

References

1. Ahn, S., Jo, Y., Lee, K., Kwon, S., Hong, I., Park, S.: Anyanomaly: Zero-shot customizable video anomaly detection with lvlm (2025), `https://arxiv.org/abs/2503.04504`, accessed 28 June 2026

2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Engin, M., Bauer, M., Baldassarre, F., Natsev, P., Houlsby, N., Pang, R., Soricut, R.: Paligemma: A versatile 3b vlm for transfer (2024), https://arxiv.org/abs/2407.07726, accessed 28 June 2026
4. Cai, Z., Li, F., Zheng, Z., Bi, H., He, L.: Headhunt-vad: Hunting robust anomaly-sensitive heads in mllm for tuning-free video anomaly detection (2025), https://arxiv.org/abs/2512.17601, accessed 28 June 2026
5. Cao, C., Lu, Y., Wang, P., Zhang, Y.: A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation (2023), https://arxiv.org/abs/2305.13611, accessed 28 June 2026
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
7. Chatterjee, D., Remelli, E., Song, Y., Tekin, B., Mittal, A., Bhatnagar, B., Camgöz, N.C., Hampali, S., Sauser, E., Ma, S., Yao, A., Sener, F.: Memory-efficient streaming videollms for real-time procedural video understanding (2025), https://arxiv.org/abs/2504.13915, accessed 28 June 2026
8. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video (2024), https://arxiv.org/abs/2406.11816, accessed 28 June 2026
9. Chen, Y., Liu, J., Fan, R., Li, Y., Chang, C., Zhao, S., Fok, W.W.T., Qi, X., Wu, Y.C.: Aligning effective tokens with video anomaly in large language models (2025), https://arxiv.org/abs/2508.06350, accessed 28 June 2026
10. Chen, Z., Wang, W., Cao, Y., Liu, Y., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024), https://arxiv.org/abs/2412.05271, accessed 28 June 2026
11. Cheng, Y., Lin, Y.H., Chen, M.H., Yang, F.E., Lai, S.H.: Vader: Towards causal video anomaly understanding with relation-aware large language models (2025), https://arxiv.org/abs/2511.07299, accessed 28 June 2026
12. Ding, X., Wu, H., Yang, Y., Jiang, S., Bai, D., Chen, Z., Cao, T.: Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition (2025), https://arxiv.org/abs/2503.06220, accessed 28 June 2026
13. Doorenbos, L., Spurio, F., Gall, J.: Video panels for long video understanding (2025), https://arxiv.org/abs/2509.23724, accessed 28 June 2026
14. 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 (2024), https://arxiv.org/abs/2405.00181, accessed 28 June 2026
15. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1705–1714 (2019)

16. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 733–742 (2016)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685, accessed 28 June 2026
18. Huang, C., Wang, B., Wen, J., Liu, C., Wang, W., Shen, L., Cao, X.: Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought (2025), https://arxiv.org/abs/2505.19877, accessed 28 June 2026
19. Huang, Y., Li, C., Zhang, H., Lin, Z., Lin, Y., Liu, H., Li, W., Liu, X., Gao, J., Huang, Y., Ding, X., Yuan, Y.: Track any anomalous object: A granular video anomaly detection pipeline (2025), https://arxiv.org/abs/2506.05175, accessed 28 June 2026
20. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: Ovbench and videochat-online (2025), https://arxiv.org/abs/2501.00584, accessed 28 June 2026
21. Joo, H.K., Vo, K., Yamazaki, K., Le, N.: Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection (2023), https://arxiv.org/abs/2212.05136, accessed 28 June 2026
22. Kang, H., Park, Y., Yoo, Y., Choi, Y., Kim, S.J.: Open-ended hierarchical streaming video understanding with vision language models (2025), https://arxiv.org/abs/2509.12145, accessed 28 June 2026
23. Karim, H., Doshi, K., Yilmaz, Y.: Real-time weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6848–6856 (2024)
24. Kim, M., Shim, K., Choi, J., Chang, S.: Infinipot-v: Memory-constrained kv cache compression for streaming video understanding (2025), https://arxiv.org/abs/2506.15745, accessed 28 June 2026
25. Kim, W., Choi, C., Lee, W., Rhee, W.: An image grid can be worth a video: Zero-shot video question answering using a vlm (2024), https://arxiv.org/abs/2403.18406, accessed 28 June 2026
26. Lee, H., Kim, H., Kim, I.J., Choi, Y.: Flashback: Memory-driven zero-shot, real-time video anomaly detection (2025), https://arxiv.org/abs/2505.15205, accessed 28 June 2026
27. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
28. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text summarization branches out pp. 74–81 (2004)
29. Lin, D., Qu, M., Han, K., Jiao, J., Jin, X., Wei, Y.: A unified reasoning framework for holistic zero-shot video anomaly analysis (2025), https://arxiv.org/abs/2511.00962, accessed 28 June 2026
30. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023), https://arxiv.org/abs/2310.03744, accessed 28 June 2026
31. Liu, J., Yu, Z., Lan, S., Wang, S., Fang, R., Kautz, J., Li, H., Alvare, J.M.: Streamchat: Chatting with streaming video (2025), https://arxiv.org/abs/2412.08646, accessed 28 June 2026

32. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6536–6545 (2018)
33. Liu, Z., et al.: Nvila: Efficient frontier visual language models. arXiv preprint arXiv:2412.04468 (2024), https://arxiv.org/abs/2412.04468, accessed 28 June 2026
34. Liu, Z., Wu, X., Wu, J., Wang, X., Yang, L.: Language-guided open-world video anomaly detection under weak supervision (2026), https://arxiv.org/abs/2503.13160, accessed 28 June 2026
35. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision. pp. 2720–2727 (2013)
36. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (2024)
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
38. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction (2025), https://arxiv.org/abs/2501.03218, accessed 28 June 2026
39. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models (2024), https://arxiv.org/abs/2405.16009, accessed 28 June 2026
40. Reiss, T., Hoshen, Y.: Video anomaly detection via semantic attributes (2024), https://openreview.net/forum?id=mhtD74Jgyw, accessed 28 June 2026
41. Shao, Y., He, H., Li, S., Chen, S., Long, X., Zeng, F., Fan, Y., Zhang, M., Yan, Z., Ma, A., Wang, X., Tang, H., Wang, Y., Li, S.: Eventvad: Training-free event-aware video anomaly detection (2025), https://arxiv.org/abs/2504.13092, accessed 28 June 2026
42. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4771–4780 (2018)
43. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
44. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
45. Wang, P., Wang, S., Fan, Z., Fan, Y., Dang, K., Ren, X., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024), https://arxiv.org/abs/2409.12191, accessed 28 June 2026
46. 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: European conference on computer vision. pp. 322–339. Springer (2020)
47. 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 (2023), https://arxiv.org/abs/2308.11681, accessed 28 June 2026

48. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge (2025), https://arxiv.org/abs/2501.13468, accessed 28 June 2026
49. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams (2025), https://arxiv.org/abs/2510.09608, accessed 28 June 2026
50. Yan, Y., Xu, J., Di, S., Liu, Y., Shi, Y., Chen, Q., Li, Z., Huang, Y., Xie, W.: Learning streaming video representation via multitask training (2025), https://arxiv.org/abs/2504.20041, accessed 28 June 2026
51. Yang, S., Feng, Y., Liu, Y., Zhang, J., Qin, J.: Monitor: Exploiting large language models with instruction for online video anomaly detection (2025), https://arxiv.org/abs/2510.21449, accessed 28 June 2026
52. Yang, Y., Zhao, Z., Shukla, S.N., Singh, A., Mishra, S.K., Zhang, L., Ren, M.: Streammem: Query-agnostic kv cache memory for streaming video understanding (2025), https://arxiv.org/abs/2508.15717, accessed 28 June 2026
53. Yang, Z., Radke, R.: Context-aware video anomaly detection in long-term datasets (2024), https://arxiv.org/abs/2404.07887, accessed 28 June 2026
54. Yang, Z., Gao, C., Shou, M.Z.: Panda: Towards generalist video anomaly detection via agentic ai engineer (2025), https://arxiv.org/abs/2509.26386, accessed 28 June 2026
55. Ye, M., Liu, W., He, P.: Vera: Explainable video anomaly detection via verbalized learning of vision-language models (2025), https://arxiv.org/abs/2412.01095, accessed 28 June 2026
56. Zanella, L., Menapace, W., Mancini, M., Wang, Y., Ricci, E.: Harnessing large language models for training-free video anomaly detection (2024), https://arxiv.org/abs/2404.01014, accessed 28 June 2026
57. Zeng, X., Qiu, K., Zhang, Q., Li, X., Wang, J., Li, J., Yan, Z., Tian, K., Tian, M., Zhao, X., Wang, Y., Wang, L.: Streamforest: Efficient online video understanding with persistent event memory (2025), https://arxiv.org/abs/2509.24871, accessed 28 June 2026
58. Zhai, X., Basilico, B.M., Oliver, A., Puyuel, J.B., Eysenbach, M., et al.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
59. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2023)
60. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Dai, J., Jin, X.: Flash-vstream: Memory-based real-time understanding for long video streams (2024), https://arxiv.org/abs/2406.08085, accessed 28 June 2026
61. Zhang, H., Xu, X., Wang, X., Zuo, J., Han, C., Huang, X., Gao, C., Wang, Y., Sang, N.: Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm (2024), https://arxiv.org/abs/2406.12235, accessed 28 June 2026
62. 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 (2025), https://arxiv.org/abs/2412.06171, accessed 28 June 2026
63. Zhang, Y., Li, B., Liu, H., Lee, Y.J., et al.: Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/ (2024), accessed 28 June 2026

64. Zhou, X., Arnab, A., Buch, S., Yan, S., Myers, A., Xiong, X., Nagrani, A., Schmid, C.: Streaming dense video captioning (2024), https://arxiv.org/abs/2404.01297, accessed 28 June 2026
65. Zhu, L., Chen, Q., Shen, X., Cun, X.: Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning (2025), https://arxiv.org/abs/2505.23504, accessed 28 June 2026