STAY SILENT, WARN TIMELY: PERSISTENT VIDEO AGENTS FOR ONLINE ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In safety-critical applications, anomalies must be detected from incoming video in time to support intervention. Recent video-language interaction models can process streaming observations and decide when to respond, offering a foundation for persistent video anomaly detection (VAD). Yet online VAD is still commonly evaluated through short-clip scores or isolated alarms, leaving the session-level behavior required for continuous monitoring largely unmeasured. A practical monitor must maintain one query, remain silent during normal activity, issue timely evidence-grounded warnings, and continue after responding. We formulate this setting as **Timely Anomaly Interaction**, in which a model observes only the available video prefix and produces a trajectory of silence and response decisions. We introduce TAI-VAD, a benchmark and training set covering single-event, multi-event, and long-video sessions. It provides temporal anomaly intervals and timestamp-aligned causal descriptions, and evaluates warning timeliness, anomaly coverage, normal-time false alarms, and post-response monitoring. Experiments with open video interaction models reveal that general interaction ability does not reliably transfer to persistent anomaly monitoring: greater anomaly sensitivity often coincides with more false alarms, while conservative models miss events. To establish a task-specific baseline, we propose PACT, which a supervised fine-tuning method to adapt a pretrained interaction model. PACT achieves the highest Single-event AUROC (0.711) and improves later-event recall over its backbone in Multi-event sessions. Together, our task formulation, benchmark, and training baseline enable systematic study of persistent video anomaly monitoring.

1 INTRODUCTION

Video anomaly detection (VAD) aims to identify abnormal events in video and serves as a key component of safety-critical systems such as intelligent video surveillance and autonomous driving Hashimoto et al. (2022); Acsintoae et al. (2022); Jia et al. (2025); Zou et al. (2026); Bogdoll et al. (2024). Existing methods are commonly categorized as offline or online according to the visual context available at inference time. Offline methods analyze complete videos through sampling frames to detect anomalies or localize abnormal events (Sultani et al., 2018; Acsintoae et al., 2022). Online methods, in contrast, operate only on the observed video prefix. Early online methods processed incoming frames of short clips to produce anomaly scores, including end-to-end real-time detectors (Karim et al., 2024). More recent approaches incorporate temporal memory and vision-language models for streaming anomaly prediction, detection, and analysis (Yang et al., 2025a;b).

As multimodal agents are increasingly deployed in open-ended environments Tang et al. (2024); Wu et al. (2024), VAD must move beyond one-shot detection or retrospective analysis toward persistent monitoring. A deployable anomaly-monitoring agent should maintain a single monitoring objective throughout a continuous video stream, remain silent during normal activity, issue timely and evidence-grounded warnings when anomalies occur, and resume monitoring after each response. Existing online methods typically focus on frame scoring, event detection, or analysis, without explicitly modeling this complete monitoring loop. Consequently, persistent monitoring over an entire video session remains insufficiently specified and has not been comprehensively evaluated.

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.

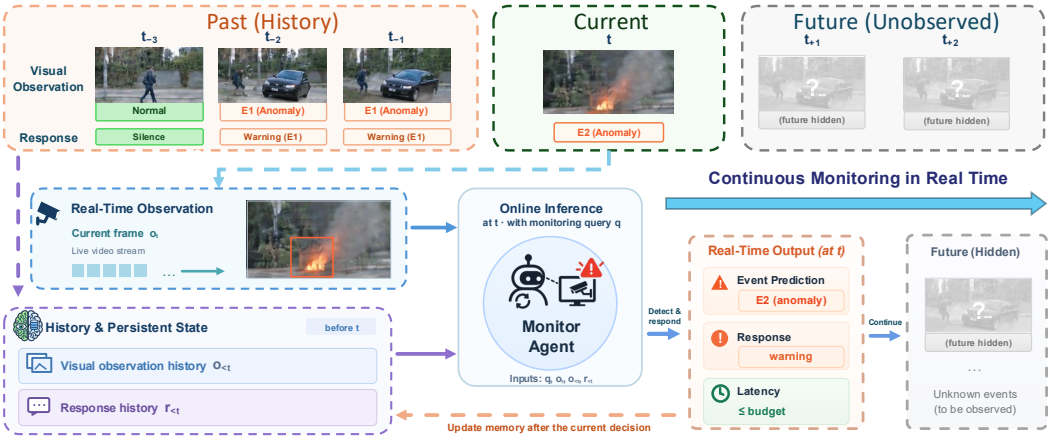

Figure 1: Persistent anomaly monitoring as causal interaction. A single query is registered once and persists throughout the session; the agent maintains state and memory, remains silent during normal activity, and issues evidence-grounded warnings for multiple events without access to future frames.

As Figure 1 shows, a persistent anomaly-monitoring agent must do more than detect a single event: it must remain silent through normal periods, on-time warning for anomaly and potentially detect for later event. Existing benchmarks cover only parts of this setting. ProactiveVideoQA (Wang et al., 2025b) includes a small anomaly-realted QA subset, while AssistPDA (Yang et al., 2025b) studies anomaly prediction, detection, and analysis. Neither of them comprehensively evaluates single-event, multi-event, and long-video monitoring under one session protocol. The detailed comparison can be found in Appendix Table 6, which summarizes how our work fill the gap.

We argue that evaluation for persistent anomaly-monitoring should capture the complete response trajectory for comprehensively evaluation over a video session, including timely warnings, sustained coverage throughout anomaly, silence during normal intervals. Existing methods evaluate only subsets of these behaviors and do not assess them jointly under continuous observation (Wang et al., 2025b; Yang et al., 2025b). To address these gaps and enable comprehensive evaluation of persistent monitoring, we formulate online VAD as **Timely Anomaly Interaction**, a session persistent monitoring task, and introduce TAI-VAD, a benchmark spanning single-event, multi-event, and long-video scenarios. We design a set of evaluation metrics to measure the full monitoring behavior instead of single detection decision.

We extensively evaluate a broad range of pretrained interaction models on this benchmark. Despite their broad semantic knowledge for video understanding, direct transfer does not guarantee persistent monitoring: these models may recognize anomalous content in during observations, yet fail to sustain silence during normal intervals, issue timely warnings, or maintain an appropriate monitoring state after responding. These limitations motivate **PACT** (*Persistent Anomaly-aware Causal Tuning*), a single-stage post-training framework initialized from JoyAI-VL-Interaction (Yao et al., 2026). The key insight is that persistent anomaly monitoring requires learning a sequential decision policy, rather than merely recognizing anomalous visual content. PACT therefore performs supervised fine-tuning, training the model to condition each response on the causally available video prefix and interaction history. The supervision teaches the model when to remain silent, when to issue a timely warning, and how to maintain the monitoring state throughout a session. In this way, PACT transforms a pretrained video-language model into an anomaly-monitoring agent that follows the desired causal action and warning protocol.

Overall, our contributions can be summarized:

- We formulate **Timely Anomaly Interaction**, a session task in which a single monitoring query governs causal observation throughout a complete video session, requiring the model to coordinate silence, warning, and continued monitoring.
- We introduce TAI-VAD, a dataset and benchmark for single-event, multi-event, and long-video monitoring. Its temporal annotations and timestamp-aligned causal descriptions support training and evaluation (Section 3).

- We extensively evaluate a broad range of pretrained video interaction models on TAI-VAD, revealing substantial gaps in persistent monitoring. Based on this analysis, we propose PACT, a post-training framework that teaches pretrained models to make causal monitoring decisions over continuous video streams. PACT improves persistent-monitoring performance to varying degrees across different tasks and evaluation metrics.

2 Related Work

2.1 Offline and online video anomaly detection

VAD is traditionally studied in offline and online settings. Offline methods receive a complete video and estimate anomaly scores or localize abnormal segments from sampled frames or clips, potentially using future context (Sultani et al., 2018; Liu et al., 2018; Yao et al., 2022; Acsintoae et al., 2022; Liu et al., 2024). Online methods seek decisions before video completion and restrict each decision to currently available observations; in practice, many still use fixed-length clips or sliding windows as the decision unit and emit local scores or thresholded alarms. Representative systems include REWARD (Karim et al., 2024) for low-latency weakly supervised detection, MoniTor (Yang et al., 2025a) for instruction-guided textual-memory scoring, and AssistPDA (Yang et al., 2025b) for streaming prediction, detection, and proactive surveillance responses. Despite these advances, online VAD remains centered on local or first-alarm outputs rather than a persistent trajectory that includes normal exposure and multiple events.

2.2 Video-language anomaly understanding and interaction models

Video-language models extend VAD with semantic descriptions, causal reasoning, and open-vocabulary interpretation through prompting and post-training (Zanella et al., 2024; Yang et al., 2024; Lv & Sun, 2024; Zhang et al., 2024; Zou et al., 2026; Huang et al., 2025; Zhu et al., 2025; Ye et al., 2025; Tian et al., 2025). Recent benchmarks evaluate contextual anomaly understanding and conversational explanation (Yuan et al., 2024; Yu et al., 2026; Gani et al., 2025; Zhang et al., 2025), while MoniTor (Yang et al., 2025a) adds textual memory to online anomaly scoring. Yet online VAD commonly decouples anomaly detection, semantic reasoning, and user-facing response, limiting causal evidence use and end-to-end causal monitoring. ProactiveVideoQA (Wang et al., 2025b), JoyAI-VL-Interaction (Yao et al., 2026), VideoChat3 (Li et al., 2026), MMDuet2 (Wang et al., 2025a), and AssistPDA (Yang et al., 2025b) provide initial studies of time-aware response, prefix-restricted observation, and stateful streaming interaction. These efforts provide useful interaction primitives, but do not establish an anomaly-specific protocol for timely warnings, false-alarm control under normal exposure, multi-event re-arming, and post-response monitoring. Our work addresses this gap with a single-query causal monitoring task, a benchmark, and post-training for evidence-grounded interaction (see detailed comparison in Appendix Table 6).

3 TAI-VAD Dataset and Benchmark

Existing VAD benchmarks primarily target video anomaly recognition or time-aware question answering, rather than persistent monitoring over complete video sessions (Sultani et al., 2018; Gani et al., 2025; Wang et al., 2025b). They therefore provide limited coverage of normal intervals, repeated anomaly segments, and monitoring after an earlier response. We introduce TAI-VAD to evaluate these behaviors across single-event, multi-event, and long-video sessions. We curate TAI-VAD exclusively from publicly available datasets and annotations; representative sources include UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), ECVA Du et al. (2024), and MSAD (Zhu et al., 2024), while the complete source inventory and per-source statistics are reported in Appendix Table 8. Temporal supervision is inherited from the source datasets, preserving their human annotations. We use an external proprietary model (OpenAI, 2025) as an auxiliary reviewer for videos with ambiguous temporal boundaries. We convert the heterogeneous source annotations into a unified temporal representation and organize each video as a full monitoring session. Detailed provenance, temporal normalization, and review procedures are provided in Appendix B.

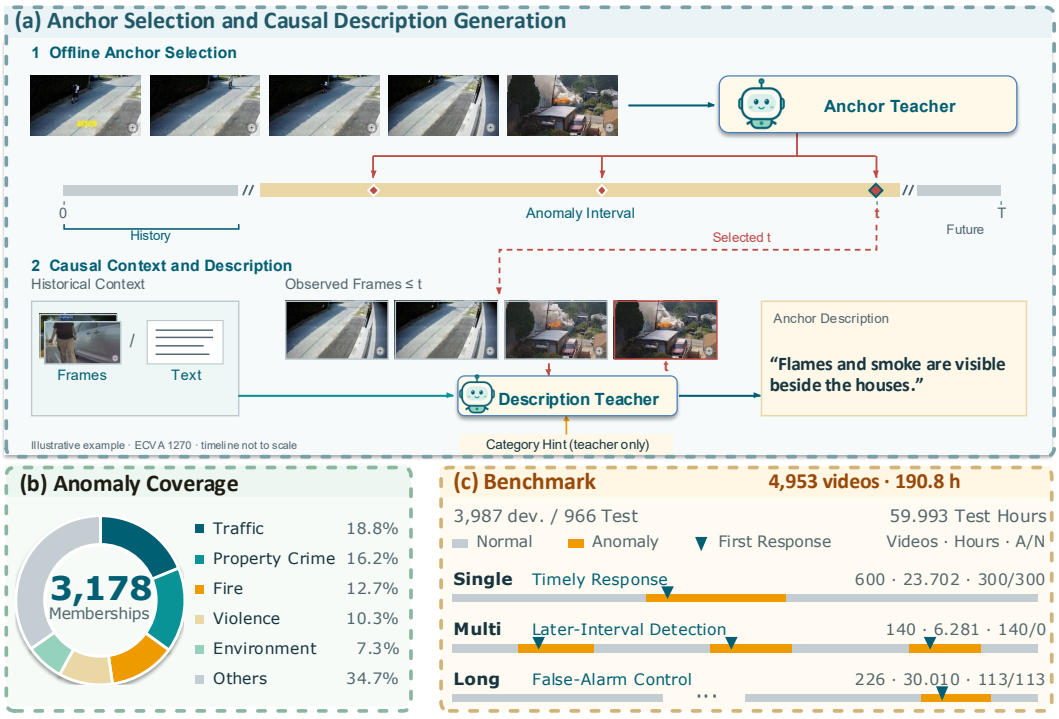

Figure 2: TAI-VAD overview. (a) Offline anchor selection and causal description generation. (b) Category distribution over 3,178 video–category memberships. (c) Single-, Multi-, and Long-video settings with schematic timelines; A/N denotes abnormal/normal, and Multi* denotes multi-event.

3.1 Dataset Curation

Causal description generation.

Existing VAD supervision primarily consists of coarse-grained anomaly labels or holistic descriptions and reasoning over an entire video. Such supervision does not capture how anomalous actions evolve as evidence becomes available during causal streaming. In our setting, the model must understand the ongoing action at timestamp from only the observations available at that moment. To provide supervision at this temporal granularity, we construct timestamp-aligned causal descriptions, as illustrated in Figure 2(a). The procedure separates *where* to annotate from *what* is observable at that moment: an offline teacher first identifies informative moments within annotated anomaly intervals, and each target is then generated using only observations available up to its timestamp. Later content may therefore guide anchor placement but cannot influence the target description. For example, before an anomalous action becomes visible, the target may state that "a person is standing beside a vehicle, with no abnormal action visible yet." Once sufficient evidence appears, it may describe that "the person breaks the vehicle window and reaches inside." Thus, each description reflects the evidence available at the current timestamp rather than summarizing the event using future content

Training on these causal trajectories provides supervision at the level of sequential monitoring decisions rather than isolated anomaly labels. It teaches the model to accumulate evidence from observed prefixes, distinguish emerging anomalies from normal activity or insufficient evidence, and issue timely yet evidence-grounded warnings. Because normal and anomalous checkpoints are interleaved within complete sessions, the model also learns when to remain silent, sustain a response as an anomaly unfolds, and maintain its monitoring state across successive observations. Statistics are summarized in Figure 2(b) and the implementation details are shown in Appendix B.4, B.5 and B.6.

3.2 Task Formulation

At a high level, a deployable VAD model should reliably monitor a continuous video stream, remain silent during normal activity, and respond to anomalies at the appropriate time. Traditional detection over single event detection is insufficient to characterize persistent anomaly monitoring. A practical

agent must also remain attentive to subsequent events and detect after long-term normal observation. Existing benchmarks, however, offer limited evaluation of post-response monitoring and prolonged normal observation, especially in long videos. We therefore define three complementary settings—Single, Multi, and Long—to evaluate the capabilities (Figure 2(c) and Appendix C.1.).

Single-event. Videos contain at most one anomalous segment, testing timely response while preserving silence during surrounding normal activity. ***Multi-event.*** Videos contain multiple separated anomaly segments, testing whether the model continues monitoring and responds to later segments. ***Long-video.*** Extended videos contain prolonged normal exposure, with or without an anomaly, testing sustained monitoring and false-alarm control. See Appendix C.1 for construction details.

Each complete video defines a monitoring session, and the ordered sequence of model decisions and responses forms its monitoring trajectory. We denote the j-th annotated anomaly segment by $[s_j, e_j)$, where s_j and e_j are its start and end timestamps, and j indexes anomaly segments in the session. These segments serve as the basic temporal units for matching model responses to ground-truth anomalies during evaluation. The model processes each session causally at one-second checkpoints. At checkpoint t, it accesses only visual observations up to time t, together with the interaction state retained from previous checkpoints; future frames and annotations remain unavailable.

Video streaming protocol. Each session contains a complete video of duration T and a fixed monitoring instruction q. At each checkpoint $t < T$, the model receives the latest observation x_t and its retained state h_{t-1}, which summarizes prior interaction, then produces a monitoring action

$$a_t \sim \pi_\theta(\cdot \mid q, x_t, h_{t-1}), \qquad a_t \in \{\textsc{silence}\} \cup \{\textsc{response}(w) : w \in \mathcal{Y}\}, \tag{1}$$

where $\mathcal{Y}$ denotes the space of warning texts. Thus, a response action contains its warning text, whereas a silence action contains no text. The state is updated after each checkpoint according to

$$h_t = F_\theta(h_{t-1}, q, x_t, a_t), \tag{2}$$

where F_θ abstracts the model-specific history or memory update. Responses are expected during anomalous activity and silence otherwise. The complete trajectory $\mathcal{A} = \{(t, a_t) : t < T\}$ is retained through the end of the video, with the monitoring state updated after every interaction. Each response is generated from only the currently available video prefix and retained interaction history, so its warning is causal and evidence-grounded. Future observations and benchmark annotations are unavailable to the model. Model details are provided in Appendix D.2.

3.3 Evaluation metrics

Conventional VAD evaluation, which often scores independent clips or only the first detection, is insufficient for persistent monitoring. They cannot distinguish a timely first response from sustained monitoring, quantify false alarms during normal observation, or assess responses to later anomalies. As shown in Figure 3, we therefore evaluate the complete monitoring trajectory along four complementary dimensions: **timely detection of anomalies**, **sustained response throughout anomalous activity**, **silence during normal observation**, and **continued monitoring after an earlier response**. All the three task scenarios, namely single, multi-event and long-video, share and evaluated uner this trajectory design, while multi-event and long-video extend it to repeated segments and prolonged observation. Detaled description are provided in Section 5 and Appendix C.2.

4 Persistent Anomaly-aware Causal Tuning

4.1 Preliminaries: Video-language interaction models

Video-language interaction models extend conventional VLMs from one-shot visual question answering to sequential interaction over continuous video streams. At each temporal checkpoint t, an interaction model receives the active query q, the newly available visual observation x_t, and a retained interaction context h_{t-1}, and then selects an action such as remaining silent or producing a response. Recent work has explored this paradigm through proactive video interaction, streaming video understanding, and stateful online assistance, including JoyAI-VL-Interaction (Yao et al., 2026), MMDuet2 (Wang et al., 2025a), VideoChat3 (Li et al., 2026), AssistPDA (Yang et al., 2025b), and streaming video-language models (Yang et al., 2026). These models differ in their backbones,

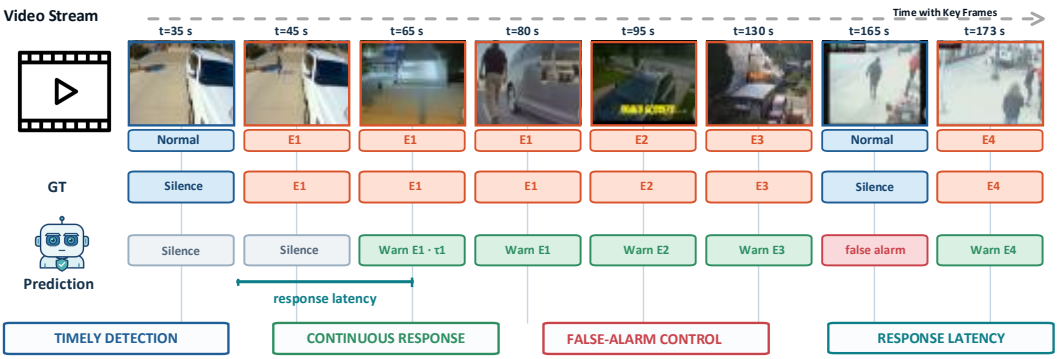

Figure 3: Trajectory evaluation on a multi-event video. Shared timelines align observations, annotations, and predictions over the monitoring session.

action spaces, memory mechanisms, and training procedures. We select JoyAI-VL-Interaction (Yao et al., 2026) as a representative interaction model because it provides native per-step silence and response decisions while preserving the underlying pretrained VLM architecture (Yao et al., 2026). Its streaming interaction interface and retained session context provide a natural starting point for persistent VAD without introducing task-specific architectural changes. We also trained a Qwen3.5-9B adaptation, but its streaming evaluation produced frequent invalid action outputs and false alarms (Appendix E.2). We use JoyAI as the primary instantiation because its native action interface supports direct per-step decisions with a four-A100 training setup; this choice does not imply an intrinsic ranking of the backbones.

4.2 PACT: Task-specific persistent causal post-training

We evaluate representative pretrained interaction models on TAI-VAD (Table 1). Their results reveal contrasting failure modes: JoyAI-VL-Interaction (Yao et al., 2026) and VideoLLM-online (Chen et al., 2024) show very low anomaly coverage (1.13% and 0%), while Qwen3-VL-4B-Instruct Bai et al. (2025) achieves higher coverage (73.93%) at the cost of a high normal-time false-positive rate (47.74%). This gap between general interaction ability and persistent anomaly monitoring motivates a task-specific adaptation. We therefore introduce PACT (*Persistent Anomaly-aware Causal Tuning*), a supervised fine-tuning framework for streaming VAD adaptation. The objective of PACT is to teach a pretrained interaction model a session monitoring policy: 1) accumulate evidence from the observed video prefix, 2) decide at each checkpoint whether to remain silent or respond, 3) ground each warning in the currently available evidence, and 4) maintain the monitoring state across successive observations. PACT therefore trains the model on complete monitoring trajectories rather than isolated observations.

Trajectory learning of persistent response behavior. We formulate supervised fine-tuning as multi-turn interaction over a complete monitoring session. Let $\tau = \{(c_t, z_t)\}_{t=1}^{T_\tau}$ denote a monitoring trajectory, where $c_t = (q, x_t, h_{t-1})$ contains the monitoring query q, the newly available observation x_t, and the interaction state retained from previous turns h_{t-1}. The target z_t is the assistant output at turn t, consisting of either a silence action or a response action followed by an evidence-grounded warning. Normal and anomalous checkpoints are both retained to supervise the complete monitoring trajectory. We optimize autoregressive supervision over assistant tokens:

$$\mathcal{L}_{\text{PACT}}(\theta) = -\frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \frac{1}{N_\tau} \sum_{t} \sum_{j=1}^{|z_t|} w_{t,j} \log p_\theta \left(z_{t,j} \mid c_t, z_{t,<j} \right), \tag{3}$$

where N_τ is the number of supervised assistant tokens in trajectory τ, and $w_{t,j}$ denotes the role-dependent token weight. This trajectory objective teaches the model to accumulate evidence, select timely responses, and maintain persistent monitoring behavior. Details of the token weighting, masking, and normalization are provided in Appendix E.1.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset and benchmark construction. TAI-VAD is collected from complete videos released by open-source benchmarks, including ECVA (Du et al., 2024), MSAD (Zhu et al., 2024), UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020), ComplexVAD (Mumcu et al., 2025), and VAD-R1 (Huang et al., 2025). We edit and consolidate them into a common temporal representation and augment the source annotations with temporal segment boundaries and timestamp-aligned causal descriptions. Each source video is retained as a complete monitoring session, so evaluation can use the full temporal context available up to every decision point. Source provenance, annotation procedures, and exact split composition are provided in Appendix B–C.1.

Evaluation task and metrics. We evaluate three complementary tracks: ***Single-event*** for isolated anomaly response, ***Multi-event*** for later-segment response and inter-segment silence, and ***Long-video*** for sustained monitoring under prolonged normal exposure. All models use the same source-disjoint test split. The main table reports six complementary measures: Binary AUROC for response/silence discrimination, Anomaly Coverage for sustained responses during anomalous activity, Segment Recall@5s for timely response, Normal-time FPR for false responses during normal exposure, Response latency for warning timing, and Avg infer time for per-decision computational cost. Multi-event and Long-video use setting-specific measures of later-event response, inter-segment silence, and prolonged normal monitoring. Specifically, Table 2 reports first-segment Recall@5s (R@5s), later-segment Recall@5s (Later@5s), the fraction of sessions in which all anomaly segments are detected within the five-second budget (All@5s), and normal-time FPR. Table 3 reports timely Recall@5s (R@5s), anomaly-time Coverage, and normal-time FPR. Together, these measures provide a comprehensive evaluation of persistent monitoring. Detailed split, protocol, and metric definitions are provided in Appendix C.1– C.2.

Baselines. We compare PACT with representative pretrained video and interaction models covering continuous, streaming, and causal-window inference, including JoyAI (Yao et al., 2026) and Qwen (Bai et al., 2025; Team, 2026) backbones, MiniCPM-o (Cui et al., 2026), MM-Duet2 (Wang et al., 2025a), Dispider (Qian et al., 2025), VideoLLM-online (Chen et al., 2024), and VideoChat3 (Li et al., 2026). All models are evaluated under the same persistent-monitoring objective and causal observation setting. Because their native state management and output interfaces differ, we retain model-specific evaluation conventions; the corresponding protocols and prediction-to-score mappings are given in Appendix D.2.

Implementation. PACT is a single-stage trajectory supervised fine-tuning procedure initialized from JoyAI-VL-Interaction (Yao et al., 2026). It adapts the model on causal multi-turn monitoring trajectories, where each trajectory contains successive observations, monitoring actions, and evidence-grounded warnings. Full-parameter training uses a single server with four A100 80GB GPUs. The training procedure follows the same video-streaming protocol used for evaluation, including one-second checkpoints and persistent interaction state. Optimization and configuration details are provided in Appendix E.1 and Appendix E.2.

5.2 MAIN RESULTS

Single-event results. PACT achieves the highest Binary AUROC (0.711), indicating the strongest balance between anomaly response and normal-time silence on the Single-event track. The pretrained baselines exhibit contrasting behaviors. Qwen3-VL-8B achieves the highest Anomaly Coverage (86.05%) and Segment Recall@5s (82.09%), but responds during 60.85% of normal checkpoints; Qwen3-VL-4B shows a similar sensitivity–false-alarm trade-off. Conversely, JoyAI, MM-Duet2, and VideoLLM-online have low false-positive rates but cover little anomalous activity. Relative to the strongest pretrained AUROC baseline, Qwen3.5-9B (0.640), PACT improves coverage and timely recall while lowering normal-time FPR from 19.07% to 13.71%. These results show that our models monitoring achieves both timely anomaly responses and restraint during normal activity.
Multi-event and Long-video results. Multi-event tests whether a model responds to later anomalies while remaining silent between events; Long-video tests these behaviors under prolonged observation. Tables 2 and 3 report the corresponding results, with additional metrics in Appendix D.1.

Table 1: Single-event comparison on the frozen 600-video test cohort under the shared causal replay protocol. The table summarizes response quality, anomaly coverage, timely detection, false-alarm control, and computational cost. Complete model-specific protocols, symbol definitions, invalid-output handling, and pending-run details are provided in Appendix D.2.

Method	Binary AUROC ↑	Anomaly Coverage (%) ↑	Segment Recall@5s (%) ↑	Normal-time FPR (%) ↓	Response latency (s) ↓	Avg infer time (s) ↓
Causal-window video models						
VideoChat3-4B (W=32)	0.51	24.57	67.23	23.10	4.19	0.95
JoyAI-Interaction continuous framework						
Qwen3-VL-8B-Instruct	0.63	86.05	82.09	60.85	1.72	1.44
Qwen3-VL-4B-Instruct	0.63	73.93	79.05	47.74	2.31	1.04
Qwen3.5-9B	0.64	47.16	58.45	19.07	4.90	1.02
JoyAI-VL-Interaction base	0.51	1.13	22.64	0.02	3.38	0.48
Other streaming interaction models						
MiniCPM-o 4.5	0.52	37.48	62.50	33.36	5.10	1.15
MMDuet2	0.51	1.91	24.32	0.26	7.41	0.69
Dispider	0.50	1.57	8.78	1.36	16.82	–
VideoLLM-online	0.50	0.00	0.00	0.00	–	0.08
Ours						
PACT	0.71	55.90	67.57	13.71	1.15	0.52

Table 2: Multi-event results: first- and later-event detection and normal FPR.

Method	R@5s (%) ↑	Later@5s (%) ↑	All@5s (%) ↑	Normal-time FPR (%) ↓
Qwen3-VL-8B	76.99	95.65	74.34	73.07
VideoChat3	2.65	10.63	0.00	1.05
MMDuet2	22.12	22.22	10.62	1.36
VideoLLM-online	0.00	0.00	0.00	0.00
JoyAI	14.16	0.00	1.77	0.07
PACT	64.60	55.07	57.52	55.08

Table 3: Long-video results: timely detection, anomaly coverage, and normal FPR.

Method	R@5s (%) ↑	Coverage (%) ↑	Normal-time FPR (%) ↓
Qwen3-VL-8B	72.13	39.46	40.37
VideoChat3	22.95	2.60	1.15
MMDuet2	2.46	0.04	0.08
VideoLLM-online	0.00	0.00	0.00
JoyAI	0.00	0.00	0.00
PACT	3.28	0.29	8.88

On Multi-event videos, PACT substantially improves over its JoyAI base model: later-segment Recall@5s rises from 0.00% to 55.07%, showing greater ability to respond after an earlier event. This gain comes with a normal-time FPR of 55.08% over 0.07% for JoyAI. Qwen3-VL-8B detects more later events (95.65%) but has an even higher FPR (73.07%). Thus, stronger multi-event detection remains closely coupled with false responses during normal periods.

Long-video results remain uneven: Qwen3-VL-8B has the highest Recall@5s (72.13%) but a 40.37% normal-time FPR, whereas VideoChat3 reduces FPR to 1.15% but covers only 2.60% of anomalous checkpoints. PACT reaches only 3.28% Recall@5s. No model combines timely and sustained warnings with reliable silence over long sessions.

Failure case analysis. Long-video monitoring reveals two opposing failures: JoyAI remains largely silent and misses anomalies, whereas Qwen3-VL-8B detects more events but generates frequent false alarms. VideoChat3 achieves a more favorable trade-off using a stateful 32-frame causal window, but its low coverage shows that local context alone cannot sustain warnings throughout an anomaly. PACT improves later-event detection yet still produces false responses during normal Multi-event intervals and misses most Long-video anomalies. Together, these patterns point to a framework-level limitation in how streaming systems maintain temporal state: a short window can discard useful history, while persistent memory may bury rare anomaly evidence among abundant normal observations. PACT adapts the response policy through SFT without changing the underlying state-management architecture, and its remaining Long-video failures show the limit of this adaptation. Designing interaction frameworks that selectively retain, update, and retrieve long-term evidence is therefore an important direction for persistent VAD.

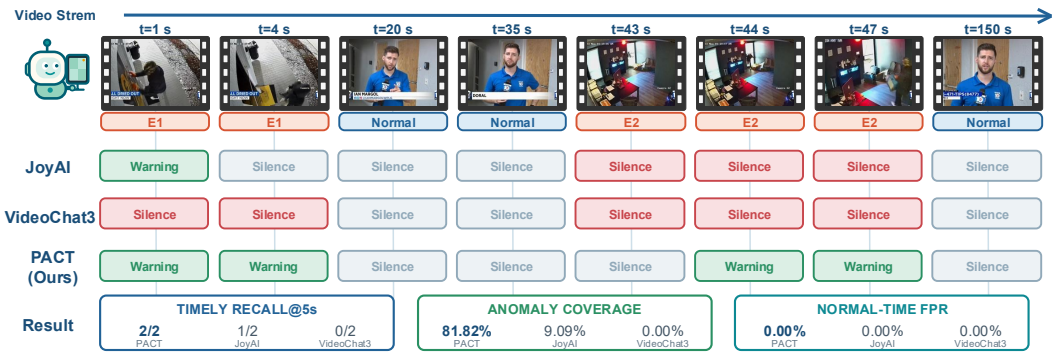

Figure 4: Multi-event case on ECVA 1478. Green and red bars indicate correct and incorrect decisions, respectively; cards show selected model actions. Metrics use all 158 one-second decisions.

SFT objective	AUROC ↑	Cov. (%) ↑	R@5s (%) ↑	FPR (%) ↓	Det. segs. ↑
PACT (weighted)	0.71	55.90	67.57	13.71	214/296
Uniform SFT	0.56	13.29	11.49	1.62	35/296

Table 4: Weighting ablation on 600 Single-event videos.

Model	Accuracy (%)
Qwen3-VL-8B-Instruct	84.16
PACT (ours)	87.70

Table 5: VAD-R1 accuracy.

5.3 Ablation

Following JoyAI-VL-Interaction (Yao et al., 2026), PACT assigns different weights to supervised assistant tokens. We ablate this design with uniform token weights. On the 600-video Single-event test set (Table 4), anomaly Coverage drops from 55.90% to 13.29% and Recall@5s from 67.57% to 11.49%; detected segments fall from 214 to 35 out of 296. Although normal-time FPR falls from 13.71% to 1.62%, Binary AUROC drops from 0.71 to 0.56. Uniform weighting thus favors silence at the expense of timely detection. Despite efforts to balance training supervision, anomalies remain sparse in continuous video, which may explain this shift toward silence. See more in Appendix F.

5.4 Offline VAD diagnostic and evidence boundary

Reliable persistent monitoring builds on strong anomaly recognition. We therefore evaluate PACT on the offline VAD-R1 test set (Huang et al., 2025) who is a comprehensive offline video anomaly reasoning benchmark covering various scenarios. PACT achieves 87.70% accuracy, compared with an author-provided 84.16% for Qwen3-VL-8B-Instruct (Table 5). This result suggests that training for silence/response interaction is compatible with strong offline VAD performance, although offline accuracy alone does not measure persistent streaming behavior.

5.5 Qualitative Result

Figure 4 illustrates a qualitative result of a multi-event session with two separated annotated anomalies. PACT warns one second after each onset and remains silent at the selected normal checkpoints between and after the events. In contrast, the JoyAI Yao et al. (2026) responds only to the first event, while VideoChat3 Li et al. (2026) remains silent throughout. The case visualizes post-warning monitoring and later-event detection; aggregate results are reported above.

6 Conclusion

We bridge the gap between conventional VAD and persistent monitoring with TAI-VAD, the first comprehensive benchmark spanning single-event, multi-event, and long-video sessions with full-trajectory evaluation. Experiments with pretrained interaction models reveal a recurring trade-off: greater anomaly sensitivity often brings more false alarms, whereas conservative models miss events. Our trajectory-level SFT baseline, PACT, achieves the highest Single-event AUROC and a favorable warning-silence balance. The limited performance across all evaluated models in Multi-event and Long-video monitoring points to a central challenge in streaming interaction framework design: maintaining useful long-term memory while filtering irrelevant normal context. Addressing this architectural/memory-design challenge is an important direction for potential future work.

AI USE STATEMENT

We used generative AI to produce auxiliary historical summaries and timestamp-aligned descriptions for SFT, and to assist targeted review of ambiguous temporal boundaries (Appendix B). AI tools also assisted with study design, code drafting and review editing, and translation. Source annotations are preserved, and generated annotations are identified separately. Reported measurements were computed from saved predictions and evaluation scripts. The authors reviewed AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

Our benchmark reuses publicly released videos and collects no new footage. Surveillance-oriented anomaly monitoring raises privacy and fairness concerns; false alarms may prompt unwarranted intervention, while missed events may delay assistance. We preserve dataset provenance and distinguish source annotations from generated descriptions.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B document data sources, annotation procedures, and evaluation splits. The streaming protocol and metrics are specified in Section 3.2 and Appendix C.2; model-specific replay and SFT settings are detailed in Appendices D.2, E.1, and E.2. Reported tables are derived from retained per-video predictions and scoring artifacts.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Daniel Bogdoll, Jan Imhof, Tim Joseph, Svetlana Pavlitska, and J Marius Zöllner. Hybrid video anomaly detection for anomalous scenarios in autonomous driving. *arXiv preprint arXiv:2406.06423*, 2024.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, and Jun Liu. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2412.07183*, 2024.

Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vanebench: Video anomaly evaluation benchmark for conversational lmms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3123–3140, 2025.

Satoshi Hashimoto, Alessandro Moro, Kenichi Kudo, Takayuki Takahashi, and Kazunori Umeda. Unsupervised video anomaly detection in traffic and crowded scenes. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pp. 870–876. IEEE, 2022.

Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. VAD-R1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *arXiv preprint arXiv:2505.19877*, 2025. URL https://arxiv.org/abs/2505.19877.

Yizhen Jia, Rong Quan, Yue Feng, Haiyan Chen, and Jie Qin. Doubly contrastive learning for source-free domain adaptive person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3949–3957, 2025.

Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. doi: 10.1109/WACV57701.2024.00670.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. VideoChat3: Fully open video MLLM for efficient and generalist video understanding, 2026. URL https://arxiv.org/abs/2607.14935.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 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.

Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7):1–38, 2024.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

Furkan Mumcu, Michael J. Jones, Yasin Yilmaz, and Anoop Cherian. ComplexVAD: Detecting interaction anomalies in video. In *IEEE Winter Conference on Applications of Computer Vision Workshops*, February 2025. doi: 10.1109/WACVW65960.2025.00122. URL https://www.merl.com/publications/TR2025-016.

OpenAI. Openai models. https://platform.openai.com/docs/models, 2025. Accessed: July 20, 2026.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10599–10609, 2025. doi: 10.1109/CVPR52734.2025.00991.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. MMDuet2: Enhancing proactive interaction of video MLLMs with multi-turn reinforcement learning, 2025a. URL https://arxiv.org/abs/2512.06810.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. ProactiveVideoQA: A comprehensive benchmark evaluating proactive interactions in video large language models, 2025b. URL https://arxiv.org/abs/2507.09313.

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 *European Conference on Computer Vision*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307. IEEE, 2024.

Shengtian Yang, Yue Feng, Yingshi Liu, Jingrou Zhang, and Jie Qin. MoniTor: Exploiting large language models with instruction for online video anomaly detection. *arXiv preprint arXiv:2510.21449*, 2025a. URL https://arxiv.org/abs/2510.21449.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pp. 304–322. Springer, 2024.

Zhenyu Yang, Kairui Zhang, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Don't pause: Streaming video-language synchrony for online video understanding. *arXiv preprint arXiv:2606.06991*, 2026.

Zhiwei Yang, Chen Gao, Jing Liu, Peng Wu, Guansong Pang, and Mike Zheng Shou. AssistPDA: An online video surveillance assistant for video anomaly prediction, detection, and analysis. *arXiv preprint arXiv:2503.21904*, 2025b. URL https://arxiv.org/abs/2503.21904.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459, 2022.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8679–8688, 2025.

Yating Yu, Congqi Cao, Zhaoying Wang, Weihua Meng, Jie Li, Yuxin Li, Zihao Wei, Zhongpei Shen, and Jiajun Zhang. Cuebench: Advancing unified understanding of context-aware video anomalies in real-world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12187–12195, 2026.

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, 2024.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 2024.

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 *Proceedings of the computer vision and pattern recognition conference*, pp. 13843–13853, 2025.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. VAU-R1: Advancing video anomaly understanding via reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2505.23504.

Shu Zou, Xinyu Tian, Lukas Wesemann, Fabian Waschkowski, Zhaoyuan Yang, and Jing Zhang. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4223–4233, 2026.

Table 6: Task-level comparison of persistent anomaly-monitoring properties. **Y** denotes an explicitly defined or evaluated property, △ partial or indirect coverage, and – an unspecified property. The final column refers to joint evaluation of warning timing, anomaly coverage, normal-time false alarms, and later-event detection over one monitoring trajectory.

	Interaction protocol			Session scope		Evaluation
Task line	Single persistent query	Causal prefix	Silence action	Multi-segment monitoring	Post-response continuation	Anomaly-trajectory evaluation
Offline / video-language VAD (Sultani et al., 2018; Liu et al., 2018; Zanella et al., 2024; Huang et al., 2025)	–	–	–	–	–	–
Streaming anomaly VAD (Karim et al., 2024; Yang et al., 2025a)	△	**Y**	–	△	–	–
Proactive video interaction (Wang et al., 2025b; Yang et al., 2025b; Li et al., 2026)	△	**Y**	△	–	△	△
Stateful interaction (Yao et al., 2026; Wang et al., 2025a)	△	**Y**	**Y**	–	△	△
TAI-VAD + PACT (ours)	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

A Comparison of Monitoring Protocols

B Dataset Construction Details

B.1 Video registry and annotation provenance

Dataset curation relies exclusively on publicly available video resources and annotations released by prior studies and open-source projects. We establish a canonical video registry by normalizing source identifiers and consolidating known copies and derived variants. Each canonical identity is associated with a primary media asset, source labels, and annotation provenance. Auxiliary language annotations from VAD-R1 and VAU-R1 are linked to the corresponding video identities without increasing the media count (Huang et al., 2025; Zhu et al., 2025). The registry comprises 4,953 canonical video identities, including 3,289 abnormal and 1,664 normal videos, with a total duration of 190.800 hours.

All videos are reused from the source releases; no additional footage is collected. Temporal annotations, including the human annotations provided by the original dataset curators, are inherited with their provenance. Our additional temporal review targets ambiguous test cases rather than replacing the source annotation process. Public availability establishes the provenance of the data; the privacy, licensing, and deployment considerations associated with surveillance data remain relevant, as discussed in the ethics statement.

B.2 Temporal annotation and session generation

Each registered video forms a complete monitoring session, retaining its normal intervals before, between, and after anomalous activity. Sessions provide the shared video and temporal annotation records from which training trajectories and evaluation inputs are derived. Training uses source intervals to assign response or silence targets at 1 Hz; during evaluation, the model generates its own monitoring trajectory. Temporal annotations and reference descriptions are kept separate from evaluation inputs.

We normalize source intervals to seconds with the half-open convention $[s, e)$, retaining the original values and conversion metadata. Overlapping or adjacent anomaly intervals are merged; each maximal continuous interval forms a *source segment*, and gaps remain normal time. Normal sessions have an empty segment set. For ComplexVAD, we take the union of annotated object frames to obtain temporal support. Object tracks do not determine independent event identities, and separated source segments may represent different stages of the same underlying event.

B.3 BOUNDARY REVIEW AND ANNOTATION ELIGIBILITY

We check temporal annotations against media duration and flag ambiguous time encodings, invalid boundaries, and uncertain visual alignment. For the targeted review of 66 test videos, we use `gpt-5.6-sol` (OpenAI, 2025), the strongest proprietary model available to us for offline video-language understanding at the time of annotation. Given timestamped frames, the model distinguishes visible anomaly, anomaly absence, and insufficient evidence, followed by boundary review and conflict adjudication. This yields 39 temporal repairs; 27 videos remain unresolved and are excluded from temporal scoring.

B.4 CAUSAL GENERATION OF HISTORICAL MEMORY

We use `gpt-5.6-sol` (OpenAI, 2025) as the external multimodal teacher throughout causal data construction. For historical memory, the teacher generates factual summaries for completed 100-second history blocks. For a block ending at time b, 16 frames are sampled uniformly from $[b - 100, b)$, and the resulting summary is introduced only at b or later. Frame selection is independent of anomaly labels. The summarizer receives historical frames and timestamps, without event labels, anomaly boundaries, response targets, or future observations. Summaries describe visible entities, actions, state changes, and unresolved conditions.

After each group of five historical summaries, a text-only request combines them with the preceding long-term summary. This rolling compression retains cumulative context while bounding the length of the memory representation. The input audit requires each selected frame's presentation timestamp to be no later than its requested time and strictly earlier than the boundary at which the memory becomes available. These generated memories condition training data. Benchmark replay instead uses the evaluated system's own state-management mechanism and does not receive teacher-generated memories.

The development pool yields 3,167 successful historical summaries from 3,168 requests and 316 successful rolling compressions. The single unresolved historical-summary request remains recorded as a failure. The reported counts measure generation completion, not the number of summaries used by the reported PACT checkpoint; the summaries have not undergone independent semantic verification.

B.5 OFFLINE ANCHOR SELECTION AND CAUSAL DESCRIPTION GENERATION

The `v0_4_reasoning` SFT partition received description augmentation. The same `gpt-5.6-sol` teacher performs offline anchor selection and causal target generation. It first views timestamped overviews of the available SFT anomaly intervals, with a source category as an attention hint, and selects three to five existing response positions within each source segment when sufficient observations are available. Segments with fewer supported positions may contribute fewer anchors or none. This stage is an offline annotation procedure and may access observations later than a selected anchor; its observations and action summaries are retained only for auditing.

A separate request generates the target description at each anchor t. Its visual input contains at most eight frames from the preceding eight seconds within the current training chunk, including the frame at t, and its historical context is the memory already available to the student. Future frames, source-segment descriptions, and the selection-stage summaries are excluded. A teacher-only category hint guides attention but is not included in student inputs. The prompt requires claims to be supported by the supplied observations and prohibits inferred future outcomes, hidden causes, or intentions. The temporal restriction applies to the visual and memory evidence; offline anchor selection and category conditioning remain explicit parts of the annotation procedure.

The teacher produces a short warning describing visible anomaly evidence or a factual state caption when the specific anomaly is uncertain. An initially empty description may receive one additional captioning request with the same causal inputs. Captions remain distinct from evidence descriptions and are not treated as confirmation of the source category. The generated text replaces only selected SFT response targets and their necessary subsequent history references. Anomaly boundaries, action targets, frames, visual memories, and non-SFT partitions are unchanged; dev and test descriptions are not generated by this augmentation procedure. Teacher-generated memories and reference de-

scriptions are used only during data construction and are never provided to the evaluated model during benchmark replay.

The completed augmentation contains 1,614 descriptions: 1,065 evidence descriptions and 549 factual state captions. These modify 395 of the 1,103 trajectories in that partition. Non-anchor responses retain their original templates, and all silence targets remain unchanged.

B.6 Dataset statistics and counting conventions

Figure 5 summarizes video resources and generated annotations. Table 8 reports exact video counts by source. The 3,987 development videos and 966 test videos partition the full registry by canonical video identity. Training samples are derived windows of at most 100 seconds, so trajectory counts are distinct from source-video counts. The `v0_4_reasoning` SFT partition contains 1,103 trajectories from 834 videos, with 69,062 turns: 13,131 responses and 55,931 silences. Of these trajectories, 453 contain response targets and 650 contain only silence; the latter include normal portions of abnormal videos. The reported PACT checkpoint instead uses a 1,114-trajectory, 63,823-turn training view. An additional 4,242 trajectories are reserved outside the reported PACT training; the development partitions contain 127 dev trajectories and 43 held-out dev-companion trajectories.

Table 7: Data and annotation statistics. Video counts refer to canonical identities; memory and description counts refer to generated text records. The SFT resource and reported PACT training use different data versions.

Scope	Quantity	Count
Video collection	Abnormal / normal videos	3,289 / 1,664
	Total duration (hours)	190.80
	Development / test videos	3,987 / 966
	Provisional coarse categories	15
Development memory	Historical / rolling summaries	3,167 / 316
SFT resource	Trajectories / source videos	1,103 / 834
	Response / silence turns	13,131 / 55,931
	Generated anchor descriptions	1,614
	Evidence descriptions / factual captions	1,065 / 549
Reported PACT training	Trajectories / turns	1,114 / 63,823

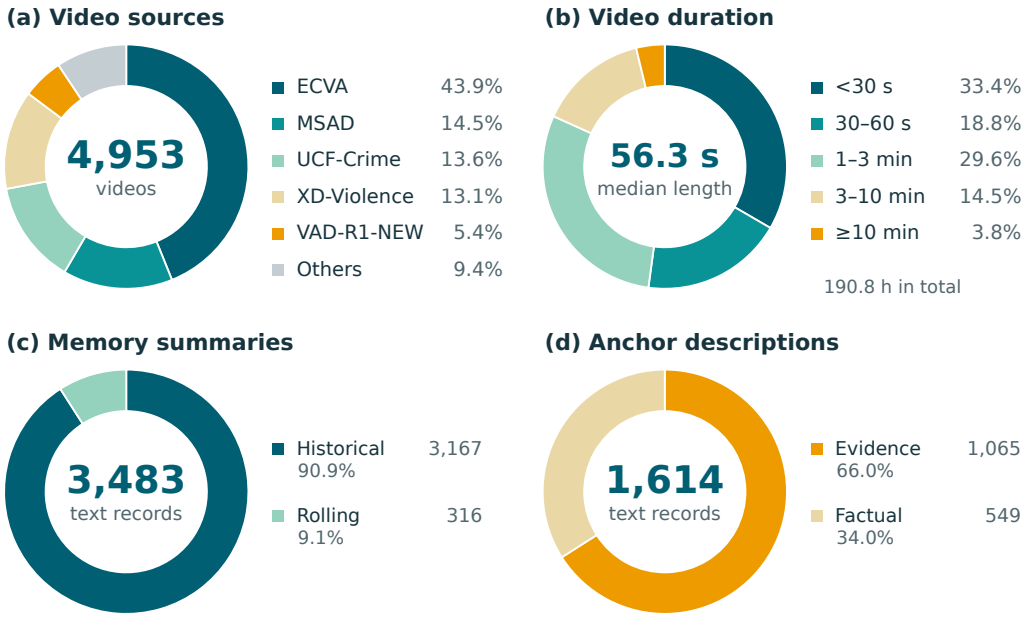

Figure 5: Dataset composition. (a) Source and (b) duration distributions over 4,953 canonical videos; sources outside the five largest are grouped as Others. (c) Historical and rolling memory summaries. (d) Evidence descriptions and factual captions at selected anchors. Sector areas represent proportions within each panel; legends in (c–d) also report absolute record counts.

Table 8: Canonical video counts by source. Copies and auxiliary language annotations are not counted as additional videos.

Source	Abnormal	Normal	Total
ECVA	2,166	8	2,174
MSAD	240	480	720
UCF-Crime	326	350	676
XD-Violence	54	596	650
VAD-R1-NEW	269	0	269
TAD	93	100	193
ComplexVAD Test	113	0	113
ShanghaiTech	11	70	81
UBnormal	17	60	77
Total	3,289	1,664	4,953

Source labels are mapped to a provisional taxonomy for analysis (Table 9). ComplexVAD lacks the required event semantics and is excluded from category counts. A video contributes once to each category it contains, so category memberships are not a count of unique videos. Source-event counts retain the original grouping and should not be interpreted as manually confirmed independent events. Generated memory, anchor descriptions, source events, and 1 Hz response turns are reported separately because they represent different annotation units.

Table 9: Provisional anomaly categories in the core source pool, excluding ComplexVAD. Video counts are per-category support; source-event counts retain source grouping.

Category	Videos	Source events
Traffic and transport safety	596	598
Theft, robbery, and property crime	516	518
Fire and explosion	403	403
Interpersonal violence and weapons	327	327
Environmental harm	232	232
Building, infrastructure, and utilities	195	195
Food and agriculture safety	162	162
Animal related	151	151
Falls and physical hazards	150	150
Public disorder and conflict	131	131
Natural hazard	122	122
Medical and public health	113	113
Water incident	55	55
Other public safety	14	14
Industrial and workplace accident	11	11
Total memberships / source events	3,178	3,182

C Evaluation Settings and Metrics

C.1 Evaluation-setting construction

Table 10 summarizes the test composition across the three settings. ***Single-event.*** The 600-video setting balances abnormal and normal video counts. It retains all 439 legacy VAD-R1 test videos and adds 161 videos, using duration strata and source, category, and relative-onset coverage to guide selection. Its abnormal subset covers 15 provisional coarse classes and 84 source-specific native families. Matching normal and abnormal durations yields a standardized mean difference of 0.010 in log duration. Complete replay exposes the model to both the anomaly and its surrounding normal activity.

Multi-event. We reserve 140 candidate videos with explicit multi-event or multi-span source annotations to test detection of later anomaly segments and silence during intervening gaps. These candidates all come from ECVA. Their separated spans may represent distinct events or different stages of one event; we therefore report later-*segment* detection without assuming semantic independence. After overlapping or adjacent spans are merged, some eligible candidates contain only one source segment; they remain in this track but do not contribute to later-segment recall.

Long-video. We select 113 anomalous videos from publicly available benchmarks and pair them with 113 normal videos from the previously assembled collection of public videos. The abnormal and normal sides provide 14.921 and 15.089 monitoring hours, respectively, with a log-duration standardized mean difference of 0.011. This setting evaluates behavior after extended observation and normal exposure. Duration matching controls exposure differences. The source composition of the two groups may differ, so this track alone cannot separate anomaly recognition from source-domain effects.

Table 10: TAI-VAD test composition. Hours cover all scheduled videos. Action and segment columns count videos with eligible annotations for the respective metrics; they do not indicate inference completion. Multi denotes source-segment candidates, pending confirmation of independent events.

Track	Abnormal / normal	Hours	Action	Segment	Segments
Single	300 / 300	23.702	600	596	296
Multi	140 / 0	6.281	113	113	320
Long	113 / 113	30.010	226	226	122
Total	553 / 413	59.993	939	935	738

Segment counts include only annotation-eligible videos. Multiple source segments may belong to the same underlying event.

C.2 Evaluation metric definitions

Timely detection. For each eligible source segment j with interval $[s_j, e_j)$, we assign responses within $[s_j - \epsilon, e_j + \epsilon]$, using $\epsilon = 2$ seconds. This tolerance accommodates response timestamps near annotated segment boundaries on the one-second decision grid. Each response belongs to at most one segment: overlapping tolerance windows are resolved by original-span membership, distance to the span, and temporal order. Let τ_j be the first assigned response time and $\tau_j = \infty$ for a missed segment. With M eligible segments, timely recall is

$$\text{Recall@}\delta = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}[\tau_j < \infty \ \wedge \ \tau_j - s_j \leq \delta] \,. \tag{4}$$

We use $\delta = 5$ seconds as the primary budget; unrestricted segment recall is retained as a supplementary diagnostic. Boundary tolerance does not extend this budget. Response latency is the mean signed delay $\tau_j - s_j$ over detected segments, retaining tolerated early responses. Missed segments enter the recall denominator but not the delay average. Delay percentiles remain supplementary diagnostics.

Continuous response and silence. Checkpoint labels are determined by the union of the original anomaly spans. Normal checkpoints within ϵ of a span are excluded from the confusion matrix as uncertain boundaries; pure-normal videos receive no such exemption. With counts pooled within each track, we report

$$\text{Coverage} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \qquad \text{tIoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{5}$$

For a complete binary action trajectory, we report Binary AUROC by using response $= 1$ and silence $= 0$ as the score:

$$\text{AUROC}_{\text{bin}} = \frac{1}{2} \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) = \frac{\text{Coverage} + 1 - \text{FPR}}{2}. \tag{6}$$

This is the AUROC of the observed binary response/silence score; it should not be interpreted as the ranking ability of an unthresholded continuous anomaly confidence. Continuous response remains necessary for Coverage because repeated responses during an anomaly interval are valid and contribute to the anomaly time numerator. Coverage does not distinguish contiguous from intermittent responses with the same number of covered checkpoints. Coverage measures sustained response during anomalies, FPR measures false responses during normal exposure, and tIoU summarizes their temporal overlap. Action F1, tIoU, and video-level false-alarm diagnostics are retained in supplementary records.

Warning-text grounding. The reported tables score when a response occurs, not whether its warning text is supported by the observations available at that time. Evidence-grounded warning is a task objective, not a verified result of the action-result tables. Causal input restriction prevents future frames from entering the model input but does not itself establish factual grounding. A separate content evaluation would show reviewers the video prefix through the response timestamp and ask

Table 11: Detailed Multi-event results. Later-segment recall uses all later segments in the denominator. FPR is the normal-time false-response rate.

Method	Action videos	Segment videos	Segments	First R@5s ↑	Later R@5s ↑	All timely ↑	Coverage ↑	Normal-time FPR ↓	tIoU ↑	Delay P50 ↓	Delay P90 ↓	Infer. ↓
Qwen3-VL-8B	113	113	320	76.99	95.65	74.34	88.40	73.07	–	-1.00	2.00	–
Qwen3.5-9B	–	–	–	–	–	–	–	–	–	–	–	–
VideoChat3-4B	113	113	320	2.65	10.63	0.00	1.56	1.05	1.55	7.00	35.80	0.53
MiniCPM-o 4.5	–	–	–	–	–	–	–	–	–	–	–	–
MMDuet2	113	113	320	22.12	22.22	10.62	1.78	1.36	1.76	1.00	23.60	0.81
VideoLLM-online	113	113	320	0.00	0.00	0.00	0.00	0.00	0.00	–	–	0.09
JoyAI-VL-Interaction	113	113	320	14.16	0.00	1.77	0.20	0.07	0.20	1.00	4.20	0.57
PACT	113	113	320	64.60	55.07	57.52	60.43	55.08	41.71	-2.00	1.00	0.64

Table 12: Detailed Long-video results. Delay percentiles are computed over matched segments; FPR is the normal-time false-response rate.

Method	Action videos	Segment videos	Segments	R@5s ↑	Coverage ↑	Normal-time FPR ↓	tIoU ↑	F1 ↑	Delay P50 ↓	Delay P90 ↓	Infer. ↓
Qwen3-VL-8B	226	226	122	72.13	39.46	40.37	6.55	12.29	-1.32	3.62	4.28
Qwen3.5-9B	–	–	–	–	–	–	–	–	–	–	–
VideoChat3-4B	226	226	122	22.95	2.60	1.15	2.27	4.44	3.33	79.24	0.52
MiniCPM-o 4.5	–	–	–	–	–	–	–	–	–	–	–
MMDuet2	226	226	122	2.46	0.04	0.08	0.04	0.07	2.12	9.94	0.81
VideoLLM-online	226	226	122	0.00	0.00	0.00	0.00	0.00	–	–	0.10
JoyAI-VL-Interaction	226	226	122	0.00	0.00	0.00	0.00	0.00	–	–	0.68
PACT	226	226	122	3.28	0.29	8.88	0.14	0.27	-0.57	1.60	0.71

them to classify factual claims in the first matched warning as supported, contradicted, or not verifiable from that prefix. It would report the fraction of warnings containing an unsupported claim alongside the action metrics. No warning-text factuality or hallucination rate is reported here.

Average inference time. For K decision calls with measured wall-clock durations ℓ_k, Avg infer time is $K^{-1}\sum_{k=1}^{K}\ell_k$. It includes calls producing silence, response, or invalid output and follows each runner's recorded call boundary. This processing cost is measured separately from video-time response latency; offline replay timing does not establish end-to-end warning latency under real-time scheduling.

Persistence and reporting. Multi-track evaluation includes later-segment Recall@5s, the fraction of sessions in which all segments are detected within budget, and normal-time FPR. All later segments enter the recall denominator, irrespective of whether the first was detected. The all-segments timely rate uses all segment-eligible Multi-event candidates as its denominator, including candidates with only one segment after interval merging. Long-track results are stratified by video duration and waiting time to segment onset. We report tracks separately with annotation eligibility and trajectory completeness; inference failures are not silently removed. These metrics evaluate actions and timing. Evidence-grounded warning evaluation requires a separate assessment of the first matched response using only its observed prefix and is not established by temporal matching alone. All delays above use video time; processing cost and end-to-end warning latency require separate wall-clock measurements.

D Evaluation Results and Protocols

D.1 Detailed persistent-monitoring results

The main text reports only the track-specific headline metrics. Tables 11 and 12 retain the full set of quantities used by the scorer, including eligibility counts, continuous-response measures, delay percentiles, and inference cost. Within each track, all rows use the same test videos and causal replay. Test video identities are disjoint from development identities. A dash denotes a result that is not yet available under the frozen protocol or is undefined for the corresponding denominator.

D.2 Model-specific evaluation protocols

All compared models use the same complete-video replay, 1 Hz decision grid, test tracks, normalized action schema, and trajectory-level scorer. Test video identities are disjoint from development identities. The following interface details distinguish persistent interaction models from causal-window baselines without changing the task or the evaluation metrics.

Stateful interaction models. For JoyAI-VL-Interaction and PACT, the same monitoring instruction remains active throughout each video. In the reported replay, its text is repeated in the per-checkpoint prompt; this does not create a new monitoring objective. The session state is reset only between videos. At each checkpoint, the adapter supplies the newly available frame together with the model's retained interaction history and runtime memory. The model's own silence/response outputs are appended to that history, and the session continues after every response until the video ends. Teacher-generated summaries and evaluator annotations are never exposed during replay. In the reported runtime, one-second turns are grouped into 100-turn chunks, history compression runs every five chunks with a ten-frame lead, and long-term memory is bounded by the runtime profile. Raw outputs and invalid generations are retained for scoring rather than converted to silence.

Pretrained Qwen continuous backbones. The newly added interaction-backbone test uses the official pretrained Qwen3-VL-8B-Instruct inside the continuous JoyAI-Interaction framework. Its Single-event run is complete on all 600 videos (85,629 decisions), with no invalid actions, and is reported in Table 1. Its Multi-event and Long-video runs are also complete and are reported in Tables 11 and 12. Long-video contains 226 complete videos and 108,169 decisions, with no runtime failures, invalid actions, or truncated generations. Of 122 eligible segments, 96 are detected and 88 are detected within five seconds. Only 8 of 113 pure-normal videos have no false alarms (7.08%). Its binary-action AUROC is 0.495: anomaly coverage (39.46%) and normal-time FPR (40.37%) are similar, despite high segment-level timely recall. Warning-text grounding has not been evaluated. The Single-event timing was measured on four H200 GPUs, with two GPUs each for the main and summary models and four concurrent video sessions. Long-video uses the same GPU allocation, retaining two completed videos at concurrency four and evaluating the remaining 224 at concurrency sixteen. The reported 4.279-second mean inference time preserves all measured call durations across both settings and is not a controlled speed comparison. The recorded earlier Qwen3-VL run used Qwen3-VL-4B-Instruct; its scores are retained under that model name and are not assigned to the 8B model. For the completed Qwen3-VL-8B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3.5-9B runs, the official pretrained backbone is inserted into the continuous JoyAI-Interaction adapter. Each 1-Hz checkpoint supplies the new frame and monitoring query while the video-level session retains model-generated history, chunk summaries, and runtime state. The shared runtime uses 100-turn chunks, a ten-frame summary lead, compression every five chunks, and a 262,144-token serving context. Each model retains its native processor and chat template, with thinking disabled and at most 128 generated tokens. Invalid generations remain in the scoring denominator. These rows are distinct from the historical $W = 32$ independent causal-window runs, which are not part of the main table.

Causal-window baselines. VideoChat3-4B follows its native causal-window runner and is kept separate from the continuous JoyAI-Interaction rows. Other baseline backends follow the same outer causal replay contract, with their native history and output policies recorded separately. The 600-video VideoChat3 run has a native `VAD_Timing` summary (success rate 67.00%, hit rate 62.00%, detection AUC 0.565, mean delay -21.234 s). Its endpoint-inclusive output contains 85,679 native steps. Binding it to the frozen half-open decision grid removes 50 out-of-duration endpoint steps across 43 videos and leaves exactly 85,629 decisions; no scheduled decision is missing. Native `curr_time`, rather than the one-second-later `window_end`, defines the shared causal checkpoint. A native `</Standby>` is mapped to silence because it does not issue an alarm; the native runner still uses it internally to raise the resolution of the next three frames. Single-event repeats the monitoring query at every checkpoint, as recorded by its effective configuration. The Multi-event/Long-video run used the same stateful 32-frame window but the runner's initial-query-only default; this cross-track query-policy difference is retained and disclosed rather than rewritten after inference. Its two completed shards contained 365/366 videos because rank 1 skipped local index 0; we reran only that missing normal video with the frozen model and effective configuration, then recovered the full per-step actions and wall times from the original logs. The merged result contains 366/366 videos and 130,850 decisions. All three tracks are scored with the unchanged `tai-vad-scoring-1.2.2` implementation.

Dispider partial-result. Dispider completed 585/600 videos. Its available output is scored with all planned videos retained; 15 timestamp failures leave 481 invalid normal checkpoints, handled by the same midpoint convention as other incomplete outputs. Dispider uses native 16-frame windows,

with triggers mapped causally to the next 1-Hz checkpoint without score interpolation. Response latency is the mean over 186 detected segments. Average per-call inference time is unavailable because saved timing values are native-window averages replicated across replay checkpoints. Its separate native mean video-window AUC is 0.5867.

Invalid-output handling and metric midpoints. Malformed or incomplete generations are retained in the causal-decision denominator and are not remapped to silence. Consequently, the scorer leaves Action F1 and tIoU undefined for the affected runs, and video-level false-alarm diagnostics are withheld when invalid normal decisions prevent an audited bound. The scorer produces deterministic lower and upper bounds for checkpoint-level FPR: the lower bound treats invalid normal decisions as non-false alarms, whereas the upper bound treats them as false alarms. For compact presentation in the main table, we report the arithmetic midpoint of each pair (all values are percentages): Qwen3-VL-4B continuous, $(47.7320 + 47.7418)/2 = 47.7369$; Qwen3.5 continuous, $(19.0225+19.1229)/2 = 19.0727$; and MiniCPM-o 4.5, $(33.2316+33.4811)/2 = 33.3564$. These midpoint values are descriptive point summaries, not statistical confidence intervals or verified actions for the invalid outputs. Anomaly Coverage retains all anomaly checkpoints in its denominator, with invalid anomaly outputs contributing no detections. Holding this conservative Coverage fixed, we propagate the FPR bounds through Equation 6 and report the resulting midpoint. Using the recorded Coverage/FPR and rounding to three decimals gives Binary AUROC 0.631, 0.640, and 0.521 for the three runs, respectively. These are conditional summaries under the stated missing-output convention, not exact AUROCs of complete observed binary trajectories. Original bounds and invalid records remain in the scorer artifacts.

E PACT Training and Experimental Details

E.1 SFT optimization details

For the per-token weights in Equation 3, an initial silence decision receives weight 1, a repeated silence receives weight 0.4, and a response decision receives weight 1.5. The action token at a transition between silence and response receives weight 4. Other supervised assistant tokens, including warning text, receive unit weight. This weighting preserves normal-time supervision for false-alarm control while preventing long silence runs from dominating the objective. Warning-text tokens are averaged with the other supervised assistant tokens within their trajectory; we do not supervise a separate reasoning head. Each trajectory is normalized by its assistant-target token count before trajectories are averaged, so long monitoring sessions do not receive disproportionate weight. User inputs, visual tokens, and other prompt tokens are excluded from supervision.

E.2 Experimental details

The reported PACT checkpoint is obtained after 140 optimizer updates from JoyAI-VL-Interaction using the `preliminary_aligned_v2_normal50_r1` training view. It contains 1,114 trajectories and 63,823 turns; the development view contains 127 trajectories and 5,775 turns. Nearly all training trajectories contain some causal descriptions and historical memory. All 1,614 newly generated `v0_4_reasoning` anchor descriptions are included; the number of generated historical summaries used has not been separately counted. We fine-tune all parameters for one epoch in BF16 on a single server with four A100 80GB GPUs, using ZeRO-3. The per-device batch size is one with two gradient-accumulation steps, giving a global batch size of eight. We use AdamW with a learning rate of 5×10^{-6}, cosine decay, a 0.05 warm-up ratio, weight decay of 0.01, maximum gradient norm of 1, and random seed 20260913. Inputs are limited to 65,536 tokens; the complete training and development sets pass the encoding audit without truncation.

Single-event evaluation contains 600 complete videos and 85,629 scheduled decisions. Four anomalous videos with multiple independent source events are excluded only from the segment-level denominator, leaving 296 eligible segments; all 600 videos remain in the action metrics. Segment matching uses the two-second temporal tolerance defined in Appendix C.2. Delay is the signed difference between the first matched response and the source-segment onset and is reported over the 214 detected segments. Tolerated early responses are retained.

The mean wall-clock time per model call is 0.516 seconds, with P50/P90 values of 0.350/0.700 seconds; response-producing calls average 0.620 seconds. These measurements come from offline causal replay and are not interpreted as end-to-end wall-clock warning latency. The current evaluation does not score the warning grounding of generated responses.

Equal-coefficient SFT ablation. The uniform-weight `checkpoint-140` assigns unit weight to all supervised assistant tokens and retains the trajectory-mean loss reduction. It uses the same training trajectories, JoyAI initialization, optimizer settings, and global batch size of eight as weighted PACT. The weighted run used four A100 GPUs with a per-device batch size of one and two accumulation steps; the uniform run used four H200 GPUs with a per-device batch size of two and one accumulation step. Its Single-event evaluation completes all 600 videos and 85,629 decisions, with no runtime failures or invalid actions. The same annotation gate and two-second tolerance leave 296 eligible segments: 35 are detected, including 34 within five seconds. The mean signed response delay is 0.290 seconds over those 35 detections, compared with 214 detections for the weighted baseline; missed segments do not enter either delay average. The mean per-call inference time is 2.526 seconds. This evaluation uses eight H200 GPUs, with four each for the main and summary models. It reuses 75 completed videos evaluated at concurrency eight and evaluates the remaining 525 at concurrency sixteen, preserving all recorded call durations. Inference times were collected under different serving settings from the weighted baseline and do not provide a controlled speed comparison.

Exploratory Qwen3.5 backbone diagnostic. We also completed a 140-update, full-parameter SFT run initialized from Qwen3.5-9B on the same monitoring trajectories, using eight A100 80GB GPUs. On the 596-video provisional subset of the 600-video Single-event replay, the checkpoint-100 and checkpoint-140 runs produced valid actions at 94.57% and 77.59% of scored checkpoints, respectively. Their normal-time FPRs are undefined because of invalid outputs; the corresponding lower bounds are 94.09% and 74.95%. These diagnostics show that the current Qwen3.5 adaptation does not offer a reliable alternative monitoring policy, while the invalid outputs prevent a clean backbone ranking.

F Exploratory GRPO Adaptation

We report an unsuccessful exploratory GRPO adaptation whose monitoring targets are closely related to those of PACT SFT: remain silent during normal activity and respond during anomalous activity, with additional emphasis on state transitions. Despite this similarity, the tested variant rapidly over-specializes toward the normal/silence response pattern and loses anomaly sensitivity. We document this failure through training trajectories and completed Single-event evaluations. The experiment compares different temporal windows within each video group and updates only the highest-reward candidate; its negative results characterize this implementation rather than GRPO in general. All PACT results in the main text remain SFT-only.

F.1 Initialization and causal rollout design

We train two branches: one initialized from weighted PACT `checkpoint-140`, and one initialized directly from the released JoyAI-VL-Interaction checkpoint without our task-specific SFT. The latter uses a Qwen3-VL backbone, but is distinct from the official Qwen3-VL-8B-Instruct baseline evaluated elsewhere in this paper. The SFT initialization is also distinct from the uniform-weight SFT ablation in Appendix E.2. GRPO update numbers below count additional optimizer updates after the specified initialization.

The fixed training plan contains 176 videos, with four distinct 16-turn windows per video, giving 704 candidate windows from the GRPO partition. Each window contains at most one ground-truth state change. The four cases are normal-to-normal, normal-to-abnormal, abnormal-to-normal, and abnormal-to-abnormal, represented by 138, 233, 168, and 165 windows, respectively. Selection encourages case diversity but does not require every video group to contain all four cases.

At 1 Hz, each candidate receives up to 16 preceding frames paired with ground-truth *action tokens only*; ground-truth warning descriptions are not supplied. Within the selected 16-turn window, the model generates each action and optional warning using its own preceding outputs. Current and

future target actions are unavailable as model inputs, and frames beyond the current decision are never exposed. Generation continues after a response. Reward covers only these 16 decisions; behavior after the window is neither rolled out nor scored in that candidate. This training context is shorter and uses a different history distribution from full-video evaluation, which maintains self-generated interaction state and compressed memory.

F.2 Reward and policy update

Let $y_t \in \{0, 1\}$ be the annotated normal/abnormal state and $\hat{y}_t$ the parsed silence/response action at turn t. For a valid 16-turn trajectory, the implemented reward is

$$R = \frac{1}{2}\frac{\sum_{t=1}^{16} w_t \, \mathbf{1}[\hat{y}_t = y_t]}{\sum_{t=1}^{16} w_t} + \frac{1}{2}\frac{\sum_{t=1}^{15} v_t \, \mathbf{1}[(\hat{y}_t, \hat{y}_{t+1}) = (y_t, y_{t+1})]}{\sum_{t=1}^{15} v_t}. \tag{7}$$

The weights are one except at a true state-change edge b: the two adjacent decisions have $w_b = w_{b+1} = 4$, and that edge has $v_b = 8$. The second term checks the complete ordered pair of states, including stable pairs; it is not merely a test for whether a change occurs. The same weighting applies to anomaly onset and offset. Training uses zero temporal tolerance. Invalid actions, truncated generations, or incomplete model trajectories receive $R = -1$; infrastructure failures are raised as errors rather than converted into negative rewards.

There is no unconditional bonus for emitting a response. A response must match the annotated state to earn action reward. Neither the warning's semantic correctness nor its visual grounding is scored, and no external judge is called. Thus this is an action- and transition-matching reward, not a learned assessment of anomaly explanations.

Relationship to SFT. The two designs share the same 1 Hz causal monitoring task, silence/response action space, and interval-derived normal/abnormal targets. Both emphasize state transitions: SFT increases the loss weight of transition action tokens, while GRPO increases the reward weights of boundary decisions and adjacent state pairs. GRPO therefore largely reformulates the action supervision already used by SFT as a trajectory reward. The optimization procedures are different, however: SFT directly fits annotated actions and warning text, whereas this GRPO variant reinforces selected self-generated trajectories using an action-only reward. The training-window coverage and history construction also differ. Similar task targets do not make the two objectives or their effective training distributions equivalent.

One trajectory is sampled for each of the four different windows of a video. Writing their rewards as R_i, the selected candidate is $i^* = \arg\max_i R_i$, with ties resolved by candidate order. Its advantage is $(R_{i^*} - \bar{R})/(s_R + 10^{-6})$, where s_R is the sample standard deviation of the four rewards. The other three candidates are excluded from the update, rather than receiving negative advantages. Equal-reward groups have zero policy advantage; a batch with no nonzero-variance group is skipped. Consequently, these groups are not four sampled answers to an identical conditioning context.

The selected trajectories use a token-level clipped policy objective with clipping parameter 0.2 and KL coefficient 0.02 relative to a frozen copy of the branch's initial model. The objective is averaged over sampled assistant tokens within each selected trajectory and then over trajectories; it includes warning-text tokens even though the scalar reward scores actions only. Each update contains 16 video groups (64 candidate trajectories and 16 selected trajectories) across eight H200 GPUs. Four candidates are generated concurrently per rank, with two video groups assigned to each rank. Full-parameter BF16 training uses AdamW at 10^{-6}, gradient clipping at one, ZeRO-2 with CPU optimizer offload, activation checkpointing, and FlashAttention 2. Rollouts use temperature one, top-p one, no top-k truncation, a 32-token generation limit per turn, and a 65,536-token context limit. The plan specifies 19 passes over the fixed groups (209 updates), seed 42, and a checkpoint every eight updates. The SFT-initialized and direct JoyAI branches were stopped after 149 and 106 updates, respectively.

F.3 Completed evaluations and negative results

At evaluation submission time, SFT-initialized updates 16 and 96 were chosen for their high candidate-mean training rewards among the saved checkpoints then available. This was not held-out validation or test-based checkpoint selection. The direct JoyAI branch was evaluated at its final

saved checkpoint, update 104. For the rollout batches associated with these three updates, candidate-mean/selected-mean rewards were $0.538/0.947$, $0.581/0.966$, and $0.394/0.926$, respectively. Candidate mean averages all 64 candidates; selected mean averages the 16 group winners. These are training-batch summaries under the pre-update behavior policy, not post-update evaluation scores or a comparison on an identical set of windows.

All three checkpoints complete the frozen 600-video Single-event test with 85,629 decisions and zero runtime failures, invalid actions, or truncated outputs. Evaluation uses the JoyAI continuous adapter with self-generated history, the shared summarizer, and the same two-second scoring tolerance as the main comparison. Each run uses two H200 GPUs, one for the main model and one for the summarizer, with video concurrency four. Action metrics aggregate all 600 videos, including the four videos excluded only from segment scoring; the segment denominator is 296. Earlier summaries limited to the 596 segment-eligible videos have different action-metric denominators and should not be substituted for the values below.

Table 13: Negative Single-event results of the exploratory video-group GRPO variant. Coverage, Recall@5s, and FPR are percentages; Detected counts segments with any matched response, including responses later than five seconds. PACT is the weighted SFT-140 reference.

Initialization	GRPO update	Binary AUROC ↑	Coverage (%) ↑	Recall@5s (%) ↑	FPR (%) ↓	Detected / 296
PACT (SFT-140 only)	–	0.711	55.90	67.57	13.71	214
SFT-140 + GRPO	16	0.606	23.66	37.16	2.38	118
SFT-140 + GRPO	96	0.506	2.11	3.04	0.84	9
JoyAI base + GRPO	104	0.500	0.03	1.35	0.00	4

Table 13 shows reduced false alarms accompanied by substantial losses in anomaly coverage and timely detection. SFT-initialized update 16 detects 118 segments, of which 110 are timely; update 96 detects only nine, all timely. The direct JoyAI branch is nearly silent, with just four positive decisions over 12,847 scored anomalous checkpoints. Its action accuracy is still 84.81%, because normal checkpoints dominate the replay; an always-silent policy would already achieve 84.81% after rounding. High action accuracy therefore does not demonstrate useful anomaly detection.

Rapid over-specialization toward normal-time silence. The training logs show that the drift appears within the first few passes, despite the close relationship to SFT's action targets. Over the first three passes through the same 704 candidate windows (updates 1–11, 12–22, and 23–33), the fraction of entirely silent trajectories rises from 22.30% to 45.17% and 67.76% in the SFT-initialized branch. The direct JoyAI branch rises from 66.05% to 87.07% and 94.18%. Only 138/704 windows (19.60%) are normal throughout, so the increasing silence also extends into windows containing anomalous activity. These are training-pass aggregates with policies changing during each pass, rather than fixed-policy evaluations. The held-out results corroborate the behavioral degradation: after only 16 additional updates, SFT-initialized Recall@5s falls from 67.57% to 37.16%, and by update 96 it reaches 3.04%. Thus the observed failure is rapid over-specialization to the normal/silence response pattern, at the expense of abnormal-event sensitivity, even though the intended supervision resembles SFT.

The logs also expose a potential weakness of the selection rule. At SFT-initialized updates 16 and 96, 13/16 and 11/16 selected winners, respectively, come from normal-to-normal windows. At direct JoyAI update 104, all 64 sampled candidates are entirely silent, yet the selected-mean reward is 0.926. Perfectly matching an easy normal window can therefore produce a high winner reward while harder abnormal windows contribute no policy update. Comparing different windows within a group also mixes policy quality with differences in window difficulty. This is a plausible contributor to the observed silence bias, but the experiment does not isolate its causal effect from the other training choices.

Other plausible contributors are the short rollout horizon, action-only ground-truth prefixes, reward mismatch with full-session coverage and recall, and the absence of warning-grounding supervision. Controlled ablations would be needed to distinguish them. In particular, a higher training reward is not evidence of improved abnormal-event sensitivity, and these results do not show that extended reasoning is necessary or that GRPO cannot work for this task. The evaluations retain the bench-

mark's provisional source-segment annotations and do not assess warning grounding; they establish a failure to improve the measured monitoring tradeoff for the tested variant.