Domain-Adaptive Video Captioning for Surveillance Videos

Fengqi Zhang, Chunrui Hua, Shuaijie Li, Wen Qi

Anker Innovations
{frank.zhang1, jimmy.hua, shuaijie.li, kevin.qi}@anker-in.com

Abstract

Surveillance video captioning aims to generate natural language descriptions for security video and plays a crucial role in applications such as anomaly detection and behavior analysis. While vision-language models (VLMs) perform well on general video captioning, they often struggle in surveillance scenarios due to domain discrepancies in visual style and content emphasis. Moreover, reliable evaluation of caption quality in this domain also remains challenging. We propose a comprehensive framework that addresses both domain adaptation and evaluation. For domain adaptation, we enhance VLMs via reinforcement fine-tuning, exploring multiple reward designs to optimize relevance and accuracy while helping to improve out-of-domain generalization. For evaluation, we develop an LLM-as-a-Judge process that assesses generated descriptions from two complementary perspectives: a local semantic score capturing fine-grained action fidelity and a global quality score evaluating linguistic fluency and temporal coherence. Experimental results show that our domain adaptation approach outperforms both zero-shot and supervised fine-tuning baselines. Ablation studies further validate the effectiveness of our reward design, training strategy, and out-of-domain generalization. In addition, empirical analyses show that the proposed LLM-as-a-Judge framework aligns closely with human evaluations and exhibits stable and robust performance across different conditions.

1 Introduction

Surveillance video captioning aims to generate natural language descriptions for surveillance footage by converting raw visual input into human-readable text. This task is essential in the security domain, where it supports downstream applications such as explainable anomaly detection (Yuan et al. 2024) and behavior analysis. Compared with general video captioning (Abdar et al. 2024; Tang et al. 2025), surveillance videos present distinctive visual characteristics, such as high-angle viewpoints, poor illumination, and lower image quality. These domain-specific characteristics are not well represented in open-domain video datasets, making it difficult for general models to perform effectively in surveillance scenarios. In addition, the primary focus of captioning in surveillance scenarios lies in describing events that are semantically relevant to security applications (Duong, Le, and Hoang 2023), rather than offering general scene-level descriptions. Effective captions should capture subtle activities, object interactions, and contextually important anomalies. Therefore, generating precise and security-aware descriptions for surveillance videos remains an underexplored research problem.

Recent vision-language models (VLMs), such as VILA (Lin et al. 2023) and Qwen-VL (Bai et al. 2025), have achieved remarkable results in bridging visual content with natural language. Pretrained on large-scale multi-modal datasets, these models exhibit strong zero-shot captioning capability across diverse domains. However, applying general-purpose VLMs to surveillance video captioning presents two key challenges. First, as discussed earlier, surveillance videos differ substantially from open-domain video content in terms of camera perspective, illumination, and event focus. As a result, general-purpose VLMs may fail to generate appropriate visual representations for such inputs, leading to inaccurate or incomplete descriptions. Second, evaluating caption quality in surveillance contexts remains an open problem. Traditional metrics such as BLEU (Papineni et al. 2002) primarily assess surface-level lexical overlap and often fail to capture semantic fidelity or task-specific relevance. While human evaluation offers greater accuracy, it is labor-intensive and difficult to scale. These limitations highlight the need for both domain-adaptive modeling and reliable evaluation process tailored to surveillance applications.

To address these challenges, we propose a comprehensive framework for domain-adaptive surveillance video captioning, which consists of two key components: (1) a reinforcement fine-tuning (RFT) approach that adapts VLMs to surveillance data, and (2) an LLM-as-a-Judge (LAJ) evaluation process that assesses caption quality in a structured and interpretable manner. Together, these components aim to improve the relevance, accuracy, and interpretability of generated descriptions in surveillance scenarios. For model adaptation, we adopt Group Relative Policy Optimization (GRPO) (Shao et al. 2024) to align VLMs with the visual characteristics and linguistic patterns of surveillance videos. Compared to standard supervised fine-tuning (SFT), the GRPO-based RFT not only enhances the model's ability to understand surveillance-specific visual and semantic information but also improves its generalization to out-of-domain scenarios. We design and compare several designs of reward functions, including rule-based, similarity-based, and LLM-based criteria. In addition, we adopt a two-stage training pipeline, where SFT first provides a stable initialization and RFT further refines the model. Experimental results demonstrate that our RFT approach outperforms both zero-shot and SFT baselines, showing its effectiveness in improving the domain adaptation capability of VLMs for surveillance video captioning.

On the other side, we explore the use of LLM-as-a-judge (Gu et al. 2024) for assessing the quality of surveillance video captions. Large language models possess extensive world knowledge and strong language understanding capabilities, making them well-suited for evaluating open-ended generative tasks such as video captioning. Unlike traditional metrics that focus on surface-level lexical overlap, LLM-based evaluation can better capture semantic relevance and contextual appropriateness. Moreover, it offers a scalable and cost-efficient alternative to human annotation while maintaining high evaluation quality. However, existing LLM-as-a-judge methods (Hu et al. 2024) are generally designed for open-domain scenarios and fail to capture the specific priorities of surveillance applications. To address this gap, we propose a domain-specific LLM-as-a-judge (LAJ) approach tailored to surveillance video captioning. Our evaluation process is structured around two complementary components: a local semantic score, which measures fine-grained action fidelity, and a global quality score, which evaluates linguistic fluency and temporal coherence. This design ensures that both event-level accuracy and overall narrative consistency are jointly assessed. Empirical results show that our LAJ method achieves high correlation with human evaluation results and maintains stable performance under temperature variations and textual perturbations, demonstrating both accuracy and robustness.

The main contributions of this work are as follows:

- We propose a comprehensive framework for domain-adaptive surveillance video captioning that jointly addresses model adaptation and evaluation.
- We introduce a RFT strategy based on GRPO to adapt vision–language models to the surveillance domain, together with the design and analysis of task-specific reward functions.
- We develop a domain-specific LAJ evaluation framework tailored to surveillance captioning, which assesses captions through two complementary components: a local semantic score and a global quality score.
- We conduct empirical studies demonstrating that our framework achieves improved caption quality, stronger out-of-domain generalization, and higher evaluation reliability, while maintaining robustness and consistency across conditions.

2 Related Work

Video Captioning

Traditional video captioning methods often rely on carefully designed architectures that combine CNN (Wang et al. 2018) or transformer-based (Seo et al. 2022) visual encoders with text decoders. For example, the CM^2 method (Kim et al. 2024) enhances dense video captioning by retrieving relevant textual features from a memory bank and integrating them with multi-scale video features through a transformer encoder-decoder. The MT method (Zhou et al. 2018) leverages temporal convolutional networks and self-attention to perform joint event localization and caption generation in an end-to-end manner. In recent years, vision-language models (VLMs) have advanced video captioning with strong zero-shot generalization. Early efforts such as Video-LLaVA (Lin et al. 2024) extended image-based models to temporal inputs. More recently, Qwen-VL (Bai et al. 2025), InternVL (Chen et al. 2024), and Video-ChatGPT (Maaz et al. 2024) have demonstrated impressive performance across a range of video-language tasks, offering scalable and versatile alternatives to task-specific architectures. Most existing approaches are trained on open-domain datasets like MSR-VTT (Xu et al. 2016), ActivityNet Captions (Krishna et al. 2017), and YouCook2 (Zhou et al. 2018). However, these datasets lack the static views, event sparsity, and security semantics typical of surveillance videos. To address this, newer datasets such as UCA (Yuan et al. 2024) and SmartHome-Bench (Zhao et al. 2025) provide more targeted benchmarks for surveillance-oriented captioning, though they remain underutilized in current research.

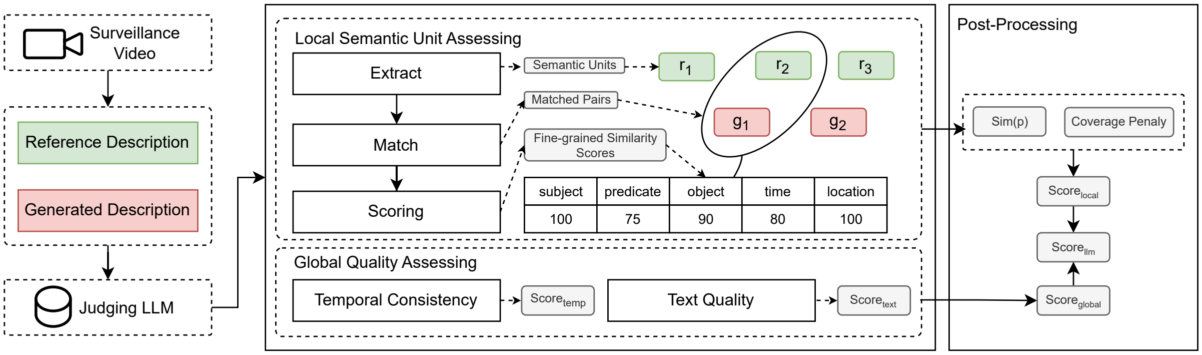
LLM-as-a-judge

Conventional evaluation metrics for open-ended generation tasks, such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE-L (LIN 2004), measure n-gram overlap but lack semantic understanding. Recently, LLM-as-a-judge (Hu et al. 2024) has emerged as a promising alternative, leveraging large language models to provide more human-aligned assessments. A representative work (Hu et al. 2024) proposes multi-dimensional evaluation criteria covering coherence, consistency, fluency, and relevance, and validates its effectiveness through correlation studies and robustness tests under input perturbations. However, applying LLM-based evaluation to surveillance video captioning remains unexplored, as existing frameworks do not account for domain-specific requirements such as temporal precision and event salience.

3 Reinforcement Fine-Tuning

We adopt GRPO (Shao et al. 2024) to enhance VLMs for surveillance video captioning. It is a variant of PPO (Schulman et al. 2017) that eliminates the need for a separate critic model, thereby significantly reducing training costs. Unlike PPO, which requires training an additional value function alongside the policy model, GRPO estimates the baseline from group scores, substantially lowering computational overhead and memory requirements while maintaining training effectiveness.

Formally, GRPO optimizes the policy model by maximizing the following objective:



subject	predicate	object	time	location
100	75	90	80	100

Figure 1: An illustration of the proposed LLM-as-a-Judge framework

$$J(\theta) = E[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(o|q)]$$
$$[\frac{1}{G}\sum_{i=1}^{G} \min(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon,$$
$$1+\epsilon)A_i) - \beta D_{KL}[\pi_\theta || \pi_{ref}]]$$

where π_θ and $\pi_{\theta_{old}}$ are the current and old policy models, A_i is the advantage calculated from relative rewards within each group, and β controls the KL divergence penalty from the reference model π_{ref}. The key innovation of GRPO lies in its group-relative advantage calculation. For each question q, GRPO samples a group of outputs $\{o_1, o_2, \ldots, o_G\}$ and uses the average reward of the group as the baseline, making the advantage computation more stable and aligned with the comparative nature of reward models.

Reward Function Design

The design of reward functions represents a fundamental challenge in reinforcement learning for text generation, as it directly shapes the optimization landscape and determines what behaviors the model learns to exhibit. In the context of surveillance video captioning, this challenge is particularly nuanced due to the domain-specific requirements for temporal precision, event salience, and factual accuracy. To address this complexity, we systematically explore three types of reward mechanisms that capture different aspects of caption quality.

Rule-based Reward We employ BLEU scores to measure lexical overlap between predicted captions and ground truth references, as BLEU is a widely-adopted standard evaluation metric for captioning tasks. The reward aggregates multiple n-gram precisions:

$$R_{rule}(o, r) = \sum_{n=1}^{4} \text{BLEU}_n(o, r).$$

where r denotes the reference description of the given video. This approach is computationally lightweight and provides stable training signals. However, its focus on exact lexical matching may constrain the model's exploration of diverse but semantically equivalent expressions.

Similarity-based Reward To capture semantic similarity beyond surface-level matching, we utilize BERT-based sentence embeddings due to their lightweight nature and effectiveness in similarity assessment. Specifically, we encode both predicted and reference captions using BERT model and compute cosine similarity:

$$R_{sim}(o, r) = \frac{\text{embed}(o) \cdot \text{embed}(r)}{||\text{embed}(o)|| \cdot ||\text{embed}(r)||}.$$

While slightly more computationally intensive than rule-based rewards, this approach enables exploration by evaluating semantic similarity rather than word-level matching, allowing the model to discover diverse yet meaningful expressions.

LLM-based Reward We leverage the LLM-as-a-Judge framework (detailed in the next section) to provide more nuanced evaluation. The LLM evaluates captions across multiple dimensions including accuracy, completeness, and temporal consistency:

$$R_{llm}(o, r) = \text{LAJ}(o, r),$$

where LAJ represents the LLM-as-a-Judge evaluation process introduced in the next section. Although computationally expensive, this approach offers high customization and can provide end-to-end guidance by defining evaluation criteria tailored to surveillance video captioning requirements.

Fine-tuning Pipeline

Similar to other successful RFT applications such as DeepSeek-R1 (Guo et al. 2025), we adopt a two-stage training pipleine to provide better stability and achieve superior performance:

Stage 1: Supervised Fine-tuning (SFT): We first perform supervised fine-tuning on surveillance video captioning data to provide a warm start for the model. This stage helps the model adapt to the domain-specific visual patterns and linguistic requirements of surveillance scenarios.

Stage 2: Reinforcement Fine-Tuning (RFT): Building upon the SFT checkpoint, we apply GRPO with our designed reward functions. This stage further optimizes the model's performance while maintaining stability through the group-relative baseline mechanism.

4 LLM-as-a-Judge

We propose a domain-specific LAJ evaluation process designed for surveillance video captioning. As illustrated in Figure 1, it comprises two complementary components: local semantic score and global quality score, which assess the quality of generated descriptions in a reference-based manner and reference-free manner, respectively.

Local Semantic Score

In this component, the objective is to establish a fine-grained basis by evaluating the semantic consistency between the generated description and the reference description. To achieve this, we first introduce the concept of a **semantic unit**, which serves as the fundamental element of comparison. A semantic unit is defined along five essential dimensions: *subject*, *predicate*, *object*, *time*, and *location*, collectively representing a concrete action described in the description. Each description can thus be decomposed into a set of semantic units, enabling structured and interpretable comparisons. Specifically, given a generated description and a reference description, the process of semantic unit extraction and matching consists of the following three steps:

- **Extraction**: Both the generated description and the reference description are decomposed and extracted into a collection of semantic units respectively: $G = \{g_1, g_2, ...g_m\}, R = \{r_1, r_2, ..., r_n\}$, where each semantic unit corresponds to a distinct action.
- **Matching**: For each semantic unit r_i extracted from the reference description, the objective of this step is to find the semantically matched unit g_j in the generated description. When such a match exists, the two units form a matched pair $p_k = (r_i, g_j)$ for subsequent evaluation, and the collection of all such pairs is denoted as $P \subseteq R \times G$. Note that unmatched units may occur in both sets. Specifically, semantic units from the reference description that remain unmatched indicate missing information in the generated description. While semantic units from the generated description correspond to hallucinated information not present in the reference description.
- **Scoring**: Finally, each matched pair is further assessed using a flexible mechanism that produces fine-grained similarity scores rather than binary judgments. Specifically, for a matched semantic unit pair $p_k \in P$, fine-grained similarity scores are assessed along the five essential dimensions of semantic units:

  $$(s_k^{subject}, s_k^{predicate}, s_k^{object}, s_k^{time}, s_k^{location})$$

  where each fine-grained similarity ranges from 0 to 100. For example, if the subjects are "a man" and "a male", the subject similarity $s^{subject}$ should be high. In contrast, if the predicates are "run quickly" and "walk slowly", the predicate similarity $s^{predicate}$ would be much lower due to the semantic divergence in action and manner.

All above steps are performed by the judeging LLM. The outputs include the extracted semantic unit sets G and R, the collection of matched pairs P, and the corresponding similarity scores for each pair. Then, the outputs are post-processed to derive the local evaluation score $score_{local}$. First, for each matched semantic unit pair, an aggregated similarity score is computed as:

$$sim(p_k) = \sum w^d * s_k^d$$

where s_k^d denotes the fine-grained similarity score along dimension d and w^d is the corresponding weight. In practice, the weights are empirically set to $(0.3, 0.3, 0.2, 0.1, 0.1)$ for the five dimensions, respectively.

Subsequently, a **coverage penalty** mechanism is introduced to account for the effects of missing and hallucinated information. This mechanism incorporates a factor inspired by the definition of F1-score:

$$recall = \frac{|P|}{|R|}, precision = \frac{|P|}{|G|}$$

$$penalty = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Based on the fine-grained similarity scores of semantic units and the penalty factor, the local semantic score $score_{local} \in [0, 100]$ is defined as:

$$score_{local} = penalty \cdot \frac{1}{|P|} \sum_{p_k \in P} sim(p_k) \qquad (1)$$

Global Quality Score

At the macro level, the judging LLM conducts a global evaluation of the generated text from the following two dimensions, with each dimension assigned a score ranging from 0 to 100:

- Temporal consistency $score_{temp}$: it measures the degree to which the sequence of events in the generated text aligns with a logical and coherent temporal progression. For instance, when the sequence of actions is described as *"A person walks to the water cooler, fills a cup, and then drinks from it"*, the temporal ordering is coherent and consistent with real-world logic. In contrast, the sequence *"A person drinks from a cup, fills it with water, and then walks to the cooler"* presents events in an implausible order, thereby reducing the temporal consistency score.
- Text quality $score_{text}$: it measures the intrinsic linguistic merit of the generated text by examining its fluency, grammatical accuracy, and readability, thereby determining whether it adheres to high standards of formal written discourse. For instance, the sentence *"He go to the park and was play basketball happily"* contains grammatical errors and awkward phrasing, thereby reducing the text's fluency and resulting in a lower text quality score.

Based on the scores obtained from the two dimensions above, the global quality score $score_{global} \in [0, 100]$ is defined as follows:

$$score_{global} = \frac{score_{temp} + score_{text}}{2} \qquad (2)$$

Finally, the overall evaluation score for a generated description is computed as a weighted combination of the local semantic score and the global quality score:

$$score_{llm} = w_{local} \cdot score_{local} + w_{global} \cdot score_{global} \quad (3)$$

where $score_{llm} \in [0, 100]$ and the weights are empirically set to $w_{local} = 0.8$ and $w_{global} = 0.2$ in practice.

5 RFT Experiments

Experimental Setup

Dataset. We conduct experiments on the UCF-Crime Annotation (UCA) dataset (Yuan et al. 2024), which comprises 15,677 surveillance videos with a resolution of 320 × 240 pixel, each paired with an annotated natural language descriptions. The dataset covers diverse surveillance scenarios including normal activities and anomalous events, making it well-suited for evaluating domain-adaptive video captioning methods. We follow the standard train/test split. During training and testing, each video is uniformly sampled into 10 frames.

Baseline Models. We evaluate our approach against zero-shot VLMs: *VILA-1.5-3B*[1] and *Qwen2.5-VL-3B*[2] without any fine-tuning.

Training Details. All training is conducted based on the *VILA-1.5-3B* model. For SFT, we use a learning rate of 1×10^{-5} and train for 2 epochs, resulting in the model *VILA-1.5-3B-SFT*. For RFT, we adopt the two-stage training pipeline consisting of 2 epochs of SFT followed by 1 epoch of GRPO-based RFT, resulting in the model *VILA-1.5-3B-RFT*. The group size is set to 8 and the KL penalty coefficient is set to $\beta = 0.2$. For the similarity-based reward, *all-MiniLM-L6-v2*[3] is used as the embedding model. For the LLM-based reward during training, *Claude-3.5-Sonnet-V2* is used as the judging LLM for higher inference efficiency.

Evaluation Metrics. We report: (1) BLEU@1 which measures the lexical overlap, serving as a traditional n-gram–based baseline; and (2) the proposed LAJ evaluation score $score_{llm}$ composed of the local semantic score $score_{local}$ and the global quality score $score_{global}$. In the evaluation process, *Claude-4-Sonnet* is used as the judging LLM. The details of the prompts used by LAJ is shown in A.

Main Results

Table 1 presents the main experimental results on the UCA dataset, comparing our model *VILA-1.5-3B-RFT* with zero-shot and SFT baselines. *VILA-1.5-3B-RFT* achieves the best overall performance, demonstrating strong domain adaptation capability for surveillance video captioning. Compared with the zero-shot baseline, our approach consistently improves performance on every metric, with a particularly large gain in the local semantic score, where our model scores 43.55 compared to 31.78 for the zero-shot VILA model, representing a 37.0% relative improvement. Furthermore, our model also surpasses the SFT approach on both BLEU@1 score and LLM evaluation score, which demonstrates the effectiveness of the proposed RFT approach in enhancing both caption accuracy and contextual understanding in surveillance scenarios.

Table 1: Main results on UCA dataset

Model	BLEU@1	$\text{Score}_{\text{local}}$	$\text{Score}_{\text{global}}$	$\text{Score}_{\text{llm}}$
Qwen2.5-VL-3B	14.83	29.06	64.20	36.09
VILA-1.5-3B	20.16	31.78	66.50	38.72
VILA-1.5-3B-SFT	24.14	41.54	**70.5**	47.33
VILA-1.5-3B-RFT(**ours**)	**27.60**	**43.55**	70.25	**48.89**

Table 2: Results of ablation study on reward function

Reward	BLEU@1	$\text{Score}_{\text{local}}$	$\text{Score}_{\text{global}}$	$\text{Score}_{\text{llm}}$
Rule	26.57	38.39	69.90	44.69
Similarity	23.04	36.22	68.00	42.57
LLM	19.97	37.69	69.55	44.07
Rule + Similarity	**28.44**	**39.81**	**71.30**	**46.11**
Rule + Similarity + LLM	26.99	38.94	68.90	44.93

Ablation Study on Reward Function

To validate the effectiveness of different reward functions in RFT, we conduct a ablation study on different reward configurations, and compare individual reward functions as well as their pairwise and three-way combinations. This ablation study is performed under the RFT-only training strategy. As shown in Table 2, different reward configurations in RFT exhibit clear performance trade-offs. Among single rewards, the Rule-based reward achieves the best performance, while the Similarity-based reward emphasizes semantic alignment but performs lower. Furthermore, the Rule + Similarity combination delivers the highest overall performance outperforming all configurations. This result indicates that jointly optimizing lexical and semantic objectives effectively enhances both local semantic fidelity and global text quality in caption generation. On the other hand, the LLM-only reward achieves comparable global results but does not bring further improvements, although we demonstrate the effectiveness of the LAJ evaluation process in the following section. The reason may be that when LAJ is used as a reward function, it exhibits relatively low reward variance (Razin et al. 2025), which limits its optimization effect. We leave this problem as future work.

Ablation Study on Training Strategy

We conduct an ablation study to analyze the effectiveness of different training strategies. Figure 2 summarizes the performance of three configurations: SFT-only training, RFT-only training, and the two-stage fine-tuning pipeline. Our results show that both SFT-only and RFT-only training improve model performance compared to the baseline. Among them, SFT-only training achieves the highest $score_{llm}$ score

[1] https://huggingface.co/Efficient-Large-Model/VILA1.5-3b

[2] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

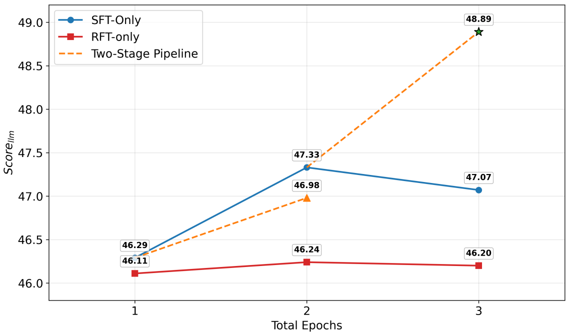


Figure 2: Results of ablation study on training strategy

Table 3: Results of generalization test on COCO Caption dataset

Model / Training	**BLEU@1**
VILA-1.5-3B	24.02
VILA-1.5-3B-SFT	43.45
VILA-1.5-3B-RFT(*ours*)	**47.06**

of 47.33, while RFT-only training performs slightly lower, reaching 46.24. By contrast, the proposed two-stage fine-tuning pipeline consistently yields further improvements, achieving the best overall performance with an $score_{llm}$ score of 48.89.

Generalization Test

To evaluate the generalization ability of our approach, we conduct an out-of-domain test using the COCO Caption dataset. As shown in Table 3, the zero-shot VILA-1.5-3B baseline achieves a BLEU@1 score of 24.02, reflecting limited transferability from surveillance-domain training. After applying SFT, performance improves substantially to 43.45, indicating that supervised fine-tuning enhances general language alignment. Our proposed RFT approach further boosts BLEU@1 to 47.06, achieving the best generalization performance. These results demonstrate that reinforcement fine-tuning not only improves in-domain performance but also enhances the model's robustness and adaptability to unseen domains.

6 Empirical Study of LLM-as-a-Judge

In this section, the default judging model is *Claude-4-Sonnet* with the temperature set to 0.0 unless otherwise specified. The details of the evaluation prompts is shown in A.

Consistency with Human Evaluation

To verify the reliability of the proposed *LLM-as-a-Judge* evaluation, we compare its scores with human judgments on a randomly selected subset of 200 samples.

Human Evaluation Protocol. The human evaluation follows the same structure as the LAJ framework, including the *local semantic score* and the *global quality score*. To reduce labeling effort for local evaluation, we simplify the process by asking annotators to assign a score from 0 to 10 based on two aspects. (1) **completeness**: whether all action information in the reference description is correctly covered in the generated text, and (2) **hallucination**: whether any actions not present in the reference appear in the generation. For global evaluation, annotators rate *temporal consistency* and *text quality* on the same 0–10 scale, following the same definitions as in the LAJ evaluation. The final human score is obtained by aggregating the weighted local and global scores, consistent with the LAJ post-processing procedure. Each generated description is independently evaluated by three annotators, and the averaged score is used as the final human rating.

Table 4: Correlation matrix between LAJ scores and human ratings

	Human Eval	Claude-4-sonnet	Gemini-2.5-Pro	GPT-4.1
Human Eval	1.000	**0.768**	0.715	0.743
Claude-4-Sonnet	0.768	1.000	0.749	0.835
Gemini-2.5-Pro	0.715	0.749	1.000	0.728
GPT-4.1	0.743	0.835	0.728	1.000

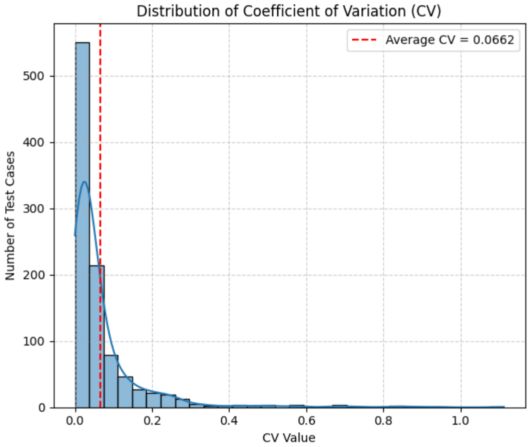


Figure 3: Distribution of score stability across temperature settings

Consistency Results. We compute the Spearman's rank correlation coefficient (SRCC) between human ratings and LAJ scores obtained from different judging LLMs, including Claude 4 Sonnet, GPT-4.1, and Gemini 2.5 Pro. As shown in Table 4, the proposed LAJ evaluation using Claude-4-Sonnet achieves the highest correlation with human ratings (SRCC = 0.768), outperforming both GPT-4.1 and Gemini-2.5-Pro. These results indicate that LAJ can reliably approximate human judgment, exhibiting strong monotonic agreement across evaluation dimensions. Moreover, the relatively high inter-LLM correlations (e.g., 0.835 between Claude 4 Sonnet and GPT-4.1) suggest that the evaluation process is stable and transferable across different LLM backbones.

Stability Analysis on Temperature

A potential concern when employing LLM-as-a-Judge evaluation lies in its sensitivity to randomness induced by the temperature parameter. Higher temperature values increase sampling diversity during generation, potentially leading to inconsistent judgments. Even when the temperature is fixed to zero, minor stochasticity can still arise due to inherent computational non-determinism, making it essential to examine whether such randomness materially affects evaluation reliability.

To quantify this effect, we perform multiple evaluations for each generated description under seven different temperature settings:$\{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. For each case, we compute the Coefficient of Variation (CV) across repeated evaluations, which measures the relative dispersion of scores. A lower CV indicates higher consistency and therefore greater stability of the evaluation process.

As illustrated in Figure 3, most test cases exhibit extremely low variability, with the vast majority of CV values concentrated below 0.1, and an overall average CV of 0.0662. This sharp concentration near zero demonstrates that temperature-induced randomness has a negligible effect on the final evaluation results. Only a small number of outlier cases exceed CV = 0.2, suggesting that the LAJ evaluation process maintains high stability and reproducibility even under varied temperature configurations. Overall, these findings show that LAJ is stable to stochastic sampling effects, ensuring reliable performance across repeated runs.

Robustness to Text Perturbations

To further verify the robustness of the LAJ evaluation process, we analyze whether the evaluation remains consistent when the generated descriptions are perturbed. Specifically, we design two contrasting types of perturbations: semantic-preserving and semantic-altering modifications.

Semantic-Preserving Perturbations. In this setting, we set the generated description as the same as the reference descriptions, and use an LLM to replace words in generated descriptions with synonyms, ensuring that the overall meaning remains unchanged. For example, the sentence *"The cars on the road continue to move forward slowly"* can be rewritten as *"Vehicles on the roadway keep proceeding ahead at a slow pace"*. We then apply LAJ to evaluate the perturbed description. Ideally, semantically equivalent descriptions should receive near-perfect similarity scores (100). Across three independent perturbation trials, the resulting scores are 94.74, 94.79, and 94.82, respectively. These consistently high values indicate that LAJ is largely invariant to superficial lexical substitutions and captures semantic equivalence rather than surface-level word matching.

Table 5: Effect of semantic-altering perturbations

Perturbation Type	$\mathrm{Score}_{\mathrm{llm}}$
No Perturbation	47.06
Subject	21.10
Subject + Predicate	15.18
Subject + Predicate + Object	12.15

Semantic-Altering Perturbations. To assess sensitivity to semantic changes, we perform controlled perturbations on the generated descriptions produced by our *VILA-1.5-3B-RFT* model. Specifically, we modify the core semantic components of each description (i.e., the *subject*, *predicate*, and *object*) to intentionally alter the meaning of the described actions. For example, the subject *"car"* can be replaced with *"dog"*, and the predicate *"move forward"* can be changed to *"run to the park"*, which substantially alters the original semantics of the sentence. In this case, we expect a notable score decrease as semantic fidelity deteriorates. As shown in Table 5, the evaluation score decreases monotonically as the degree of semantic distortion increases, from 47.06 with no perturbation to only 12.15 when all core elements are modified. This consistent downward trend demonstrates that the LAJ evaluation process is capable of accurately distinguishing genuine semantic changes from superficial variations, thereby suggesting its semantic sensitivity and robustness.

7 Conclusion

In this work, we presented a comprehensive framework for domain-adaptive surveillance video captioning that jointly addresses model adaptation and evaluation. For model adaptation, we employed a RFT approach based on GRPO, incorporating multiple reward designs to align caption generation with surveillance semantics while improving out-of-domain generalization. For evaluation, we developed a domain-specific LAJ process that assesses generated descriptions from two complementary perspectives: the local semantic score, capturing fine-grained action fidelity, and the global quality score, evaluating linguistic fluency and temporal coherence.

Our experiments demonstrated the effectiveness of our approach. The proposed RFT method outperformed both zero-shot and SFT baselines, while ablation studies showed the benefits of our reward design and two-stage training strategy. Moreover, the LAJ evaluation framework showed high correlation with human judgments and exhibited stable and robust performance across temperature variations and text perturbations, validating its reliability as an automatic evaluation process for surveillance captioning.

References

Abdar, M.; Kollati, M.; Kuraparthi, S.; Pourpanah, F.; McDuff, D.; Ghavamzadeh, M.; Yan, S.; Mohamed, A.; Khosravi, A.; Cambria, E.; et al. 2024. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Duong, H.-T.; Le, V.-T.; and Hoang, V. T. 2023. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, 23(11): 5024.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, X.; Gao, M.; Hu, S.; Zhang, Y.; Chen, Y.; Xu, T.; and Wan, X. 2024. Are LLM-based Evaluators Confusing NLG Quality Criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9530–9570.

Kim, M.; Kim, H. B.; Moon, J.; Choi, J.; and Kim, S. T. 2024. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13894–13904.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5971–5984.

LIN, C. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004*.

Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; and Han, S. 2023. VILA: On Pre-training for Visual Language Models. arXiv:2312.07533.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12585–12602.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Razin, N.; Wang, Z.; Strauss, H.; Wei, S.; Lee, J. D.; and Arora, S. 2025. What makes a reward model a good teacher? an optimization perspective. *arXiv preprint arXiv:2503.15477*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Seo, P. H.; Nagrani, A.; Arnab, A.; and Schmid, C. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17959–17968.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7622–7631.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; and Jiao, Z. 2024. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22052–22061.

Zhao, X.; Zhang, C.; Guo, P.; Li, W.; Chen, L.; Zhao, C.; and Huang, S. 2025. SmartHome-Bench: A Comprehensive Benchmark for Video Anomaly Detection in Smart Homes Using Multi-Modal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3975–3985.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8739–8748.

Appendix

A LLM-as-a-Judge Prompt

The prompt used for the LLM-as-a-Judge evaluation is shown below. Note that this prompt does not directly produce the final $score_{llm}$; instead, the score is obtained through post-processing according to Equations (1)(2)(3).

LLM-as-a-Judge Prompt

You are an AI evaluation assistant responsible for assessing the similarity between a predicted description and a reference text. Your evaluation must be strictly based on the provided inputs: the predicted description (to be evaluated) and the reference text (ground truth). The evaluation consists of two parts:

Part 1: Local Semantic Unit Assessing
Objective: Extract and match semantic units to assess how well the predicted description captures the key events from the reference text, focusing on intent and outcomes rather than literal, granular actions.
Step 1: Extraction
Extract semantic units from both the reference text and the predicted description. A semantic unit should represent a distinct event, intent, or state.
Rule of Thumb: Focus on Intent over Action. If a series of micro-actions logically combine to serve a single, higher-level purpose, you should extract that single, consolidated purpose.
Examples:
Example 1: Merging actions into a single intent.
Text: ”A person picks up a cup, opens their mouth, and pours water into their mouth.”
Rationale: The individual actions (picking up, opening mouth, pouring) are all sub-components of the single, high-level event ”drinking water”. You must extract the core intent.
Correct Extraction: [”A person drinks water”]
Example 2: Separating distinct intents.
Text: ”A man runs to the bus stop, and then reads a book while waiting.”
Rationale: Running and reading are two distinct core events that are not sub-components of each other. They should be extracted separately.
Correct Extraction: [”A man runs to the bus stop”, ”The man reads a book while waiting”]
Step 2: Matching & Scoring
Matching process: For each reference group, find the most semantically similar group among all description groups. If the best match is sufficiently similar (i.e., highly related in meaning or intent), pair them and proceed to scoring. If no meaningful match is found, mark the reference group as a ”missing group” (indicating under-generation or omission). All unmatched description groups are marked as ”hallucinated groups” (indicating over-generation or fabrication).
Scoring rules: Your task is not to calculate the final score, but to provide detailed dimension scores for each matched group pair. Final calculation will be handled by code. For each matched group pair, evaluate their similarity across the following five dimensions: subject, predicate, object, time, and location.
Dimension scoring criteria: For each dimension, provide a score_rate from 0 to 100, where 100 means a perfect match and 0 means completely mismatched. If a dimension is naturally absent in both the reference and description groups (e.g., no location information), set its value to null in the output so that subsequent code can ignore it.
Note on Semantic Equivalence: Use common sense when matching. A description of a series of actions that logically result in a single event should be considered a strong match for that event. For instance, if the reference group is ”a person drinks water” and the prediction group (after extraction) is ”a person sips from a bottle”, these are a very strong match. Similarly, if the extraction resulted in ”a person pours liquid into their mouth”, this is also a very strong match for ”drinking water”. Be strict with core elements only when the overall intent differs. In your explanation, you must clarify this conceptual link.

Part 2: Global Quality Assessing
Objective: Directly score two dimensions without detailed extraction.
Dimension A: Temporal Consistency (100 points) Assess whether the sequence of events in the video description logically matches the reference text (e.g., if event A occurs before event B in the reference, the description should not reverse them).
Scoring: 100: Completely consistent; no errors in event order. 70 - 90: Slight inconsistencies with minimal impact. 40 - 60: Moderate inconsistencies affecting coherence. 0 - 30: Severe inconsistencies or illogical order.
Provide a brief explanation (1-2 sentences) for the score.
Dimension B: Text Quality (100 points)
Independently assess the language quality of the video description (do not compare with the reference). Focus on:
Fluency: Natural expression and readability. Grammar: Correct syntax and sentence structure. Spelling: No typos or spelling errors. Semantics: No ambiguous or meaningless phrases.
Scoring: 100: Perfect; professional, error-free language. 70 - 90: Good; minor issues that do not affect understanding. 40 - 60: Fair; noticeable errors affecting clarity. 0 - 30: Poor; frequent errors or nonsensical text.
Provide a brief explanation (1-2 sentences) for the score.

Present the results as a structured JSON object. For transparency, include all details. JSON structure:
JSON

```
{{
    "part1": {{
        "reference_groups": [Extracted semantic units from the reference text,
            each as {{"id": int, "info": "..."}}],
        "description_groups": [Extracted semantic units from the description,
            same format],
        "matching_groups": [
            {{
                "reference_id": int,
                "description_id": int,
                "dimension_scores": {{
                    "subject": {{ "score_rate": int (0-100) or null }},
                    "predicate": {{ "score_rate": int (0-100) or null }},
                    "object": {{ "score_rate": int (0-100) or null }},
                    "time": {{ "score_rate": int (0-100) or null }},
                    "location": {{ "score_rate": int (0-100) or null }},
                    "explanation": "Reason for the scores, noting any conceptual
                        equivalence."
                }}
            }}
        ],
    }},
    "part2": {{
        "temporal_consistency": {{
            "score": int (0 - 100),
            "explanation": "string"
        }},
        "text_quality": {{
            "score": int (0 - 100),
            "explanation": "string"
        }}
    }}
}}
```

Note: The key ”info” in the JSON structure is kept for consistency with your original format, but it now represents a ”semantic unit”.

Execution Instructions

Always evaluate in English, even if the input text is in another language (translate if necessary). Explanations should be objective, consistent, and concise. If either the video description or reference text is missing, return an error. Now, please evaluate the similarity between the provided video description and reference text.

Prediction description: {prediction}

Ground Truth description: {ground_truth}