
Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought

Chao Huang¹ Benfeng Wang¹ Wei Wang^{1*} Jie Wen² Chengliang Liu³
Li Shen^{1,4} Xiaochun Cao¹

¹Shenzhen Campus of Sun Yat-sen University ²Harbin Institute of Technology, Shenzhen

³Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University

⁴Shenzhen Loop Area Institute

{huangch253, wangwei29, caoxiaochun}@mail.sysu.edu.cn wangbf23@mail2.sysu.edu.cn
wenjie@hit.edu.cn liucl1996@163.com mathshenli@gmail.com

Abstract

Recent advancements in reasoning capability of Multimodal Large Language Models (MLLMs) demonstrate its effectiveness in tackling complex visual tasks. However, existing MLLM-based Video Anomaly Detection (VAD) methods remain limited to shallow anomaly descriptions without deep reasoning. In this paper, we propose a new task named Video Anomaly Reasoning (VAR), which aims to enable deep analysis and understanding of anomalies in the video by requiring MLLMs to think explicitly before answering. To this end, we propose Vad-R1, an end-to-end MLLM-based framework for VAR. Specifically, we design a Perception-to-Cognition Chain-of-Thought (P2C-CoT) that simulates the human process of recognizing anomalies, guiding the MLLMs to reason about anomalies step-by-step. Based on the structured P2C-CoT, we construct Vad-Reasoning, a dedicated dataset for VAR. Furthermore, we propose an improved reinforcement learning algorithm AVA-GRPO, which explicitly incentivizes the anomaly reasoning capability of MLLMs through a self-verification mechanism with limited annotations. Experimental results demonstrate that Vad-R1 achieves superior performance, outperforming both open-source and proprietary models on VAD and VAR tasks. Codes and datasets will be released at <https://github.com/wbfwonderful/Vad-R1>.

1 Introduction

Video Anomaly Detection (VAD) focuses on identifying abnormal events in videos, and has been widely applied in a range of domains like surveillance systems [53] and autonomous driving [39, 79]. Traditional VAD methods typically fall into two paradigms: semi-supervised and weakly-supervised VADs. The semi-supervised VAD methods [79, 34, 20, 36, 19, 17, 22] aim at modeling the features of normal events, while there are only video-level annotations available for weakly-supervised VAD methods [71, 53, 18, 17, 86, 26, 21, 70]. With the development of vision-language models, some studies introduce semantic information into VAD [64, 73, 72, 80, 7, 23]. However, traditional VAD methods only remain at the level of detection, lacking understanding and explanation of anomalies.

Recently, the reasoning capability of large language models has emerged as a key frontier [44, 9, 58]. Unlike daily dialogue, reasoning requires models to think before answering, enabling them to perform causal analysis and further understanding. In particular, DeepSeek-R1 demonstrates the effectiveness of Reinforcement Learning (RL) in stimulating and refining reasoning capability [9]. Concurrently, parallel efforts have begun to extend reasoning to the multimodal domain [57, 60].

*Corresponding author.

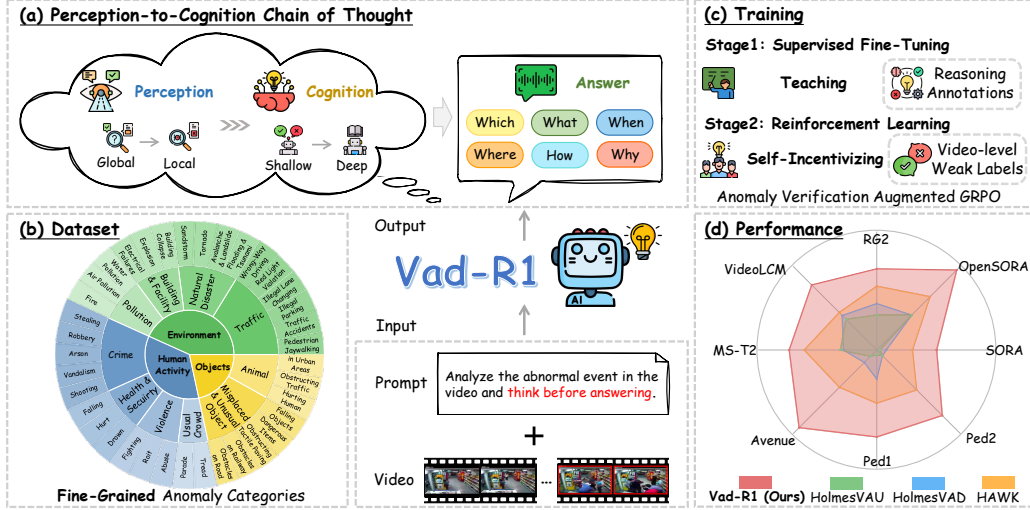


Figure 1: **Overview of Vad-R1.** Vad-R1 is an end-to-end framework for video anomaly reasoning. A structured Perception-to-Cognition Chain-of-Thought is introduced to guide Vad-R1 in performing step-by-step reasoning. Based on the structured CoT, a new dataset for video anomaly reasoning is constructed, including fine-grained anomaly categories. Then, A two-stage training pipeline is adopted to progressively enhance the reasoning capability of Vad-R1. Finally, Vad-R1 outperforms existing MLLMs-based VAD methods with a great margin on VANE benchmark.

Despite the growing interest in reasoning capability, existing Multimodal Large Language Models (MLLMs) based VAD methods still fall short in this regard. Those methods can be divided into two categories based on the role of MLLMs. Some methods regard MLLMs as auxiliary modules [38, 88, 89, 11], where MLLMs provide supplementary explanation after the classifier predicts the anomaly confidence. In this context, anomaly understanding is a step after detection, and the output of MLLMs does not directly promote anomaly detection. Subsequently, although some methods utilize MLLMs to directly perform anomaly detection and understanding [54, 40, 77, 83, 13, 12], MLLMs only generate anomaly descriptions or perform simple anomaly question answering based on video content, lacking thinking and analytical abilities. Thus, reasoning remains underexplored in VAD.

To bridge this gap, we propose a new task: Video Anomaly Reasoning (VAR), which aims to empower MLLMs with the ability to perform structured, step-by-step reasoning about anomalous events in the videos. Compared with existing video anomaly detection or understanding tasks, VAR targets a deeper level of analysis by mimicking the human cognitive process, enabling contextual understanding, behavior interpretation, and norm violation analysis. To this end, we propose Vad-R1, the first end-to-end MLLM-based framework for VAR, which explicitly performs reasoning before generating its response. However, enabling reasoning in video anomaly tasks presents two major challenges. Firstly, existing VAD datasets lack structured reasoning annotations, making them unsuitable for training and evaluating anomaly reasoning models. Secondly, how to effectively train models to acquire reasoning capability remains an open challenge. Unlike tasks with clearly defined objectives, open-ended VAR requires models to perform multi-step causal reasoning, making it difficult to define clear training objectives and directly guide the reasoning process.

For the first challenge, we design a structured Perception-to-Cognition Chain-of-Thought (P2C-CoT) for video anomaly reasoning, as shown in Figure 1(a). Inspired by the process of human understanding the anomalies in the videos, the proposed P2C-CoT first guides the model to perceive from the global environment of the video to the suspicious clips of the video. After perception, the model will make cognition based on visual clues from shallow to deep level. Finally, the model gives the analysis result as answer, including the anomaly category, the anomaly description, the temporal range of anomaly, the approximate spatial position of the anomaly and so on. Then based on the CoT, we construct Vad-Reasoning, a specially designed dataset for VAR, which includes fine-grained anomaly categories as shown in Figure 1(b). Vad-Reasoning consists of two complementary subsets. One subset contains videos with P2C-CoT annotations, which are generated by proprietary models step-by-step. The other subset contains a larger number of videos, where there are only video-level

weak labels available due to high annotation costs. **For the second challenge**, inspired by the success of DeepSeek-R1, we propose a training pipeline with two stages as shown in Figure 1(c). In the first stage, Supervised Fine-Tuning (SFT) is performed to equip the base MLLM with fundamental anomaly reasoning capability. In the second stage, RL is employed to further incentivize the reasoning capability with the proposed Anomaly Verification Augmented Group Relative Policy Optimization (AVA-GRPO) algorithm, an extension of original GRPO [50] specifically designed for VAR. During RL training, the model first generates a group of completions. Based on these completions, the original videos are temporally trimmed and the trimmed videos are then fed back to the model to generate new completions. The two sets of completions are subsequently compared, and an additional anomaly verification reward is assigned if a predefined condition is satisfied. Finally, AVA-GRPO promotes MLLM’s video anomaly reasoning capability through this self-verification mechanism with limited annotations. In summary, the contributions of this paper are threefold:

- We propose a new task named Video Anomaly Reasoning (VAR), which extends traditional VAD from surface-level recognition to deeper cognitive understanding. To this end, we develop Vad-R1, a novel end-to-end MLLM-based framework tailored for VAR, which aims at further analysis and understanding of anomalies in the video.
- We design a structured Perception-to-Cognition Chain-of-Thought to guide the model in performing step-by-step structured reasoning. Building upon this paradigm, we construct Vad-Reasoning, a specially designed dataset for video anomaly reasoning with two subsets. Furthermore, we propose an improved reinforcement learning algorithm AVA-GRPO, which incentivizes the reasoning capability of MLLMs through a self-verification way.
- The experimental results demonstrate that the proposed Vad-R1 achieves consistently superior performance across multiple evaluation scenarios, surpassing both open-source and proprietary models in video anomaly detection and reasoning tasks.

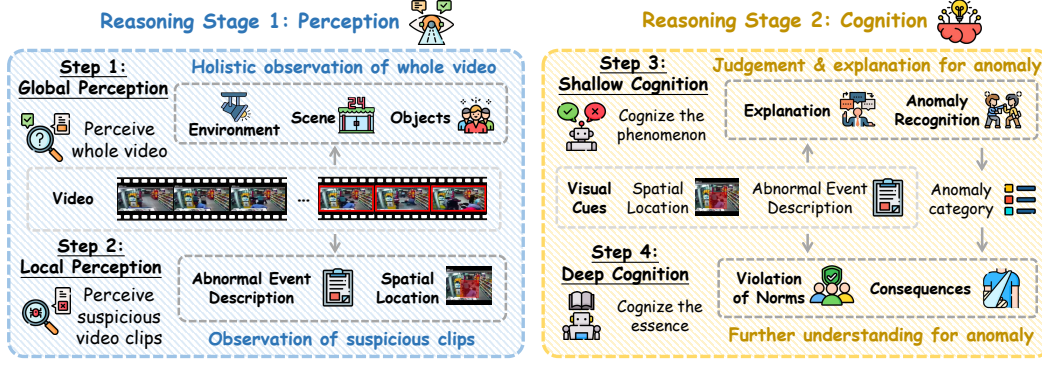
2 Related Works

Video Anomaly Detection and Dataset Video anomaly detection aims at localizing the abnormal events in the videos. Based on the training data, traditional VAD methods typically fall into two paradigms, the semi-supervised VAD [79, 34, 20, 36, 19, 17, 48] and weakly supervised VAD [71, 53, 18, 17, 26, 21]. Furthermore, some studies try to introduce text description to enhance detection [64, 73, 72, 80, 7, 8]. Recently, there has been growing interest in integrating MLLMs into VAD to improve understanding and explanation [38, 54, 40, 77, 83, 88, 89, 11, 13, 12]. However, current studies remain at shallow understanding with MLLMs, lacking in-depth exploration of reasoning capability. In this paper, we propose an end-to-end framework to explore the enhancement of reasoning capability for video anomaly tasks.

Furthermore, the existing VAD datasets primarily provide coarse-grained category labels [53, 71, 39, 1] or abnormal event description [13, 12, 54, 82], lacking annotation of reasoning process. To address this gap, we propose a structured Perception-to-Cognition Chain-of-Thought and a dataset specially designed for video anomaly reasoning, providing step-by-step CoT annotations.

Video Multimodal Large Language Model The video multimodal large models provide an interactive way to understand video content. Early works integrate visual encoders into large language models by aligning visual and textual tokens via mapping networks or adapter layers [27, 32, 41, 87, 91]. Compared with static images, videos inherently contain rich temporal dynamics and redundant visual information. To address this, some studies explore token compression mechanism to handle longer contexts and reduce computational overhead [31, 76, 90, 25]. In addition, recent works have explored online video stream understanding [6, 10, 78, 74]. Nevertheless, these methods remain at the level of video understanding and lack exploration of reasoning capability.

Multimodal Large Language Model with Reasoning Capability Enhancing the reasoning capability of MLLMs has become a major research focus. Early studies have explored multi-stage reasoning frameworks and large-scale CoT datasets to enhance the reasoning capability of MLLMs [75, 63, 35]. Recently, DeepSeek-R1 [9] demonstrates the potential of reinforcement learning in enhancing the reasoning capability, inspiring subsequent efforts to reproduce its success in multimodal domains [24, 84]. In the field of video understanding, several studies also utilize RL



(a) Illustration of the proposed structured Chain-of-Thought, including two stages: perception and cognition.



Figure 2: Overview of the proposed Perception-to-Cognition CoT and Vad-Reasoning dataset.

to improve spatial reasoning [30], temporal reasoning [68] and general causal reasoning [14, 92]. Building upon these advances, this paper focuses on the video anomaly reasoning task.

3 Method: Vad-R1

Overview In this section, we introduce Vad-R1, a novel end-to-end MLLM-based framework for VAR. The reasoning capability of Vad-R1 is derived from a two-stage training strategy: SFT with high quality CoT annotated videos and RL based on AVA-GRPO algorithm. We begin by introducing the proposed P2C-CoT in Section 3.1. Based on the P2C-CoT, we construct Vad-Reasoning, a new dataset as detailed in Section 3.2. Then, we introduce the improved RL algorithm AVA-GRPO in Section 3.3. Finally, we introduce the training pipeline of Vad-R1 in Section 3.4.

3.1 Perception-to-Cognition Chain-of-Thought

When humans interpret a video, they typically first observe the events that occur in the video, and then develop a deeper understanding based on visual observation. Motivated by this, we design a structured Perception-to-Cognition Chain-of-Thought (P2C-CoT) for video anomaly reasoning, which gradually transitions from **Perception** to **Cognition** consisting of 2 stages with 4 steps as shown in Figure 2(a), and concludes with a concise answer as shown in Figure 2(b).

Perception When watching a video, humans typically begin with a holistic observation of the scene and environment, and then shift attention to specific objects or events that appear abnormal. In line with this pattern, the perception stage of the proposed P2C-CoT reflects a transition from global observation to focused local observation. The model initially focuses on the whole environment, describes the scenes and recognizes the objects in the video. This step requires the model to have a comprehensive understanding of the normality in the video. Building upon this holistic understanding of the normality, the model then focuses on the events that deviate from the established normality, identifies what happens, when and where the event happens.

Cognition After observing the video content, humans typically identify abnormal events based on visual cues, and then proceed to reason about the potential consequences. Similarly, the cognitive stage of the proposed P2C-CoT reflects a progression from shallow cognition to deep cognition. The model first assesses the abnormality of the event and explains why it is considered anomalous with

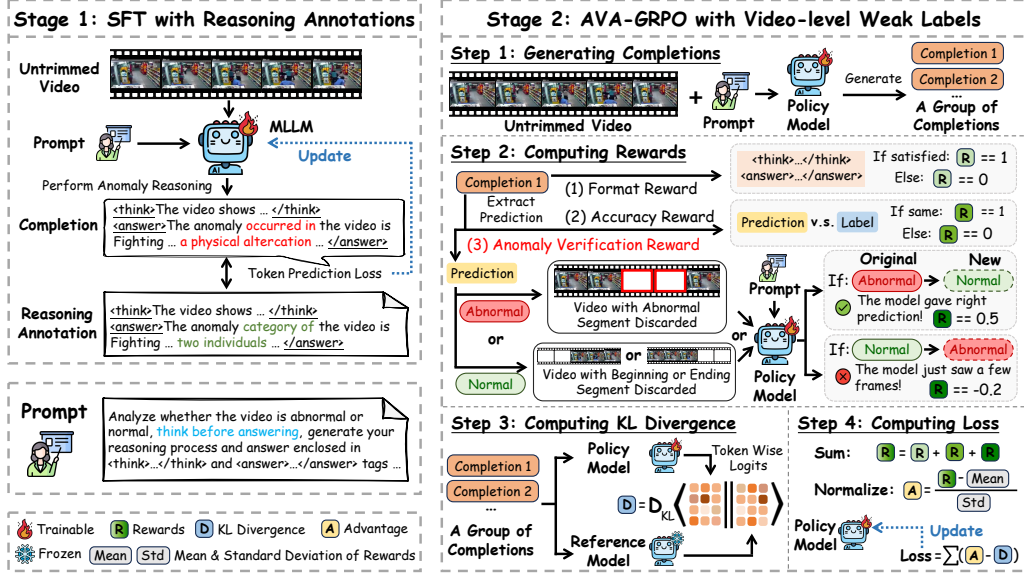


Figure 3: **Illustration of the two-stage training pipeline for Vad-R1.** Stage 1 enables the model to acquire basic reasoning capability with CoT annotated video. Stage 2 further enhances the model’s reasoning capability through reinforcement learning.

relevant visual signals. It then engages in higher-level cognition to reason the underlying causes, the violated social expectations, and the possible consequences of the abnormal event.

Answer As shown in Figure 2(b), after completing the reasoning process, the model is expected to provide a concise summary of its judgment about the given video. The final answer consists of several key dimensions related to the anomaly, including category (**Which**), description of the event (**What**), spatio-temporal localization (**When & Where**), the reason **Why** it is identified as an anomaly and the potential influence (**How**). Notably, for normal videos, the corresponding P2C-CoT is simplified into two steps. Please refer to Appendix B for more details.

3.2 Dataset: Vad-Reasoning

Video Collection The existing VAD datasets generally lack the annotation of reasoning process. To construct a more suitable dataset for VAR, we take the following two aspects into consideration. On the one hand, we aim for the proposed dataset to cover a wide range of real-life scenarios. Similar to HAWK [54], we collect videos from current VAD datasets. The video scenarios include crimes under surveillance (UCF-Crime [53]), violent events under camera (XD-Violence [71]), traffic (TAD [39]), campus (ShanghaiTech [34]) and city (UBnormal [1]). Besides, we also collect videos from ECVA [12], a multi-scene benchmark. On the other hand, we strive to broaden the coverage of anomaly categories. To this end, we define a taxonomy of anomalies comprising three main types: Human Activity Anomaly, Environments Anomaly, and Objects Anomaly. Each type is categorized into several main categories, which are further divided into fine-grained subcategories. Then, we collect additional videos from the internet based on the existing dataset to expand the categories of anomalies. In total, the proposed Vad-Reasoning dataset contains 8203 videos for training and 438 videos for test. As shown in Figure 2(c), the training set of Vad-Reasoning is split into two subsets: Vad-Reasoning-SFT which contains 1755 videos annotated with high-quality reasoning process, and Vad-Reasoning-RL which contains 6448 videos with video-level weak labels.

Annotation To construct the proposed Vad-Reasoning dataset, we design a multi-stage annotation pipeline with two proprietary models Qwen-Max [59] and Qwen-VL-Max [61]. In order to ensure that the P2C-CoT annotation covers all key information in the video, we follow the principle of high frame information density [81]. Specifically, we first prompt Qwen-VL-Max to generate dense

description of video frames. These frame-level descriptions are then fed into Qwen-Max to generate the CoT step-by-step with different prompts. Please refer to Appendix B for more details.

3.3 AVA-GRPO

The original GRPO shows great effectiveness in text-based reasoning tasks. However, as mentioned above, the multimodal tasks like VAR are inherently more complex. In addition, there are only video-level weak labels available for RL stage due to high annotation costs, making it difficult to evaluate output quality based solely on accuracy and format reward. To address this challenge, we propose Anomaly Verification Augmented GRPO (AVA-GRPO), which introduces an additional reward through a self-verification mechanism, as illustrated in the right part of Figure 3.

Overview of GRPO We begin by reviewing the original GRPO [50]. GRPO discards the value model and aims at maximizing the relative advantages among a group of generated answers. Given a question q , the model will first generate a group of candidate completions $O = \{o_i\}_{i=0}^G$. Subsequently, a set of rewards $R = \{r_i\}_{i=0}^G$ are computed based on the predefined reward functions. The rewards are then normalized within the group to compute the relative advantages as

$$A_i = \frac{r_i - \text{mean}(R)}{\text{std}(R)}, \quad (1)$$

where A_i is the advantage score of o_i , which provides more effective assessment of both individual answer quality and relative comparisons within the group. What’s more, to prevent the current policy π_θ from drifting excessively from the reference one π_{ref} , GRPO introduces a KL divergence regularization term. The final objective function of GRPO is formulated as

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{q, O\}} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_\theta(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)} A_i, \text{clip} \left(\frac{\pi_\theta(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right) \right], \quad (2)$$

where the ratio $\frac{\pi_\theta(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)}$ quantifies the relative change between the current policy and the old one, and the $\text{clip}(\cdot, 1 - \epsilon, 1 + \epsilon)$ operation constrains the ratio within a range.

Anomaly Verification Reward GRPO replaces the value model with group relative scores, reducing the memory usage and training time. However, simple accuracy and format rewards are insufficient to evaluate the quality of answers for video anomaly reasoning task. To address this, we propose AVA-GRPO, an extension of GRPO that incorporates a novel anomaly verification reward. As shown in the right part of Figure 3, for each completion o_i , the predicted category of the video is first extracted. The video is then temporally trimmed based on the extracted prediction, and the trimmed video is fed into the model to generate a new answer. Additional anomaly verification rewards are assigned by comparing the original and regenerated answers.

On the one hand, if the video is initially classified as abnormal, the predicted temporal range of the abnormal event is extracted, and the corresponding segment is discarded from the original video to create a new trimmed video containing only normal segments. Then the trimmed video is re-fed into the model. If the trimmed video is subsequently predicted as normal, it suggests that the discarded segment is indeed abnormal and the model’s initial prediction was correct. In this situation, a positive reward will be assigned to reinforce the model’s original prediction.

On the other hand, inspired by Video-UTR [81], we consider the phenomenon of *temporal hacking* for video-MLLMs, where the models tend to generate predictions by relying only on a few frames, typically the beginning or ending of the video, instead of comprehensively processing the entire video sequence, which is detrimental to the recognition of anomaly events. As a consequence, if the video is initially predicted as normal, we randomly discard either the beginning or the ending segment of the video and feed the trimmed video into the model again. If the trimmed video is then

Table 1: Effectiveness of anomaly reasoning.

Method	Strategy	Answer		Detection	
		BLEU-2	METEOR	Recall	F1
Qwen2.5-VL-7B [61]	Direct Answer	0.184	0.339	0.431	0.597
	Random Reasoning	0.179	0.328	0.377	0.540
	Structured Reasoning	0.198 (+0.019)	0.352 (+0.013)	0.696 (+0.265)	0.730 (+0.133)
Qwen3-8B [62]	Direct Answer	0.038	0.184	0.368	0.534
	Random Reasoning	0.040	0.191	0.554	0.655
	Structured Reasoning	0.043 (+0.005)	0.193 (+0.009)	0.681 (+0.313)	0.686 (+0.153)
Vad-R1	Direct Answer	0.268	0.441	0.838	0.861
	Structured Reasoning	0.293 (+0.025)	0.487 (+0.046)	0.843 (+0.005)	0.862 (+0.001)

predicted as abnormal, it suggests the model made its original prediction only based on insufficient visual evidence, which is not expected. Therefore, a negative reward is assigned in this case.

3.4 Training Pipeline

To address the challenge of lacking large-scale annotation, we design a two-phase training pipeline as shown in Figure 3. For the first stage, supervised fine-tuning is performed on the Vad-Reasoning-SFT dataset, in which videos are annotated with high-quality Chain-of-Thought (CoT) as described before. In this stage, the model’s capability is gradually shifted from general multimodal understanding to video anomaly understanding, and it is enabled to acquire basic anomaly reasoning capability. In the second stage, training is continued on the Vad-Reasoning-RL dataset with the proposed AVA-GRPO reinforcement learning algorithm, which evaluates the quality of model responses in a self-verification manner with only video-level weak labels available. This stage aims at moving the model beyond pattern-matching tendencies from SFT, enabling it to develop more flexible, transferable anomaly reasoning capability. Please refer to Appendix C for more details.

4 Experiments

Implementation Details Vad-R1 is trained with two stages based on Qwen-2.5-VL-7B [61]. For the first stage, SFT is performed with Vad-Reasoning-SFT dataset for four epochs. For the second stage, RL is performed with AVA-GRPO for one epoch, where there are only video-level weak labels available for Vad-Reasoning-RL dataset. All experiments are conducted with 4 NVIDIA A100 (80GB) GPUs. Please refer to Appendix C for more details.

Evaluation Metrics and Baselines We first evaluate Vad-R1 on the test set of Vad-Reasoning, focusing on two aspects: anomaly reasoning and anomaly detection. For anomaly reasoning, we assess the text quality of reasoning process with BLEU [46], METEOR [3] and ROUGE [33] metrics. Besides, we also adopt LLM-as-judge evaluations [54]. For anomaly detection, we report accuracy, precision, recall and f1 scores for anomaly classification, along with mIoU and R@K for anomaly temporal grounding. Besides, to further explore the capabilities of Vad-R1, we also conduct experiments on VANE [15], a video anomaly benchmark for MLLMs, where the MLLMs are asked to answer multi-choice questions. In this case, we report the accuracy of every category. We compare Vad-R1 with general video MLLMs [27, 32, 41, 87, 91], reasoning video MLLMs [30, 68, 14, 92] and some proprietary

Table 2: LLM-as-judge evaluations.

Method	R	D	C
<i>Open-Source video MLLMs</i>			
InternVideo2.5 [69]	0.580	0.517	0.487
InternVL3 [96]	0.692	0.608	0.586
VideoChat-Flash [29]	0.367	0.292	0.356
VideoLLaMA3 [85]	0.549	0.449	0.497
LLaVA-NeXT-Video [94]	0.541	0.452	0.491
Qwen2.5-VL [61]	0.638	0.555	0.542
<i>Open-Source video reasoning MLLMs</i>			
Open-R1-Video [67]	0.411	0.307	0.338
Video-R1 [14]	0.390	0.414	0.243
VideoChat-R1 [30]	0.634	0.559	0.528
<i>MLLM-based VAD methods</i>			
Holmes-VAD [88]	0.388	0.275	0.343
Holmes-VAU [89]	0.385	0.301	0.375
HAWK [54]	0.218	0.185	0.115
<i>Proprietary MLLMs</i>			
Claude3.5-Haiku [2]	0.711	0.637	0.611
QVQ-Max [60]	0.690	0.639	0.521
GPT-4o [43]	0.724	0.679	0.542
Vad-R1 (Ours)	0.734	0.659	0.662

Table 3: Performance comparison of anomaly reasoning and detection on Vad-Reasoning dataset.

Method	Params.	Anomaly Reasoning			Anomaly Detection				
		BLEU-2	METEOR	ROUGE-2	Acc	F1	mIoU	R@0.3	R@0.5
Open-Source video MLLMs									
InternVideo2.5 [69]	8B	0.110	0.264	0.109	0.715	0.730	0.417	0.458	0.424
InternVL3 [96]	8B	0.124	0.286	0.116	0.779	0.756	0.550	0.613	0.540
VideoChat-Flash [29]	7B	0.012	0.084	0.047	0.683	0.487	0.536	0.538	0.358
VideoLLaMA3 [85]	7B	0.066	0.200	0.092	0.665	0.624	0.425	0.451	0.419
LLaVA-NeXT-Video [94]	7B	0.094	0.238	0.104	0.651	0.423	0.576	0.601	0.585
Qwen2.5-VL [61]	7B	0.113	0.264	0.116	0.761	0.730	0.567	0.610	0.563
Open-Source video reasoning MLLMs									
Open-R1-Video [67]	7B	0.060	0.179	0.084	0.793	0.790	0.559	0.642	0.540
Video-R1 [14]	7B	0.135	0.317	0.132	0.624	0.694	0.334	0.392	0.328
VideoChat-R1 [30]	7B	0.128	0.287	0.123	0.793	0.790	0.559	0.642	0.540
MLLM-based VAD methods									
Holmes-VAD [88]	7B	0.003	0.074	0.027	0.565	0.120	-	-	-
Holmes-VAU [89]	2B	0.077	0.182	0.075	0.490	0.371	-	-	-
HAWK [54]	7B	0.042	0.156	0.042	0.513	0.648	-	-	-
Proprietary MLLMs									
Claude3.5-Haiku [2]	-	0.097	0.253	0.098	0.580	0.354	0.518	0.543	0.524
GPT-4o [43]	-	0.154	0.341	0.133	0.711	0.760	0.472	0.565	0.476
Gemini2.5-Flash [55]	-	0.133	0.308	0.120	0.624	0.707	0.370	0.437	0.358
Proprietary reasoning MLLMs									
Gemini2.5-pro [56]	-	0.145	0.356	0.137	0.829	0.836	0.636	0.722	0.638
QVQ-Max [60]	-	0.142	0.318	0.121	0.702	0.747	0.430	0.503	0.412
o4-mini [45]	-	0.106	0.263	0.109	0.884	0.875	0.644	0.736	0.631
Vad-R1 (Ours)	7B	0.233	0.406	0.194	0.875	0.862	0.713	0.770	0.706

models [60, 43, 56, 55]. Furthermore, we also consider MLLM-based VAD methods [54, 89, 88]. Due to space limitations, please refer to the Appendix D for more experimental results.

4.1 Main Results

Does reasoning improve anomaly detection? Table 1 demonstrates the effectiveness of anomaly reasoning. On the one hand, we evaluate the performance of Qwen2.5-VL [61] and Qwen3 [62]. As shown in the first two rows of Table 1, compared with directly answering, prompting models to reason according to the proposed perception-to-cognition chain-of-thought will gain greater performance. In the meanwhile, we evaluate the effect of random reasoning. In this case, the performance improvement is minimal, even inferior to direct answering. Notably, Qwen3 is a hybrid reasoning model that supports both reasoning and non-reasoning modes for the same task. The consistent performance gap across different settings further highlights the effectiveness of the proposed P2C-CoT for anomaly reasoning and detection. On the other hand, We compare the performance of Vad-R1 trained with the full P2C-CoT versus training with only the final answer portion of the P2C-CoT as shown in the third row of Table 1. When Vad-R1 is trained with only the final answer, it exhibits a performance drop.

How well does Vad-R1 perform in anomaly reasoning and detection? Table 2 and Table 3 shows the performance comparison on the test set of Vad-Reasoning. On the one hand, Table 2 shows the results of LLM-as-judge evaluations. Following HAWK [54], we evaluate the models’ outputs based on Reasonability(R), Detail(D) and Consistency(C). We observe that Vad-R1 achieves the best performance among all open-source methods and even surpasses proprietary MLLMs such as GPT-4o, particularly in terms of Reasonability and Consistency. On the other hand, Table 3 shows the results of anomaly reasoning and detection tasks. Vad-R1 achieves great performance on both text quality of anomaly reasoning process and the accuracy of anomaly detection. It is worth noting that Vad-R1 significantly outperforms existing proprietary reasoning MLLMs Gemini2.5-Pro, QVQ-Max and o4-mini on anomaly reasoning capability, with BLEU score improvements of 0.088, 0.091, and

Table 4: Performance comparison on VANE.

Method	SORA	OpenSORA	RG2	VideoLCM	MS-T2	Avenue	Ped1	Ped2
<i>Open-Source MLLMs</i>								
Video-LLaMA [87]	11.59	18.00	16.00	10.57	10.41	30.00	16.66	5.55
VideoChat [27]	10.74	28.00	4.00	17.64	20.83	32.25	13.33	13.88
Video-ChatGPT [41]	26.47	22.00	12.00	18.26	16.66	39.39	40.00	19.44
Video-LLaVA [32]	10.86	18.00	16.00	19.23	16.66	3.03	2.77	6.06
MovieChat [51]	8.69	10.00	16.00	14.42	6.25	18.18	6.66	11.11
LLaMA-VID [31]	7.97	14.00	20.00	19.23	14.58	27.27	6.66	19.44
TimeChat [47]	21.73	26.00	28.00	22.11	20.83	24.20	27.58	11.11
<i>MLLM-based VAD methods</i>								
Holmes-VAU [89]	2.17	34.00	24.00	29.81	25.00	6.06	3.33	5.56
Holmes-VAD [88]	6.52	34.00	32.00	33.56	22.92	12.12	20.00	5.56
HAWK [54]	24.64	52.00	44.00	36.54	50.00	36.36	36.67	38.89
Vad-R1 (ours)	41.30	78.00	56.00	63.46	60.42	75.76	60.00	63.89

Table 5: Comparison of different training strategies for Vad-R1.

Strategy	Anomaly Reasoning				Anomaly Detection				
	BLEU-2	ROUGE-1	ROUGE-2	ROUGE-L	Prec.	mIoU	R@0.3	R@0.5	R@0.7
Qwen2.5-VL	0.113	0.505	0.199	0.477	0.768	0.567	0.610	0.563	0.526
+SFT	0.219	0.456	0.196	0.429	0.712	0.612	0.677	0.599	0.535
+AVA-GRPO	0.143	0.513	0.207	0.486	0.810	0.675	0.736	0.661	0.606
+SFT+AVA-GRPO	0.233	0.530	0.238	0.501	0.882	0.713	0.770	0.706	0.651

0.127, respectively. Besides, compared with existing MLLM-based VAD methods, Vad-R1 also exhibits greater advantages in anomaly reasoning and detection.

Table 4 shows the performance comparison on VANE benchmark. Vad-R1 achieves significant improvements and outperforms all baselines, including both general video MLLMs and MLLM-based VAD methods. Although the VANE benchmark is formulated as a multiple-choice question-answering task, its abnormal-related options inherently require the models to perform abnormal behavior recognition, semantic understanding, and reasoning. To better align with our anomaly reasoning setting, we also require the models to think before answering. Therefore, the evaluation on VANE serves as a complementary validation of Vad-R1’s reasoning capability, demonstrating its superior reasoning ability under complex abnormal event understanding tasks.

4.2 Ablation Studies

How to obtain the capability of reasoning? Table 5 shows the effectiveness of different training strategies. When directly performing RL to the base model without prior SFT, the performance improvement is limited. This suggests that, without fundamental reasoning capability, the model struggles to benefit from RL training with video-level weak labels. In contrast, applying SFT leads to a more significant performance improvement, indicating that the structured Chain-of-Thought annotations effectively equip the model with basic anomaly reasoning capability. Notably, the combination of SFT and RL gains the best performance. The results align with the conclusion of DeepSeek-R1 [9], which suggests that SFT stage provides fundamental reasoning capability for the model, while RL stage further enhances its reasoning capability. Table 6 shows the effectiveness of the proposed AVA-GRPO. Compared with original GRPO, AVA-GRPO includes an additional anomaly verification reward, which incentivizes the anomaly reasoning capability of MLLM with only video-level weak labels. In addition, we add a length reward to control the length of output. For both 16-frame and 32-frame settings, AVA-GRPO outperforms the original GRPO across video reasoning and detection tasks. In contrast, using only one reward leads to limited or unstable improvement. These results demonstrate that the combination of length and anomaly rewards is essential for improving the overall reasoning and detection performance.

Table 6: Ablation results of different reward strategies.

Frames	Strategy	Reasoning		Detection				
		ROUGE-L	ROUGE-2	Precision	mIoU	R@0.3	R@0.5	R@0.7
16	GRPO	0.502	0.475	0.861	0.712	0.770	0.699	0.640
	GRPO+len_reward	0.529	0.501	0.856	0.710	0.770	0.697	0.642
	GRPO+ano_reward	0.496	0.467	0.866	0.707	0.765	0.695	0.638
	AVA-GRPO	0.530	0.501	0.882	0.713	0.770	0.706	0.651
32	GRPO	0.495	0.468	0.831	0.695	0.761	0.692	0.624
	GRPO+len_reward	0.528	0.499	0.849	0.701	0.770	0.695	0.631
	GRPO+ano_reward	0.494	0.467	0.842	0.699	0.763	0.686	0.629
	AVA-GRPO	0.533	0.504	0.900	0.726	0.786	0.715	0.661

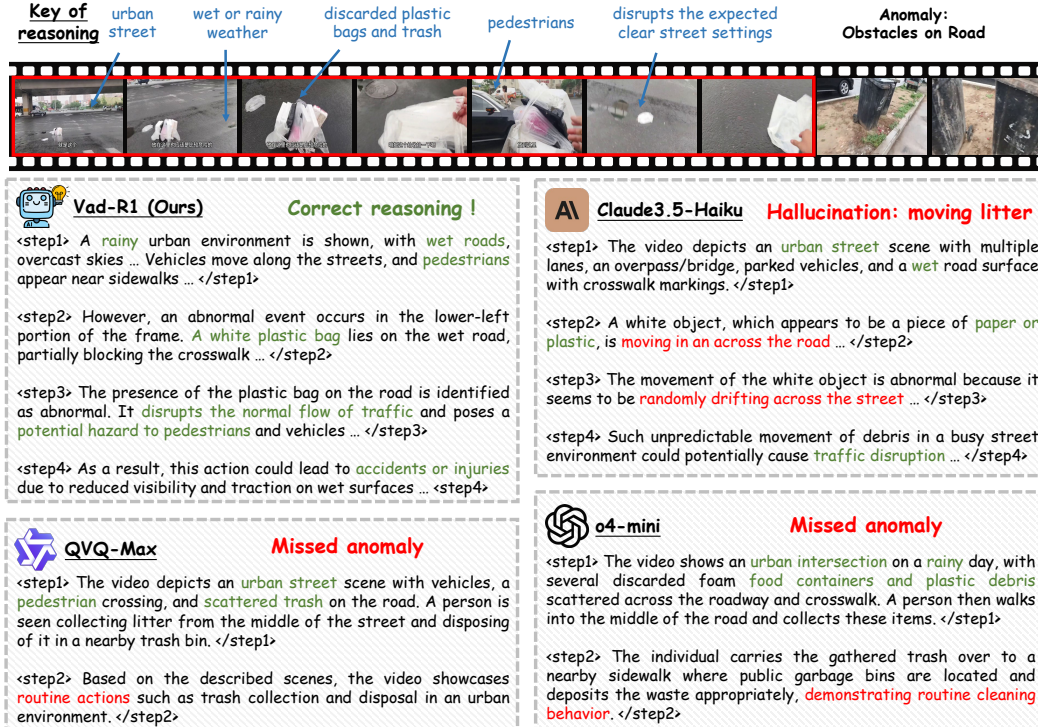


Figure 4: Qualitative result for an abnormal video.

4.3 Qualitative Analyses

As shown in Figure 4, Vad-R1 correctly performs anomaly reasoning and identifies the white plastic bag as an anomaly. In contrast, although Claude identifies the plastic bag as abnormal, it defines the cause of the abnormality as moving plastic bag, rather than the plastic bag acting as an obstacle. Besides, QVQ-Max and o4-mini also identify the white plastic bag, they do not treat it as an anomaly.

5 Conclusion

In this paper, we present Vad-R1, a novel end-to-end MLLM-based framework for video anomaly reasoning which aims to enable deep analysis and understanding of anomalies in videos. Vad-R1 performs structured anomaly reasoning process through a structured Chain-of-Thought that progresses gradually from perception to cognition. The anomaly reasoning capability of Vad-R1 is derived from a two-stage training strategy, combining supervised fine-tuning on CoT-annotated videos and reinforcement learning with an anomaly verification mechanism. Experimental results demonstrate that Vad-R1 achieves superior performance on anomaly detection and reasoning tasks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2022ZD0119200), the National Natural Science Foundation of China (No.62301621, No.62306343), Shenzhen Science and Technology Program (No. 20231121172359002, 2023A008), Shenzhen General Research Project (No. JCYJ20241202125904007), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398, No. 2025A1515011322), the CIE-Smartchip research fund (No.2024-08), and China Postdoctoral Science Foundation (No. 2025T180435, No. 2024M753741).

References

- [1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022. 3, 5, 26
- [2] Anthropic. Claude 3.5 haiku, 2024. URL <https://www.anthropic.com/claude/haiku>. 7, 8, 32
- [3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024. 28
- [5] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20392–20401, June 2023. 26
- [6] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418, 2024. 3
- [7] Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18319–18329, 2024. 1, 3
- [8] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023. 3
- [9] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 3, 9
- [10] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wangui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540*, 2025. 3
- [11] Zongcan Ding, Haodong Zhang, Peng Wu, Guansong Pang, Zhiwei Yang, Peng Wang, and Yanning Zhang. Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model. *arXiv preprint arXiv:2504.10320*, 2025. 2, 3
- [12] Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, and Jun Liu. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2412.07183*, 2024. 2, 3, 5, 25, 26

- [13] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18793–18803, 2024. 2, 3, 25, 26
- [14] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 4, 7, 8, 31, 32
- [15] Hanan Gani, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3123–3140, 2025. 7, 28
- [16] HPCAI Tech. Open-sora: Democratizing efficient video production for all. <https://github.com/hpcaitech/Open-Sora>, 2024. 28
- [17] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 18(8):5171–5179, 2021. 1, 3
- [18] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 54(5):3197–3210, 2022. 1, 3
- [19] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 34(11):9389–9403, 2022. 1, 3
- [20] Chao Huang, Jie Wen, Chengliang Liu, and Yabo Liu. Long short-term dynamic prototype alignment learning for video anomaly detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 866–874, 2024. 1, 3
- [21] Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, and Bob Zhang. Multi-modal evidential learning for open-world weakly-supervised video anomaly detection. *IEEE Transactions on Multimedia*, 2025. 1, 3
- [22] Chao Huang, Qianyi Li, and Bob Zhang. Prototype-guided and dynamic-aware video anomaly detection. *Neural Networks*, page 107583, 2025. 1
- [23] Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-vad: Explainable fine-grained video anomaly detection based on visual-language models. In *Forty-second International Conference on Machine Learning*, 2025. 1
- [24] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 3
- [25] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 3
- [26] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023. 1, 3
- [27] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 7, 9
- [28] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 26, 28

- [29] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhao Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 7, 8, 31, 32
- [30] 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. 4, 7, 8, 31, 32
- [31] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 3, 9
- [32] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3, 7, 9
- [33] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7
- [34] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 1, 3, 5, 26
- [35] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025. 3
- [36] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021. 1, 3
- [37] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 26, 28
- [38] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024. 2, 3
- [39] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021. 1, 3, 5, 26
- [40] 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. In *Proceedings of the ACM on Web Conference 2025*, pages 4004–4013, 2025. 2, 3
- [41] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3, 7, 9
- [42] Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2722–2732, 2023. 29
- [43] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7, 8, 29, 32
- [44] OpenAI. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 1
- [45] OpenAI. Openai o3 and o4-mini system card, 2025. URL <https://openai.com/index/o3-o4-mini-system-card/>. 8, 32
- [46] 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 of the Association for Computational Linguistics*, pages 311–318, 2002. 7

- [47] 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*, pages 14313–14323, 2024. 9
- [48] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15984–15995, 2024. 3
- [49] Runway Research. Gen-2: The next step forward for generative ai. <https://research.runwayml.com/gen2>, 2024. 28
- [50] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3, 6
- [51] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 9
- [52] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024. URL <https://arxiv.org/abs/2409.10173>. 30
- [53] 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*, pages 6479–6488, 2018. 1, 3, 5, 25, 26, 28
- [54] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Yingcong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024. 2, 3, 5, 7, 8, 9, 26, 31, 32
- [55] Gemini Team. Gemini 2.5 flash preview model card, 2025. URL <https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash-preview.pdf>. 8, 32
- [56] Gemini Team. Gemini 2.5 pro preview model card, 2025. URL <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf>. 8, 32
- [57] Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. 1
- [58] Qwen Team. QwQ: Reflect deeply on the boundaries of the unknown, 2024. URL <https://qwenlm.github.io/blog/qwq-32b-preview/>. 1
- [59] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 5
- [60] Qwen Team. QVQ-Max: Think with evidence, 2025. URL <https://qwenlm.github.io/blog/qvq-max-preview/>. 1, 7, 8, 32
- [61] Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 7, 8, 31, 32
- [62] Qwen Team. Qwen3: Think deeper, act faster, 2025. URL <https://qwenlm.github.io/blog/qwen3/>. 7, 8
- [63] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. 3
- [64] Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21017–21025, 2025. 1, 3

- [65] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. 28
- [66] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model, 2023. 28
- [67] Xiaodong Wang and Peixi Peng. Open-r1-video, 2025. URL <https://github.com/Wang-Xiaodong1899/Open-R1-Video>. 7, 8, 31, 32
- [68] Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *arXiv preprint arXiv:2503.13377*, 2025. 4, 7
- [69] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 7, 8, 31, 32
- [70] Lian Wu, Chao Huang, Shuping Zhao, Jinkai Li, Jianchuan Zhao, Zhongwei Cui, Zhen Yu, Yong Xu, and Min Zhang. Robust fall detection in video surveillance based on weakly supervised learning. *Neural networks*, 163:286–297, 2023. 1
- [71] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 1, 3, 5, 25, 26
- [72] 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*, pages 18297–18307, 2024. 1, 3
- [73] 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, pages 6074–6082, 2024. 1, 3
- [74] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025. 3
- [75] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3
- [76] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 3
- [77] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pages 304–322. Springer, 2024. 2, 3
- [78] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025. 3
- [79] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459, 2022. 1, 3
- [80] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. *arXiv preprint arXiv:2412.01095*, 2024. 1, 3
- [81] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025. 5, 6, 24

- [82] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22052–22061, 2024. 3, 26
- [83] 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*, pages 18527–18536, 2024. 2, 3
- [84] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025. 3
- [85] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 7, 8, 31, 32
- [86] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 29:1197–1201, 2022. 1
- [87] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 7, 9
- [88] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 2024. 2, 3, 7, 8, 9, 26, 31, 32
- [89] 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. *arXiv preprint arXiv:2412.06171*, 2024. 2, 3, 7, 8, 9, 26, 31, 32
- [90] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3
- [91] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 7
- [92] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller llms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025. 4, 7
- [93] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. 30
- [94] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7, 8, 31, 32
- [95] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023. 29
- [96] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 7, 8, 31, 32

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction accurately state our contributions.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations are discussed in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The instructions for replicating the results are provided in Section 4.1 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset construction process is provided in Section 3.2 and Appendix. The code and dataset will be publicly released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental settings are provided in Section 4.1 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include the results of significance statistics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Details of computing resource are provided in Section 4.1 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research adheres to the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The potential impacts are discussed in Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The users are required to follow specific usage guidelines, and the dataset and code can only be used for academic research.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All original sources have been properly credited, and the licenses and terms of use associated with the external assets utilized in this paper have been strictly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: All new assets presented in this paper will be released with appropriate metadata and usage documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The training pipeline, dataset and parameters of MLLM are detailed in Section 3.4, Section 4.1 and Appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Summary of Appendix

This appendix provides supplementary information for the main paper. Firstly, we provide detailed information about the proposed Vad-Reasoning dataset, including the construction process, statistical analysis, and some examples. Then, we provide more experimental details covering prompts, settings, parameters, and computing resources. Furthermore, we provide more experimental results as well as visualizations. Finally, we discuss the potential impact and limitation.

B The proposed Vad-Reasoning Dataset

B.1 Annotation Pipeline

The training set of Vad-Reasoning consists of two subsets: Vad-Reasoning-SFT and Vad-Reasoning-RL. For Vad-Reasoning-RL, we retain the original dataset annotations and collapse them into video-level weak labels (Abnormal or Normal). For Vad-Reasoning-SFT, we design a multi-stage annotation process based on the proposed P2C-CoT, as shown in Figure 5.

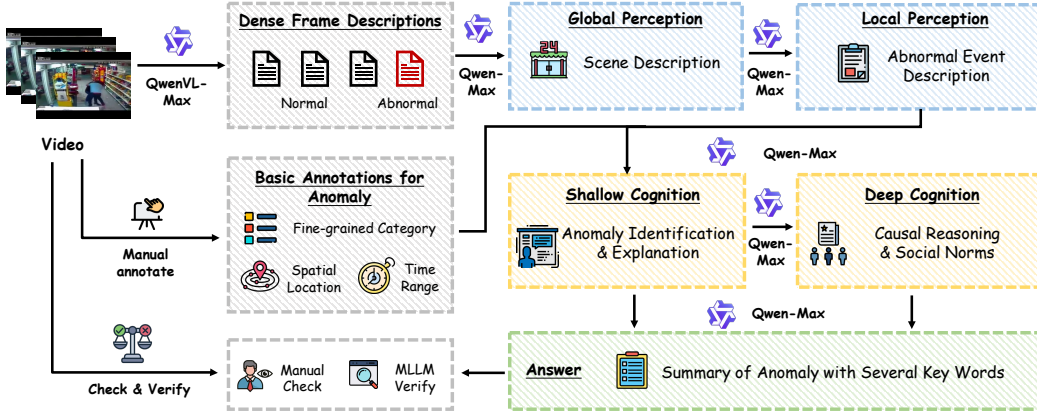
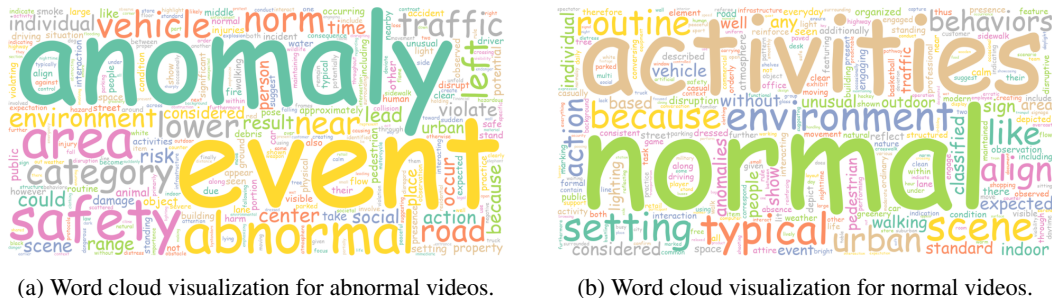


Figure 5: Illustration of multi-stage annotation process of Vad-Reasoning-SFT dataset.

Firstly, we manually annotate the fine-grained anomaly categories of the video, the duration of the anomaly events, and the approximate spatial location where the anomaly events occurred. Then, in order to ensure that the P2C-CoT annotation covers all key information in the video, we follow the principle of high frame information density [81]. Specifically, the video is decomposed into separate frames with a frame interval of 16. The extracted frames are then fed into Qwen-VL-Max to generate detailed dense frame descriptions. Considering the redundancy and high density of video information, directly prompting the model to generate annotations for the entire video would result in high annotation costs and information loss. In comparison, we first perform dense sampling at certain intervals, and then require Qwen-VL-Max to describe the video frames in detail, which can greatly preserve the key information in the video frames and reduce information loss.

After obtaining detailed dense video frame descriptions, we employ Qwen-Max to summarize and generalize the proposed P2C-CoT from these descriptions. Specifically, to ensure the structure of the reasoning process, we generate each step of P2C-CoT separately with different prompts, instead of generating them all at once. This design allows each step to focus on a specific task (e.g. step 1 focuses on scene description while step 2 focuses on abnormal event description). Each step will receive the output of previous step as input to ensure logical coherence. After generating all the reasoning steps, the model then generate a short summary about the anomaly in the video as final answer. In addition, we explicitly define anomaly categories to align model outputs with human-defined semantics, while enforcing constraints on relevance, objectivity, and neutrality. To further ensure the correctness of the generated CoT, all annotations are firstly verified by large language models and subsequently checked by human experts.



B.2 Statistical Analysis and Comparison

We compare Vad-Reasoning with existing video anomaly detection and understanding datasets in Table 7 and Table 8. Vad-Reasoning consists of a total of 8641 videos, covering 34 million frames and over 360 hours of duration, making it one of the largest datasets among video anomaly understanding benchmarks. Besides, Vad-Reasoning-SFT provides fine-grained Chain-of-Thought (CoT) annotations, explicitly simulating human reasoning over abnormal events, with an average annotation length of 260 words. For the annotations, the recent video anomaly understanding datasets like CUA [13] and ECVA [12] contain the description about the cause and effect of the anomaly. However, their corresponding annotations are isolated and disjointed, lacking a systematic structure and logical progression. In contrast, the proposed Vad-Reasoning-SFT dataset provides structured and coherent anomaly reasoning annotation.

Figure 6 illustrates word cloud visualizations generated from the annotations of our Vad-Reasoning-SFT dataset. For abnormal videos, terms such as anomaly, event, and safety frequently appear, emphasizing that the CoT annotations of abnormal videos focus on irregular situations and potential risks. In contrast, normal videos are characterized by high-frequency words like activities, normal, and environment, which highlight everyday scenarios and typical human behaviors. Our Vad-Reasoning-SFT dataset not only captures descriptive content but also encourages models to engage in deeper causal interpretation of abnormal events. These observations indicate that our dataset is purposefully constructed to promote both recognition and reasoning in complex real-world contexts.

Figure 7 presents a comprehensive statistical overview of the proposed Vad-Reasoning dataset. The overall distribution of video length is relatively even as shown in Figure 7(a) and (b). Most of the videos in the Vad-Reasoning dataset are collected from UCF-Crime [53] and XD-Violence [71] as shown in Figure 7(c) and (d). And we collect additional 10 percent of videos from the internet. The proportion of normal and abnormal videos in the two subsets is basically balanced as shown in Figure 7(e). Finally, the fine-grained anomaly distributions are shown in Figure 7(f)-(h).

B.3 Examples

We provide two examples of the proposed Vad-Reasoning dataset in Figure 8 and Figure 9. Notably, the CoT of normal videos will be simplified into two steps, the simple perception and cognition.

C Implementation Details

C.1 Prompt

The prompt used for performing video anomaly reasoning is shown in Figure 10. The prompt is composed of three parts, **Task Definition**, **Output Specification** and **Format Requirements**. Firstly, the **Task Definition** outlines the overall goal of video anomaly reasoning and explicitly require the model to think before answering. Secondly, the **Output Specification** provides detailed guidelines on the reasoning process and the expected answer. Finally, the **Format Requirements** presents concrete output examples with explicitly defined tags (e.g., `<think></think>` and `<answer></answer>`).

Table 7: Basic metadata comparison of datasets. Here "Mixture" indicates that the dataset is composed by integrating videos from multiple existing datasets.

Dataset	Source	Videos	Frames	Duration	Resolution	FPS
Traditional Video Anomaly Detection Datasets						
UCF-Crime [53]	Surveillance	1900	13,741,393	128h	320 × 240	Multiple
XD-Violence [71]	Multiple	4754	18,714,328	217h	Multiple	24
ShanghaiTech [34]	Campus	437	317,398	-	856 × 480	-
UCSD Ped1 [28]	Campus	70	14,000	-	238 × 158	-
UCSD Ped2 [28]	Campus	28	4,560	-	360 × 240	-
CUHK Avenue [37]	Campus	37	30,652	0.3h	640 × 360	25
TAD [39]	Traffic	518	540,212	-	Multiple	-
UBnormal [1]	Generation	543	236,902	2.2h	Multiple	30
NWPU Campus [5]	Campus	547	1,466,073	16.3h	Multiple	25
Video Anomaly Understanding Datasets						
UCA [82]	Surveillance	1854	13,163,270	121.9h	320 × 240	Multiple
CUVA [13]	Multiple	986	3,345,097	32.5h	Multiple	Multiple
ECVA [12]	Multiple	2127	19,042,560	88.2h	Multiple	Multiple
VAD-Instruct50k [88]	Mixture	6654	32,455,721	345h	Multiple	Multiple
HIVAU-70k [89]	Mixture	6654	32,455,721	345h	Multiple	Multiple
HAWK [54]	Mixture	7898	14,878,233	142.5h	Multiple	Multiple
Vad-Reasoning-SFT	Mixture	2193	8,680,615	88.3h	Multiple	Multiple
Vad-Reasoning-RL	Mixture	6448	25,495,729	272.2h	Multiple	Multiple
Vad-Reasoning	Mixture	8641	34,173,344	360.5h	Multiple	Multiple

Table 8: The annotation type comparison of datasets. * denotes that the videos in Vad-Reasoning-RL are only labeled with video-level labels (Abnormal or Normal).

Dataset	Anomalies	Text Annotation	Reasoning
Traditional Video Anomaly Detection Datasets			
UCF-Crime [53]	13	Anomaly class	-
XD-Violence [71]	6	Anomaly class	-
ShanghaiTech [34]	13	-	-
UCSD Ped1 [28]	5	-	-
UCSD Ped2 [28]	5	-	-
CUHK Avenue [37]	5	-	-
TAD [39]	7	-	-
UBnormal [1]	22	-	-
NWPU Campus [5]	28	-	-
Video Anomaly Understanding Datasets			
UCA [82]	13	Event descriptions	-
CUVA [13]	42	Anomaly description, cause, effect	Isolated
ECVA [12]	100	Anomaly description, cause, effect	Isolated
VAD-Instruct50k [88]	13	Clip caption & QA	-
HIVAU-70k [89]	13	Clip/Event/Video-level Caption & QA	Isolated
HAWK [54]	-	Anomaly description & QA	-
Vad-Reasoning-SFT	37	Chain-of-Thought	Structured & coherent
Vad-Reasoning-RL	1*	Video-level label	-
Vad-Reasoning	37	Hybrid annotation	-

C.2 Training Process of AVA-GRPO

The core of the proposed AVA-GRPO is the additional anomaly verification reward as shown in Algorithm 1. Besides, we additionally consider a length reward. We first separately calculate the length of the reasoning text for abnormal videos and normal videos in Vad-Reasoning-SFT. During RL training, if the length of output satisfies the corresponding range, a length reward will be assigned. Notably, for each completion, the model will be only updated once. Consequently, the objective function of AVA-GRPO is simplified as

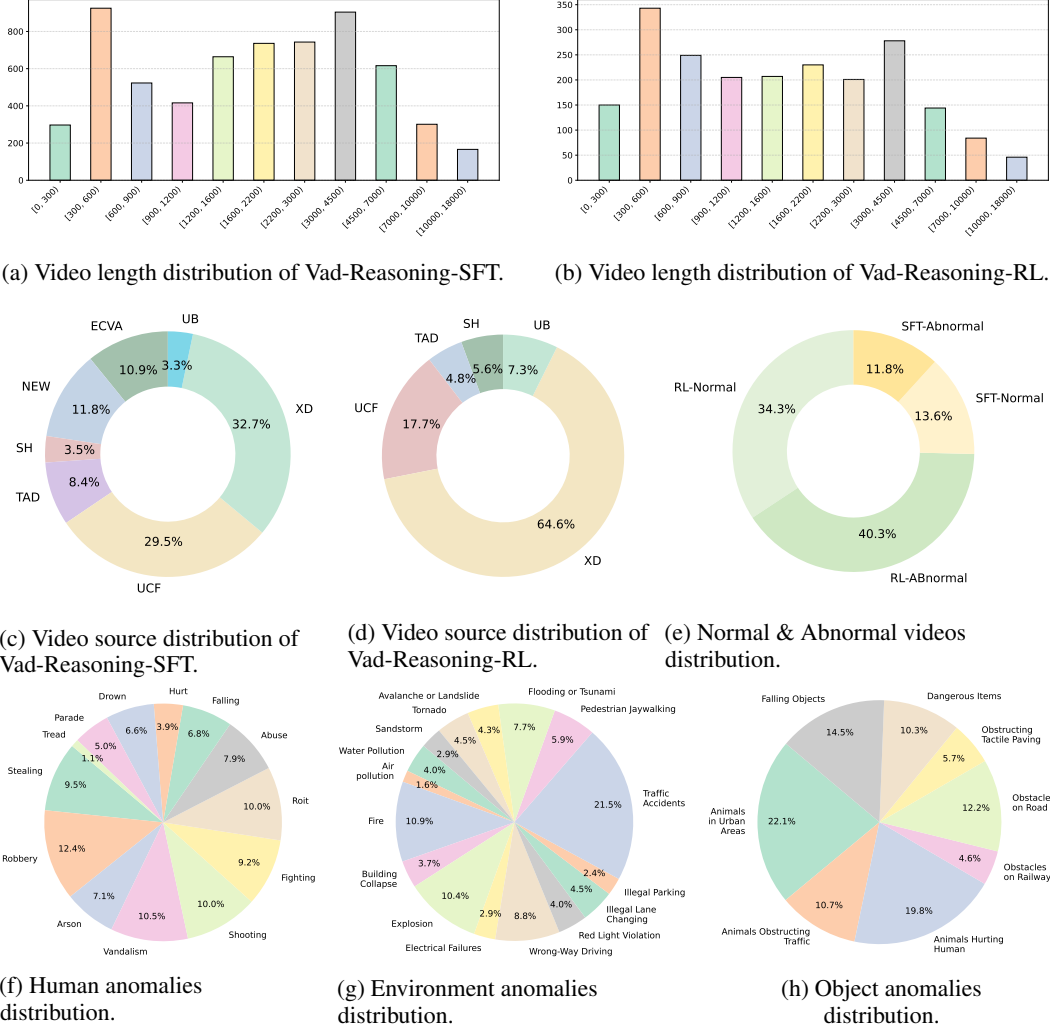


Figure 7: Statistical analyses of the proposed Vad-Reasoning dataset.

$$\mathcal{L}_{\text{AVA-GRPO}}(\theta) = \mathbb{E}_{\{q, O\}} \left[\frac{1}{G} \sum_{i=1}^G \left(\frac{\pi_{\theta}(o_i | q)}{\pi_{\theta_{\text{no_grad}}}(o_i | q)} A_i - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right) \right], \quad (3)$$

where $\pi_{\theta_{\text{no_grad}}}$ is equivalent to π_{θ} . Finally, the training process of AVA-GRPO is shown in Algorithm 2.

C.3 More Experimental Details

All experiments are conducted on 4 NVIDIA A100 (80GB) GPUs. For supervised fine-tuning stage, we train the base MLLM on Vad-Reasoning-SFT dataset for four epochs, taking approximately 6 hours. For reinforcement learning stage, we continue to train the model on the Vad-Reasoning-RL dataset for one epoch, taking about 26 hours. For efficiency, we uniformly normalize the video to 16 frames, and the maximum number of pixels per frame is limited to $128 \times 28 \times 28$ during training. The learning rates for both stages are set to 1×10^{-6} . The number of completions generated in a group is set to 4. The hyperparameter β in Equation 3 is set as 0.04. AVA-GRPO includes five types of rewards. The specific values and meanings are shown in Table 9. For normal videos, the length range of reasoning process is set as $[140, 261]$, while it is set as $[233, 456]$ for abnormal videos.



Figure 8: An abnormal example of Vad-Reasoning.

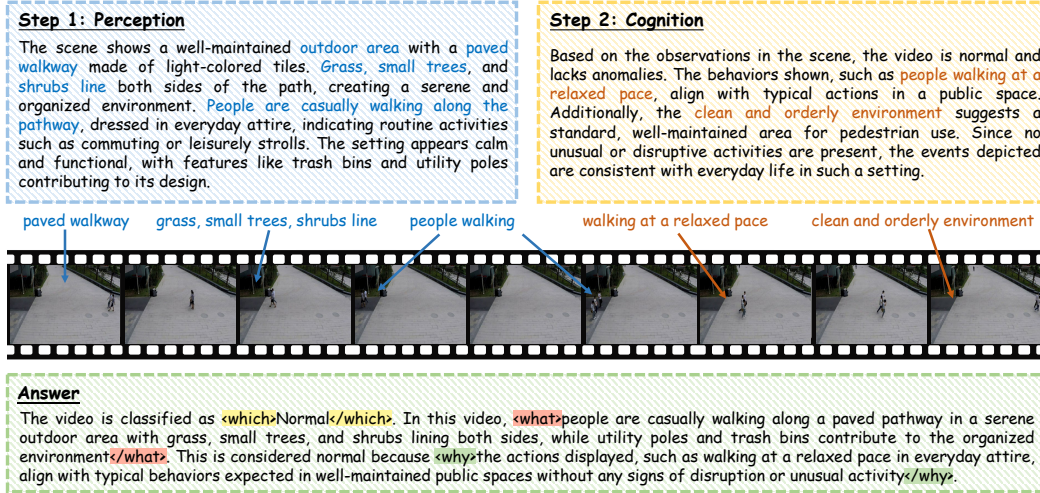


Figure 9: A normal example of Vad-Reasoning.

C.4 Evaluation on VANE Benchmark

VANE [15] is a benchmark designed to evaluate the ability of video-MLLMs to detect anomalies in the video. It consists of 325 video clips and 559 question-answer pairs, covering both real-world surveillance and AI-generated video clips, and are categorized into nine anomaly types. For real-world anomalies, VANE collect 128 videos clips from existing video anomaly detection datasets (e.g., CUHK Avenue [37], UCSD-Ped1/Ped2 [28], and UCF-Crime [53]). For AI-generated anomalies, VANE includes 197 clips videos generated with SORA [4], OpenSora [16], Runway Gen2 [49], ModelScopeT2V [65] and VideoLCM [66]. We report the performance of Vad-R1 and other MLLM-based VAD methods on different categories. Notably, since Vad-R1 is trained with the proposed

<p>Your task is to analyze whether the given video is abnormal or normal. Think before answering, and generate:</p> <ol style="list-style-type: none"> 1. A structured reasoning process enclosed in <code><think>/<think></code> tags 2. A final explanation enclosed in <code><answer>/<answer></code> tags
<p>For abnormal videos, the reasoning should be based on a structured 4-step process:</p> <p><code><think></code> must include the following four steps enclosed in corresponding tags:</p> <p><code><step1></code>: Scene Description — Provide an objective overview of the environment and normal behaviors, without mentioning any abnormal activity or speculation.</p> <p><code><step2></code>: Abnormal Event Description — Describe the abnormal event and its approximate spatial location (e.g., bottom left of the frame), without explaining why it is abnormal.</p> <p><code><step3></code>: Abnormal Event Recognition — Explain why this event is considered abnormal compared to normal patterns or expectations.</p> <p><code><step4></code>: Causal Reasoning and Social Norms — Analyze potential negative consequences and explain how this behavior violates social norms or expectations.</p> <p>Answer must be a single, coherent paragraph in natural language, which includes exactly the following five tags:</p> <p><code><which></code>: Define the video as <code>"Abnormal"</code></p> <p><code><what></code>: What happened (describe the anomalous event)</p> <p><code><when></code>: When it happened, in normalized frame indices (e.g., <code><when>[0.25, 0.45]</when></code>)</p> <p><code><where></code>: Where it happened (use approximate spatial descriptions)</p> <p><code><why></code>: Why it is considered abnormal</p> <p><code><how></code>: How this behavior could cause harm or violate norms</p> <p>For normal videos, the reasoning should be simplified to just two steps:</p> <p><code><think></code> must include only the following two steps:</p> <p><code><step1></code>: Scene and Object Description — Provide a concise and objective overview of the environment and typical behaviors, without mentioning anomalies.</p> <p><code><step2></code>: Normal Event Explanation — Explain why the video is considered normal.</p> <p>Answer must be a single, coherent paragraph in natural language, which includes the following three tags:</p> <p><code><which></code>: Define the video as <code>"Normal"</code></p> <p><code><what></code>: A concise description of the event in the video.</p> <p><code><why></code>: Why it is considered normal.</p>
<p>Example Output for Abnormal Videos:</p> <pre> <think> <step1>The video shows ...</step1> <step2>Next, we observe an abnormal event ...</step2> <step3>Based on these observations ...</step3> <step4>As a result, this behavior ...</step4> </think> <answer> The video is classified as <which>Abnormal</which>. In this video, <what>a pedestrian ...</what>, occurring during the time range <when>[0.121, 0.826]</when>. The event takes place approximately in the <where>lower-left area of the frame</where>. This is considered abnormal because <why>pedestrians are expected to ...</why>. As a result, <how>such behavior could ...</how>. </answer> </pre> <p>Example Output for Normal Videos:</p> <pre> <think> <step1>The video shows scenes ...</step1> <step2>Based on the described scenes ...</step2> </think> <answer> The video is classified as <which>Normal</which>. In this video, <what>people are ...</what>. This is considered normal because <why>...</why> </answer> </pre>

Figure 10: Prompt template for performing video anomaly reasoning.

Table 9: Reward types and the corresponding values.

Type	Meaning	Value
Accuracy	Evaluate classification result	1
Format	Evaluate format of output	1
Anomaly verification: Abnormal	Evaluate correctness of videos predicted as abnormal	0.5
Anomaly verification: Normal	Evaluate correctness of videos predicted as normal	-0.2
Length	Evaluate length of output	0.2

Vad-Reasoning dataset, which incorporates videos from UCF-Crime, we exclude the corresponding UCF-Crime subset from VANE benchmark.

D More Experimental Results

D.1 LLM-Guided Evaluation

In this section, we provide additional LLM-guided evaluations [42, 95] to further assess the reasoning quality of Vad-R1. Table 10 demonstrates the results of pair-wise comparison [95]. Given the output of the our Vad-R1 and other models, we require GPT-4o [43] to compare the two answers and choose the better one. We also consider the impact of the order of answers on the judgment. The results show that GPT-4o overwhelmingly trends to choose Vad-R1 in both comparison orders, confirming the superior reasoning quality, contextual alignment, and interpretability of its responses. This consistent preference demonstrates that Vad-R1 not only produces semantically accurate descriptions but also exhibits a more human-like reasoning process. Table 11 shows the results of double-right [42], which jointly evaluates whether the model can provide a reasonable thinking process and correctly

Algorithm 1 Anomaly verification reward

Input: Current video v , policy model π_θ , generated completions $O = \{o_i\}_{i=1}^G$.

Output: Anomaly verification reward R_{ano} .

```
1: Init anomaly verification reward:  $R_{ano} = \{r_i\}_{i=1}^G$ , where  $r_i = 0$ 
2: for each  $o_i \in O$  do
3:   Extract prediction  $p$  of  $v$  from completion  $o_i$ 
4:   if  $p == \text{Normal}$  then
5:     Randomly discard either the beginning or the ending segment of  $v$ 
6:   else
7:     Discard the predicted abnormal segment of  $v$ 
8:   end if
9:   Obtain a trimmed video  $\tilde{v}$ 
10:  Generate a new completion  $\tilde{o} \sim \pi_\theta(\cdot \mid p, \tilde{v})$ 
11:  Extract new prediction  $\tilde{p}$  of  $\tilde{v}$  from new completion  $\tilde{o}$ 
12:  if  $p == \text{Abnormal}$  and  $\tilde{p} == \text{Normal}$  then
13:    Assign positive reward  $r_i \leftarrow 0.5$ 
14:  else if  $p == \text{Normal}$  and  $\tilde{p} == \text{Abnormal}$  then
15:    Assign negative reward  $r_i \leftarrow -0.2$ 
16:  end if
17: end for
18: return  $R_{ano} = \{r_i\}_{i=1}^G$ 
```

Algorithm 2 AVA-GRPO

Input: Vad-Reasoning-RL dataset $\mathcal{D} = \{(v_j, Y_j)\}_{j=1}^N$, initial policy model $\pi_{\theta_{\text{init}}}$.

Output: Updated policy model π_θ .

```
1: Init policy model:  $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$ 
2: Init reference model:  $\pi_{\text{ref}} \leftarrow \pi_\theta$ 
3: for  $e \in \{1, \dots, E\}$  do
4:   for  $(v_j, Y_j) \in \mathcal{D}$  do
5:     Generate a group of completions  $O = \{o_i\}_{i=1}^G \sim \pi_\theta(\cdot \mid p, v_j)$ 
6:     Compute accuracy reward  $R_{acc} = \{r_i\}_{i=1}^G$ 
7:     Compute format reward  $R_f = \{r_i\}_{i=1}^G$ 
8:     Compute anomaly verification reward  $R_{ano} \leftarrow \text{Algorithm 1}$ 
9:     Compute length reward  $R_{len} = \{r_i\}_{i=1}^G$ 
10:    Compute sum  $R = R_{acc} + R_f + R_{ano} + R_{len}$ 
11:    Compute advantages  $A = \frac{R - \text{mean}(R)}{\text{std}(R)}$ 
12:    Update  $\pi_\theta$  with Equation 3
13:  end for
14: end for
15: return  $\pi_\theta$ 
```

identify the anomaly, thereby offering a more comprehensive assessment of the anomaly reasoning capability. In this setting, we adopt semantic similarity to evaluate the correctness of the reasoning process. We adopt Qwen3-embedding [93] and Jina-embedding-V3 [52] to calculate semantic similarity separately. We observe that Vad-R1 achieves the highest RR score of 66.29% and 70.16% under different embedding models, indicating that its reasoning outputs are most aligned with the predictions.

D.2 Experiments on More Input Tokens

During both training and inference, the video is uniformly sampled into 16 frames as input, with a maximum pixel count of $128 \times 28 \times 28$ per frame. In this section, we increase the number of frames to 32 and 64 per video, and the maximum pixel to $256 \times 28 \times 28$ per frame. The results are shown in Table 12. On the one hand, We observe that increasing the number of frame from 16 to 64 yields improvement across both anomaly reasoning and detection, showing that the extra frames

Table 10: Pair-wise comparison between Vad-R1 and other models.

Model		Rate			Model		Rate		
A	B	A_win	B_win	Tie	A	B	A_win	B_win	Tie
Open-Source video MLLMs									
Vad-R1	InternVideo2.5 [69]	83.56	12.79	3.65	InternVideo2.5 [69]	Vad-R1	11.42	70.09	18.49
Vad-R1	InternVL3 [96]	78.82	17.31	3.87	InternVL3 [96]	Vad-R1	24.15	50.11	25.74
Vad-R1	VideoChat-Flash [29]	90.64	7.76	1.60	VideoChat-Flash [29]	Vad-R1	11.64	74.89	13.47
Vad-R1	VideoLLaMA3 [85]	85.93	10.09	3.98	VideoLLaMA3 [85]	Vad-R1	13.50	81.90	4.60
Vad-R1	LLaVA-NeXT-Video [94]	84.97	11.16	3.87	LLaVA-NeXT-Video [94]	Vad-R1	14.35	81.55	4.10
Vad-R1	Qwen2.5-VL [61]	83.83	11.16	5.01	Qwen2.5-VL [61]	Vad-R1	17.54	70.39	12.07
Open-Source video reasoning MLLMs									
Vad-R1	Open-R1-Video [67]	90.21	7.74	2.05	Open-R1-Video [67]	Vad-R1	5.24	93.62	1.14
Vad-R1	Video-R1 [14]	82.23	12.07	5.69	Video-R1 [14]	Vad-R1	8.88	86.10	5.01
Vad-R1	VideoChat-R1 [30]	79.95	14.81	5.24	VideoChat-R1 [30]	Vad-R1	18.68	70.62	10.71
MLLM-based VAD methods									
Vad-R1	Holmes-VAD [88]	80.28	5.57	14.15	Holmes-VAD [88]	Vad-R1	5.92	80.87	13.21
Vad-R1	Holmes-VAU [89]	78.54	8.68	12.79	Holmes-VAU [89]	Vad-R1	8.45	79.45	12.10
Vad-R1	HAWK [54]	77.45	4.33	18.22	HAWK [54]	Vad-R1	4.10	77.90	18.00

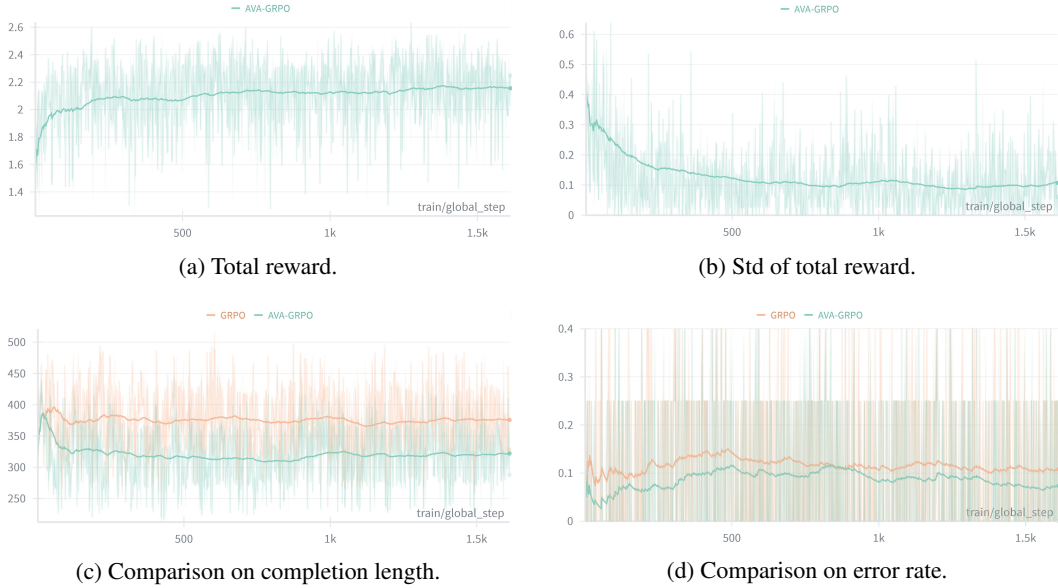


Figure 11: RL training curves of Vad-R1.

provide more useful visual evidence. On the other hand, the benefit of a higher resolution depends on the number of input frames. When increasing the max number of pixels to $256 \times 28 \times 28$ with 16 frames, the model gains small but consistent performance improvement, suggesting that high resolution details compensate for the short clip. In contrast, the performance will drop if we increase the max pixels for 32 frames, possibly due to token redundancy. Consequently, increasing frames is more useful, whereas higher resolution might lead to information overload.

D.3 Training Curves

Figure 11 demonstrates the key training curves of Vad-R1 during RL stage. Figure 11(a) shows the total reward of AVA-GRPO, which increases steadily and converges after approximately 1000 steps, indicating consistent improvement in the degree of matching policy for the output of Vad-R1. Figure 11(b) illustrates the standard deviation of total reward, which decreases rapidly in the early stage and stabilizes below 0.1, suggesting that the output quality of Vad-R1 gradually improves as the training progresses. Figure 11(c) compares the completion length between GRPO and AVA-GRPO.

Table 11: Results of double-right metrics with different embedding models.

Model	Qwen3-embedding				Jina-embedding-V3			
	RR↑	RW↓	WR↓	WW↓	RR↑	RW↓	WR↓	WW↓
<i>Open-Source video MLLMs</i>								
InternVideo2.5 [69]	37.59	25.51	2.05	34.85	41.91	21.18	7.97	28.94
InternVL3 [96]	50.57	27.33	0.23	21.87	56.72	21.18	1.14	20.96
VideoChat-Flash [29]	4.56	63.78	0.00	31.66	30.07	38.27	0.00	31.66
Video-LLaMA3 [85]	49.54	21.10	0.00	29.36	55.35	15.29	0.00	29.36
LLaVA-NeXT-Video [94]	41.23	23.92	0.23	34.62	52.62	12.53	0.91	33.94
Qwen2.5-VL [61]	50.34	25.74	0.00	23.92	56.49	19.59	1.37	22.55
<i>Open-Source video reasoning MLLMs</i>								
Open-R1-Video [67]	26.42	37.59	0.00	35.99	38.95	25.06	0.68	35.31
Video-R1 [14]	11.85	50.57	0.23	37.35	30.75	31.66	6.83	30.76
VideoChat-R1 [30]	48.52	30.75	0.00	20.73	53.99	25.28	1.82	18.91
<i>MLLM-based VAD methods</i>								
Holmes-VAD [88]	10.93	45.56	0.91	42.60	23.05	33.94	1.14	42.37
Holmes-VAU [89]	7.76	27.63	1.14	63.47	20.78	14.61	2.97	61.64
HAWK [54]	5.47	25.97	0.46	68.10	6.15	25.28	1.82	66.75
<i>Proprietary MLLMs</i>								
Claude3.5-Haiku [2]	44.65	36.90	0.00	18.45	51.94	29.61	1.37	17.08
GPT-4o [43]	48.97	22.10	0.00	28.93	56.49	14.58	1.14	27.79
Gemini2.5-Flash [55]	34.02	28.54	0.00	37.44	39.73	22.83	2.05	35.39
<i>Proprietary reasoning MLLMs</i>								
Gemini2.5-Pro [56]	32.65	50.46	0.00	16.89	60.27	22.83	0.00	16.90
QvQ-Max [60]	42.11	28.38	0.00	29.51	50.11	20.37	0.69	28.83
o4-mini [45]	<u>57.37</u>	31.34	0.00	11.29	<u>66.36</u>	22.35	0.23	11.06
Vad-R1 (Ours)	66.29	21.18	0.00	12.53	70.16	17.31	1.14	11.39

Table 12: Performance comparison of different numbers of input frames and spatial resolutions.

Frames	Max Pixels	Anomaly Reasoning			Anomaly Detection				
		BLEU-2	METEOR	ROUGE-2	Acc	F1	mIoU	R@0.3	R@0.5
16	128 × 28 × 28	0.233	0.406	0.194	0.875	0.862	<u>0.713</u>	0.770	<u>0.706</u>
	256 × 28 × 28	0.238	0.412	0.198	0.886	0.878	<u>0.713</u>	<u>0.772</u>	0.702
32	128 × 28 × 28	<u>0.242</u>	<u>0.416</u>	<u>0.201</u>	0.900	<u>0.891</u>	0.726	0.786	0.715
	256 × 28 × 28	0.238	0.413	0.198	0.888	0.883	0.708	0.772	0.695
64	128 × 28 × 28	0.244	0.420	0.203	0.895	0.892	0.709	0.777	0.695

We observe that the reasoning process generated by AVA-GRPO during the training process is shorter in length, presenting more concise and focused responses. Figure 11(d) presents the error rate comparison, where AVA-GRPO shows a clear downward trend and stabilizes at a lower level than vanilla GRPO. This indicates that AVA-GRPO effectively reduces incorrect reasoning and leads to more accurate and reliable responses during training.

D.4 More Qualitative Results

We provide more qualitative results in Figure 13 and Figure 12. Vad-R1 demonstrates stable anomaly reasoning and detection capabilities. On the one hand, Figure 12 shows a normal example in Vad-Reasoning dataset. The video shows a normal scene in campus. Vad-R1 accurately describes the content in the video and identifies its normality. In comparison, GPT-4o and QVQ-Max demonstrate hallucination, pointing out anomalies that do not exist. And although Gemini2.5-Flash considers reverse walking as a potential abnormal event, it does not take into account the specific scenario of pedestrian walkways. On the other hand, Figure 13 shows results on VANE benchmark. When facing unseen videos, Vad-R1 still demonstrates great reasoning capability in complex environments and



Figure 12: Qualitative result for a normal video.

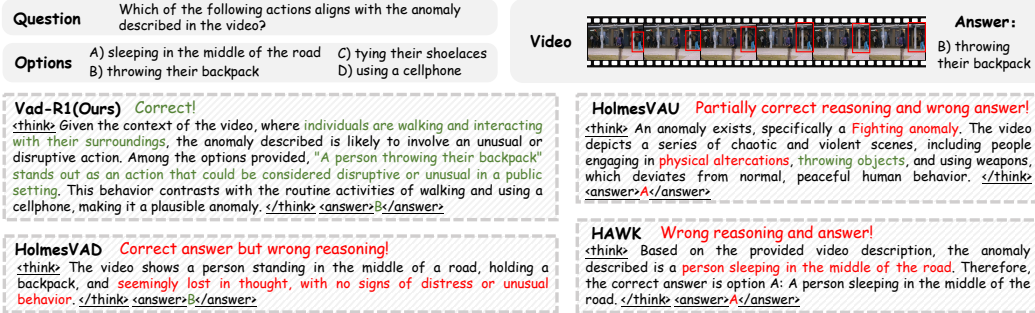


Figure 13: Qualitative performance on VANE benchmark.

correctly identifies anomalies in the video. In comparison, the reasoning process of HolmesVAU is partially correct, resulting in incorrect judgment, while HolmesVAD makes correct judgment but incorrect reasoning process.

E Impact and Limitation

In this paper, we propose a new task: Video Anomaly Reasoning, which enables MLLM to perform deep analysis and further understanding of the anomalies in the video. We hope our work can contribute to the video anomaly researches.

However, the inference speed of Vad-R1 remains a limitation, as the multi-step reasoning process introduces additional computational overhead.