RELIABILITY-AWARE TEST-TIME REINFORCEMENT LEARNING FOR ANOMALOUS VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomalous video understanding aims not only to identify abnormal events in videos, but also to interpret their objects, actions, interactions, and high-level semantics. Recent Video Large Language Models (Video-LLMs) have demonstrated promising zero-shot capabilities for this task, yet frozen models remain limited by diverse anomaly patterns and evolving deployment environments. Test-time reinforcement learning offers a promising solution by enabling models to adapt from self-generated feedback without additional human annotations. However, directly applying it to anomalous video understanding faces three reliability bottlenecks: (1) pseudo-labels may be unstable across semantically equivalent queries, introducing unreliable self-supervision; (2) coarse consensus rewards ignore generation uncertainty, weakening reward informativeness; and (3) unanimous rollout groups receive identical rewards, causing group-relative advantages to collapse and eliminating effective policy-gradient signals. To address these challenges, we propose a reliability-aware online test-time reinforcement learning framework that improves reliability at the sample, generation, and optimization levels. Specifically, dual-query consistency filtering retains samples with cross-query consistent pseudo-labels, an entropy-aware consensus reward combines answer agreement with generation uncertainty, and virtual-anchor advantage estimation preserves informative relative advantages for unanimous rollout groups. Experiments on VAU-Bench demonstrate consistent improvements over frozen and supervised baselines. Notably, on the ECVA subset with thinking, our method improves accuracy from 75.81% to 90.18% over the frozen backbone.

1 INTRODUCTION

Anomalous video understanding (AVU) Ye et al. (2025) aims not only to identify abnormal events, but also to understand their objects, actions, interactions, and high-level semantics. This capability is crucial for applications such as public surveillance, industrial monitoring, and emergency response, where anomalies are often rare, diverse, and context-dependent.

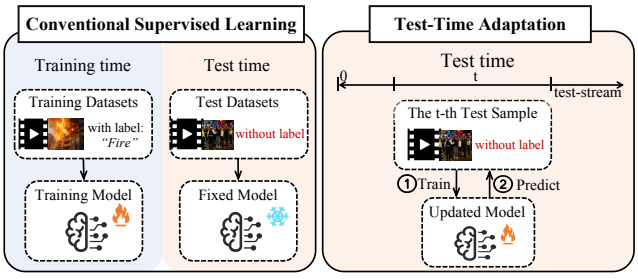

Figure 1: Comparison between conventional supervised learning and test-time adaptation for anomalous video understanding.

Recent Video Large Language Models (Video-LLMs) show promising zero-shot capabilities (Lin et al., 2024; Zhang et al., 2025b), but deploying a frozen model remains challenging under changing environments and diverse anomaly patterns. Models trained offline may suffer from distribution shifts, while repeated supervised retraining is impractical because anomalous events are rare and fine-grained video annotation is costly (Sultani et al., 2018; Zhang et al., 2025a). This motivates models that can continuously adapt to evolving test distributions without human supervision.

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.

Test-time adaptation (TTA) Sun et al. (2020); Wang et al. (2021; 2022); Niu et al. (2022) provides a natural solution by updating model parameters using incoming unlabeled samples and carrying the adapted model forward through the test stream. As illustrated in Figure 1, TTA exploits deployment-time data without additional labels. Although widely studied for discriminative vision models and recently extended to vision-language models, its use for anomalous video understanding with generative Video-LLMs remains largely unexplored.

A promising approach is test-time reinforcement learning, which generates multiple responses for an unlabeled sample, derives pseudo-supervision from their agreement, and optimizes consensus-based rewards (Zuo et al., 2025; Singh et al., 2026; Jahagirdar et al., 2026; Wang et al., 2026). However, this strategy is fragile for AVU because abnormal events are often temporally sparse, semantically ambiguous, and dependent on subtle causal or temporal cues. This exposes three reliability bottlenecks. First, at the **sample level**, pseudo-labels may be unstable across semantically equivalent queries, causing unreliable samples to reinforce confirmation bias. Second, at the **generation level**, consensus rewards usually consider only discrete answer agreement and ignore predictive uncertainty, making generations with different confidence indistinguishable. Third, at the **optimization level**, unanimous rollout groups receive identical rewards, causing group-relative advantages to collapse and eliminating effective policy-gradient signals (Shao et al., 2024). These problems are especially harmful in online adaptation because erroneous updates can propagate to subsequent samples in next steps.

To address them, we propose a **reliability-aware online test-time reinforcement learning framework** for AVU. Our key idea is to improve reliability at the sample, generation, and optimization levels. Specifically, **dual-query consistency filtering** retains samples with consistent pseudo-labels across semantically equivalent queries; an **entropy-aware consensus reward** combines answer agreement with token-level uncertainty to distinguish generations with different confidence; and **virtual-anchor advantage estimation** introduces a virtual negative reference for unanimous groups to preserve informative relative advantages. The retained samples are optimized with Group Relative Policy Optimization (GRPO), enabling continuous adaptation over the unlabeled test stream.

We evaluate our method on three VAU-Bench subsets under both *w/o Think* and *w/ Think* settings. Using Qwen2.5-VL-3B as the backbone, our method consistently outperforms the frozen model across all six QA settings. Moreover, although adaptation uses only QA-style supervision, the adapted models also improve anomaly classification and generally benefit fine-grained anomaly recognition without classification-specific labels or updates, demonstrating transfer beyond multiple-choice QA.

Our main contributions are summarized as follows:

- We formulate anomalous video understanding as an online test-time reinforcement learning problem and develop a framework that continuously adapts Video-LLMs from unlabeled test streams without ground-truth supervision.
- We identify three reliability bottlenecks, sample-level pseudo-label instability, generation-level uncertainty neglect, and optimization-level advantage collapse, and address them with dual-query consistency filtering, entropy-aware consensus rewards, and virtual-anchor advantage estimation.
- Extensive experiments on VAU-Bench demonstrate consistent gains across QA settings and cross-task transfer to anomaly classification and fine-grained anomaly recognition.

2 Related Work

Video Large Language Models. Video-LLMs extend large language models with visual encoders for video-conditioned dialogue and reasoning. Video-LLaVA aligns image and video representations in a shared language space, while InternVideo2 scales multimodal pre-training for general video recognition and video-language tasks (Lin et al., 2024; Wang et al., 2024). TimeChat further introduces timestamp-aware temporal modeling for long-video understanding (Ren et al., 2024). Later studies improve video instruction tuning with large-scale synthetic data and enhance reasoning through structured reasoning distillation (Zhang et al., 2025b; Shi et al., 2025). However, most Video-LLMs are trained offline and frozen during inference, limiting their adaptation to rare anomaly patterns and deployment shifts.

Anomalous Video Understanding. Traditional video anomaly detection mainly identifies whether and when abnormal events occur through weakly supervised multiple-instance learning or temporal feature modeling (Sultani et al., 2018; Tian et al., 2021). Recent methods introduce vision–language models for stronger semantic generalization. VadCLIP adapts CLIP for weakly supervised detection, while open-vocabulary and language-guided open-world methods generalize to unseen anomaly categories and changing anomaly definitions (Wu et al., 2024b;a; Liu et al., 2026). LAVIDA further uses pseudo-anomaly exposure and MLLM-based semantic modeling for zero-shot detection (Dai et al., 2026). Training-free and agentic approaches include LAVAD, AnomalyRuler, and PANDA, which respectively leverage LLM-based temporal aggregation, scene-specific rules, and agentic planning with reflection and memory (Zanella et al., 2024; Yang et al., 2024; 2025). Beyond binary detection, Holmes-VAU and VERA support long-term or explainable anomaly understanding, while VAU-R1 and Vad-R1 enhance anomaly reasoning through reinforcement fine-tuning and structured Chain-of-Thought supervision (Zhang et al., 2025a; Ye et al., 2025; Zhu et al., 2025; Huang et al., 2025). However, these methods mainly rely on offline training or inference-time prompting and cannot continuously update model parameters from incoming unlabeled videos to adapt to evolving scenes, environments, or anomaly types.

Test-time Reinforcement Learning. Test-time adaptation addresses deployment shifts by updating models with unlabeled test inputs. Early methods rely on self-supervised objectives or entropy minimization, while later studies improve continual adaptation, sample reliability, and resistance to forgetting (Sun et al., 2020; Wang et al., 2021; 2022; Niu et al., 2022). In video understanding, ViTTA performs feature-level adaptation for action recognition, while COPRA generates instance-conditioned parameters for video anomaly detection (Lin et al., 2023; Jacob et al., 2026). More recently, test-time reinforcement learning enables large models to learn from self-generated feedback. TTRL derives pseudo-rewards from response agreement, TTRV extends this paradigm to vision–language models, and TTA-Vid further exploits multiple visual views for video reasoning (Zuo et al., 2025; Singh et al., 2026; Jahagirdar et al., 2026). SCOPE improves consensus-based optimization through confidence-weighted pseudo-label estimation (Wang et al., 2026). Despite these advances, reliable self-supervision for anomalous video understanding remains underexplored: consensus can be unstable across semantically equivalent queries, aggregate consensus may obscure generation-level uncertainty, and unanimous rollout groups can produce degenerate group-relative advantages. Our work addresses these issues through reliability-aware test-time reinforcement learning.

3 METHOD

3.1 OVERVIEW

We consider test-time reinforcement learning for anomalous video understanding, instantiated as multiple-choice QA following existing benchmarks. At inference, given a sequence of test samples $\{x_t\}_{t=1}^{T}$, let $x_t = (v_t, q_t, \mathcal{O}_t)$ denote the t-th sample, where v_t, q_t, and $\mathcal{O}_t$ represent the input video, original question, and multiple-choice options. The ground-truth answer is unavailable during test-time adaptation. Given the current Video-LLM π_{θ_t} and test sample x_t, the goal is to adapt the model on this sample and predict the final option $\widehat{a}_t$. The adapted parameters are carried forward, enabling the model to progressively improve its anomalous video understanding over the test stream. The framework can also be extended to other anomalous video understanding tasks.

To obtain reliable and informative self-supervision while preserving effective optimization on unlabeled test samples, we design the four-stage framework illustrated in Figure 2. First, to filter unreliable pseudo-labels, the original question is rewritten into a semantically equivalent form while retaining the original multiple-choice options. The current Video-LLM then samples multiple responses to both question forms, retaining only samples with consistent cross-query consensus. Second, for each retained sample, we construct an entropy-aware consensus reward that jointly considers answer agreement and generation uncertainty. Third, a virtual negative anchor introduces reward variation into unanimous rollout groups, yielding non-zero group-relative advantages and effective optimization signals. Finally, the model is updated using GRPO, performs deterministic inference on the current sample, and is carried forward.

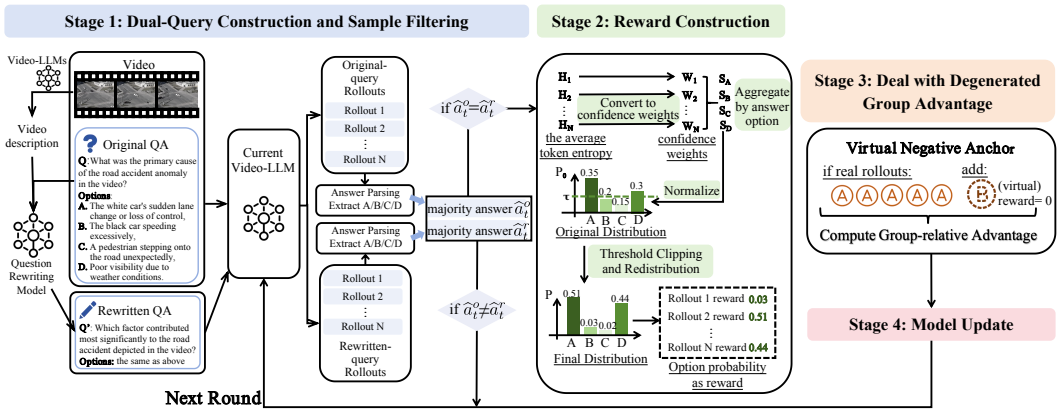

Figure 2: Overview of the proposed online test-time reinforcement learning framework. Original and reformulated questions are used to identify reliable test samples. Retained samples are optimized using an entropy-aware consensus reward, and virtual-anchor-enhanced advantage estimation.

3.2 Dual-Query Construction and Sample Filtering

3.2.1 Semantics-Preserving Query Reformulation

To assess whether self-generated consensus is robust to linguistic variation, we construct a semantically equivalent query view for each test question. A video-LLM first generates a description of the input video. The video description, original question, and answer options are then provided to a question-rewriting model, which is instructed to rewrite only the question while preserving its original meaning. The answer options and their order remain unchanged. The original and reformulated inputs are denoted as $x_t^o = (v_t, q_t, \mathcal{O}_t)$ and $x_t^r = (v_t, \widetilde{q}_t, \mathcal{O}_t)$, where $\widetilde{q}_t$ is the reformulated question.

3.2.2 Sampling and Consensus-Based Sample Filtering

To select more reliable samples for test-time optimization, we retain only those with consistent consensus across both query views. The current policy π_{θ_t} independently generates N stochastic responses for each view: $\mathcal{Y}_t^o = \{y_{t,1}^o, \ldots, y_{t,N}^o\}$ and $\mathcal{Y}_t^r = \{y_{t,1}^r, \ldots, y_{t,N}^r\}$. For each view $u \in \{o, r\}$, we extract an answer label from each response using

$$a_{t,n}^u = \mathrm{Extract}\big(y_{t,n}^u\big), \qquad a_{t,n}^u \in \mathcal{A} \cup \{\varnothing\}, \tag{1}$$

where $\mathcal{A} = \{A, B, C, D\}$ is the candidate answer-option set and $\varnothing$ denotes a response from which no valid option can be parsed. For each view $u \in \{o, r\}$, the support count for option a is then computed as

$$c_t^u(a) = \sum_{n=1}^{N} \mathbf{1}\left[a_{t,n}^u = a\right], \qquad a \in \mathcal{A}. \tag{2}$$

The consensus answer for each view $u \in \{o, r\}$ is

$$\widehat{a}_t^u = \arg\max_a c_t^u(a), \qquad a \in \mathcal{A}. \tag{3}$$

A sample is retained only when both views have a strict majority and their consensus answers are identical:

$$s_t = \mathbf{1}\left[\max_a c_t^o(a) > \frac{N}{2}\right] \mathbf{1}\left[\max_a c_t^r(a) > \frac{N}{2}\right] \cdot \mathbf{1}\left[\widehat{a}_t^o = \widehat{a}_t^r\right]. \tag{4}$$

For example, when $N = 4$, suppose the extracted answers for the original and reformulated queries are (A, A, A, B) and (A, A, B, C), respectively. Although both views yield the same consensus answer, $\widehat{a}_t^o = \widehat{a}_t^r = A$, their support counts are $c_t^o(A) = 3$ and $c_t^r(A) = 2$. Since the reformulated query does not satisfy the strict-majority condition $c_t^r(A) > N/2$, the sample is discarded with $s_t = 0$.

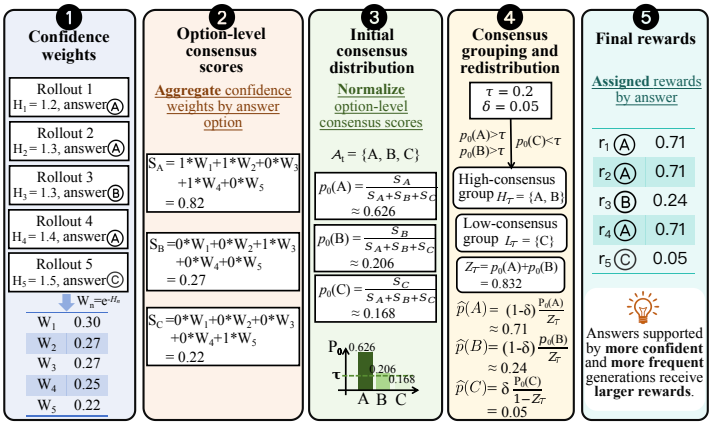

Figure 3: An example of the entropy-aware consensus reward computation with five rollouts retained from the first stage.

Samples with $s_t = 0$ do not contribute to parameter update. For samples with $s_t = 1$, only the original-question rollouts are used for reward construction, while the reformulated-question rollouts are used solely for reliability assessment.

3.3 Entropy-Aware Consensus Reward

To provide more informative supervision than a binary majority reward, we construct an entropy-aware consensus reward for each retained test sample. We use only the rollouts generated from the original query for reward estimation and policy optimization. Figure 3 provides an example of the complete entropy-aware consensus reward computation process.

Let L_n denote the number of generated tokens in the n-th original-query rollout, and let $P_{n,\ell}(v)$ be the probability of vocabulary token v at generation position ℓ. The mean token entropy of the rollout is

$$H_n = -\frac{1}{L_n} \sum_{\ell=1}^{L_n} \sum_{v \in \mathcal{V}} P_{n,\ell}(v) \log P_{n,\ell}(v), \tag{5}$$

and its corresponding confidence weight is

$$w_n = \exp(-H_n), \tag{6}$$

A rollout with lower mean token entropy therefore receives a larger confidence weight.

To obtain an option-level consensus score, we first group the rollouts according to their extracted answer labels and then sum the confidence weights within each group. Specifically, for each answer label $a \in \mathcal{A}$,

$$S_t(a) = \sum_{n=1}^{N} \mathbf{1}\left[a_{t,n}^o = a\right] w_n \tag{7}$$

Thus, $S_t(a)$ measures the confidence-weighted support received by option a: it increases when more rollouts select a or when those rollouts are generated with lower uncertainty.

We denote $\mathcal{A}_t$ as the set of answer labels that actually appear among the N original-query rollouts. The option-level scores are normalized into an initial consensus distribution:

$$p_0(a \mid x_t^o) = \frac{S_t(a)}{\sum_{b \in \mathcal{A}_t} S_t(b)}, \qquad a \in \mathcal{A}_t. \tag{8}$$

To prevent weakly supported answers from introducing noisy optimization signals, we apply a consensus threshold to divide the observed answer labels into high- and low-consensus groups:

$$\mathcal{H}_{\tau,t} = \{a \in \mathcal{A}_t : p_0(a \mid x_t^o) > \tau\}, \qquad \mathcal{L}_{\tau,t} = \mathcal{A}_t \setminus \mathcal{H}_{\tau,t}, \tag{9}$$

where τ is the consensus threshold. The total probability assigned to the high-consensus group is

$$Z_{\tau,t} = \sum_{a \in \mathcal{H}_{\tau,t}} p_0(a \mid x_t^o). \tag{10}$$

Rather than assigning zero reward to low-consensus answers, we retain a small probability mass to avoid overly hard supervision and preserve plausible alternatives:

$$\widehat{p}(a \mid x_t^o) = \begin{cases} (1-\delta)\dfrac{p_0(a \mid x_t^o)}{Z_{\tau,t}}, & a \in \mathcal{H}_{\tau,t}, \\ \delta\dfrac{p_0(a \mid x_t^o)}{1 - Z_{\tau,t}}, & a \in \mathcal{L}_{\tau,t}. \end{cases} \tag{11}$$

Here, τ determines which answer labels receive high-consensus status, whereas δ controls the small probability mass retained for the remaining observed labels. If no low-consensus label exists, we retain the initial distribution p_0.

Finally, each rollout receives the probability assigned to its extracted answer, or zero if no valid answer is parsed:

$$r_n = \begin{cases} \widehat{p}(a_{t,n}^o \mid x_t^o), & a_{t,n}^o \in \mathcal{A}_t, \\ 0, & a_{t,n}^o = \varnothing. \end{cases} \tag{12}$$

Thus, rollouts supporting the same answer receive the same reward, while unparseable responses receive no positive reinforcement.

3.4 Virtual-Anchor Advantage Estimation

GRPO computes group-relative advantages by normalizing the rollout rewards:

$$A_n = \frac{r_n - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}} + \epsilon}, \tag{13}$$

where $\mathcal{R} = r_1, \ldots, r_N$, and $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ denote its mean and standard deviation, respectively.

When all rollouts are parsed into the same answer and receive identical rewards, $\sigma_{\mathcal{R}} = 0$, causing their group-relative advantages to vanish and eliminating effective policy-gradient signals. To prevent unanimous rollout groups from yielding zero group-relative advantages and ineffective gradients, inspired by virtual-sample augmentation (He et al., 2026), we introduce a virtual negative anchor and temporarily augment the reward set as $\widetilde{\mathcal{R}} = r_1, \ldots, r_N, r_{\mathrm{va}}$, where $r_{\mathrm{va}} = 0$ is a virtual reward. The mean and standard deviation in Equation (13) are then computed over $\widetilde{\mathcal{R}}$ rather than $\mathcal{R}$, with $\epsilon = 10^{-12}$. Only the advantages of the N real rollouts are retained. The virtual negative anchor affects optimization solely through these normalization statistics and is excluded from the GRPO loss computation.

3.5 Optimization

Finally, the model parameters are updated by minimizing the standard GRPO loss using the resulting advantages, with a KL penalty. The adapted model then performs deterministic inference on the current sample, and its parameters are carried forward to the next test sample, forming a continuous online adaptation process.

4 Experiments

4.1 Experimental Setup

Dataset and metrics. We evaluate on the test split of VAU-Bench (Zhu et al., 2025), which integrates MSAD, UCF-Crime, and ECVA and provides annotations for multiple-choice QA and anomaly classification over 19 major anomaly categories. Only unlabeled test samples are used for test-time adaptation, with ground-truth labels accessed solely for evaluation. We report accuracy for QA as well as binary and multi-class anomaly classification. For the main QA experiments, each setting is repeated with random seeds 42, 43, and 44, and we report the mean and standard deviation. Following VAU-R1, results are reported under *w/o Think* and *w/ Think* prompting settings.

Baselines. We compare with the frozen Qwen2.5-VL-3B-Instruct backbone and VAU-R1. The results for both baselines are taken from Zhu et al. (2025). VAU-R1 uses the same backbone and is trained on the VAU-Bench training split through supervised reinforcement learning across multiple anomaly-understanding tasks. All methods follow the same prompting and evaluation protocols.

Implementation details. We use Qwen2.5-VL-3B-Instruct as the trainable policy backbone, with frozen instances of the same model for video description and question reformulation. We perform full-parameter adaptation with a learning rate of 1×10^{-6}, batch size 2, and one update per batch. GRPO samples $N = 2$ rollouts with temperature 1.0 and top-p 1.0, with $\beta = 0.04$, $\tau = 0.2$, and $\delta = 0.05$. The maximum completion length is 20 tokens for w/o Think and 512 for w/ Think. Videos are sampled at 1 FPS, capped at 60 frames for QA and uncapped for classification. Each subset is independently initialized from the original backbone and processed in shuffled test order. For each batch, the model adapts on unlabeled inputs before deterministic inference on the same batch, and the updated parameters are retained for subsequent batches. Samples rejected by dual-query filtering are excluded from updates. For cross-task evaluation, the QA-adapted model is directly evaluated on binary and multi-class classification without further adaptation or supervision. Experiments are conducted on NVIDIA A800 and H100 GPUs. Additional hyperparameter analyses of batch size and the number of sampled generations are provided in Appendix C.1.

4.2 Main Results

Dataset	Model	w/o Think	w/ Think
MSAD	Qwen2.5-VL-3B	85.83	82.50
	VAU-R1	88.33	87.08
	Ours	**91.94**$^{\pm 0.64}$	**88.47**$^{\pm 0.48}$
UCF-Crime	Qwen2.5-VL-3B	91.63	83.27
	VAU-R1	92.03	**91.63**
	Ours	**93.63**$^{\pm 0.69}$	**91.63**$^{\pm 0.40}$
ECVA	Qwen2.5-VL-3B	85.58	75.81
	VAU-R1	89.53	86.51
	Ours	**92.16**$^{\pm 0.27}$	**90.18**$^{\pm 0.23}$

Table 1: Multiple-choice QA accuracy (%) on VAU-Bench. Superscripts denote standard deviations. Best results are highlighted in bold.

Dataset	Model	w/o Think		w/ Think	
		Bin. Acc.	Multi. Acc.	Bin. Acc.	Multi. Acc.
MSAD	Qwen2.5-VL-3B	79.17	69.58	73.33	56.67
	VAU-R1	82.08	71.25	74.58	60.83
	Ours	**87.50**	**73.75**	**78.33**	**62.50**
UCF-Crime	Qwen2.5-VL-3B	64.54	58.57	62.55	52.19
	VAU-R1	62.55	57.77	62.15	**56.97**
	Ours	**78.88**	**62.55**	**67.33**	54.58
ECVA	Qwen2.5-VL-3B	52.83	30.16	49.89	22.00
	VAU-R1	49.66	30.61	55.78	**31.07**
	Ours	**64.84**	**34.02**	**56.85**	26.03

Table 2: Binary and multi-class anomaly classification accuracy (%). Bin.Acc. and Multi.Acc. denote binary and multi-class accuracy, respectively.

Dataset	Reward	w/o Think	w/ Think
MSAD	Qwen2.5-VL-3B	85.83	82.50
	Binary $0/1$ reward	87.08	85.42
	Entropy-Aware Reward	**88.75**	**86.25**
UCF-Crime	Qwen2.5-VL-3B	91.63	83.27
	Binary $0/1$ reward	91.63	86.06
	Entropy-Aware Reward	**92.03**	**88.05**
ECVA	Qwen2.5-VL-3B	85.58	75.81
	Binary $0/1$ reward	86.07	85.16
	Entropy-Aware Reward	**88.58**	**87.44**

Table 3: Comparison of reward designs.

Configuration	MSAD	UCF-Crime	ECVA
	w/o T / w/ T	w/o T / w/ T	w/o T / w/ T
Qwen2.5-VL-3B	85.83 / 82.50	91.63 / 83.27	85.58 / 75.81
+ ER	88.75 / 86.25	92.03 / 88.05	88.58 / 87.44
+ ER + VA	90.42 / 87.92	92.83 / 90.44	90.87 / 88.58
+ ER + VA + DQ	**91.94 / 88.47**	**93.63 / 91.63**	**92.16 / 90.18**

Table 4: Cumulative component analysis. Here, w/o T and w/ T abbreviate w/o Think and w/ Think, respectively. ER, VA, and DQ denote the entropy-aware reward, virtual negative anchor, and dual-query filtering.

4.2.1 Performance on Multiple-Choice QA Task

Table 1 reports the three-run mean QA accuracy on the three VAU-Bench subsets. Compared with the frozen Qwen2.5-VL-3B backbone, our method improves accuracy by 6.11, 2.00, and 6.58 percentage points under *w/o Think*, and by 5.97, 8.36, and 14.37 points under *w/ Think* on MSAD, UCF-Crime, and ECVA, respectively. Our method also matches or outperforms VAU-R1 in all six settings. It is worth noting that VAU-R1 is trained with ground-truth supervision on a substantially larger labeled dataset, whereas our method adapts exclusively using unlabeled test samples. The largest gain is observed on ECVA under *w/ Think*, improving from 75.81% to 90.18%. Across all datasets, *w/ Think* remains lower than *w/o Think*, a trend also observed in VAU-R1. Recent work attributes such degradation to visual forgetting, where prolonged multimodal reasoning progressively reduces attention to visual tokens and weakens perceptual grounding (Tian et al., 2026). Nevertheless, our larger gains under *w/ Think*, particularly on UCF-Crime and ECVA, indicate that test-time adaptation effectively mitigates this more challenging reasoning regime.

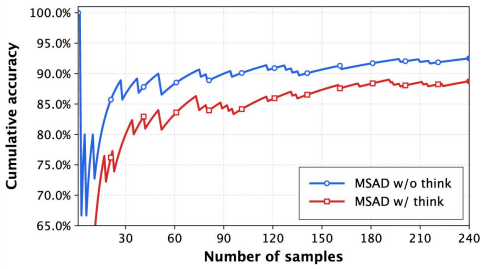

Figure 4: Cumulative accuracy along the test stream on MSAD under *w/o Think* and *w/ Think* settings.

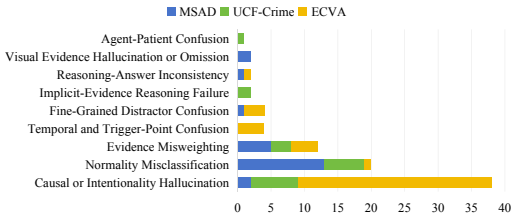

Figure 5: Distribution of model failure modes across MSAD, UCF-Crime, and ECVA.

4.2.2 Performance on Anomaly Classification Task

Table 2 evaluates whether knowledge acquired through multiple-choice QA adaptation transfers to anomaly classification. Without additional parameter updates or classification supervision, our method consistently outperforms the frozen Qwen2.5-VL-3B backbone across all datasets, prompting settings, and classification tasks. For binary classification, the gains range from 4.78 to 14.34 percentage points, with the largest improvement on UCF-Crime under the *w/o Think* setting. For fine-grained multi-class classification, our method also improves the frozen backbone in all six settings, indicating useful cross-task transfer from QA adaptation to category-level anomaly recognition. VAU-R1 is additionally reported as a supervised reference, as it is jointly trained with labeled anomaly-understanding tasks, including classification. Despite using no classification supervision during adaptation, our method matches or surpasses VAU-R1 in 10 of the 12 classification settings, while remaining lower only for multi-class classification on UCF-Crime and ECVA under *w/ Think*. These results further support the effectiveness of the learned test-time representations.

4.3 Insightful Analysis

4.3.1 Ablation of Reward Designs

Table 3 compares our entropy-aware consensus reward with a TTRL-style binary $0/1$ reward under the same basic reinforcement learning pipeline, without dual-query filtering or the virtual negative anchor. Our reward achieves higher accuracy in all six settings, improving over the binary reward by 1.67, 0.40, and 2.51 percentage points under *w/o Think*, and by 0.83, 1.99, and 2.28 points under *w/ Think* on MSAD, UCF-Crime, and ECVA, respectively. These results show that incorporating generation-level uncertainty provides a more informative optimization signal than binary consensus.

4.3.2 Component Ablation

Table 4 presents a cumulative ablation of the three proposed components. The entropy-aware reward improves the frozen backbone across all six settings, with particularly large gains under *w/ Think*. Adding the virtual negative anchor further improves every setting, showing the benefit of restoring optimization signals for unanimous rollout groups. However, consensus alone does not guarantee correctness: Figure D.1 shows that optimization can reinforce the current consensus regardless of whether it is correct. Dual-query filtering addresses this issue by improving pseudo-label quality (Table C.2), and yields further gains in all six settings. Overall, the experiments demonstrate the complementarity of the three components.

4.3.3 Performance Along the Test Stream

Figure 4 further examines model behavior along the test stream. Although cumulative accuracy fluctuates at the early stage due to the limited number of observed samples, it gradually improves and stabilizes as more test samples are processed under both prompting settings. Notably, no progressive degradation is observed despite continuously carrying the adapted parameters forward. This trend indicates that the accumulated updates remain stable over the stream and can provide sustained benefits beyond isolated per-batch optimization.

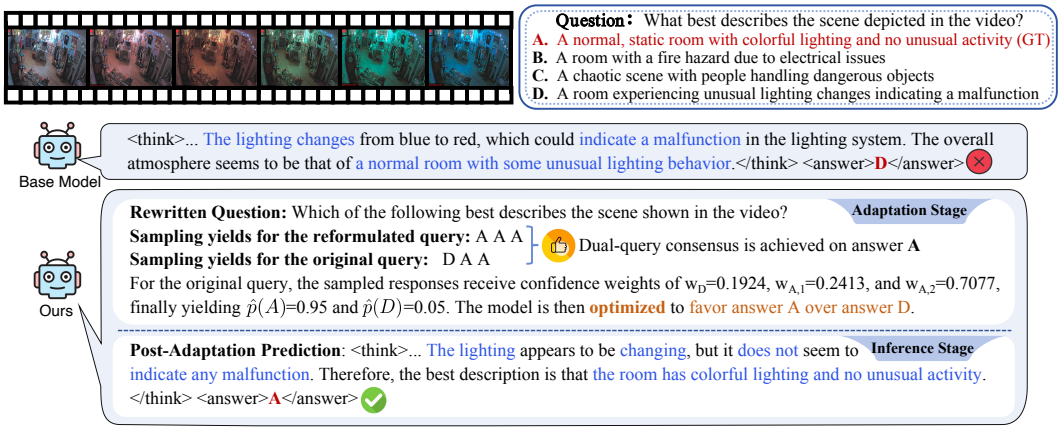

Figure 6: Qualitative case study of multiple-choice QA. The base model Qwen2.5-VL-3B misinterprets lighting changes as a malfunction, whereas our method uses dual-query consistency and entropy-aware rewards to correct the prediction from D to A.

4.3.4 Error Analysis

We categorize the 85 incorrect predictions under the *w/ Think* setting, including 42 from ECVA, 24 from MSAD, and 19 from UCF-Crime. Causal or intentionality hallucination is the dominant failure mode (38/85, 44.71%), reflecting unsupported explanations beyond visual evidence. Normality misclassification ranks second (20/85, 23.53%), followed by evidence misweighting (12/85, 14.12%). Overall, the main bottlenecks lie in evidence-grounded causal reasoning, normality calibration, and discriminative evidence selection rather than basic anomaly recognition.

4.4 Case Study

Figure 6 illustrates how our method corrects an error caused by unreliable evidence prioritization. We use $N = 3$ rollouts per query view and random seed 42. This setting allows us to illustrate reward allocation across different answers within a retained rollout group. The base model incorrectly selects option D by over-interpreting the changing illumination as evidence of a malfunction, despite the absence of explicit anomalous activity. In contrast, stochastic sampling yields (D, A, A) for the original query and (A, A, A) for its semantically equivalent reformulation. The consistent cross-query consensus on A indicates that the erroneous prediction D is unstable, whereas A reflects a more robust interpretation. Our entropy-aware reward further assigns a consensus probability of 0.95 to A and 0.05 to D, producing positive relative advantages for the A-supporting rollouts and a negative advantage for D. After adaptation, the model correctly recognizes that the lighting changes alone do not indicate a malfunction and predicts A. This case illustrates how our method recovers latent correct knowledge from stochastic generations and converts it into an effective self-supervised optimization signal.

5 Conclusion

We presented a test-time reinforcement learning framework for anomalous video understanding, enabling Video-LLMs to continuously adapt from unlabeled test videos. The framework addresses unreliable pseudo-supervision, insufficient reward informativeness, and degenerate group-relative advantages through dual-query consistency filtering, an entropy-aware consensus reward, and a virtual negative anchor, respectively. These components are integrated with GRPO for stable online adaptation along the unlabeled test stream. Experiments on VAU-Bench show consistent improvements under both *w/o Think* and *w/ Think* settings. The adapted models also transfer effectively to anomaly classification without additional classification-specific updates, consistently outperforming the frozen Qwen2.5-VL-3B backbone. Future work will explore more efficient adaptation and better preservation of fine-grained class boundaries during continuous test-time learning.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and proofreading, including improving grammar, clarity, and consistency of terminology. We also used these tools to flag potential discrepancies in cross-references and bibliographic information for author verification. These uses supported manuscript preparation and quality control; AI-generated suggestions do not constitute independent validation of the experimental results. The authors retain responsibility for verifying the reported results and references, deciding on revisions, and ensuring the accuracy and integrity of the final manuscript.

REFERENCES

Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No Need For Real Anomaly: MLLM Empowered Zero-Shot Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35648–35658, 2026.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage Collapse in Group Relative Policy Optimization: Diagnosis and Mitigation, 2026. Accepted at ICML 2026.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Darryl Cherian Jacob, Xinyu Liu, Kai Wang, and Pan He. COPRA: Conditional Parameter Adaptation with Reinforcement Learning for Video Anomaly Detection. *arXiv preprint arXiv:2605.15325*, 2026.

Soumya Shamarao Jahagirdar, Edson Araujo, Anna Kukleva, Muhammad Jehanzeb Mirza, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Rogerio Feris, James R. Glass, and Hilde Kuehne. TTA-Vid: Generalized Test-Time Adaptation for Video Reasoning. 2026.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video Test-Time Adaptation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22952–22961, 2023.

Zihao Liu, Xiaoyu Wu, Jianqin Wu, Xuxu Wang, and Linlin Yang. Language-Guided Open-World Video Anomaly Detection under Weak Supervision. In *International Conference on Learning Representations*, 2026.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient Test-Time Model Adaptation without Forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024.

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.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8523–8533, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33153–33163, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More Thought, Less Accuracy? On the Dual Nature of Reasoning in Vision-Language Models. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4975–4986, 2021.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual Test-Time Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, 2022.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. Beyond Majority Voting: Towards Fine-grained and More Reliable Reward Signal for Test-Time Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37251–37265, San Diego, California, United States, 2026. Association for Computational Linguistics.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In *Computer Vision – ECCV 2024*, pp. 396–416. Springer, 2024.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-Vocabulary Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18297–18307, 2024a.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024b.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models. In *Computer Vision – ECCV 2024*, pp. 304–322. Springer, 2024.

Zhiwei Yang, Chen Gao, and Mike Zheng Shou. PANDA: Towards Generalist Video Anomaly Detection via Agentic AI Engineer. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Muchao Ye, Weiyang Liu, and Pan He. VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688, 2025.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing Large Language Models for Training-free Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *Transactions on Machine Learning Research*, 2025b.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. VAU-R1: Advancing Video Anomaly Understanding via Reinforcement Fine-Tuning, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

A Additional Implementation Details

A.1 Online Test-Time Adaptation Procedure

Each VAU-Bench subset is processed independently. Before adapting to a new subset, the trainable policy is reinitialized from the original Qwen2.5-VL-3B-Instruct backbone. The complete online adaptation procedure is summarized in Algorithm 1.

Algorithm 1: Online test-time reinforcement learning

Require: Test stream $\mathcal{D}_{\text{test}} = \{(v_t, q_t, \widetilde{q}_t, \mathcal{O}_t)\}_{t=1}^{T}$ with precomputed reformulations $\widetilde{q}_t$; initial policy π_{θ_0}; rollout count N; coefficients τ, δ, β; virtual-anchor reward r_{va} and scale s_{va}
Ensure: Post-adaptation predictions $\{\widehat{a}_t\}_{t=1}^{T}$ and adapted policy π_θ

$\theta \leftarrow \theta_0$
$\pi_{\text{ref}} \leftarrow \pi_{\theta_0}$ and freeze π_{ref}
for $t = 1, \dots, T$ **do**
  $(v, q, \widetilde{q}, \mathcal{O}) \leftarrow (v_t, q_t, \widetilde{q}_t, \mathcal{O}_t)$
  Sample $y_i \sim \pi_\theta(\cdot \mid v, q, \mathcal{O}), i = 1, \dots, N$
  Parse $a_i \leftarrow \text{ParseOption}(y_i) \in \{A, B, C, D, \varnothing\}$
  Sample $\widetilde{y}_i \sim \pi_\theta(\cdot \mid v, \widetilde{q}, \mathcal{O}), i = 1, \dots, N$
  Parse $\widetilde{a}_i \leftarrow \text{ParseOption}(\widetilde{y}_i)$
  $a^\star \leftarrow \text{StrictMajority}(a_{1:N})$
  $\widetilde{a}^\star \leftarrow \text{StrictMajority}(\widetilde{a}_{1:N})$ {Ignore invalid labels; require count $> N/2$}
  $s_t \leftarrow \mathbf{1}[a^\star \neq \varnothing]\, \mathbf{1}[\widetilde{a}^\star \neq \varnothing]\, \mathbf{1}[a^\star = \widetilde{a}^\star]$
  if $s_t = 1$ **then**
    for $i = 1, \dots, N$ **do**
      $H_i \leftarrow \dfrac{1}{T_i^-} \displaystyle\sum_{k=1}^{T_i^-} \mathcal{H}[\pi_\theta(\cdot \mid v, q, \mathcal{O}, y_{i,<k})]$ {T_i^- contains positions strictly before EOS}
    end for
    $r_{1:N} \leftarrow \text{EntropyConsensusReward}(a_{1:N}, H_{1:N}, \tau, \delta)$ {Invalid completions receive reward 0}
    if $a_1 = \cdots = a_N \neq \varnothing$ **and** $\text{AllClose}(r_{1:N})$ **then**
      $\mu, \sigma \leftarrow \text{MeanSampleStd}(r_1, \dots, r_N, r_{\text{va}})$
      for $i = 1, \dots, N$ **do**
        $A_i \leftarrow s_{\text{va}} \dfrac{r_i - \mu}{\sigma + 10^{-12}}$
      end for{The virtual anchor has no completion or token loss}
    else
      $\mu, \sigma \leftarrow \text{MeanSampleStd}(r_{1:N})$
      for $i = 1, \dots, N$ **do**
        $A_i \leftarrow \dfrac{r_i - \mu}{\sigma + 10^{-12}}$
      end for
    end if
    for $i = 1, \dots, N$ **do**
      for $k = 1, \dots, T_i^+$ **do**
        $\ell_{i,k}^{\theta} \leftarrow \log \pi_\theta(y_{i,k} \mid v, q, \mathcal{O}, y_{i,<k})$
        $\ell_{i,k}^{\text{ref}} \leftarrow \log \pi_{\text{ref}}(y_{i,k} \mid v, q, \mathcal{O}, y_{i,<k})$
        $d_{i,k} \leftarrow \exp\Big(\ell_{i,k}^{\text{ref}} - \ell_{i,k}^{\theta}\Big) - \Big(\ell_{i,k}^{\text{ref}} - \ell_{i,k}^{\theta}\Big) - 1$
        $\rho_{i,k} \leftarrow \exp\Big(\ell_{i,k}^{\theta} - \text{sg}[\ell_{i,k}^{\theta}]\Big)$
      end for
    end for
    $\mathcal{L}_t \leftarrow \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \frac{1}{T_i^+} \sum_{k=1}^{T_i^+} [-\rho_{i,k} A_i + \beta d_{i,k}]$ {T_i^+ includes EOS}
    Update θ using $\nabla_\theta \mathcal{L}_t$
  else
    $\mathcal{L}_t \leftarrow 0 \cdot \displaystyle\sum_{i=1}^{N} \sum_{k=1}^{T_i^+} \log \pi_\theta(y_{i,k} \mid v, q, \mathcal{O}, y_{i,<k})$ {Graph-connected zero loss; no parameter update}
  end if
  $\widehat{y}_t \leftarrow \text{GreedyDecode}(\pi_\theta \mid v, q, \mathcal{O})$
  $\widehat{a}_t \leftarrow \text{ParseOption}(\widehat{y}_t)$ {Prediction uses the current, optionally updated policy}

43: **end for**
44: **return** $\{\widehat{a}_t\}_{t=1}^{T}$ and π_θ

A.2 Answer Parsing and Invalid Responses

We use a rule-based parser to map each generated response to $\mathcal{A} = \{A, B, C, D\}$ or the invalid label $\varnothing$. The parser first searches for an option enclosed by `<answer>` tags. If no such match is found, it searches for a standalone option tag, such as `<B>` or `<B/>`. It then examines the text following the final `</think>` tag and selects the last standalone option letter in that segment. As a final fallback, it selects the last standalone option letter in the complete response. An option letter is considered standalone only when it is not immediately adjacent to an alphanumeric character. If none of these rules yields a valid option, the response is mapped to $\varnothing$.

Generated response	Parsed answer
`<answer>B</answer>`	B
`<C/>`	C
Reasoning about the video `</think>` Therefore, the answer is D.	D
The possible answers are A and B, but the final choice is C.	C
The response does not contain a valid option letter.	$\varnothing$

Table A.1: Examples of answer parsing.

Invalid responses do not contribute to majority support or entropy-aware consensus estimation and are assigned a reward of zero. Nevertheless, a strict majority requires more than $N/2$ votes among all N rollouts. If the sample passes the cross-query consistency filter, invalid rollouts remain generated sequences and participate in group-wise advantage normalization and the token-level policy loss. The virtual anchor is activated only when all rollouts are valid and predict the same option.

A.3 Dataset Preprocessing and Test-Stream Construction

Before test-time adaptation, each video was decoded using the video-loading pipeline employed in our experiments. The MSAD, UCF-Crime, and ECVA subsets are processed independently. For each subset, the trainable model is reset to the original Qwen2.5-VL-3B-Instruct checkpoint before processing the corresponding test stream. Within each batch, the model first adapts using the retained unlabeled samples and then performs deterministic inference on the same batch.

B Prompt Templates

This section reports the complete prompts used in all stages. Unless otherwise specified, the answer options and their original order are preserved exactly.

B.1 Video Description Prompt

The following prompt is used to obtain a structured, visually grounded description of each test video.

```
[SYSTEM MESSAGE]
You are a video understanding assistant specializing in
anomalous event analysis.  Describe only events that are
supported by the visible content of the video.  Do not
introduce unsupported identities, intentions, causes, or
background information.

[USER MESSAGE]
Carefully analyze the video and provide a concise but
informative description of its content.  Your description
should include:
```

```
1. the scene and the main people, objects, or vehicles;
2. the key actions in their temporal order;
3. the event or behavior that departs from normal activity;
4. the observable cause and consequence of the event, when
they can be determined from the video; and
5. the visual evidence supporting the description.
Output only the video description.
```

B.2 Question Reformulation Prompt

```
[SYSTEM MESSAGE]
You are a precise English paraphrasing assistant for
multiple-choice video anomaly understanding questions.
Your task is to rewrite only the question so that the
wording is clearly different while the meaning stays exactly
the same.
You must preserve:
- the original intent, scope, and difficulty
- the same entities, events, causal relations, temporal
relations, and polarity
- compatibility with all four options and the description
You must not:
- answer the question
- change what is being asked
- introduce new facts, assumptions, or hints
- remove important qualifiers such as primary, key, main,
most likely, or eventual
- change the question into a statement
Output JSON only in this exact format:
{"rewritten_question": "..."}

[USER MESSAGE]
Rewrite the following multiple-choice question in natural
English.
Only rewrite the question. Use the options and description
only to preserve meaning exactly.
Hard requirements:
1. Keep the rewritten text as a question.
2. Keep the meaning exactly the same.
3. Make the wording noticeably different from the original
when possible.
4. Do not change which option is correct.
5. Do not mention that you are paraphrasing.
6. Return JSON only.
Question: {question}
Option 1: {option_1}
Option 2: {option_2}
Option 3: {option_3}
Option 4: {option_4}
Description: {description}
```

B.3 Multiple-Choice QA Prompt without Explicit Reasoning

This is the prompt used by the test-time adaptation implementation for the w/o Think experiments.

```
[USER MESSAGE]
You are a helpful AI assistant performing video anomaly
detection reasoning.
Below is a video description and a reasoning process.  Based
on this, answer the following multiple-choice question
correctly.
### Question:  [QUESTION]
### Options:  "A. "[OPTION_A] "B. "[OPTION_B] "C.
"[OPTION_C] "D. "[OPTION_D]
Please only output one option letter (A, B, C, or D). Please
do not output other words.
```

B.4 Multiple-Choice QA Prompt with Explicit Reasoning

The following prompt is used by the test-time adaptation implementation for the w/ Think experiments.

```
[USER MESSAGE]
You are a helpful AI assistant performing video anomaly
detection reasoning.
Below is a video description and a reasoning process.  Based
on this, answer the following multiple-choice question
correctly.
### Question:  [QUESTION]
### Options:  "A. "[OPTION_A] "B. "[OPTION_B] "C.
"[OPTION_C] "D. "[OPTION_D]
When the user asks a question, you should first carefully
reason through the problem internally, and then present the
final option.
The expected format is:  <think> your detailed reasoning
process here </think> <answer> Please only output one option
letter!  (A, B, C, or D) </answer>.
```

B.5 Joint Binary and Fine-Grained Anomaly Classification Prompt

Binary anomaly detection and fine-grained anomaly classification are performed in a single inference call. The label set contains the `Normal` class and 19 fine-grained anomaly classes. Consequently, predicting `Normal` indicates that no anomaly is detected, whereas predicting any other label simultaneously indicates an anomalous video and specifies its fine-grained anomaly category.

Prompt without explicit reasoning.

```
[SYSTEM MESSAGE]
You are a helpful assistant.

[USER MESSAGE]
Answer the question:  "What type of anomaly appears in the
video?  If no anomaly appears, please answer 'Normal'."
according to the content of the video.
Select one of the answer from:
['Abuse', 'Agriculture', 'Arrest', 'Arson', 'Assault',
'Explosion', 'Fighting', 'Fire', 'Food Safety', 'Medical
Abnormality', 'Natural Hazard', 'Normal', 'Object_falling',
'People_falling', 'Robbery', 'Shooting', 'Stealing',
'Traffic_accident', 'Vandalism', 'Water_incident'].
The expected format is:  <answer> Output the corresponding
option here </answer>
```

Batch Size	Acc.
1	**92.92**
2	91.25
4	90.42
8	88.75

(a) Effect of batch size.

Number of Generations	Acc.
2	**92.50**
3	91.25
4	89.17

(b) Effect of the number of generations.

Table C.1: Hyperparameter analysis of batch size and the number of sampled generations on the MSAD subset of VAU-Bench under the w/o Think setting.

Prompt with explicit reasoning.

```
[SYSTEM MESSAGE]
You are a helpful assistant.

[USER MESSAGE]
Answer the question: "What type of anomaly appears in the
video? If no anomaly appears, please answer 'Normal'."
according to the content of the video.
Select one of the answer from:
['Abuse', 'Agriculture', 'Arrest', 'Arson', 'Assault',
'Explosion', 'Fighting', 'Fire', 'Food Safety', 'Medical
Abnormality', 'Natural Hazard', 'Normal', 'Object_falling',
'People_falling', 'Robbery', 'Shooting', 'Stealing',
'Traffic_accident', 'Vandalism', 'Water_incident'].
When the user asks a question, you should first carefully
reason through the problem internally, and then present the
final option.
The expected format is: <think> Your detailed reasoning
process here </think> <answer> Output the corresponding
option here </answer>
```

C Effectiveness of Reliable Test-Time Optimization Mechanisms

C.1 Hyperparameter Analysis

Table C.1 investigates the effects of batch size and sampled generations N on MSAD under the w/o Think setting. For the batch-size study, the number of generations is fixed at 3, whereas for the generation-number study, the batch size is fixed at 2. All experiments in this table use random seed 42. As shown in Table C.1a, a batch size of 1 achieves the highest accuracy, while a batch size of 2 yields a comparable result. Increasing the batch size to 4 or 8 causes performance degradation. This suggests that smaller batches are more suitable for test-time adaptation, as they allow more frequent updates and reduce interference from jointly optimizing heterogeneous test samples. Regarding the number of generations, Table C.1b shows that $N = 2$ achieves the highest accuracy, whereas further increasing N provides no additional gains. Larger rollout groups may introduce more low-quality or uncertain responses, thereby weakening the consensus signal. Overall, these results indicate that test-time adaptation benefits from compact batches and rollout groups rather than simply increasing either quantity.

C.2 Effect of Dual-Query Construction and Sample Filtering

Table C.2 evaluates the effectiveness of dual-query consistency filtering. Across the three independently adapted datasets, only a small proportion of samples fail to form a strict majority under w/o Think, whereas the no-consensus rates generally increase under w/ Think, particularly on UCF-Crime. This suggests that longer reasoning generations introduce greater response variability and make consensus less stable.

More importantly, among samples whose original and reformulated queries yield different consensus answers, the original-query consensus is incorrect in 50.00%–75.00% of cases across all set-

tings. The proportion reaches 73.33% on ECVA without thinking, 75.00% on MSAD with thinking, and 75.00% on UCF-Crime without thinking. These results indicate that cross-query disagreement is strongly associated with unreliable pseudo-labels. By excluding samples with missing or inconsistent consensus, dual-query filtering prevents a substantial fraction of unstable and frequently incorrect pseudo-supervision from entering test-time optimization.

Dataset	Prompt	No Consensus Original	No Consensus Reformulated	Consensus Mismatch	Incorrect Original Consensus among Mismatches (%)
ECVA	w/o Think	1	4	15	73.33
ECVA	w/ Think	7	7	9	66.67
MSAD	w/o Think	2	6	6	50.00
MSAD	w/ Think	8	4	12	75.00
UCF-Crime	w/o Think	2	9	4	75.00
UCF-Crime	w/ Think	11	12	18	55.56

Table C.2: Effectiveness of dual-query consistency filtering. "Consensus mismatch" denotes samples for which the original and reformulated queries produce different strict-majority answers. The last column reports the proportion of mismatched samples whose original-query consensus is incorrect.

D Additional Case Studies

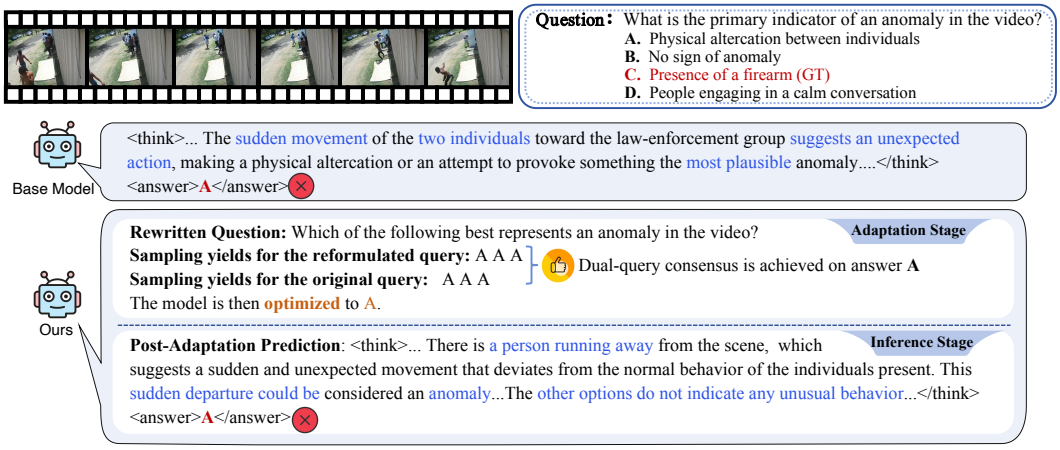

Figure D.1: Qualitative case study of a persistent failure after test-time adaptation.

The main paper presents a successful correction case in which the model changes its prediction from an unsupported malfunction interpretation to the correct normal-scene description. We provide an additional case to illustrate the limitations of the proposed framework.Figure D.1 presents a case in which the frozen model makes an incorrect prediction and the error remains after adaptation.

Failure analysis. Although test-time adaptation produces a more coherent and context-specific reasoning process, the prediction remains incorrect because the decisive visual cue—the firearm—is not recognized. Before adaptation, the model focuses on sudden movement toward law-enforcement personnel and a potential confrontation. After adaptation, it develops a clearer behavioral rationale by associating a person's sudden departure with anomalous behavior, yet still incorrectly selects physical altercation despite acknowledging that no altercation is visible. This indicates improved organization of perceived scene information and coherent causal reasoning. However, the reasoning is still grounded in an incorrect visual representation: the model focuses on secondary motion cues while missing the firearm and introducing potentially unsupported scene details. This case exposes

a fundamental limitation of our method. Dual-query consistency and entropy-aware rewards can improve reasoning over perceived evidence, but they cannot recover critical visual evidence that the backbone fails to encode. When both query views share the same perceptual blind spot, an incorrect answer may remain consistently supported and therefore cannot be corrected through self-generated supervision.