VAD-TTRL: PROPOSE, VERIFY, REINFORCE FOR LABEL-FREE TEST-TIME VIDEO ANOMALY DETECTION AND UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have shown strong capabilities in video understanding, yet video anomalies are often scene- and deployment-dependent, limiting the reliability of fixed pretrained models across environments. Conventional adaptation typically requires task-specific annotations and expensive post-training. Test-time reinforcement learning offers an appealing alternative by adapting models directly on unlabeled test videos using self-generated supervision. However, existing approaches commonly treat self-consensus among sampled predictions as supervision, which can reinforce systematic errors. We introduce **VAD-TTRL**, a label-free test-time reinforcement learning framework based on a *propose–verify–reinforce* paradigm. Multiple stochastic reasoning trajectories first form a structured consensus hypothesis, which is then verified using temporal and spatial consistency together with counterfactual visual evidence. Only sufficiently supported hypotheses contribute to reliability-weighted reinforcement signals, while uncertain predictions are excluded from adaptation. These rewards are optimized with Group Relative Policy Optimization (GRPO), enabling adaptation without ground-truth anomaly labels or task-specific annotations. Experiments across multiple MLLM backbones show that VAD-TTRL improves anomaly detection and temporal localization while largely preserving reasoning quality. On Qwen2.5-VL-7B, it improves accuracy from 0.761 to 0.829 and mIoU from 0.567 to 0.657; when applied on top of sota VAD-R1 (Huang et al., 2025), it further improves accuracy from 0.875 to 0.886 and mIoU from 0.713 to 0.729, demonstrating the value of evidence-verified consensus for label-free test-time video anomaly understanding.

1 INTRODUCTION

Video anomaly detection (VAD) aims to identify unusual or harmful events in videos, such as violence, accidents, and theft (Sultani et al., 2019). Unlike conventional action recognition, anomalies are often sparse, temporally localized, and highly context-dependent (Saligrama et al., 2010). Traditional VAD methods learn visual appearance and motion patterns under fully supervised (Wang & Miao, 2010; Li et al., 2022a), weakly supervised (Sultani et al., 2019; Tian et al., 2021; Lv et al., 2023; Li et al., 2022b), or unsupervised settings (Liu et al., 2018; Gong et al., 2019; Wang & Cherian, 2019; Thakare et al., 2022; Tur et al., 2023; Yang et al., 2023). Despite strong detection performance, these approaches primarily produce anomaly scores and provide limited understanding of *what* happened, *when* and *where* it occurred, and *why* it is abnormal.

Recent VLM- and MLLM-based methods move toward semantic and explainable anomaly understanding by combining visual evidence with linguistic knowledge (Yuan et al., 2023; Wu et al., 2023; Chen et al., 2023; Yang et al., 2024; Tang et al., 2024; Ma et al., 2025; Yang et al., 2025). Methods such as Holmes-VAD (Zhang et al., 2024a) and Holmes-VAU (Zhang et al., 2025b) demonstrate this potential, but still lack structured thinking and rely largely on task-specific fine-tuning or instruction tuning over curated datasets, requiring additional supervision and retraining for new environments. Reinforcement learning (RL) has recently emerged as a powerful mechanism for improving structured reasoning in language and multimodal models (OpenAI, 2026; Guo et al., 2025; Zheng et al., 2025; Team, 2025a; Bai et al., 2025). Within video anomaly understanding, VAU-

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.

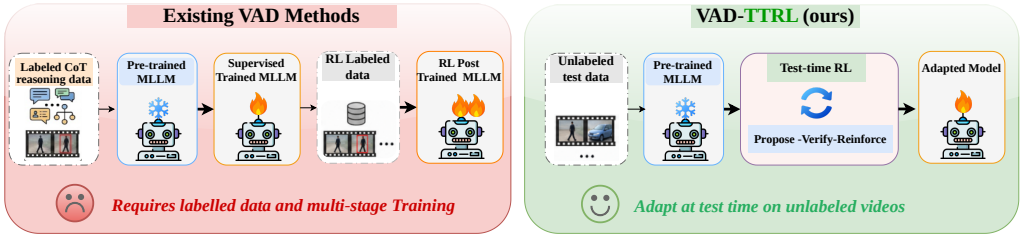

Figure 1: **Training pipeline comparison.** Existing RL-based VAD methods rely on labeled reasoning data, supervised fine-tuning, and RL post-training, whereas VAD-TTRL adapts a pretrained MLLM directly on unlabeled videos through evidence-verified test-time reinforcement learning.

R1 (Zhu et al., 2025) and Vad-R1 (Huang et al., 2025) use reinforcement fine-tuning to improve fine-grained anomaly reasoning, with Vad-R1 combining structured perception-to-cognition reasoning with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). As illustrated in Fig. 1, these methods typically rely on labeled reasoning data, supervised fine-tuning, and subsequent RL post-training. While effective, such pipelines require specialized annotations and multi-stage offline training. Constructing temporal labels, spatial descriptions, reasoning traces, and explanatory supervision at scale is expensive, and repeating this process for each deployment domain limits adaptability.

This motivates a fundamental question: *Can a pretrained MLLM improve its video anomaly detection and understanding directly from unlabeled test videos, without task-specific ground-truth supervision?* Test-Time Reinforcement Learning (TTRL) (Zuo et al., 2025) offers a promising direction by using majority agreement among sampled predictions as surrogate supervision on unlabeled test data. TTRV (Singh et al., 2025) extends this idea to vision-language models. However, directly applying majority-based reinforcement to video anomaly understanding is risky. Unlike tasks with short discrete answers, VAD requires a structured interpretation involving anomaly category, temporal extent, spatial evidence, event description, and rationale.

Moreover, correlated samples may agree on the same incorrect interpretation, so self-consistency does not necessarily imply visual correctness. Treating consensus as supervision without verification can therefore reinforce systematic errors rather than correct them.

To address this limitation, we introduce **VAD-TTRL**, an evidence-verified test-time reinforcement learning framework for label-free video anomaly detection and understanding. Our key principle is that *consensus should form a proposal rather than an immediate training target*. VAD-TTRL follows a *propose–verify–reinforce* paradigm: multiple stochastic trajectories first construct a structured consensus hypothesis; the hypothesis is then evaluated using temporal and spatial consistency together with counterfactual visual evidence; and only sufficiently supported hypotheses contribute to the policy update.

For an abnormal proposal, VAD-TTRL tests whether the predicted temporal segment actually supports the anomaly decision: suppressing that segment should reduce abnormal confidence, whereas retaining it should largely preserve the decision. We also perturb a matched non-anomalous segment to check that the response is specific to the proposed anomaly interval rather than a generic effect of modifying the video. For a Normal proposal, we instead test whether the prediction remains stable under multiple temporal perturbations and whether any competing abnormal evidence emerges. The resulting evidence-weighted rewards are optimized with GRPO (Shao et al., 2024), without using ground-truth anomaly labels, temporal boundaries, spatial annotations, or reference explanations during test-time learning. Our contributions are:

- We introduce **VAD-TTRL**, a label-free test-time reinforcement learning framework that updates pretrained MLLMs directly from unlabeled videos using self-generated and visually verified supervision.
- We propose an **evidence-verified propose–verify–reinforce** strategy that treats self-consensus as a hypothesis rather than direct supervision, and performs policy updates only when the hypothesis is supported by structured consistency and counterfactual visual evidence.

- We show across multiple MLLM backbones and model scales that VAD-TTRL improves anomaly detection and temporal localization while maintaining competitive anomaly reasoning performance, and can further enhance an already post-trained anomaly reasoning model.

2 Related Work

Conventional Video Anomaly Detection. Video Anomaly Detection (VAD) identifies frames or temporal segments that deviate from expected activity (Hasan et al., 2016), with applications in surveillance and video understanding (Sultani et al., 2019; Wu et al., 2020). Conventional methods primarily model visual appearance and motion under different supervision regimes. Unsupervised approaches learn normal patterns and detect deviations (Liu et al., 2018; Gong et al., 2019; Wang & Cherian, 2019; Thakare et al., 2022; Tur et al., 2023; Yang et al., 2023), while fully supervised methods exploit fine-grained annotations for localization (Wang & Miao, 2010; Liu & Ma, 2019). Weakly supervised methods instead learn from video-level labels (Sultani et al., 2019; Tian et al., 2021; Lv et al., 2023), typically through Multiple Instance Learning, with subsequent work improving instance weighting, feature modeling, temporal efficiency, and robustness to noisy supervision (Ilse et al., 2018; Chen et al., 2022; Gao & Liu, 2025; Zhou et al., 2023b;a). Despite strong detection performance, these approaches largely produce anomaly scores and provide limited semantic understanding of *what*, *when*, *where*, and *why* an event is anomalous.

Multimodal and Explainable VAD. Recent work incorporates vision-language models (VLMs) and multimodal large language models (MLLMs) to move beyond anomaly scoring toward semantic and explainable video understanding (Yuan et al., 2023). Holmes-VAD (Zhang et al., 2024a) and Holmes-VAU (Zhang et al., 2025b) combine temporal modeling with VLM-based reasoning to produce interpretable anomaly descriptions, while VAD-LLaMA (Lv & Sun, 2024) employs instruction tuning and temporal memory for anomaly detection and explanation. CLIP-based methods such as VadCLIP (Wu et al., 2023) exploit frozen vision-language representations together with adaptive prompting and MIL objectives. VERA (Ye et al., 2025) further performs coarse-to-fine anomaly reasoning through scene-temporal integration. Training-free approaches have also emerged (Zanella et al., 2024; Wani et al., 2026; 2027), leveraging pretrained multimodal models without task-specific optimization. However, they typically use multi-stage pre/post-processing around a fixed inference model and therefore do not adapt the model itself using unlabeled test videos.

Reinforcement Learning for Multimodal Reasoning. Recent advances in reasoning-oriented language models have motivated the use of reinforcement learning to improve structured multimodal reasoning. Several works develop chain-of-thought supervision, temporal reasoning pipelines, and large-scale reasoning datasets for MLLMs (Thawakar et al., 2025; Arnab et al., 2025; Hao Fei, 2024). The success of reinforcement-based reasoning (Shao et al., 2024) has further inspired video reasoning models such as Video-R1 (Feng et al., 2025) and DeepVideo-R1 (Park et al., 2026). Within video anomaly understanding, VAU-R1 (Zhu et al., 2025) and Vad-R1 (Huang et al., 2025) use reinforcement learning to improve temporal, spatial, and explanatory reasoning. These approaches demonstrate the benefits of RL for structured video understanding but typically rely on task-specific reasoning annotations and multi-stage post-training, including supervised fine-tuning followed by reinforcement learning. Such pipelines remain expensive to construct and difficult to repeat when adapting to new domains.

Test-Time Reinforcement Learning. Earlier test-time adaptation methods update models at inference through entropy minimization (Wang et al., 2021), robust continual pseudo-labeling (Wang et al., 2022), or external CLIP-based feedback for VLMs (Zhao et al., 2024). TTRL (Zuo et al., 2025) instead adapts language models directly on unlabeled test data using agreement among sampled outputs as self-supervision, with subsequent extensions to vision-language (Singh et al., 2025) and video understanding (Jahagirdar et al., 2026). However, existing test-time RL methods largely rely on response-level consensus or majority-based pseudo-labels, which is restrictive for video anomaly understanding where predictions jointly encode anomaly class, temporal extent, spatial evidence, and rationale, and correlated generations may agree on the same error. Our work addresses this limitation with VAD-TTRL. Rather than treating self-consensus as direct supervision, VAD-TTRL uses consensus only to *propose* a structured hypothesis, explicitly *verifies* it against temporal, spatial, semantic, and counterfactual visual evidence, and then *reinforces* evidence-supported trajectories through test-time RL.

3 METHODOLOGY

We consider label-free test-time adaptation of a pretrained multimodal large language model (MLLM) for video anomaly detection and understanding. Given an unlabeled adaptation set, VAD-TTRL updates the policy using only model-generated and visually verified supervision; no ground-truth anomaly labels, temporal boundaries, spatial annotations, or reference explanations are used. We evaluate the adapted policy both on the adaptation set and, without further updates, on unseen videos to assess transfer beyond transductive optimization.

Let v denote an unlabeled video and π_θ the current MLLM policy. We sample a total of N stochastic reasoning trajectories $\mathcal{O}(v) = \{o_i\}_{i=1}^{N}$, where each $o_i = (y_i, d_i, I_i, s_i, r_i)$ contains the predicted anomaly class, event description, temporal interval, spatial location, and explanatory rationale .

A naive test-time RL approach could directly reinforce the majority prediction, but correlated samples may agree on the same error; self-consistency therefore does not guarantee visual correctness. VAD-TTRL instead follows a *Propose–Verify–Reinforce* paradigm: consensus first forms a structured hypothesis, visual evidence is used to verify the hypothesis, and only evidence-supported trajectories contribute to the GRPO policy update. An overview is shown in Fig. 2.

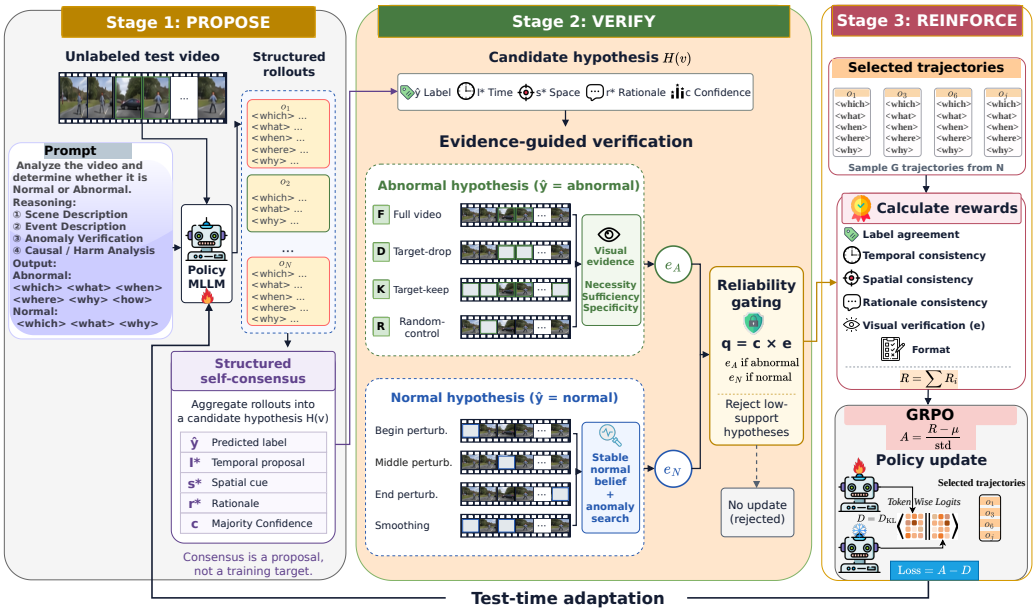

Figure 2: Overview of VAD-TTRL, a self-supervised test-time reinforcement learning framework for video anomaly detection and understanding. Given an unlabeled test video and a prompt, the policy MLLM first generates multiple structured reasoning trajectories to form a consensus proposal, then verifies the proposed hypothesis using spatio-temporal and counterfactual visual evidence, and finally performs a reliability-weighted GRPO update.

3.1 PROPOSE: STRUCTURED SELF-CONSENSUS

The PROPOSE stage transforms the sampled trajectories into a structured hypothesis describing the model's current interpretation of the video. We aggregate the predicted anomaly classes across the N sampled trajectories to obtain the candidate anomaly class and its majority-consensus confidence,

$$\hat{y} = \mathrm{mode}(y_1, \ldots, y_N), \qquad c = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[y_i = \hat{y}]. \tag{1}$$

Importantly, $\hat{y}$ is treated as a *proposal* rather than a pseudo-label that is immediately used for training. Importantly, $\hat{y}$ is treated as a *proposal* rather than a pseudo-label that is immediately used for training. For an Abnormal proposal, temporal predictions associated with the majority class are

compared through pairwise overlap, and a representative interval is selected as the temporal consensus,

$$I^* = \arg\max_{I_i} \frac{1}{|\mathcal{I}| - 1} \sum_{I_j \in \mathcal{I},\, j \neq i} \mathrm{IoU}(I_i, I_j), \tag{2}$$

where $\mathcal{I}$ contains temporal proposals associated with $\hat{y}$. Similarly, free-form spatial predictions are mapped to coarse canonical regions and aggregated to obtain a representative spatial hypothesis s^*. Rationales associated with the candidate class are compared semantically, and a representative rationale r^* is selected as the semantic consensus.

The proposal stage therefore constructs the hypothesis $\mathcal{H}(v) = (\hat{y}, I^*, s^*, r^*, c)$. where the temporal and spatial components are included only for an Abnormal proposal. This hypothesis summarizes *what the model currently believes*; whether that belief is supported by the video is determined in the VERIFY stage.

3.1.1 VERIFICATION OF ABNORMAL PROPOSALS

For an Abnormal proposal, the consensus interval $I^\star$ indicates where the model believes the anomalous evidence occurs. We verify this hypothesis by constructing four controlled views of the same video:

$$\mathcal{V}_A = \{F, D, K, R\}, \tag{3}$$

where F is the original video, D attenuates the proposed anomaly region, K preserves the proposed region while attenuating surrounding context, and R applies a matched intervention to a non-target temporal region. Let $p_A(X)$ denote the model's probability of predicting *Abnormal* for view X. The four views allow us to test three complementary properties of the proposed interval.

Necessity. If $I^\star$ contains evidence responsible for the Abnormal decision, removing it should reduce the model's Abnormal confidence:

$$\Delta_{\mathrm{nec}} = \max\left(0,\, p_A(F) - p_A(D)\right). \tag{4}$$

Sufficiency. If $I^\star$ contains sufficient evidence for the decision, retaining this region while suppressing the surrounding context should preserve a prediction close to that of the full video:

$$S_{\mathrm{suf}} = \max\left(0,\, 1 - |p_A(F) - p_A(K)|\right). \tag{5}$$

Matched control. A confidence change caused by perturbing $I^\star$ is meaningful only if it is not simply a consequence of perturbing any part of the video. We therefore require the prediction to remain stable when an equally strong intervention is applied to an off-target region:

$$C_{\mathrm{ctrl}} = \max\left(0,\, 1 - |p_A(F) - p_A(R)|\right). \tag{6}$$

The final Abnormal evidence score combines the model's original confidence with necessity, sufficiency, and intervention specificity:

$$e_A = p_A(F) \left(\frac{\Delta_{\mathrm{nec}} + S_{\mathrm{suf}}}{2} \right) C_{\mathrm{ctrl}}. \tag{7}$$

A high e_A therefore indicates that the proposed interval is supported by the model's response to targeted visual interventions rather than by self-consensus alone.

3.1.2 VERIFICATION OF NORMAL PROPOSALS

Normal proposals require a different verification strategy because there is no predicted anomalous interval to remove. Instead, we test whether the Normal decision remains stable under mild temporal perturbations and whether any of these perturbations exposes strong competing evidence for abnormality.

We construct a set $\mathcal{U}$ of five altered video views: light masking of the beginning, middle, and ending portions of the video, together with sparse and dense temporal smoothing. Let $p_N(v)$ and $p_A(v)$ denote the model's Normal and Abnormal probabilities for an altered view v, respectively.

We first measure the average stability of the Normal prediction,

$$\mathrm{Stab}_N = \frac{1}{|\mathcal{U}|} \sum_{v \in \mathcal{U}} p_N(v), \tag{8}$$

and the strongest competing Abnormal response,

$$\mathrm{Risk}_A = \max_{v \in \mathcal{U}} p_A(v). \tag{9}$$

The resulting Normal evidence score is

$$e_N = \mathrm{clip}_{[0,1]} \left(\mathrm{Stab}_N - \lambda_N \, \mathrm{Risk}_A \right), \tag{10}$$

where λ_N controls how strongly emerging Abnormal evidence is penalized. Thus, e_N is high when the Normal interpretation remains stable across perturbations and no altered view reveals strong evidence for an anomaly.

3.1.3 Reliability-Gated Pseudo-Supervision

The class-conditional visual evidence score is

$$e = \begin{cases} e_A, & \hat{y} = \text{Abnormal}, \\ e_N, & \hat{y} = \text{Normal}. \end{cases} \tag{11}$$

A proposal is retained only when its visual evidence satisfies $e \geq \tau_{\text{vis}}$. Otherwise, VAD-TTRL abstains from learning from that video and performs no policy update.

For an accepted proposal, we combine its majority-consensus confidence c with its visual evidence score e to obtain the verified pseudo-label quality $q = c\, e$.

Thus, reliable pseudo-supervision requires both agreement among sampled trajectories and support from the underlying video evidence. VAD-TTRL therefore performs *selective test-time learning*, reinforcing only sufficiently supported hypotheses.

3.2 Reinforce: Reliability-Weighted Test-Time Optimization

Once a hypothesis passes verification, the Reinforce stage converts the verified structure into trajectory-level learning signals. Rather than optimizing all proposal trajectories, we sample a smaller group $\mathcal{G} \subset \mathcal{O}$ from the original proposal pool, allowing both majority-consistent and inconsistent trajectories to participate in the relative ranking.

For each selected trajectory o_i, we construct the structured reward

$$R_i = \lambda_y R_i^y + \lambda_t R_i^t + \lambda_s R_i^s + \lambda_r R_i^r + \lambda_v R_i^v + R_i^{\text{constraint}}, \tag{12}$$

where R_i^y measures agreement with the verified anomaly decision, R_i^t temporal consistency, R_i^s spatial agreement, R_i^r rationale consistency, and R_i^v transfers the visual-evidence score to trajectories consistent with the accepted hypothesis. $R_i^{\text{constraint}}$ promotes valid structured outputs and penalizes degenerate predictions. Crucially, these rewards are conditioned on the *verified* hypothesis rather than raw majority voting; agreement with an unsupported consensus therefore provides no learning signal. We optimize the resulting rewards using GRPO (Shao et al., 2024), which normalizes rewards within each sampled group and updates the policy according to their relative advantages. Full reward definitions and optimization details are provided in Appendix E.

4 Experiments

Implementation Details. We instantiate VAD-TTRL with Qwen2.5-VL-7B (Bai et al., 2025) and perform a single stage of GRPO-based (Shao et al., 2024) test-time adaptation directly on unlabeled videos. Unlike supervised post-training approaches (Huang et al., 2025; Zhu et al., 2025), no ground-truth anomaly labels, temporal boundaries, spatial annotations, or reference rationales are used during adaptation. We set the majority-consensus threshold to $\tau_{\text{maj}} = 0.75$ and the visual-evidence threshold to $\tau_{\text{vis}} = 0.20$, and optimize for one epoch with a learning rate of 1×10^{-6}. The

reward combines label, temporal, spatial, rationale-consistency, and visual-verification signals, with rationale similarity measured using BERTScore (Zhang et al., 2020). Prompt design, full hyperparameters, reward weights, and computational details are provided in Appendix B.

Evaluation Metrics. We evaluate VAD-TTRL on anomaly classification, temporal localization, and reasoning. Classification is measured using Accuracy, Precision, Recall, and F1; temporal localization using mIoU and Recall@K; and generated explanations using BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE (Lin, 2004). We additionally evaluate anomaly understanding on VANE-Bench (Bharadwaj et al., 2025) using accuracy across benchmark subsets.

Baselines. We compare against open-source video MLLMs, video reasoning models, proprietary MLLMs, and MLLM-based VAD methods, covering both general-purpose and anomaly-specific approaches. To assess backbone generalization, we further apply VAD-TTRL to multiple QwenVL, InternVL, and anomaly-reasoning backbone models and compare each backbone before and after adaptation.

Evaluation Questions. Our experiments examine whether VAD-TTRL (i) improves detection, localization, and reasoning over existing methods, (ii) transfers across different backbones using unlabeled videos, (iii) benefits from the individual *Propose–Verify–Reinforce* components, and (iv) improves detection and localization without degrading anomaly reasoning.

4.1 Main Results

Vad-Reasoning Benchmark. We evaluate on the 439-video Vad-Reasoning test set introduced by Vad-R1 (Huang et al., 2025), containing annotations for anomaly classification, temporal localization, spatial understanding, and reasoning. VAD-TTRL uses these annotations only for evaluation; adaptation accesses use only the unlabeled videos and task prompt.

Detection, Localization, and Reasoning. Table 1 compares VAD-TTRL with general-purpose, reasoning-oriented, and anomaly-specific MLLMs. Despite using no ground-truth supervision during adaptation, VAD-TTRL achieves strong detection and temporal localization performance. Vad-R1 (Huang et al., 2025) remains stronger overall, but relies on supervised fine-tuning followed by task-specific reinforcement learning, whereas ours, VAD-TTRL, performs a single stage of label-free test-time adaptation.

Table 1: Performance comparison on the Vad-Reasoning benchmark. We report anomaly reasoning, classification, and temporal localization performance. VAD-TTRL is evaluated as a label-free test-time adaptation approach. † denotes a task-specific model trained with supervised fine-tuning and reinforcement learning on curated anomaly-reasoning data.

Method	Params.	Anomaly Detection & Localization					Anomaly Reasoning		
		Acc	F1	mIoU	R@0.3	R@0.5	BLEU-2	METEOR	ROUGE-2
Open-Source Video MLLMs									
VideoChat-Flash (Li et al., 2024b)	7B	0.683	0.487	0.536	0.538	0.358	0.012	0.084	0.047
VideoLLaMA3 (Zhang et al., 2025a)	7B	0.665	0.624	0.425	0.451	0.419	0.066	0.200	0.092
LLaVA-NeXT-Video (Zhang et al., 2024b)	7B	0.651	0.423	0.576	0.601	0.585	0.094	0.238	0.104
Qwen2.5-VL (Bai et al., 2025)	7B	0.761	0.730	0.567	0.610	0.563	0.113	0.264	0.116
Open-Source Video Reasoning MLLMs									
Open-R1-Video (Wang et al., 2026)	7B	0.793	0.790	0.559	0.642	0.540	0.060	0.179	0.084
Video-R1 (Feng et al., 2025)	7B	0.624	0.694	0.334	0.392	0.328	0.135	0.317	0.132
VideoChat-R1 (Li et al., 2025)	7B	0.793	0.790	0.559	0.642	0.540	0.128	0.287	0.123
Proprietary MLLMs									
Claude3.5-Haiku (Anthropic, 2024)	–	0.580	0.354	0.518	0.543	0.524	0.097	0.253	0.098
GPT-4o (Hurst et al., 2024)	–	0.711	0.760	0.472	0.565	0.476	**0.154**	**0.341**	0.133
Gemini2.5-Flash (Gemini Team, 2025)	–	0.624	0.707	0.370	0.437	0.358	0.133	0.308	0.120
MLLM-based VAD Methods									
Holmes-VAD (Zhang et al., 2024a)	7B	0.565	0.120	–	–	–	0.003	0.074	0.027
Holmes-VAU (Zhang et al., 2025b)	2B	0.490	0.371	–	–	–	0.077	0.182	0.075
HAWK (Tang et al., 2024)	7B	0.513	0.648	–	–	–	0.042	0.156	0.042
Vad-R1 (Huang et al., 2025)	7B	0.875†	0.862†	0.713†	0.770†	0.706†	0.233†	0.406†	0.194†
Test-Time Reinforcement Learning									
TTRL (Zuo et al., 2025)	7B	0.649	0.412	0.585	0.606	0.588	0.112	0.250	0.111
VAD-TTRL (Ours)	7B	**0.829**	**0.802**	**0.657**	**0.729**	**0.656**	0.131	0.280	**0.156**

Generalization Across MLLM Backbones. To examine whether VAD-TTRL depends on a particular pretrained model, we apply the same *Propose–Verify–Reinforce* procedure to six models spanning the Qwen and InternVL families as well as the task-specific Vad-R1 model. For each backbone, we compare the original checkpoint with its test-time adapted counterpart under the same unlabeled adaptation protocol. Table 2 shows that VAD-TTRL consistently improves anomaly detection and temporal localization across all evaluated backbones, indicating that the proposed adaptation strategy is not specific to Qwen2.5-VL.

Table 2: Effectiveness of VAD-TTRL across different MLLM backbones on the Vad-Reasoning benchmark. For each backbone, we compare the original model with its test-time label-free adapted counterpart. **Bold** indicates the better result within each backbone pair. Arrows report the relative percentage **%** change from the corresponding base model.

Backbone	Adaptation	Detection & Localization					Anomaly Reasoning		
		Acc	F1	mIoU	R@0.3	R@0.5	BLEU-2	METEOR	ROUGE-2
Qwen Models									
Qwen2.5-VL-7B (Bai et al., 2025)	Base	0.761	0.730	0.567	0.610	0.563	0.113	0.264	0.116
Qwen2.5-VL-7B	+ VAD-TTRL	**0.829** ↑ 8.9	**0.802** ↑ 9.9	**0.657** ↑ 15.9	**0.729** ↑ 19.5	**0.656** ↑ 16.5	**0.131** ↑ 15.9	**0.280** ↑ 6.1	**0.156** ↑ 34.5
Qwen3-VL-8B (Team, 2025b)	Base	0.770	0.779	0.524	0.590	0.528	0.163	**0.352**	0.148
Qwen3-VL-8B	+ VAD-TTRL	**0.836** ↑ 8.6	**0.814** ↑ 4.5	**0.621** ↑ 18.5	**0.681** ↑ 15.4	**0.629** ↑ 19.1	**0.164** ↑ 0.6	0.334 ↓ 5.1	**0.161** ↑ 8.8
Qwen3.5-4B (Qwen Team, 2026)	Base	0.726	0.612	0.526	0.529	0.520	0.078	0.226	0.075
Qwen3.5-4B	+ VAD-TTRL	**0.790** ↑ 8.8	**0.791** ↑ 29.2	**0.551** ↑ 4.8	**0.608** ↑ 14.9	**0.551** ↑ 6.0	**0.112** ↑ 43.6	**0.248** ↑ 9.7	**0.144** ↑ 92.0
Anomaly-Reasoning Model									
Vad-R1(SFT+RL) (Huang et al., 2025)	Base	0.875	0.862	0.713	0.770	0.706	**0.233**	**0.406**	0.194
Vad-R1	+ VAD-TTRL	**0.886** ↑ 1.3	**0.869** ↑ 0.8	**0.729** ↑ 2.2	**0.779** ↑ 1.2	**0.720** ↑ 2.0	0.232 ↓ 0.4	0.405 ↓ 0.2	**0.215** ↑ 10.8
InternVL Models									
InternVL2.5-8B (Chen et al., 2025)	Base	0.703	0.732	0.417	0.451	0.414	0.106	0.260	0.103
InternVL2.5-8B	+ VAD-TTRL	**0.743** ↑ 5.7	**0.769** ↑ 5.1	**0.482** ↑ 15.6	**0.522** ↑ 15.7	**0.437** ↑ 5.6	**0.116** ↑ 9.4	**0.284** ↑ 9.2	**0.112** ↑ 8.7
InternVL3_5-8B (Wang et al., 2025)	Base	0.749	0.703	0.600	0.601	0.558	0.131	0.316	0.118
InternVL3_5-8B-Instruct	+ VAD-TTRL	**0.763** ↑ 1.9	**0.705** ↑ 0.3	**0.623** ↑ 3.8	**0.663** ↑ 10.3	**0.585** ↑ 4.8	**0.135** ↑ 3.1	**0.326** ↑ 3.2	0.118 =

Performance Comparison on VANE-Bench. We further evaluate the adapted VAD-TTRL policy, without additional parameter updates, on VANE-Bench (Bharadwaj et al., 2025), a challenging benchmark for video anomaly understanding based on diverse multiple-choice questions across AI-generated and real-world videos. As shown in Table 3, VAD-TTRL consistently outperforms both general-purpose MLLMs and specialized MLLM-based VAD approaches. In particular, it achieves the best performance on nearly all evaluated subsets. This demonstrates that the *learned adaptation* transfers beyond the Vad-Reasoning videos used for optimization.

Table 3: Performance comparison on VANE-Bench (Bharadwaj et al., 2025). Accuracy (%) is reported across synthetic and real-world videos. **Bold** and underline indicate the best and second-best results, respectively.

Method	SORA	OpenSORA	RG2	VideoLCM	MS-T2	Avenue	Ped1	Ped2
Open-Source MLLMs								
Video-LLaMA (Zhang et al., 2023)	11.59	18.00	16.00	10.57	10.41	30.00	16.66	5.55
VideoChat (Li et al., 2024a)	10.74	28.00	4.00	17.64	20.83	32.25	13.33	13.88
Video-ChatGPT (Maaz et al., 2024)	26.47	22.00	12.00	18.26	16.66	39.39	40.00	19.44
Video-LLaVA (Lin et al., 2024)	10.86	18.00	16.00	19.23	16.66	3.03	2.77	6.06
MovieChat (Song et al., 2024)	8.69	10.00	16.00	14.42	6.25	18.18	6.66	11.11
LLaMA-VID (Li et al., 2023)	7.97	14.00	20.00	19.23	14.58	27.27	6.66	19.44
TimeChat (Ren et al., 2024)	21.73	26.00	28.00	22.11	20.83	24.20	27.58	11.11
MLLM-based VAD (Training + Instruction-Tuned)								
Holmes-VAU (Zhang et al., 2025b)	2.17	34.00	24.00	29.81	25.00	6.06	3.33	5.56
Holmes-VAD (Zhang et al., 2024a)	6.52	34.00	32.00	33.56	22.92	12.12	20.00	5.56
HAWK (Tang et al., 2024)	24.64	52.00	44.00	36.54	50.00	36.36	36.67	38.89
Vad-R1(SFT+RL) (Huang et al., 2025)	41.30	78.00	56.00	63.46	60.42	75.76	60.00	63.89
Reasoning-Based Training-Free VAD								
Cog-VADU (Wani et al., 2026)	24.64	80.00	60.00	50.00	50.00	72.73	56.67	72.22
Test-Time Reinforcement Learning								
VAD-TTRL (Ours)	**53.62**	**90.00**	**64.00**	**67.31**	**81.25**	**81.82**	**80.00**	**72.22**

5 ABLATIONS

5.1 FROM CONSENSUS PROPOSALS TO EVIDENCE-VERIFIED UPDATES

We isolate the effects of majority-confidence filtering, hard visual verification, and evidence-weighted reinforcement. P0 directly reinforces the consensus proposal; P1 adds confidence filtering; P2 and P3 extend P1 with hard evidence gating and soft evidence weighting, respectively; and P4 combines both mechanisms. All other data, decoding, reward, and optimization settings are fixed. Ground-truth annotations are used only for the retrospective audit below and are never available during adaptation.

Table 4: **Propose–Verify–Reinforce ablation on Vad-Reasoning.** PL Acc. denotes retrospective pseudo-label accuracy and Cov. the fraction of videos contributing to policy updates. P4 is the full VAD-TTRL model.

	Enabled Mechanism			**Update Set**		**Post-Adaptation Detection**			
Stage	**Conf.**	**Gate**	**Weight**	**Cov.**	**PL Acc.**	**Acc.**	**Prec.**	**Rec.**	**F1**
P0 Direct	–	–	–	100.0	71.3	64.9	93.1	26.5	41.2
P1 Confidence	✓	–	–	85.0	74.5	72.0	84.0	49.0	61.9
P2 Gate	✓	✓	–	82.2	74.2	75.4	92.9	51.0	65.8
P3 Weight	✓	–	✓	65.8	86.5	76.5	71.1	**83.3**	76.7
P4 Full	✓	✓	✓	72.7	**89.0**	**82.9**	86.9	74.5	**80.2**

Consensus alone is insufficient. Directly reinforcing consensus proposals (P0) yields only 71.3% pseudo-label accuracy and strongly favors Normal predictions, resulting in only 26.5% anomaly recall. Confidence filtering (P1) improves both the quality of the update set and downstream performance, but consensus strength still does not guarantee correctness: even in the full run, 23 of 230 proposals with unanimous $8/8$ agreement are incorrect. Thus, agreement is useful for proposing a hypothesis but is insufficient as supervision by itself.

Hard and soft verification play complementary roles. Hard evidence gating (P2) produces a conservative policy with 92.9% precision but only 51.0% recall, whereas evidence weighting (P3) increases recall to 83.3% at the cost of lower precision (71.1%). This exposes two distinct effects of visual evidence: gating suppresses unreliable updates, while weighting preserves useful but uncertain anomaly evidence.

Combining both gives the strongest balance. P4 achieves the best overall Accuracy (82.9%) and F1 (80.2%), together with 89.0% accepted pseudo-label accuracy at 72.7% update coverage. Within the full run, verification rejects 34.0% of eligible incorrect proposals while retaining 99.0% of eligible correct proposals. These results support the *Propose–Verify–Reinforce* design: consensus proposes the hypothesis, verification determines whether it should be trusted, and evidence weighting controls the strength of reinforcement.

Additional ablations on reward-component weights, rollout configurations, and threshold sensitivity are provided in Appendix F. Qualitative results are presented in Appendix G, while limitations and future directions are discussed in Appendix H.

6 CONCLUSION

We presented **VAD-TTRL**, a label-free test-time reinforcement learning framework for video anomaly detection and understanding. VAD-TTRL follows a *propose–verify–reinforce* paradigm: multiple stochastic trajectories first form a structured consensus hypothesis, the hypothesis is then verified using temporal and spatial consistency together with counterfactual visual evidence, and only sufficiently supported predictions are reinforced. This enables test-time adaptation without ground-truth anomaly labels or task-specific annotations. Across multiple MLLM backbones, VAD-TTRL improves anomaly detection and temporal localization while largely preserving reasoning quality, highlighting the effectiveness of evidence-verified consensus for reliable test-time video anomaly understanding.

AI USE STATEMENT

Generative AI tools were used for editorial assistance, including grammar correction, prose polishing, formatting, and assistance with the preparation of figures, tables, and LaTeX code. They were also used for software implementation support, including CUDA environment setup and debugging. All AI-assisted code and content were reviewed and verified by the authors. Generative AI was not used to develop the research methodology, design experiments, generate or modify experimental results, or alter the datasets. All research ideas, methodological decisions, experimental designs, data, results, and scientific conclusions were developed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide evaluation protocols, prompts, reward construction, hyperparameters, rollout settings, and computational resources in the main paper and Appendix. The supplementary material further includes the complete training procedure and additional ablations. Upon acceptance, we will release the source code, configuration files, and model checkpoints required to reproduce the reported experiments.

REFERENCES

Anthropic. Claude 3.5 haiku. https://anthropic.com, 2024. Large language model.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames, 2025. URL https://arxiv.org/abs/2507.02001.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. URL https://api.semanticscholar.org/CorpusID:7164502.

Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vanebench: Video anomaly evaluation benchmark for conversational lmms, 2025. URL https://arxiv.org/abs/2406.10326.

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5549–5559, 2023. doi: 10.1109/CVPRW59228.2023.00587.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection, 2022. URL https://arxiv.org/abs/2211.15098.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. URL https://arxiv.org/abs/2503.21776.

Andrew Gao and Jun Liu. Stead: Spatio-temporal efficient anomaly detection for time and compute sensitive applications, 2025. URL https://arxiv.org/abs/2503.07942.

Gemini Team. Gemini 2.5 flash preview model card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash-preview.pdf, 2025.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection, 2019. URL https://arxiv.org/abs/1904.02639.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Sun, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Wei Ji Hanwang Zhang Meishan Zhang Mong-Li Lee Wynne Hsu Hao Fei, Shengqiong Wu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceeding of the ICML*, 2024.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences, 2016. URL https://arxiv.org/abs/1604.04574.

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, 2025. URL https://arxiv.org/abs/2505.19877.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alexandre Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, A. Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, A. Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia,

Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, B. Ghorbani, Ben Leimberger, Ben Rossen, Benjamin Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, B. Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, C. Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, C. Gibson, Christina Kim, Christine Choi, C. McLeavey, Chris Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mély, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Phong Duc Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, E. Brevdo, Evan Mays, Farzad Khorasani, F. Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Hai-Biao Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Pondé de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, I. Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, İ. Okuyucu, Ikai Lan, Ilya Kostrikov, I. Sutskever, I. Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, J. Pachocki, James Aung, James Betker, James Crooks, James Lennon, J. Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, J. Q. Candela, Joe Beutler, Joe Landers, Joel Parish, Jo hannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Joshua Gross, Josh Kaplan, Josh Snyder, Josh Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, K. Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, L. Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Ouyang Long, Louis Feuvrier, Lu Zhang, Lukasz Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Made laine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, M. Hudnall, Marvin Zhang, Marwan Aljubeh, Ma teusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, M. Yatbaz, Mengxue Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Castro, Mikhail Pavlov, M. Brundage, Miles Wang, Mina Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, N. Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nickolas Stathas, Nick Turley, N. Tezak, Niko Felix, Nithanth Kudige, N. Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, O. Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, P. Welinder, Phil Tillet, Philip Pronin, Phil Tillet, Prafulla Dhariwal, Qim ing Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Raphael Gontijo Lopes, Raul Puri, Reah Miyara, R. Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, R. Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam S. Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, S. Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal A. Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. In *arXiv.org*, 2024.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018. URL https://arxiv.org/abs/1802.04712.

Soumya Shamarao Jahagirdar, Edson Araujo, Anna Kukleva, M. Jehanzeb Mirza, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Rogerio Feris, James R. Glass, and Hilde Kuehne. Ttavid: Generalized test-time adaptation for video reasoning, 2026. URL https://arxiv.org/abs/2604.00696.

Pilsung Kang. Auditing question-order effects in large language models with the qq equality: Mechanism characterization and a saturation caveat, 2026. URL https://arxiv.org/abs/2607.17219.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 333–350, Cham, 2022a. Springer Nature Switzerland. ISBN 978-3-031-19772-7.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024a. URL https://arxiv.org/abs/2305.06355.

S. Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022b. URL https://api.semanticscholar.org/CorpusID:248982052.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyun Zeng, Yu-Han Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *ArXiv*, abs/2501.00574, 2024b. URL https://api.semanticscholar.org/CorpusID:275212875.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023. URL https://arxiv.org/abs/2311.17043.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. URL https://arxiv.org/abs/2311.10122.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pp. 1490–1499, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350998. URL https://doi.org/10.1145/3343031.3350998.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline, 2018. URL https://arxiv.org/abs/1712.09867.

Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models, 2024. URL https://arxiv.org/abs/2401.05702.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection, 2023. URL https://arxiv.org/abs/2303.12369.

Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive llm, 2025. URL https://arxiv.org/abs/2502.18863.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2024. URL https://arxiv.org/abs/2306.05424.

OpenAI. Openai o1 system card, 2026. URL https://arxiv.org/abs/2412.16720.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo, 2026. URL https://arxiv.org/abs/2506.07464.

Qwen Team. Qvq-max: Think with evidence, March 2025. URL https://qwenlm.github.io/blog/qvq-max-preview/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024. URL https://arxiv.org/abs/2312.02051.

Venkatesh Saligrama, Janusz Konrad, and Pierre-marc Jodoin. Video anomaly identification. *IEEE Signal Processing Magazine*, 27(5):18–33, 2010. doi: 10.1109/MSP.2010.937393.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M. Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models, 2025. URL https://arxiv.org/abs/2510.06783.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. URL https://arxiv.org/abs/2307.16449.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos, 2019. URL https://arxiv.org/abs/1801.04264.

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, 2024. URL https://arxiv.org/abs/2405.16886.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms, 2025a. URL https://arxiv.org/abs/2501.12599.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Kamalakar Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network, 2022. URL https://arxiv.org/abs/2211.00882.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025. URL https://arxiv.org/abs/2501.06186.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, 2021. URL https://arxiv.org/abs/2101.10030.

Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection, 2023. URL https://arxiv.org/abs/2304.05841.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. URL https://arxiv.org/abs/2006.10726.

Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection, 2019. URL https://arxiv.org/abs/1908.05884.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation, 2022. URL https://arxiv.org/abs/2203.13591.

Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. pp. 1220 – 1223, 11 2010. doi: 10.1109/ICOSP.2010.5655356.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xiaodong Wang, Zhirong Wu, Langling Huang, Yuxi Zheng, and Peixi Peng. Incentivizing versatile video reasoning in mllms via data-efficient reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5444–5454, 2026.

Mohd Ubaid Wani, Sara Atito, Josef Kittler, and Muhammad Awais. Cog-VADU: A training-free cognitive reasoning framework for video anomaly detection and understanding. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=QcuSMNG7J8.

Mohd Ubaid Wani, Sara Atito, Srinivasa Rao Nandam, Josef Kittler, and Muhammad Awais. Cozsr-vad: Contextual zero-shot reasoning for video anomaly detection. In Maria De Marsico, Tin Kam Ho, Frederic Jurie, Cheng-Lin Liu, Daniel Lopresti, Ingela Nyström, Jean-Marc Ogier, Arun Ross, and Liang Wang (eds.), *Pattern Recognition*, pp. 687–702, Cham, 2027. Springer Nature Switzerland. ISBN 978-3-032-31920-3.

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, 2020. URL https://arxiv.org/abs/2007.04687.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection, 2023. URL https://arxiv.org/abs/2308.11681.

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

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection, 2023. URL https://arxiv.org/abs/2304.05112.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection, 2024. URL https://arxiv.org/abs/2404.08531.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models, 2025. URL https://arxiv.org/abs/2412.01095.

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, 2023. URL https://arxiv.org/abs/2309.13925.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection, 2024. URL https://arxiv.org/abs/2404.01014.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. URL https://arxiv.org/abs/2306.02858.

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, 2024a. URL https://arxiv.org/abs/2406.12235.

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, 2025b. URL https://arxiv.org/abs/2412.06171.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models, 2024. URL https://arxiv.org/abs/2305.18010.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection, 2023a. URL https://arxiv.org/abs/2302.05160.

Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Hengtao Shen. Batchnorm-based weakly supervised video anomaly detection, 2023b. URL https://arxiv.org/abs/2311.15367.

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.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084.

A Appendix Overview

This appendix provides supplementary details and analyses for VAD-TTRL. We first describe the construction of the reinforcement signals used during test-time adaptation, followed by ablations of the individual reward components. We then provide additional implementation details, including prompts, rollout generation, optimization settings, hyperparameters, and computational resources. Finally, we report additional quantitative and qualitative results, followed by a discussion of limitations and broader impact.

B Additional Implementation Details

Training is performed on eight NVIDIA A100 GPUs with 64 GB memory. For evaluation, we use vLLM (Kwon et al., 2023) for efficient inference on a single NVIDIA RTX 3090 GPU.

B.1 Prompt Design

Figure 3 shows the structured prompt used by VAD-TTRL. It contains three parts: *Task Definition*, *Output Specification*, and *Format Requirements*. The model predicts the anomaly class together with structured fields describing *what* occurs, *when* and *where* it occurs, and *why* it is anomalous. Abnormal outputs use `<which>`, `<what>`, `<when>`, `<where>`, `<why>`, and `<how>`, whereas Normal outputs use `<which>`, `<what>`, and `<why>`. The fixed output format enables automatic rollout parsing for consensus construction and reward computation.

C Additional Experimental Results

C.1 LLM-Guided Evaluation

Token-overlap metrics such as BLEU, METEOR, and ROUGE provide only a partial measure of explanation quality. We therefore additionally adopt the same LLM-guided evaluation protocol used by Vad-R1 (Huang et al., 2025), measuring *Reasonability*, *Detail*, and *Consistency*. Reasonability evaluates the coherence and logical validity of the explanation, Detail measures its specificity and informativeness, and Consistency measures agreement with the reference event information. Each criterion is scored in $[0, 1]$.

VAD-TTRL is evaluated on all 439 test videos using GPT-4o (`gpt-4o-2024-08-06`) (Hurst et al., 2024) as the judge. As shown in Table 5, VAD-TTRL achieves the highest Reasonability score among all compared methods, while maintaining competitive Detail and Consistency. More notably, applying VAD-TTRL to the already post-trained Vad-R1 Huang et al. (2025) model further improves Reasonability from 0.734 to 0.859, Detail from 0.659 to 0.721, and Consistency from 0.662 to 0.667, achieving the highest scores across all three criteria. These results indicate that evidence-verified test-time reinforcement can complement task-specific post-training and further improve the semantic quality of anomaly explanations.

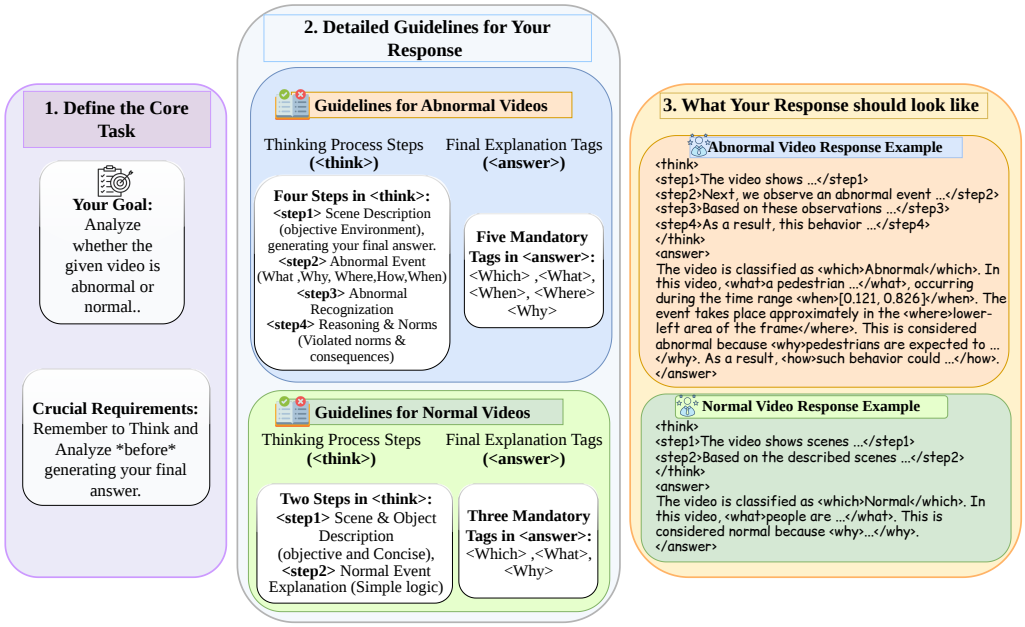

Figure 3: **Prompt design for VAD-TTRL.** The prompt defines the task, structured reasoning fields, and output format used to generate parseable trajectories for consensus construction and reward computation.

Table 5: LLM-guided evaluation of anomaly reasoning quality on the Vad-Reasoning benchmark. We report Reasonability, Detail, and Consistency scores following the evaluation protocol of Vad-R1 (Huang et al., 2025). Best and second-best results are shown in **bold** and underlined, respectively.

Method	Params.	LLM-Guided Anomaly Reasoning		
		Reasonability	Detail	Consistency
Open-Source Video MLLMs				
InternVideo2.5	8B	0.580	0.517	0.487
Flash (Li et al., 2024b)	7B	0.367	0.292	0.356
VideoLLaMA3 (Zhang et al., 2025a)	7B	0.549	0.449	0.497
LLaVA-NeXT-Video (Zhang et al., 2024b)	7B	0.541	0.452	0.491
Qwen2.5-VL (Bai et al., 2025)	7B	0.638	0.555	0.542
Open-Source Video Reasoning MLLMs				
Open-R1-Video (Wang et al., 2026)	7B	0.411	0.307	0.338
Video-R1 (Feng et al., 2025)	7B	0.390	0.414	0.243
VideoChat-R1 (Li et al., 2025)	7B	0.634	0.559	0.528
MLLM-based VAD Methods				
Holmes-VAD (Zhang et al., 2024a)	7B	0.388	0.275	0.343
Holmes-VAU (Zhang et al., 2025b)	2B	0.385	0.301	0.375
HAWK (Tang et al., 2024)	7B	0.218	0.185	0.115
Vad-R1 (Huang et al., 2025)	7B	0.734	0.659	0.662
Proprietary MLLMs				
QVQ-Max (Qwen Team, 2025)	–	0.690	0.639	0.521
GPT-4o (Hurst et al., 2024)	–	0.724	0.679	0.542
Qwen2.5-VL + VAD-TTRL (Ours)	7B	0.766	0.605	0.566
Vad-R1 + VAD-TTRL (Ours)	7B	**0.859**	**0.721**	**0.667**

Table 6: Generalization beyond the VAD-TTRL adaptation set on additional VAD-Reasoning videos. The adapted model is evaluated without further parameter updates. Gray rows denote VAD-TTRL adaptation.

Subset	Method	Acc.	Prec.	F1	mIoU	R@0.3	R@0.5	R@0.7
500	Qwen2.5-VL	.754	.715	.746	.536	.594	.532	.492
	Qwen2.5-VL + VAD-TTRL	**.806**	**.836**	**.776**	**.634**	**.696**	**.650**	**.600**
1000	Qwen2.5-VL	.766	.730	.758	.559	.620	.566	.515
	Qwen2.5-VL + VAD-TTRL(Ours)	**.819**	**.843**	**.794**	**.649**	**.724**	**.661**	**.608**

C.2 Generalization Beyond the Adaptation Set

To examine whether the gains from VAD-TTRL extend beyond the videos used during test-time optimization, we evaluate the adapted Qwen2.5-VL-7B model on two additional subsets of 500 and 1,000 videos sampled from the Vad-Reasoning SFT split. These subsets contain 268/232 and 535/465 Normal/Abnormal videos, respectively, and are not used during VAD-TTRL adaptation. The adapted policy is evaluated directly on these videos without any further parameter updates.

We compare only against the original Qwen2.5-VL-7B backbone. Models trained with supervised Vad-Reasoning data are excluded from this analysis to avoid training-set exposure and ensure a clean evaluation of transfer beyond the VAD-TTRL adaptation set.

Generalization to unseen videos. VAD-TTRL consistently improves the underlying Qwen2.5-VL model on videos that were not used during test-time optimization. On the 500-video subset, accuracy improves from 0.754 to 0.806, F1 from 0.746 to 0.776, and mIoU from 0.536 to 0.634. Similar gains are observed on the 1,000-video subset, where accuracy increases from 0.766 to 0.819, F1 from 0.758 to 0.794, and mIoU from 0.559 to 0.649. Temporal localization also improves consistently across all IoU thresholds. These results show that the benefits of VAD-TTRL are not restricted to the videos used during adaptation and transfer to previously unseen videos without additional parameter updates.

D Training Process of VAD-TTRL

VAD-TTRL performs a single stage of reinforcement-based test-time adaptation directly on unlabeled videos. As summarized in Algorithms 1–3, each video undergoes three stages: Propose, Verify, and Reinforce. The policy first generates multiple stochastic reasoning trajectories and constructs a structured consensus hypothesis. The hypothesis is then tested against counterfactual visual evidence. Only sufficiently supported hypotheses are retained for GRPO (Shao et al., 2024) optimization. No ground-truth anomaly label, temporal boundary, spatial annotation, or reference rationale is used during this process.

For a verified hypothesis, its reliability is

$$q = c\, e, \tag{13}$$

where c denotes consensus confidence and e denotes visual evidence. Proposals that fail either the consensus or visual-verification gate do not contribute to the policy update.

For each trajectory $o_i \in \mathcal{G}$, the completion-level reward combines label agreement, temporal consistency, spatial consistency, rationale consistency, visual verification, and structural constraints:

$$R_i = \lambda_y R_i^y + \lambda_t R_i^t + \lambda_s R_i^s + \lambda_r R_i^r + \lambda_v R_i^v + R_i^{\text{constraint}}. \tag{14}$$

The group-relative advantage is computed as

$$A_i = \frac{R_i - \overline{R}}{\text{Std}(\{R_j : o_j \in \mathcal{G}\}) + \epsilon}, \tag{15}$$

where $\overline{R}$ denotes the mean reward within the sampled training group $\mathcal{G}$.

Algorithm 1 PROPOSE: Structured Consensus Construction

Require: Unlabeled video v, prompt p, policy π_θ, number of rollouts N
Ensure: Consensus hypothesis $\mathcal{H}(v)$, confidence c, rollouts $\mathcal{O}(v)$

Generate $\mathcal{O}(v) = \{o_i\}_{i=1}^{N} \sim \pi_\theta(\cdot \mid p, v)$
Parse each o_i into $(y_i, d_i, I_i, s_i, r_i)$: class, event description, temporal interval, spatial location, and rationale
$\hat{y} \leftarrow \mathrm{mode}(y_1, \ldots, y_N)$
$c \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[y_i = \hat{y}]$
if $c < \tau_{\mathrm{maj}}$ **then**
    return UNKNOWN
end if
$\mathcal{M} \leftarrow \{o_i \in \mathcal{O}(v) : y_i = \hat{y}\}$
if $\hat{y} =$ ABNORMAL **then**
    $I^\star \leftarrow$ temporal medoid of $\{I_i : o_i \in \mathcal{M}\}$ using pairwise IoU
    $s^\star \leftarrow$ modal canonical spatial region in $\mathcal{M}$
else
    $I^\star, s^\star \leftarrow \varnothing$
end if
$r^\star \leftarrow$ semantic medoid of $\{r_i : o_i \in \mathcal{M}\}$
$\mathcal{H}(v) \leftarrow (\hat{y}, I^\star, s^\star, r^\star, c)$
return $\mathcal{H}(v), c, \mathcal{O}(v)$

Algorithm 2 VERIFY: Evidence-Guided Hypothesis Verification

Require: Video v, hypothesis $\mathcal{H}(v) = (\hat{y}, I^\star, s^\star, r^\star, c)$, policy π_θ
Ensure: Visual evidence score e

if $\hat{y} =$ ABNORMAL **then**
    Obtain anomaly probability $p_A(F)$ from the full video $F = v$
    Construct target-drop view D by attenuating frames inside $I^\star$
    Construct target-keep view K by preserving $I^\star$ and attenuating surrounding content
    Construct matched off-target control view R
    Evaluate $p_A(D), p_A(K), p_A(R)$ using π_θ
    $\Delta_{\mathrm{nec}} \leftarrow [p_A(F) - p_A(D)]_+$
    $S_{\mathrm{suf}} \leftarrow [1 - |p_A(F) - p_A(K)|]_+$
    $C_{\mathrm{ctrl}} \leftarrow [1 - |p_A(F) - p_A(R)|]_+$
    $e_A \leftarrow p_A(F) \left(\frac{\Delta_{\mathrm{nec}} + S_{\mathrm{suf}}}{2}\right) C_{\mathrm{ctrl}}$
    $e \leftarrow e_A$
else
    Construct the set of temporally perturbed views $\mathcal{U}$
    for $\tilde{v} \in \mathcal{U}$ **do**
        Obtain $p_N(\tilde{v})$ and $p_A(\tilde{v})$
    end for
    $\mathrm{Stab}_N \leftarrow \frac{1}{|\mathcal{U}|} \sum_{\tilde{v} \in \mathcal{U}} p_N(\tilde{v})$
    $\mathrm{Risk}_A \leftarrow \max_{\tilde{v} \in \mathcal{U}} p_A(\tilde{v})$
    $e_N \leftarrow \mathrm{clip}_{[0,1]} (\mathrm{Stab}_N - \lambda_N \mathrm{Risk}_A)$
    $e \leftarrow e_N$
end if
return e

E REWARD CONSTRUCTION

This section provides the complete formulation of the label-free rewards used to optimize VAD-TTRL. We follow the notation of Sec. 3: for an unlabeled video v, the policy generates $\mathcal{O}(v) = \{o_i\}_{i=1}^{N}$, with $o_i = (y_i, d_i, I_i, s_i, r_i)$, and the PROPOSE stage produces the hypothesis $\mathcal{H}(v) = (\hat{y}, I^\star, s^\star, r^\star, c)$. Only hypotheses that pass visual verification contribute to the policy update. No

Algorithm 3 VAD-TTRL: Propose–Verify–Reinforce

Require: Prompt p, unlabeled test videos $\mathcal{D} = \{v_j\}_{j=1}^{M}$, initial policy π_{θ_0}
Ensure: Adapted policy π_θ

$\pi_\theta \leftarrow \pi_{\theta_0}$
Initialize reference policy π_{ref}
for $v_j \in \mathcal{D}$ **do**
  $\mathcal{H}(v_j), c, \mathcal{O}(v_j) \leftarrow \textsc{Propose}(v_j, p, \pi_\theta)$
  if $\mathcal{H}(v_j) = \textsc{Unknown}$ **then**
    continue
  end if
  $e \leftarrow \textsc{Verify}(v_j, \mathcal{H}(v_j), \pi_\theta)$
  if $e < \tau_{\text{vis}}$ **then**
    continue
  end if
  $q \leftarrow c\, e$
  Uniformly sample $\mathcal{G} \subset \mathcal{O}(v_j)$
  for $o_i \in \mathcal{G}$ **do**
    Compute label-agreement reward R_i^y
    if $\hat{y} = \textsc{Abnormal}$ **then**
      Compute temporal reward R_i^t
      Compute spatial reward R_i^s
    end if
    Compute rationale-consistency reward R_i^r
    Compute visual-verification reward R_i^v
    Compute structural constraint reward $R_i^{\text{constraint}}$
    Combine rewards:

$$R_i = \lambda_y R_i^y + \lambda_t R_i^t + \lambda_s R_i^s + \lambda_r R_i^r + \lambda_v R_i^v + R_i^{\text{constraint}}$$

  end for
  Compute group-relative advantages

$$A_i = \frac{R_i - \text{mean}(R)}{\text{std}(R) + \epsilon}$$

  Update π_θ using the GRPO objective with reference-policy KL regularization
end for
return π_θ

ground-truth anomaly labels, temporal intervals, spatial annotations, or reference explanations are used in reward computation.

E.1 Consensus Pseudo-Targets

The reward targets are constructed from model-generated trajectories supporting the consensus label $\hat{y}$ and never from dataset annotations.

Temporal target. For an Abnormal proposal, let $\mathcal{I}$ denote the temporal intervals predicted by trajectories supporting $\hat{y}$. The consensus interval $I^\star$ is the temporal medoid defined in Sec. 3.1. We additionally measure its support as

$$\kappa_I = \frac{1}{|\mathcal{I}|} \sum_{I_i \in \mathcal{I}} \mathbb{I}[\text{IoU}(I_i, I^\star) \geq \tau_I]\,. \tag{16}$$

An Abnormal proposal is rejected when the predicted intervals provide insufficient support for $I^\star$, preventing mutually inconsistent temporal predictions from forming a reliable target.

Spatial target. Free-form spatial descriptions are mapped to a canonical 3×3 grid,

$$\mathcal{S} = \{\text{top-left}, \text{top-center}, \ldots, \text{bottom-right}\}.$$

The spatial target $s^\star$ is the modal canonical region among trajectories supporting $\hat{y}$. Graded spatial agreement is

$$d_{\mathrm{sp}}(s_i, s^\star) = \begin{cases} 1, & s_i = s^\star, \\ 0.5, & s_i \text{ and } s^\star \text{ are adjacent regions}, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where adjacency is determined using Chebyshev distance on the grid.

Rationale target. Let $\mathcal{R}$ denote the rationales associated with trajectories supporting $\hat{y}$. Semantic similarity is measured using BERTScore F_1 (Zhang et al., 2020), denoted by $B(r_i, r_j)$. The rationale target is the semantic medoid

$$r^\star = \arg\max_{r_i \in \mathcal{R}} \frac{1}{|\mathcal{R}| - 1} \sum_{\substack{r_j \in \mathcal{R} \\ j \neq i}} B(r_i, r_j). \tag{18}$$

Using a medoid ensures that $r^\star$ remains a complete explanation generated by the policy.

E.2 Counterbalanced Visual Verification

Consensus alone may reinforce a correlated model error. VAD-TTRL therefore checks whether the proposed label is supported by visual interventions.

For each video view X, the verifier is evaluated using two counterbalanced prompts:

$$\begin{aligned} &\text{Order 1:} \quad A = \text{Normal}, \quad B = \text{Abnormal}, \\ &\text{Order 2:} \quad A = \text{Abnormal}, \quad B = \text{Normal}. \end{aligned}$$

Let $\phi_A^{(1)}(X), \phi_B^{(1)}(X)$ and $\phi_A^{(2)}(X), \phi_B^{(2)}(X)$ denote the corresponding softmax probabilities over the single-token responses A and B. Following prompt-order counterbalancing (Kang, 2026), we compute

$$p_N(X) = \frac{1}{2}\left(\phi_A^{(1)}(X) + \phi_B^{(2)}(X)\right), \tag{19}$$

$$p_A(X) = \frac{1}{2}\left(\phi_B^{(1)}(X) + \phi_A^{(2)}(X)\right). \tag{20}$$

This reduces bias due to answer ordering or preference-token position.

Visual interventions softly suppress a selected set of frames $\mathcal{T}$. If $\bar{v}$ is the temporal mean frame, the intervened view is

$$\widetilde{v}_t = \begin{cases} \gamma v_t + (1-\gamma)\bar{v}, & t \in \mathcal{T}, \\ v_t, & t \notin \mathcal{T}. \end{cases} \tag{21}$$

Thus, the intervention attenuates the selected evidence while preserving the video tensor shape and avoiding destructive video trimming.

E.2.1 Verification of Abnormal Proposals

For an Abnormal proposal, four meaningful views are evaluated:

- F: the full video;
- D: target-drop, where frames inside $I^\star$ are suppressed;
- K: target-keep, where frames outside $I^\star$ are suppressed;
- R: a matched off-target control covering the same number of frames while minimizing overlap with $I^\star$.

The necessity, sufficiency, and matched-control scores are

$$\Delta_{\mathrm{nec}} = \max\left(0, p_A(F) - p_A(D)\right), \tag{22}$$

$$S_{\mathrm{suf}} = \max\left(0, 1 - |p_A(F) - p_A(K)|\right), \tag{23}$$

$$C_{\mathrm{ctrl}} = \max\left(0, 1 - |p_A(F) - p_A(R)|\right). \tag{24}$$

Necessity is high when suppressing the proposed interval reduces abnormal confidence. Sufficiency is high when retaining the proposed interval primarily preserves the full-video decision. The matched control requires the prediction to remain stable under an equally strong off-target intervention.

The resulting Abnormal evidence score is

$$e_A = p_A(F)\left(\frac{\Delta_{\text{nec}} + S_{\text{suf}}}{2}\right) C_{\text{ctrl}}. \tag{25}$$

E.2.2 Verification of Normal Proposals

A Normal prediction should remain stable under mild perturbations and should not reveal a strong hidden anomaly under any tested view. We construct five altered views: light masking of the beginning, middle, and ending portions of the video, together with sparse and dense temporal smoothing.

For the set of altered views $\mathcal{U}$, we compute

$$\text{Stab}_N = \frac{1}{|\mathcal{U}|}\sum_{\tilde{v}\in\mathcal{U}} p_N(\tilde{v}), \tag{26}$$

$$\text{Risk}_A = \max_{\tilde{v}\in\mathcal{U}} p_A(\tilde{v}). \tag{27}$$

The Normal evidence score is

$$e_N = \text{clip}_{[0,1]}\left(\text{Stab}_N - \lambda_N \,\text{Risk}_A\right). \tag{28}$$

This formulation rewards stability while explicitly penalizing the strongest Abnormal response observed across the intervention set.

The visual-evidence score is

$$e = \begin{cases} e_A, & \hat{y} = \text{Abnormal}, \\ e_N, & \hat{y} = \text{Normal}. \end{cases} \tag{29}$$

A proposal is retained only when

$$e \geq \tau_{\text{vis}}. \tag{30}$$

Otherwise, VAD-TTRL performs no policy update for that video. The verified proposal quality is

$$q = c\,e, \tag{31}$$

combining consensus confidence with visual evidence.

E.3 Per-Completion Reward Components

After verification, a smaller group of training trajectories $\mathcal{G} \subset \mathcal{O}(v)$ is sampled uniformly from the original proposal pool. The trajectories are not selected according to their reward or agreement with the consensus, allowing both consensus-consistent and inconsistent outputs to participate in the relative GRPO ranking.

For each $o_i \in \mathcal{G}$, define

$$a_i = \mathbb{I}[y_i = \hat{y}]. \tag{32}$$

Label agreement reward.

$$R_i^y = a_i q = a_i c e. \tag{33}$$

Thus, agreement with the proposed label receives a strong reward only when the consensus is both confident and visually supported.

Temporal reward. For an Abnormal hypothesis,

$$R_i^t = a_i e\ \text{IoU}(I_i, I^\star). \tag{34}$$

Spatial reward. For an Abnormal hypothesis,

$$R_i^s = a_i e\, d_{\text{sp}}(s_i, s^\star). \tag{35}$$

For a Normal hypothesis, $R_i^s = 0$ because no anomaly location is required.

Causal-semantic reward. Semantic agreement alone can reward a fluent but disconnected explanation. We therefore combine BERTScore (Zhang et al., 2020) agreement with lexical grounding between the event description u_i and explanation z_i:

$$R_i^{\text{cause}} = a_i e\, B(z_i, z^\star) \left(\frac{1 + \text{F1}_{\text{tok}}(u_i, z_i)}{2} \right). \tag{36}$$

Here, F1_{tok} is computed from the multiset overlap of lowercase alphanumeric tokens. This factor discourages explanations that are semantically similar to the consensus but unrelated to the event described by the same completion.

Visual-verification reward.

$$R_i^v = a_i e. \tag{37}$$

This term directly rewards trajectories that agree with the visually verified hypothesis.

Structural constraint reward. The structural constraint reward, also called format reward, checks whether o_i follows the required output schema:

$$R_i^{\text{constraint}} = \mathbb{I}[o_i \text{ follows the required output schema}]. \tag{38}$$

The Normal schema requires the `which`, `what`, and `why` fields, whereas the Abnormal schema additionally requires `when`, `where`, and `how`. This term evaluates structural compliance only and does not use any ground-truth annotation.

E.4 Total Evidence-Weighted Reward

For each trajectory $o_i \in \mathcal{G}$ associated with a verified hypothesis, the total reward is

$$\begin{aligned} R_i = &\lambda_y R_i^y + \lambda_t R_i^t + \lambda_s R_i^s \\ &+ \lambda_r R_i^r + \lambda_v R_i^v + R_i^{\text{constraint}}. \end{aligned} \tag{39}$$

Hypotheses that fail either the consensus or visual-verification gate do not enter the GRPO update and therefore receive no optimization signal.

For the default configuration used in our experiments,

$$\begin{aligned} &\lambda_y = 1.00, \qquad \lambda_t = 0.35, \qquad \lambda_s = 0.15, \\ &\lambda_r = 0.25, \qquad \lambda_v = 0.50. \end{aligned} \tag{40}$$

The structural constraint term is instantiated as

$$R_i^{\text{constraint}} = 0.20\, R_i^{\text{fmt}}, \tag{41}$$

where R_i^{fmt} measures compliance with the required output schema. For Abnormal proposals, temporal pseudo-targets are constructed only when multiple valid intervals form a sufficiently coherent cluster. Short or uninformative explanations are excluded from causal-target construction.

If any proposal or verification gate fails, $g(x) = 0$, all completion rewards are set to zero, and the complete group is excluded from policy optimization. This abstention mechanism is essential: disagreement or insufficient evidence does not become negative supervision for either class.

E.5 Group-Relative Advantages

Let G denote the number of trajectories in the uniformly sampled training group $\mathcal{N}$. Their rewards are normalized within the group as

$$\overline{R} = \frac{1}{G} \sum_{o_i \in \mathcal{G}} R_i, \tag{42}$$

$$A_i = \frac{R_i - \overline{R}}{\text{Std}(\{R_j : o_j \in \mathcal{G}\}) + \epsilon}. \tag{43}$$

Trajectories with positive relative advantage are reinforced, whereas those with negative advantage are suppressed. The policy is optimized using GRPO with KL regularization against the frozen reference policy π_{ref}. Proposals rejected by either the consensus or visual verification gate do not enter the optimization and therefore produce no policy update.

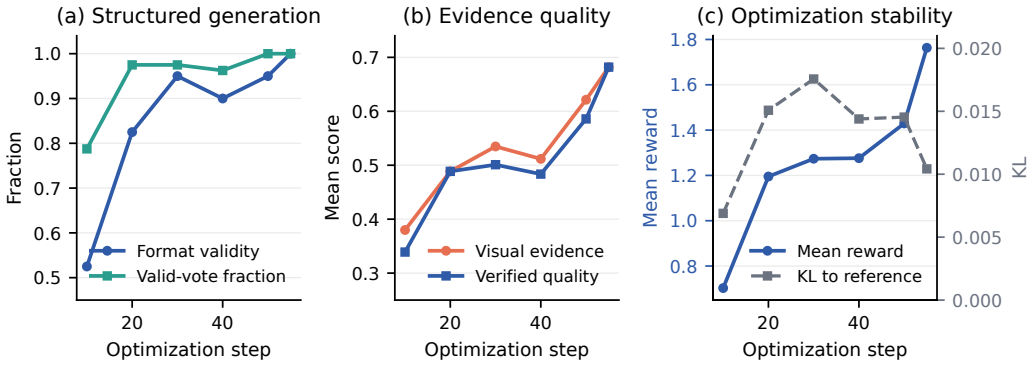

Figure 4: **Training dynamics of VAD-TTRL.** Structured-output validity and the fraction of schema-valid rollouts improve during adaptation. Visual-evidence strength and verified proposal quality also increase, while the KL divergence from the frozen reference policy remains below 0.018. Values are averaged over 10-step logging windows, with the final point at step 55 averaging the last 5 optimization steps.

E.6 Training Dynamics

Figure 4 summarizes the optimization behavior of VAD-TTRL. Structured-output validity, visual-evidence strength, and verified proposal quality generally improve during adaptation, while the policy remains close to the reference model.

F Additional Ablations

We further examine the sensitivity of VAD-TTRL to rollout configuration, consensus and verification thresholds, reward weighting, and individual reward components. Unless otherwise stated, all remaining proposal, verification, and optimization settings are fixed to the default configuration.

F.1 Effect of Rollout Configuration

We study the effect of the number of stochastic trajectories used during test-time reinforcement learning. Here, N denotes the total number of trajectories generated per video, while G denotes the number uniformly sampled into the training group $\mathcal{G}$ for the GRPO update. All remaining training and verification settings are fixed.

Table 7: Effect of rollout configuration on Vad-Reasoning. N denotes the number of generated trajectories and G the number sampled for the GRPO update. The default VAD-TTRL configuration is highlighted in gray.

Rollouts		**Detection & Localization**								**Reasoning**		
N	G	**Acc.**	**Prec.**	**Rec.**	**F1**	**mIoU**	**R@.3**	**R@.5**	**R@.7**	**B-2**	**METEOR**	**R-2F**
4	2	.804	.904	.647	.754	.608	.645	.599	.576	.107	.254	.152
8	4	**.829**	.869	.745	**.802**	**.657**	**.729**	**.656**	**.615**	**.131**	**.280**	**.156**
8	8	.585	.923	.118	.209	.553	.556	.551	.551	.093	.232	.135
16	4	.749	.927	.500	.650	.623	.656	.631	.601	.121	.268	.146
16	8	.554	**.933**	.049	.093	.541	.544	.542	.540	.094	.230	.133

Rollout sensitivity. Table 7 shows that rollout configuration substantially affects adaptation. The default $(N, G) = (8, 4)$ setting provides the strongest overall balance, reaching 0.829 accuracy, 0.802 F1, and 0.657 mIoU. Using fewer trajectories, $(4, 2)$, remains competitive but yields lower recall and localization performance, while increasing the proposal pool to $N = 16$ does not provide further gains.

Increasing the GRPO training group to $G = 8$ substantially reduces anomaly recall, from 0.745 at $(8, 4)$ to 0.118 at $(8, 8)$ and to 0.049 at $(16, 8)$. In contrast, using $G = 4$ yields considerably stronger recall for both $N = 8$ and $N = 16$. These results indicate that reinforcing a smaller uniformly sampled subset of the proposal pool is more effective than using a larger training group, potentially limiting reinforcement of correlated trajectories. Reasoning metrics show a similar trend, with $(8, 4)$ achieving the strongest overall performance.

F.2 Consensus and Verification Thresholds

VAD-TTRL updates the policy only when a proposal passes both the majority-consensus threshold τ_{maj} and the visual-evidence threshold τ_{vis}. We vary each threshold independently while holding the other at its default value.

Table 8: Sensitivity to the consensus and visual-evidence thresholds on Vad-Reasoning. When varying one threshold, all remaining settings are kept fixed. The default configuration is highlighted in gray.

Parameter	Value	Acc.	Prec.	Rec.	F1	mIoU	R@0.3	R@0.5	R@0.7
τ_{maj}	0.50	0.822	0.846	**0.755**	0.798	0.619	0.667	0.615	0.583
	0.75	**0.829**	**0.869**	0.745	**0.802**	**0.657**	**0.729**	**0.656**	**0.615**
	1.00	0.768	0.782	0.695	0.735	0.596	0.647	0.597	0.560
τ_{vis}	0.00	0.631	**0.977**	0.211	0.347	0.549	0.563	0.556	0.549
	0.20	**0.829**	0.869	0.745	**0.802**	**0.657**	**0.729**	**0.656**	**0.615**
	0.40	0.795	0.705	0.711	0.701	0.603	0.667	0.590	0.528
	1.00	0.736	0.680	**0.814**	0.741	0.512	0.567	0.501	0.465

Effect of consensus confidence . A moderate majority threshold provides the strongest overall performance. Increasing τ_{maj} from 0.50 to 0.75 improves mIoU from 0.619 to 0.657 and R@0.5 from 0.615 to 0.656, suggesting that filtering weak consensus hypotheses improves the quality of the resulting adaptation signal. Requiring unanimous agreement is more restrictive and reduces temporal localization performance.

Effect of visual verification. Visual verification has a larger effect. Without evidence filtering ($\tau_{\mathrm{vis}} = 0$), the model becomes highly conservative, achieving high precision (0.977) but only 0.211 anomaly recall. The default $\tau_{\mathrm{vis}} = 0.20$ substantially improves the precision–recall balance and achieves the highest mIoU and temporal recall. Stronger filtering eventually reduces localization quality, indicating that excessive selectivity can remove useful adaptation examples.

F.3 Reward-Weight Sensitivity

After a hypothesis passes verification, VAD-TTRL combines label agreement, temporal consistency, spatial consistency, rationale consistency, and visual verification. We vary the weights $\lambda_y, \lambda_t, \lambda_s$, and λ_r while keeping λ_v, the structural constraint, and all proposal, verification, and optimization settings fixed.

Table 9: Reward-weight configurations used in the sensitivity analysis. "H" denotes *heavy* and "Red." denotes *reduced*. The symbols λ_y, λ_t, λ_s, and λ_r correspond to the label, temporal, spatial, and rationale rewards, respectively. The default configuration is highlighted in gray.

Weight	Default	Eq.-0.25	Eq.-0.50	Red.-Lab.	Lab.-H	Temp.-H	Spa.-H	Rat.-H	Temp.+Rat.-H
λ_y	1.00	0.25	0.50	0.50	1.00	0.50	0.50	0.50	0.25
λ_t	0.35	0.25	0.50	0.25	0.25	0.50	0.25	0.25	0.50
λ_s	0.15	0.25	0.50	0.25	0.25	0.25	0.50	0.25	0.25
λ_r	0.25	0.25	0.50	0.25	0.25	0.25	0.25	0.50	0.50

Overall reward balance. The default weighting yields the strongest joint detection and localization performance, reaching 0.829 accuracy, 0.802 F1, and 0.657 mIoU. Setting all four varied reward weights to 0.50 does not improve overall performance: Equal-0.50 raises anomaly recall to 0.877,

Table 10: Sensitivity to reward weighting on Vad-Reasoning. The default weighting provides the strongest overall detection and temporal localization performance, while alternative configurations expose different precision-recall trade-offs.

Configuration	Detection & Localization								Reasoning		
	Acc.	Prec.	Rec.	F1	mIoU	R@0.3	R@0.5	R@0.7	B-2	METEOR	R-2F
Default	**.829**	**.869**	.745	**.802**	**.657**	**.729**	**.656**	**.615**	.131	.280	**.156**
Equal-0.25	.797	.786	.775	.780	.577	.617	.565	.542	.123	.277	.150
Equal-0.50	.720	.646	**.877**	.744	.479	.538	.467	.435	.129	.283	.145
Reduced-label	.800	.826	.721	.770	.599	.642	.588	.565	.117	.267	.152
Label-heavy	.786	.755	.799	.776	.549	.585	.538	.508	.124	.277	.151
Temporal-heavy	.774	.738	.799	.767	.559	.610	.551	.522	.124	.277	.149
Spatial-heavy	.774	.740	.794	.766	.562	.617	.551	.524	**.132**	**.286**	.150
Rationale-heavy	.786	.835	.672	.745	.597	.638	.601	.560	.122	.272	.150
Temporal+rationale-heavy	.738	.668	.868	.755	.495	.556	.487	.451	.126	.280	.147

but reduces precision to 0.646 and mIoU to 0.479. Similarly, emphasizing individual components changes the precision-recall trade-off without matching the localization quality of the default configuration.

Reasoning sensitivity. Text-generation metrics remain comparatively stable across reward configurations. Spatial-heavy gives slightly higher BLEU-2 and METEOR, whereas the default configuration obtains the highest ROUGE-2 F1. This suggests that the selected weighting primarily improves the joint detection-localization objective while preserving reasoning quality rather than maximizing any single textual metric.

F.4 Reward Component Ablation

We ablate each reward component individually while keeping all remaining training, proposal, and verification settings fixed. Specifically, λ_y, λ_t, λ_s, λ_r, and λ_v denote the label, temporal, spatial, rationale-consistency, and visual-verification rewards, respectively. Setting a coefficient to zero removes only the corresponding reward term.

Importantly, $\lambda_v = 0$ removes the *explicit verification reward* but does not disable visual verification: visual evidence is still used to authorize pseudo-labels and to scale the remaining reward components.

Table 11: Ablation of individual reward components on Vad-Reasoning. Each row removes one reward term while keeping all remaining proposal, verification, and optimization settings fixed.

Configuration	Detection & Localization								Reasoning		
	Acc.	Prec.	Rec.	F1	mIoU	R@.3	R@.5	R@.7	B-2	METEOR	R-2F
Full VAD-TTRL	**.829**	.869	.745	**.802**	.657	**.729**	**.656**	**.615**	.131	**.280**	**.156**
$\lambda_y = 0$ (w/o label)	.704	.930	.392	.552	.596	.624	.595	.579	.104	.247	.143
$\lambda_t = 0$ (w/o temporal)	.790	.912	.608	.729	.625	.667	.622	.590	.128	.279	.150
$\lambda_s = 0$ (w/o spatial)	.749	.943	.490	.645	.625	.656	.624	.606	.110	.259	.146
$\lambda_r = 0$ (w/o rationale)	.797	.778	**.789**	.783	**.688**	.615	.581	.581	.112	.256	.145
$\lambda_v = 0$ (w/o explicit verification reward)	.718	**.944**	.417	.578	.616	.649	.617	.595	.112	.276	.153

G Qualitative Results

We present qualitative comparisons among Qwen2.5-VL-7B, Vad-R1, and their TTRL-adapted counterparts. As shown in Fig. 6, the examples from the Vad-Reasoning benchmark (Huang et al., 2025) illustrate how test-time reinforcement learning improves anomaly recognition and reasoning reliability. For *Normal_Videos_360*, both Qwen2.5-VL-7B (Bai et al., 2025) and Vad-R1 (Huang et al., 2025) produce false-positive anomaly predictions, whereas VAD-TTRL and Vad-R1+TTRL correctly identify the video as normal by grounding their predictions in routine retail activities. For *Robbery043*, VAD-TTRL corrects the missed anomaly of Qwen2.5-VL-7B and localizes the anomalous interaction with substantial temporal overlap, while Vad-R1+TTRL preserves Vad-R1's correct anomaly prediction with similarly accurate temporal localization.



"Question": "What unusual action is the man in the white jacket performing in the video?
A) Reading a newspaper B) Dancing C) Talking on the phone D) Throwing items"

HolmesVAU

"The unusual action being performed by the man in the white jacket is shooting a gun, which is an unusual and potentially dangerous behavior in a public setting. <answer>D".

Vad-R1

"<think>Let's analyze the video step by step:\n\n1. The video shows an outdoor scene with a paved area and some greenery.\n2. In the background, there are other people walking or standing, some of whom are dressed in uniforms.\n3. A man is seen standing in the foreground, holding what appears to be a paper or a similar object.\n4. He is not performing any aggressive actions, and there is no indication of fighting or conflict. \n\nGiven these observations, the most accurate description of the events in the video is:\n\nA) Reading a newspaper</think>\n<answer>A</answer>".

HAWK

"The man in the white jacket is walking and reading a newspaper. <answer>A".

Cog-VADU

"OBSERVATION: The video shows a man in a white jacket standing outside a building, pointing towards the entrance. There are several people entering and exiting the building. The man's actions seem to be normal. There is also a staircase leading down to an underground station.\nANOMALY_DETECTED: No, \nREASONING: The man's action of pointing does not fall under any of the listed anomaly types as it is a normal behavior. There are no visual artifacts, unnatural movements observed.\nANSWER: A".

VAD-TTRL (OURS)

"<answer>B</answer>\n<evidence>\n<step1>Scene: A man in a white jacket is standing outside a building, possibly a train station, with other people in the background. The man is moving his arms in a rhythmic manner, suggesting he is dancing.</step1>\n<step2>Visible action: The man in the white jacket is moving his arms in a rhythmic manner, which is unusual for the setting and context.
</step2>\n<step3>Temporal: The man's arm movements are consistent across the frames, indicating a continuous action.
</step3>\n<step4>Comparison: The man's arm movements are unusual for the setting and context, making option B the most supported. Options A, C, and D do not match the visible evidence.</step4>\n<step5>Conclusion: The man in the white jacket is performing an unusual action of dancing, as indicated by his rhythmic arm movements.</step5>".

Figure 5: **Qualitative comparison on VANE (Bharadwaj et al., 2025).** While existing methods incorrectly interpret the observed action as reading a newspaper or shooting a gun, VAD-TTRL correctly identifies the man as *dancing* and grounds its prediction in the rhythmic arm movements observed across consecutive frames.

Qualitative results on VANE (Bharadwaj et al., 2025) Benchmark . On VANE, VAD-TTRL demonstrates stronger visual grounding and action-level reasoning compared with existing methods. As shown in Fig. 5, several sota baselines like (Huang et al., 2025), (Wani et al., 2026) ,(Tang et al., 2024) incorrectly interpret the man's behavior as reading a newspaper, whereas VAD-TTRL correctly identifies the action as *dancing*. Its reasoning is grounded in the observed rhythmic arm movements across consecutive frames, reducing unsupported or hallucinated interpretations.

H Limitations and Future Work

This work provides an initial exploration of test-time reinforcement learning for video anomaly detection, localization, understanding, and reasoning. Our results show that a single test-time adaptation framework can support these tasks using self-generated supervision, where pseudo-labels are not accepted directly but are first filtered through consensus and visual-evidence verification before model updates are performed. Moreover, the proposed approach generalizes across different MLLM backbones and also demonstrates the potential for cross-dataset adaptation, suggesting that test-time reinforcement learning can provide a flexible mechanism for adapting video anomaly models beyond a single architecture or benchmark.

Despite these encouraging results, several limitations remain. First, the gains in anomaly reasoning are still limited compared with the improvements observed in detection and localization, indicating that reliable semantic and causal reasoning remains challenging. Second, the effectiveness of the



Ground truth: Normal

Qwen2.5-VL-7B

> The video is classified as <which>Abnormal</which>. In this video, <what>a pedestrian suddenly enters the store from the left side of the frame and walks directly towards the counter, disrupting the normal flow of activity.</what> This event occurs during the time range <when>[0.121, 0.826]</when> and happens approximately in the <where>left side of the frame</where>. This is considered abnormal because <why>pedestrians are expected to enter the store through the main entrance and not directly towards the counter.</why> As a result, <how>this behavior could interfere with the work of the employees and disrupt the customer experience, potentially causing harm or violating norms.</how>

VAD-TTRL(Ours)

> The video is classified as <which>Normal</which>. In this video, <what>employees are working behind the counter, scanning items, and interacting with customers.</what>. This is considered normal because <why>the actions and environment are consistent with typical retail operations.</why>

Vad-R1

> The video is classified as <which>Abnormal</which>. In this video, <what>a person is observed holding a weapon and pointing it at another individual near the checkout counter in a retail environment</what>, occurring during the time range <when>[0.0, 1.0]</when>. The event takes place approximately in the <where>lower-left area of the frame</where>. This is considered abnormal because <why>the use of a weapon for intimidation deviates from the routine and peaceful interactions typically expected in such settings</why>. As a result, <how>such behavior could lead to physical harm, psychological distress, and disruption of normal operations, violating social norms and safety protocols that prioritize non-violence</how>.

Vad-R1+TTRL(Ours)

> The video is classified as <which>Normal</which>. In this video, <what>employees are performing routine tasks like interacting with customers and handling cash, while customers browse products and make purchases in a well-organized retail environment</what>. This is considered normal because <why>the actions and interactions align with the expected behaviors in a retail setting, where both staff and patrons conduct their activities in an orderly and secure manner without any disruptions</why>.



Ground truth: Abnormal (Robbery), interval [0.000, 0.954]

Qwen2.5-VL-7B

> > The video is classified as <which>Normal</which>. In this video, <what>people are engaged in typical banking activities, such as interacting with the teller and handling documents.</what>. This is considered normal because <why>the individuals are performing standard banking procedures, which are expected in a bank setting.</why>>

VAD-TTRL(Ours) temporal IoU=0.7914

> The video is classified as <which>Abnormal</which>. In this video, <what>a customer suddenly grabs the teller's arm and pulls them off their chair</what>, occurring during the time range <when>[0.101, 0.856]</when>. The event takes place approximately in the <where>middle of the frame, near the teller counter</where>. This is considered abnormal because <why>customers are expected to interact with the teller in a polite and professional manner</why>. As a result, <how>this behavior could lead to a disruption of the banking process and potentially cause harm to the teller or other customers</how>.

Vad-R1 temporal IoU=0.954

> The video is classified as <which>Abnormal</which>. In this video, <what>a person in the lower-left area of the frame is observed holding a weapon while standing near counters and interacting with equipment</what>, occurring during the time range <when>[0.0, 1.0]</when>. The event takes place approximately in the <where>lower-left area of the frame</where>. This is considered abnormal because <why>the presence of a weapon introduces a threatening element into an otherwise peaceful environment, deviating from routine activities</why>. As a result, <how>such behavior could lead to physical harm, property damage, and violate safety norms in public spaces, creating a dangerous situation for everyone involved</how>.

Vad-R1+TTRL(Ours) temporal IoU=0.955

> The video is classified as <which>Abnormal</which>. In this video, <what>two individuals are observed in a workspace where one person points a weapon at another individual</what>, occurring during the time range <when>[0.0, 0.999]</when>. The event takes place approximately in the <where>lower-left area of the frame</where>. This is considered abnormal because <why>the presence of a weapon in a public setting deviates from expected behaviors and creates a threatening atmosphere</why>. As a result, <how>such an act could lead to physical harm, psychological trauma, or financial loss, while also violating norms of safety and trust in shared environments</how>

Figure 6: **Qualitative results of VAD-TTRL.** Additional examples demonstrating the anomaly detection, localization, and reasoning capability of VAD-TTRL across diverse scenarios and different base models on (Huang et al., 2025) benchmark we consider normal as well as abnormal examples.

framework depends on several hyperparameters, including the consensus threshold, visual-evidence threshold, reward weighting, and adaptation settings, which may require careful tuning across different models and datasets. In addition, the current verification mechanism can still be affected by incorrect but highly consistent model predictions, particularly when the underlying model lacks sufficient visual or temporal understanding.

In future work, we plan to improve the reasoning capability of the model by developing stronger evidence-verification and reward mechanisms that better capture temporal, spatial, and causal relationships in anomalous events. Another promising direction is to reduce sensitivity to manually selected hyperparameters through adaptive or self-calibrating test-time objectives. Extending the framework to longer videos, more diverse anomaly categories, and stronger multimodal backbones would also help evaluate the generality and scalability of test-time reinforcement learning for video anomaly understanding.