THINKING WITH CAMERAS: ACTIVE VISUAL REASONING VIA DYNAMIC VIEWPOINT CONTROL FOR SURVEILLANCE VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have recently achieved remarkable progress in general-purpose video understanding. However, their application to real-world surveillance remains challenging due to the lack of large-scale domain-specific datasets and the limitation of passive observation from fixed viewpoints. In surveillance scenarios, critical visual evidence can be easily missed when targets are distant, small, occluded, or move beyond the current camera view. In this work, we introduce CamVLM, a new framework for Thinking with Cameras, which enables LVLMs to actively acquire visual evidence in real-world surveillance by continuously controlling camera viewpoints. We first construct CCTV-Anomaly, a large-scale surveillance video understanding dataset containing 14,133 videos across 10 anomaly categories, with detailed captions and event annotations. We further formulate viewpoint control as an active visual perception problem and build CamTrack-53K, an object-centric viewpoint trajectory dataset for learning camera actions. Moreover, we propose a reinforcement learning based viewpoint policy optimization framework, which models camera control as a sequential decision-making process and learns long-horizon observation strategies beyond supervised trajectory imitation. Extensive experiments demonstrate that CamVLM achieves state-of-the-art performance under both passive observation and dynamic viewpoint settings, validating the effectiveness of active camera-based reasoning for surveillance video understanding. We will release the datasets, models, and code to facilitate future research.

1 INTRODUCTION

Surveillance cameras are widely deployed in real-world applications, including anomaly detection, behavior analysis, and traffic perception. Unlike general videos collected for human viewing, streaming surveillance videos usually cover wide fields of view and contain multiple moving and interacting objects. Due to long capture distances, target objects often occupy only small regions of the entire frame and may gradually move beyond the current camera viewpoint, leading to incomplete and unreliable visual evidence. Therefore, effective surveillance video understanding requires not only recognizing what happens in a video, but also actively determining where to observe next by continuously adjusting the camera viewpoint to acquire informative visual evidence.

Recent advances in large vision-language models (LVLMs) have significantly improved visual perception and reasoning. Early video understanding approaches (Feng et al., 2025; Li et al., 2025; Wang et al., 2025a) typically encode visual contents globally and perform subsequent language reasoning, which often overlooks fine-grained visual evidence. Inspired by OpenAI-o3 (OpenAI, 2025), recent "thinking with images" approaches (Khayatkhoei et al., 2025; Zheng et al., 2026; Shen et al., 2025) introduce iterative evidence acquisition, where models actively retrieve informative regions and interleave visual perception with reasoning. Extending this idea, recent "thinking with videos" methods (Yang et al., 2026; Zhang et al., 2026; Zeng et al., 2026) further explore dynamic visual evidence acquisition from videos through iterative frame sampling or tool-based interaction. However, existing thinking-based approaches still operate on fixed visual streams. They can decide which frame or clip to inspect, but cannot determine how the camera itself should move to acquire better observations. This limitation is particularly critical in real-world surveillance sce-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

narios, where objects are often small, partially occluded, or follow unpredictable trajectories. As shown in Figure 1(a), once important objects leave the current viewpoint, passive video retrieval strategies cannot recover the missing fine-grained visual evidence.

To address this challenge, we propose Thinking with Cameras, a new active visual reasoning paradigm that enables LVLMs to continuously reason through camera actions in real-world surveillance. However, training LVLMs to control real-world cameras is hindered by the lack of large-scale viewpoint-action trajectories collected from physical surveillance systems. To overcome this limitation, we introduce a virtual camera formulation by treating each full video frame as a panoramic observation space and a local image region as the camera's current field of view. Specifically, the model starts from an initial viewpoint within the frame and learns to perform camera-like operations, including panning and zooming, to acquire informative visual evidence. Unlike treating surveillance videos as immutable visual streams, our formulation transforms visual understanding into an active perception process, where the model dynamically determines where to observe next. Unlike conventional visual reasoning, dynamic viewpoint control is inherently a sequential decision-making problem, since each camera action changes future observations and influences subsequent reasoning outcomes. Therefore, we further introduce a viewpoint policy optimization framework based on reinforcement learning, which enables LVLMs to learn long-horizon camera control strategies beyond supervised trajectory imitation. By jointly optimizing task completion and visual evidence acquisition, the learned policy can explore more effective viewpoint trajectories and continuously maintain informative objects within the field of view. As shown in Figure 2, the learned camera control policy exhibits adaptive viewpoint selection and on-demand camera actions, suggesting that efficient surveillance understanding benefits from active and evidence-driven viewpoint reasoning.



Figure 1: Motivation of CamVLM. The full frame represents the virtual panoramic space, while the bounding box denotes the camera viewpoint. (a) Passive viewpoints may miss critical cues. (b,c) Under dynamic viewpoint control, existing methods such as Video-o3 (Zeng et al., 2026) exhibit unstable visual focus, whereas CamVLM reliably captures fine-grained visual evidence.

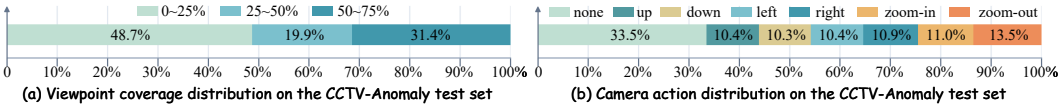


Figure 2: Behavior statistics of the learned camera control policy on the CCTV-Anomaly test set. (a) Distribution of selected viewpoint coverage. Nearly half of the viewpoints remain within 25% of the full frame, indicating that the model avoids unnecessary zoom-out and adaptively adjusts the viewing scale according to scene context. (b) Distribution of camera control actions. The high proportion of none actions and the balanced directional movements and zooming operations demonstrate an efficient on-demand control strategy without directional or zooming bias.

To support this paradigm, we construct two large-scale datasets. First, we introduce CCTV-Anomaly, a high-quality surveillance video understanding dataset containing 14,133 videos across 10 anomaly categories and 234 hours, with detailed captions and event category annotations. Second, we build CamTrack-53K, an object-centric viewpoint trajectory dataset for learning dynamic camera control. Based on these datasets and the proposed viewpoint policy optimization framework, we develop CamVLM, which actively controls camera viewpoints during reasoning and captures comprehensive fine-grained object-level spatiotemporal evidence, as illustrated in Figure 1(b,c).

Overall, our contributions are summarized as follows:

- We propose Thinking with Cameras, a new active visual reasoning paradigm that enables LVLMs to continuously control camera viewpoints and actively acquire visual evidence in real-world surveillance.
- We formulate dynamic viewpoint control as a sequential decision-making problem and introduce a reinforcement learning based viewpoint policy optimization framework, which learns long-horizon camera control strategies by jointly optimizing reasoning accuracy and visual evidence acquisition.
- We construct CCTV-Anomaly and CamTrack-53K, two large-scale datasets for surveillance video understanding and dynamic viewpoint control, respectively.
- We develop CamVLM, an LVLM with dynamic viewpoint control capability that actively determines where to observe next and achieves more reliable fine-grained surveillance video understanding.

2 Related Work

Thinking with Images and Videos. Recent advances in LVLM reasoning have explored iterative evidence acquisition beyond conventional text-centric reasoning. Inspired by OpenAI-o3 (OpenAI, 2025), "thinking with images" methods (Lai et al., 2026; Khayatkhoei et al., 2025; Zheng et al., 2026; Shen et al., 2025) interleave visual perception with language reasoning by repeatedly retrieving fine-grained image evidence, improving detailed visual understanding. Extending this idea to videos, recent "thinking with videos" approaches (Fu et al., 2025; Yang et al., 2026; Zhang et al., 2026; Zeng et al., 2026) enable LVLMs to iteratively access informative frames or clips through tool use and dynamic perception. However, these methods still operate on fixed visual streams and focus on selecting existing visual evidence rather than actively controlling the camera viewpoint, making them insufficient for real-world surveillance scenarios where targets may be distant, occluded, or leave the current field of view.

Surveillance Video Datasets. Existing surveillance video datasets mainly focus on anomaly detection, such as Subway Entrance, Subway Exit (Adam et al., 2008), UCSDPed1/2 (Li et al., 2014), Avenue (Lu et al., 2013), and UCF-Crime (Sultani et al., 2018). However, these datasets are limited in scale, realism, or annotation richness, providing insufficient supervision for modern LVLM-based surveillance understanding. Recent UDVideoQA (Vishal et al., 2026) explores surveillance video question answering in traffic scenarios but contains limited video duration and domain coverage. In contrast, our CCTV-Anomaly provides 14,133 high-quality surveillance videos across diverse anomaly categories, with both detailed captions and event-level annotations to support comprehensive surveillance video understanding.

Active Visual Perception. Active perception aims to improve visual understanding by dynamically adjusting sensing strategies according to task requirements. Existing works mainly focus on robotic perception (Zhang et al., 2025; Liu et al., 2026), object tracking (Wang et al., 2025b; Sun et al., 2025), and embodied agents (Qin et al., 2024; Liu et al., 2025), where sensors or cameras are controlled through predefined policies or reinforcement learning strategies. However, applying active perception to LVLM-based surveillance understanding remains largely unexplored. Different from existing approaches that passively analyze captured videos, we introduce a "thinking with cameras" paradigm, where LVLMs actively control viewpoints and learn long-horizon camera policies through reinforcement learning to acquire task-relevant visual cues.

3 Method

3.1 Overview

As illustrated in Figure 3, we propose CamVLM, a vision-language model that actively controls camera viewpoints to acquire visual evidence in real-world surveillance. Different from conventional LVLMs that passively analyze fixed visual streams, CamVLM treats viewpoint adjustment as

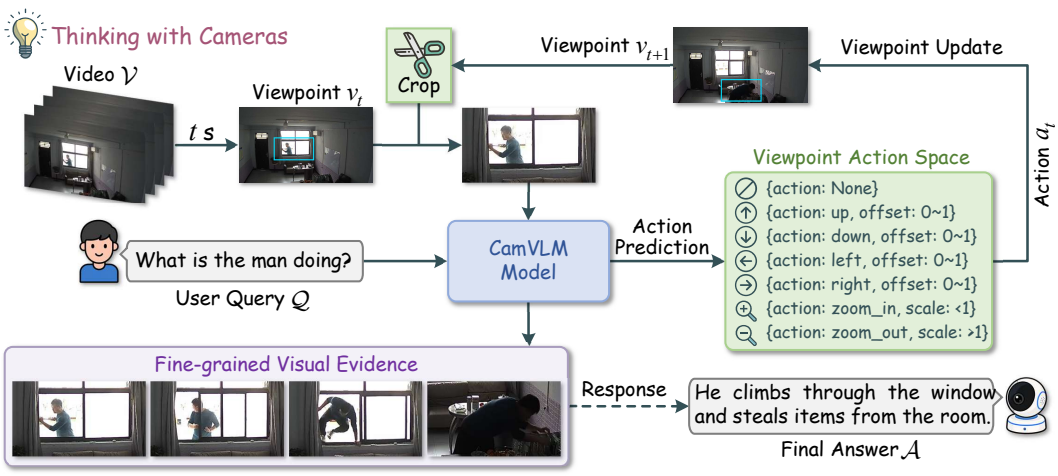


Figure 3: Overview of CamVLM. Given a surveillance video and a user query, CamVLM iteratively predicts viewpoint actions based on the positions of task-relevant objects, updates the viewpoint to keep them clearly visible, accumulates fine-grained visual evidence, and produces the final answer.

an integral part of the reasoning process and actively acquires informative visual evidence through camera actions. Given a video $\mathcal{V}$ and a user query $\mathcal{Q}$, the model maintains a dynamic viewpoint state v_t at each timestep t. Based on the current visual observation and task requirements, CamVLM predicts a viewpoint action a_t to determine how the camera should move or zoom. The executed action updates the viewpoint and provides a new observation v_{t+1}, enabling the model to progressively collect fine-grained evidence. After multiple interaction steps, the accumulated visual evidence is integrated to generate the final response $\mathcal{A}$. Formally, the viewpoint interaction process can be represented as:

$$a_t = \pi_\theta(\mathcal{Q}, v_t), \qquad v_{t+1} = \mathcal{T}(v_t, a_t), \tag{1}$$

where π_θ denotes the viewpoint policy and $\mathcal{T}$ represents the viewpoint transformation operation. The complete interaction trajectory is defined as:

$$\tau = (\mathcal{Q}, v_0, a_0, v_1, a_1, \ldots, v_{T-1}, a_{T-1}, \mathcal{A}). \tag{2}$$

The following subsections introduce the proposed dynamic viewpoint control mechanism, the construction of CamTrack-53K for learning viewpoint policies, and the training strategy including reinforcement learning based viewpoint policy optimization.

3.2 Dynamic Viewpoint Control

Dynamic viewpoint control aims to enable LVLMs to actively determine where to observe next in real-world surveillance in order to acquire task-relevant visual evidence. Ideally, such capability requires large-scale interaction data collected from physical cameras operating in real-world environments, which is difficult to obtain. Therefore, we simulate camera control within surveillance videos by treating each full-frame image as a virtual panoramic space and a local region as the current camera observation. For each input video frame, the central $1/9$ region is initially selected as the camera viewpoint v_0. This setting provides a suitable balance between observation coverage and control necessity: larger viewpoints reduce the need for active adjustment, while smaller ones may miss sufficient target information.

During inference, the model interacts with the video at one-second intervals. At timestep t, given the current viewpoint v_t and query $\mathcal{Q}$, the viewpoint policy predicts an action a_t according to the spatial state of task-relevant objects and the current visual evidence. The action space contains both spatial movement and scale adjustment operations, allowing the model to actively explore the surrounding visual space. After executing the predicted action, the viewpoint is updated through the transformation function:

$$v_{t+1} = \mathcal{T}(v_t, a_t). \tag{3}$$

Through iterative viewpoint adjustment, CamVLM can continuously acquire complementary observations in real-world surveillance when objects become distant, partially visible, or leave the initial field of view. Unlike passive video sampling methods, the proposed mechanism enables the model to actively decide which visual evidence to acquire before producing the final answer.

3.3 CamTrack-53K: Viewpoint Trajectory Generation

Training dynamic viewpoint policies requires large-scale supervision of camera-object interactions. To this end, we construct CamTrack-53K, an object-centric viewpoint trajectory dataset for learning active camera control. Specifically, we collect videos from A2D (Xu et al., 2015), MeViS (Ding et al., 2023), and Refer-YouTube-VOS (Seo et al., 2020) in VideoRefer-700K (Yuan et al., 2025). These datasets provide temporally aligned object masks and object-level question-answer annotations, enabling the construction of task-oriented viewpoint trajectories. For each video, object masks are first converted into bounding boxes. The viewpoint action space is defined as: $\{\texttt{none}, \texttt{up}, \texttt{down}, \texttt{left}, \texttt{right}, \texttt{zoom_in}, \texttt{zoom_out}\}$.

Panning actions are parameterized by offsets relative to the full frame, while zooming actions are represented by scale changes relative to the current viewpoint. Starting from the initial central $1/9$ viewpoint, we generate trajectories at one-second intervals based on object-viewpoint relationships:

- When the target object is outside the current viewpoint, the camera performs zoom-out and viewpoint adjustment actions to reacquire the object.
- When the object is visible but deviates from the optimal observation region, the camera performs panning and zooming actions to obtain a more informative view.
- When the object is already well observed, the camera maintains the current viewpoint.

The generated trajectories provide an initial policy prior for supervised learning, while subsequent reinforcement learning further refines the viewpoint policy beyond predefined trajectory patterns.

3.4 Training Pipeline

Supervised Fine-Tuning. We adopt a mixed supervised fine-tuning (SFT) strategy on the constructed CCTV-Anomaly and CamTrack-53K datasets to jointly enhance the model's surveillance video understanding and dynamic viewpoint control capabilities. For CCTV-Anomaly, the model is trained under the passive full-frame observation setting, where only the final event classification and video captioning responses are supervised. For CamTrack-53K, we follow the dynamic viewpoint interaction process described in Section 3.2. Specifically, the model is supervised to predict both the viewpoint control actions and the final reasoning response according to the generated viewpoint trajectories. The action space follows the definition in Section 3.3. When multiple viewpoint adjustments are required, the model predicts an ordered action sequence, such as $\{\texttt{action}:\texttt{up}, \texttt{offset}:\texttt{0.1}; \texttt{action}:\texttt{zoom_in}, \texttt{scale}:\texttt{0.8}\}$.

Viewpoint Policy Optimization. Dynamic viewpoint control is inherently a sequential decision-making problem. Each camera action determines the visual observations available at subsequent reasoning steps, making its quality dependent on long-term visual evidence acquisition rather than immediate supervision. Although supervised fine-tuning enables the model to imitate expert viewpoint trajectories, it cannot optimize viewpoint policies for downstream understanding, since multiple camera trajectories may lead to equally informative observations. To address this limitation, we further optimize CamVLM on CamTrack-53K using reinforcement learning (RL) with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), allowing the model to actively explore more effective viewpoint control strategies.

Unlike conventional reinforcement learning for large language models, where each sampled response consists only of textual outputs, each response in CamVLM contains an entire camera-control trajectory together with the final reasoning result. Therefore, policy optimization jointly evaluates both visual exploration behaviors and downstream reasoning performance. Specifically, given an input video $\mathcal{V}$ and a user query $\mathcal{Q}$, we follow the interaction process described in Section 3.2 to generate a complete camera-control trajectory

$$\tau = \{a_1, a_2, \cdots, a_T, y\}, \tag{4}$$

where a_t denotes the viewpoint action at time step t, and y is the final answer.

For each training sample $(\mathcal{V}, \mathcal{Q}) \in \mathcal{D}$, we sample a group of G trajectories $\{o_1, o_2, \cdots, o_G\}$ from the policy model π_θ, compute their corresponding rewards $\{r_1, r_2, \cdots, r_G\}$, and optimize the policy using GRPO:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{(\mathcal{V},\mathcal{Q})\sim\mathcal{D}}\mathbb{E}_{o_i\sim\pi_\theta}\left[A_i \log \pi_\theta(o_i|\mathcal{V}, \mathcal{Q})\right], \tag{5}$$

where the relative advantage is computed as

$$A_i = \frac{r_i - \text{mean}(\{r_1, \cdots, r_G\})}{\text{std}(\{r_1, \cdots, r_G\})}. \tag{6}$$

To jointly optimize surveillance understanding and active visual perception, we design a task-oriented reward composed of a task completion reward and a visual evidence reward:

$$\mathcal{R} = \mathcal{R}_{\text{task}} + \lambda\mathcal{R}_{\text{visual}}, \tag{7}$$

where λ balances the two objectives. The task completion reward evaluates the correctness of the final result. For multiple-choice question answering, a correct prediction receives 1, while an incorrect prediction receives 0:

$$\mathcal{R}_{\text{task}} = \begin{cases} 1, & \text{if the predicted answer is correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The visual evidence reward evaluates the quality of the observations acquired by the predicted camera policy. Instead of directly encouraging imitation of reference trajectories, it measures how well the predicted viewpoints continuously cover informative target objects throughout the interaction:

$$\mathcal{R}_{\text{visual}} = \frac{1}{T}\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\frac{|v_{t+1} \cap b_{t+1}^{(i)}|}{|b_{t+1}^{(i)}|}, \tag{9}$$

where T is the number of interaction steps, N is the number of target objects, v_{t+1} is the predicted camera viewpoint at the next step, and $b_{t+1}^{(i)}$ is the ground-truth bounding box of the i-th object.

By rewarding observation quality rather than exact camera trajectories, the proposed policy optimization allows multiple valid viewpoint strategies to emerge during training. Consequently, CamVLM learns to actively determine where to observe next, continuously maintain informative objects within the field of view, and acquire richer visual evidence for downstream reasoning.

4 CCTV-Anomaly Dataset

Existing surveillance anomaly datasets mainly provide coarse event labels, which are insufficient for fine-grained LVLM reasoning. To address this limitation, we construct CCTV-Anomaly, a large-scale surveillance video understanding dataset containing 14,133 real-world videos across 10 event categories. Each video is annotated with both an event category and a detailed natural language caption, enabling joint evaluation of event classification and video captioning. Table 1 compares CCTV-Anomaly with existing datasets, and Figure 4 presents representative examples.

Table 1: Comparison of CCTV-Anomaly with existing surveillance anomaly video datasets.

Dataset	#Videos	Duration	Caption
Subway Entrance	1	1.5 h	✗
Subway Exit	1	1.5 h	✗
UCSD Ped1	70	5 min	✗
UCSD Ped2	28	5 min	✗
Avenue	37	30 min	✗
UCF-Crime	1,900	128 h	✗
CCTV-Anomaly	**14,133**	**234 h**	✓

Dataset Construction and Split. CCTV-Anomaly is built from both existing public surveillance datasets and newly collected real-world surveillance videos, covering diverse environments and viewpoints. Each video is annotated using an LLM-assisted pipeline followed by human verification. In addition to event categories and captions, we further annotate the test set with structured semantic labels, including location, time of day, subjects, activities, and objects, to support fine-grained evaluation. More details are provided in the Appendix C.

The dataset is divided into 13,133 training videos and 1,000 testing videos, with 100 testing videos for each event category. CCTV-Anomaly supports two benchmark tasks: event classification and video captioning. Detailed evaluation protocols are described in Section 5.



Figure 4: Examples of different anomalies in our CCTV-Anomaly dataset.

5 Experiments

5.1 Experimental Setup

Implementation Details. We adopt Qwen3-VL-8B-Instruct (Bai et al., 2025) as the base LVLM. During training and evaluation, videos are uniformly sampled at 2 FPS, with a maximum of 14,336 video tokens and 448 frames. The visual encoder is frozen, while all remaining parameters are optimized. For supervised fine-tuning, we use a global batch size of 128 with a learning rate of 1×10^{-5}. For reinforcement learning, we employ GRPO-based viewpoint policy optimization with a batch size of 8 and 8 sampled trajectories per prompt. The learning rate is set to 1×10^{-6}, the KL coefficient β is set to 0, and the viewpoint alignment reward weight λ is set to 0.6. All experiments are conducted on 8 NVIDIA A800 GPUs.

Benchmarks and Evaluation Metrics. We evaluate surveillance video understanding on two benchmarks: CCTV-Anomaly and UDVideoQA (Vishal et al., 2026). For CCTV-Anomaly, we report event classification accuracy and video captioning accuracy. Classification accuracy is obtained by extracting the predicted category using regular expressions and comparing it with the ground-truth label. Captioning accuracy is evaluated over five semantic components (location, time of day, subjects, activities, and objects), each contributing 20 points to a total score of 100. We use GPT-5.5 (OpenAI, 2026a) to semantically evaluate whether the generated caption includes each ground-truth component, followed by manual verification, and report the average score over the test set. For UDVideoQA, since the official test split is unavailable, we construct UDVideoQA* with permission from the original authors using Set 03 and Set 20 from the released dataset, which provide the greatest question diversity. The resulting benchmark contains 1,190 QA pairs. Following the official evaluation protocol, we report accuracy across five question types and the weighted overall accuracy, where more complex reasoning types receive larger weights. The weights are 1.0 for Basic Understanding (BU), 1.2 for Attribution (Atr), 1.3 for Event Reasoning (ER) and Reverse Reasoning (RR), and 1.5 for Counterfactual Inference (CI).

Evaluation Settings. To evaluate the effectiveness of active camera reasoning, we consider two complementary settings: passive viewpoint understanding and dynamic viewpoint control.

Passive Viewpoint Setting. The model receives the complete video frames or a fixed viewpoint without any camera adjustment. This setting evaluates conventional surveillance video understanding under passive observation.

Dynamic Viewpoint Setting. The model starts from the central $1/9$ region of each frame and actively adjusts the viewpoint through predicted camera actions. Only the visual regions acquired through the dynamic viewpoint trajectory are provided for final answer generation. This setting evaluates whether the model can actively acquire informative visual evidence and improve surveillance video understanding through camera control.

5.2 Experiment Results

Results on CCTV-Anomaly. Table 2 compares surveillance video understanding on CCTV-Anomaly under full-frame observation. CamVLM achieves the best performance on both event classification and video captioning, demonstrating strong surveillance understanding capability. Under dynamic viewpoints, CamVLM outperforms passive-viewpoint methods on both tasks (Table 3), highlighting the effectiveness of active viewpoint adjustment. Compared with the passive baseline, dynamic viewpoint control improves event classification by 33.8%, while CamVLM achieves further gains by capturing fine-grained object-centric spatiotemporal evidence.

Table 2: Performance comparisons on CCTV-Anomaly under the conventional full-frame passive observation setting. The best results are in bold.

Methods	Sizes	CCTV-Anomaly	
		Cls Avg	Cap Avg
Gemini-3.1-Flash-Lite-Preview (Google, 2026)	-	59.3	63.0
GPT-5.6-Luna (OpenAI, 2026b)	-	57.6	61.8
Doubao-Seed-2.0-Lite (Seed, 2026)	-	56.5	65.4
Video-R1 (Feng et al., 2025)	7B	43.9	54.0
VideoZoomer (Ding et al., 2026)	7B	49.0	53.2
MiMo-VL (Yue et al., 2025)	7B	52.8	57.5
Video-o3 (Zeng et al., 2026)	7B	53.2	60.0
LLaVA-OV-2 (An et al., 2026)	8B	45.3	66.6
Qwen3-VL (Bai et al., 2025)	8B	46.0	61.8
CamVLM (Ours)	8B	**75.6**	**76.3**

Table 3: Performance comparisons on CCTV-Anomaly and UDVideoQA* under passive and dynamic viewpoint settings. The terms "Passive" and "Dynamic" denote a passive viewpoint fixed to the central $1/9$ region and a dynamic viewpoint initialized from the same region and adjusted over time, respectively. The best results are in bold.

Methods	Sizes	Settings	CCTV-Anomaly		UDVideoQA*					
			Cls Avg	Cap Avg	BU	Atr	ER	RR	CI	Avg
Gemini-3.1-Flash-Lite-Preview (Google, 2026)	-	Passive	45.0	52.3	61.3	17.6	44.1	67.2	79.0	54.9
GPT-5.6-Luna (OpenAI, 2026b)	-	Passive	42.4	51.4	56.3	13.9	38.7	59.7	32.8	39.7
Doubao-Seed-2.0-Lite (Seed, 2026)	-	Passive	37.2	51.0	55.9	21.0	48.3	51.3	66.8	49.3
Video-R1 (Feng et al., 2025)	7B	Passive	37.0	46.1	52.5	11.8	36.6	41.2	69.8	43.3
VideoZoomer (Ding et al., 2026)	7B	Passive	37.8	46.7	56.3	10.1	35.3	29.0	63.9	39.3
MiMo-VL (Yue et al., 2025)	7B	Passive	38.8	47.1	55.5	10.5	39.9	52.9	51.3	42.2
Video-o3 (Zeng et al., 2026)	7B	Passive	41.4	50.2	56.7	15.5	42.0	39.9	56.3	42.3
LLaVA-OneVision-2 (An et al., 2026)	8B	Passive	37.3	55.6	57.6	14.3	40.8	55.9	45.0	42.5
Qwen3-VL (Bai et al., 2025)	8B	Passive	36.7	51.3	57.1	10.5	41.2	46.2	51.3	41.3
Qwen3-VL (Bai et al., 2025)	8B	Dynamic	44.0	57.2	67.2	17.2	44.5	51.3	51.3	45.9
CamVLM (Ours)	8B	Dynamic	**70.5**	**69.7**	**70.6**	**29.4**	**46.2**	**64.7**	**72.7**	**57.0**

Generalization Results on UDVideoQA*. Table 3 reports zero-shot results on UDVideoQA*. CamVLM achieves the best performance across all question types under dynamic viewpoints. Compared with passive observation, dynamic viewpoint control improves the baseline by 15.7% overall, demonstrating strong generalization to unseen surveillance scenarios.

5.3 Ablation Studies

To validate the effectiveness of our design, we conduct comprehensive ablation studies on UDVideoQA*. Unless otherwise specified, passive viewpoints are fixed to the central $1/9$ region, while dynamic viewpoints are initialized from the same region. All models use the 8B configuration.

Effect of Dynamic Viewpoint Control. Table 4 compares passive and dynamic viewpoint settings. Dynamic viewpoints improve both the baseline model and CamVLM, with substantially larger gains for CamVLM. This demonstrates that viewpoint control effectively directs the model to informative regions, while CamVLM better exploits the resulting object-centric spatiotemporal evidence. We further evaluate CamVLM with viewpoints fixed to the central $1/9$ region during inference. The significant

Table 4: Ablation study of dynamic viewpoint control on UDVideoQA*. The best results are in bold. The default configuration is in gray.

Methods	Settings	UDVideoQA*					
		BU	Atr	ER	RR	CI	Avg
Qwen3-VL	Passive	57.1	10.5	41.2	46.2	51.3	41.3
Qwen3-VL	Dynamic	67.2	17.2	44.5	51.3	51.3	45.9
CamVLM	Passive	59.2	22.3	41.6	52.1	60.4	47.4
CamVLM	Dynamic	**70.6**	**29.4**	**46.2**	**64.7**	**72.7**	**57.0**

performance drop confirms that the improvements mainly stem from effective viewpoint control rather than the video understanding capability learned during training.

Effect of Different Viewpoint Sizes. We further evaluate different viewpoint sizes by training CamVLM from scratch with initial viewpoints covering the central $1/4$ and $1/6$ of the full frame, and comparing it with the corresponding passive and dynamic baselines. As shown in Table 5, CamVLM consistently outperforms the baselines across all settings. Larger viewpoints often capture the target immediately, whereas smaller viewpoints may initially miss it. Nevertheless, SFT enables the model to first zoom out for target search and then perform precise tracking, recovering complete fine-grained visual evidence.

Table 5: Ablation study of different viewpoint sizes on UDVideoQA*. The best results are in bold. The default configuration is in gray.

Methods	Settings	UDVideoQA*					
		BU	Atr	ER	RR	CI	Avg
Qwen3-VL	1/4, Passive	68.9	17.8	45.1	50.4	64.7	49.4
Qwen3-VL	1/4, Dynamic	75.6	22.3	46.2	53.4	71.0	53.7
CamVLM	1/4, Dynamic	**76.2**	**28.3**	**47.4**	**66.0**	**75.2**	**58.8**
Qwen3-VL	1/6, Passive	60.3	10.5	42.9	49.6	55.8	43.9
Qwen3-VL	1/6, Dynamic	64.4	19.8	45.4	53.0	58.9	48.3
CamVLM	1/6, Dynamic	**67.6**	**30.7**	**47.8**	**67.2**	**76.9**	**58.6**
Qwen3-VL	1/9, Passive	57.1	10.5	41.2	46.2	51.3	41.3
Qwen3-VL	1/9, Dynamic	67.2	17.2	44.5	51.3	51.3	45.9
CamVLM	1/9, Dynamic	**70.6**	**29.4**	**46.2**	**64.7**	**72.7**	**57.0**

Effect of Different Training Pipelines. Table 6 compares different training pipelines. SFT alone already achieves strong performance, indicating that it effectively learns viewpoint control and surveillance video understanding. Applying RL directly to Qwen3-VL-8B-Instruct also brings substantial gains, showing that these capabilities can emerge through autonomous exploration. Combining SFT with RL achieves the best performance, demonstrating that SFT provides a strong initialization for subsequent RL refinement.

Table 6: Ablation study of different training pipelines on UDVideoQA*. The best results are in bold. The default configuration is in gray.

Methods	Settings	UDVideoQA*					
		BU	Atr	ER	RR	CI	Avg
CamVLM (SFT)	Dynamic	65.1	27.7	38.2	63.0	72.4	53.7
CamVLM (RL)	Dynamic	**71.8**	25.2	**54.2**	54.6	**72.7**	56.0
CamVLM (SFT+RL)	Dynamic	70.6	**29.4**	46.2	**64.7**	**72.7**	**57.0**

5.4 Qualitative Analysis

Figure 5 provides qualitative examples of CamVLM. In simple interaction scenarios, CamVLM accurately captures complete action sequences by actively tracking the target through dynamic viewpoint control, maintaining informative visual evidence over time. In more challenging cases involving subtle motion differences and unexpected events, the model may still miss critical actions or confuse similar motions, such as describing running as walking and overlooking falls. These examples demonstrate that dynamic viewpoint control improves object-centric evidence acquisition and action understanding, while highlighting remaining challenges in fine-grained temporal reasoning.



Figure 5: Qualitative analysis of CamVLM on fine-grained human action understanding. CamVLM accurately recognizes complete action sequences in simple scenarios (top), but may miss subtle motion semantics and critical events in challenging videos (bottom), highlighting remaining challenges in fine-grained temporal reasoning.

6 Conclusion

In this work, we introduce CamVLM, an active visual reasoning framework that enables LVLMs to dynamically control camera viewpoints to actively acquire visual evidence in real-world surveillance. By constructing CCTV-Anomaly and CamTrack-53K, and formulating viewpoint control as a sequential decision-making problem, CamVLM learns effective camera policies. Extensive experiments demonstrate that active camera-based reasoning improves fine-grained understanding of real-world surveillance by acquiring more informative visual evidence.

AI USE STATEMENT

In this work, we used Large Language Models (LLMs) to improve the clarity, grammar, and readability of the manuscript. After using these tools, we carefully reviewed and edited the content as needed and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies video understanding in surveillance scenarios, where visual data may contain individuals and potentially sensitive events. Our research focuses on event-level understanding and active visual perception rather than personal identification. We acknowledge the potential privacy and misuse risks associated with surveillance technologies and encourage responsible use of the proposed datasets and models in accordance with applicable data licenses, privacy requirements, and ethical guidelines.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the model architecture, training pipeline, viewpoint control mechanism, dataset construction, evaluation protocols, and implementation settings in the main paper and appendix. We also provide the prompt templates used for training and evaluation. In addition, we have submitted the complete code and representative samples from both datasets as supplementary material. The full datasets, models, and code will be released to support reproduction of our results and future research.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20143–20153. IEEE, 2022.

Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

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

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2025.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025.

Google. Introducing Gemini 3.1 Flash-Lite. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/`, 2026.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Seongho Kim, Sejong Ryu, Hyoukjun You, and Je Hyeong Hong. Gta-crime: A synthetic dataset and generation framework for fatal violence detection with adversarial snippet-level domain adaptation. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 97–102. IEEE, 2025.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

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, 2014.

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.

Xiaotian Liu, Ali Pesaranghader, Jaehong Kim, Tanmana Sadhu, Hyejeong Jeon, and Scott Sanner. Activevoo: Value of observation guided active knowledge acquisition for open-world embodied lifted regression planning. *Advances in Neural Information Processing Systems*, 38:106621–106647, 2025.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. *arXiv preprint arXiv:2601.08325*, 2026.

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*, pp. 2720–2727, 2013.

OpenAI. Thinking with images. `https://openai.com/index/thinking-with-images`, 2025.

OpenAI. Introducing GPT-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 2026a.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026b.

Mauricio Perez, Alex C Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2662–2666. IEEE, 2019.

Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16307–16316. IEEE, 2024.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pp. 208–223. Springer, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Haowei Sun, Jinwu Hu, Zhirui Zhang, Haoyuan Tian, Xinze Xie, Yufeng Wang, Xiaohua Xie, Yun Lin, Zhuliang Yu, and Mingkui Tan. Open-world drone active tracking with goal-centered rewards. *Advances in Neural Information Processing Systems*, 38:147712–147747, 2025.

Joseph Raj Vishal, Nagasiri Poluri, Katha Naik, Rutuja Patil, Kashyap Hegde Kota, Krishna Vinod, Prithvi Jai Ramesh, Mohammad Farhadi, Yezhou Yang, and Bharatesh Chakravarthi. Udvideoqa: A traffic video question answering dataset for multi-object spatio-temporal reasoning in urban dynamics. *arXiv preprint arXiv:2602.21137*, 2026.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2025a.

Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025b.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pp. 322–339. Springer, 2020.

Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2264–2273, 2015.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980. IEEE, 2025.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026.

Xiao Zhang, Lu Zou, Tao Lu, Yuan Yao, Zhangjin Huang, and Guoping Wang. Instance-adaptive keypoint learning with local-to-global geometric aggregation for category-level object pose estimation. *arXiv preprint arXiv:2504.15134*, 2025.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Appendix

CONTENTS

A Limitations **14**

B Evaluation Details on UDVideoQA* **14**

C Details of CCTV-Anomaly Construction **14**

C.1 Video Collection 14

C.2 Multi-level Video Annotation 15

D Details of CamTrack-53K Construction **15**

D.1 Data Curation 15

D.2 Viewpoint Representation 15

D.3 Viewpoint Construction 16

D.4 Viewpoint Trajectory Visualization 16

E Prompt Template **16**

A LIMITATIONS

Virtual-to-Physical Camera Control. Large-scale viewpoint-action trajectories from physical cameras operating in real-world, real-time surveillance are difficult to collect. We therefore adopt a virtual camera simulation, treating each full video frame as a panoramic observation space and a local region as the current camera viewpoint. Within this space, CamVLM learns camera control policies through directional and zooming actions with corresponding parameters, focusing on how and where the camera should move rather than device-specific actuation. For real-world deployment, the predicted actions and parameters could be translated into physical camera commands according to camera calibration and hardware specifications. Direct validation on physical camera systems remains an important direction for future work.

Current viewpoint trajectories still rely on predefined rules and fixed adjustment intervals, which limits adaptability in complex scenarios. Moreover, the framework focuses on single-camera control, while real-world surveillance requires more autonomous and collaborative perception. Future work will explore physical camera deployment, adaptive trajectory generation, dynamic viewpoint scheduling, and multi-camera coordination for more intelligent visual reasoning systems.

B EVALUATION DETAILS ON UDVIDEOQA*

Since the original test split of UDVideoQA (Vishal et al., 2026) is no longer available, we construct UDVideoQA* with permission and approval from the original authors using Set 03 and Set 20 from the released dataset, as these two subsets provide the greatest question diversity. The resulting benchmark contains five question types: Basic Understanding (BU), Attribution (Atr), Event Reasoning (ER), Reverse Reasoning (RR), and Counterfactual Inference (CI). Each question type contains 238 QA pairs, resulting in a total of 1,190 QA pairs.

Following the standard evaluation protocol of UDVideoQA, we employ an external large language model (LLM) as a semantic matching evaluator. Specifically, we use GPT-5.5 (OpenAI, 2026a) as a text-only evaluator. To mitigate potential evaluation bias, all model judgments are further verified by human annotators, who independently assess the correctness of the generated answers. Each response is assigned a binary score $S_{\mathrm{LLM},i} \in \{0, 1\}$, indicating whether the predicted answer is semantically consistent with the ground-truth answer.

We further adopt the weighted scoring scheme introduced by UDVideoQA to account for differences in cognitive complexity across reasoning categories. More challenging reasoning types are assigned larger weights: 1.0 for BU, 1.2 for Atr, 1.3 for ER and RR, and 1.5 for CI. The final weighted score W is computed as

$$W = \frac{1}{N} \sum_{i=1}^{N} S_{\mathrm{LLM},i} \cdot W_{c_i}, \tag{10}$$

where $S_{\mathrm{LLM},i}$ denotes the correctness score of the i-th response, and W_{c_i} is the weight assigned to its reasoning category c_i. This formulation places greater emphasis on successful predictions for cognitively demanding question types, particularly counterfactual inference, which is designed to reduce hallucinations and encourage deeper causal understanding. We report the accuracy for each of the five question types together with the weighted overall accuracy.

For all evaluations, we set the temperature to 0 and adopt greedy decoding to ensure deterministic outputs. All reported results are obtained from a single evaluation run using the final trained model checkpoint.

C DETAILS OF CCTV-ANOMALY CONSTRUCTION

C.1 VIDEO COLLECTION

CCTV-Anomaly is constructed from two complementary sources: publicly available surveillance anomaly datasets and newly collected real-world surveillance videos. To ensure data quality and suitability for LVLM training, we perform extensive data filtering to remove duplicate videos, manually edited clips, severely occluded scenes, and samples with ambiguous event descriptions. After

filtering, the public subset contains 5,832 videos collected from existing datasets, including XD-Violence (Wu et al., 2020), CCTV-Fights (Perez et al., 2019), GTA-Crime (Kim et al., 2025), and UBnormal (Acsintoae et al., 2022). The newly collected subset contains 8,301 videos obtained from diverse real-world surveillance resources, covering various environments, viewpoints, and event scenarios.

C.2 Multi-level Video Annotation

To support fine-grained surveillance video understanding, we develop a multi-level annotation pipeline consisting of event category annotation, detailed caption annotation, and structured evaluation annotation. The annotation process adopts an LLM-assisted strategy with human verification to ensure annotation reliability. We use Doubao-Seed-2.0-Pro (Seed, 2026) as the automatic annotator.

Event Category Annotation. Each video is assigned one event category according to our predefined taxonomy. For public datasets with existing labels, we inherit the original annotations when they are semantically consistent with our categories. For unlabeled public videos and newly collected samples, event categories are generated through video analysis and subsequently verified by human annotators.

Event Caption Annotation. Each video is further annotated with a concise yet informative caption describing the key semantic information of the event, including involved subjects, major actions, and event outcomes. Instead of exhaustively describing all visual contents, the captions focus on information relevant to surveillance reasoning. We adopt an LLM-assisted caption generation pipeline, followed by human inspection to correct inaccurate descriptions and ensure consistency with the corresponding event category.

Structured Evaluation Annotation. To enable fine-grained evaluation of LVLM outputs, we randomly select 1,000 videos as the test set and construct structured semantic annotations from verified captions. Specifically, each video is annotated with five semantic components: location, time of day, subjects, activities, and objects. All structured annotations are manually reviewed to ensure consistency and completeness.

D Details of CamTrack-53K Construction

D.1 Data Curation

We construct CamTrack-53K from the A2D (Xu et al., 2015), MeViS (Ding et al., 2023), and Refer-YouTube-VOS (Seo et al., 2020) subsets of VideoRefer-700K (Yuan et al., 2025). The resulting dataset contains 53,059 samples, including 26,221 from A2D, 8,459 from MeViS, and 18,379 from Refer-YouTube-VOS.

D.2 Viewpoint Representation

The viewpoint is represented as a normalized rectangular region:

$$v_t = [x_{1,t}, y_{1,t}, x_{2,t}, y_{2,t}], \tag{11}$$

where t denotes the viewpoint step, and $(x_{1,t}, y_{1,t})$ and $(x_{2,t}, y_{2,t})$ denote the top-left and bottom-right coordinates of the viewpoint at step t, respectively. All coordinates are normalized to $[0, 1]$.

For our main experiments, the viewpoint is initialized at the center of the frame and covers $1/9$ of the full-frame area:

$$v_0 = \left[\frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3}\right]. \tag{12}$$

We additionally construct variants with initial viewpoint area ratios of $1/6$ and $1/4$ for ablation studies.

D.3 VIEWPOINT CONSTRUCTION

Each clip is associated with a viewpoint action derived from the spatial configuration of the task-relevant objects. For clips involving multiple objects, their bounding boxes are merged into the minimum enclosing box:

$$b_t = \left[\min_i x_{1,t}^{(i)}, \min_i y_{1,t}^{(i)}, \max_i x_{2,t}^{(i)}, \max_i y_{2,t}^{(i)}\right]. \tag{13}$$

Here, t denotes the viewpoint step, and i indexes the task-relevant objects. $(x_{1,t}^{(i)}, y_{1,t}^{(i)})$ and $(x_{2,t}^{(i)}, y_{2,t}^{(i)})$ denote the top-left and bottom-right coordinates of the bounding box of the i-th object, respectively. The resulting b_t represents the aggregated target region at step t.

The target viewpoint is designed to fully cover the task-relevant objects while preserving sufficient surrounding context. Let W and H denote the width and height of the video frame, respectively, and let $w(b_t)$, $h(b_t)$, and $A(b_t)$ denote the width, height, and area of the target bounding box. We adaptively determine the normalized side length of the viewpoint according to the size of the target box relative to the full frame:

$$s_{v_t} = \mathrm{clip}\left(\max\left\{\sqrt{\frac{A(b_t)}{\rho WH}}, \frac{\kappa w(b_t)}{W}, \frac{\kappa h(b_t)}{H}\right\}, s_{\min}, s_{\max}\right), \tag{14}$$

where $\rho = 0.2$ controls the desired proportion of the viewpoint occupied by the target, and $\kappa = 1.3$ is a padding factor that preserves visual context around the target. The parameters $s_{\min} = 0.2$ and $s_{\max} = 0.8$ constrain the minimum and maximum viewpoint scales, respectively. Consequently, smaller targets are associated with more compact viewpoints, whereas larger targets require broader viewpoints.

The target viewpoint is centered on the target bounding box and appropriately shifted when approaching the frame boundaries, ensuring that it always remains within the valid image region. When the target is already fully covered, approximately centered, and observed at a suitable scale, the current viewpoint is maintained to avoid unnecessary viewpoint adjustments.

D.4 VIEWPOINT TRAJECTORY VISUALIZATION

We visualize representative temporally ordered viewpoint trajectories from CamTrack-53K in Figure 6, where the yellow boxes denote object bounding boxes and the blue boxes indicate the current viewpoints. As shown, the constructed viewpoints continuously and stably cover task-relevant objects throughout their motion, while dynamically adjusting their positions and spatial extents according to object scale and spatial distribution. Meanwhile, they suppress excessive background content while preserving sufficient surrounding context and interaction cues, thereby yielding compact, coherent, and informative fine-grained visual evidence.

E PROMPT TEMPLATE

Prompt Template for Training and Passive Viewpoint Evaluation on CCTV-Anomaly. Figure 7 (top) illustrates the prompt template for training and passive viewpoint evaluation on CCTV-Anomaly.

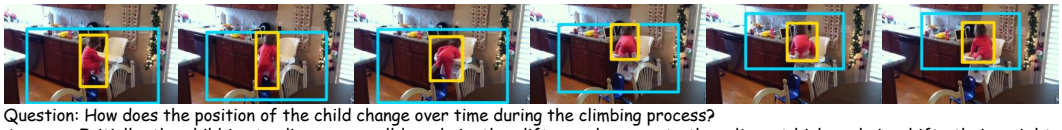
Prompt Template for Dynamic Viewpoint Evaluation on CCTV-Anomaly. Figure 7 (bottom) illustrates the prompt template for dynamic viewpoint evaluation on CCTV-Anomaly.

Prompt Template for LLM-Based Semantic Evaluation of Video Captioning on CCTV-Anomaly. Figure 8 illustrates the prompt template for LLM-based semantic evaluation of video captioning on CCTV-Anomaly.

Prompt Template for Training on CamTrack-53K and Dynamic Viewpoint Evaluation on UDVideoQA*. Figure 9 (top) illustrates the prompt template for training on CamTrack-53K and dynamic viewpoint evaluation on UDVideoQA*.

Prompt Template for Passive Viewpoint Evaluation on UDVideoQA*. Figure 9 (bottom) illustrates the prompt template for passive viewpoint evaluation on UDVideoQA*.

Prompt Template for LLM-Based Semantic Evaluation on UDVideoQA*. Figure 10 illustrates the prompt template for LLM-based semantic evaluation on UDVideoQA*.



Question: How does the position of the child change over time during the climbing process?
Answer: Initially, the child is standing on a small low chair, then lifts one leg up onto the adjacent higher chair, shifts their weight, pulls their full body onto the chair, and eventually settles to sit securely on the higher chair's seat.



Question: How does the competing player in the red jersey move as he plays this table tennis point?
Answer: He makes dynamic, fast full-body movements across the court: he sprints, lunges, and dives low to the ground in an attempt to reach and return the ball.



Question: How does the interaction between the brown dog and the white-footed dog change over time?
Answer: Initially, the brown dog and the white-footed dog are mating, with one dog mounted on the other. After some time, they separate from the mating position, turn to face each other, and remain close to one another while continuing to interact closely.



Question: What interaction are the two tigers engaged in, and what visible physical difference exists between the pair?
Answer: The two tigers are actively fighting, swiping at each other with their front paws. One of the tigers is noticeably larger than the other during this physical interaction.



Question: How do the two pandas interact with each other?
Answer: The two pandas engage in playful wrestling, with the panda that started on the elevated wooden platform pulled down to the ground by the other panda, and they then roll around together and playfully bite each other.

Figure 6: Visualization of viewpoint trajectories in CamTrack-53K. Yellow and blue bounding boxes denote object bounding boxes and current viewpoints, respectively.

Prompt for Training and Passive Viewpoint Evaluation on CCTV-Anomaly

Task:
Analyze the video and:
1. Provide a concise but complete description of the main events in the video.
2. Identify ONE single best category that most accurately describes the primary event.

Classification Rules:
- Choose exactly ONE category only.
- Select the category based on the dominant or most important event in the video.
- Prioritize safety/security incidents over normal activity.
- If multiple events occur, choose the most severe or highest-risk event.
- Only use the category names exactly as listed below.
- If none of the risk categories apply, use "normal activity".

Categories:
- break-in: Illegal entry or attempted intrusion into a property, home, room, or enclosed area.
- theft: Taking or attempting to take any property without permission, including packages, car, or belongings.
- suspicious behavior: Loitering, face occlude, peeping, casing a location, or other abnormal behavior suggesting intent to commit a crime.
- violence: Physical assault, fighting, armed threat, visible weapon, shooting, or any aggressive act against a person.
- vandalism: Intentional damage, graffiti, arson, scratching or defacement of property.
- fire hazard: Visible fire, smoke, explosion, or any fire-related danger.
- personal emergency: A person slips, falls, drowning, or shows urgent danger.
- wild animal: Presence of a wild or potentially dangerous animal (e.g., wolf, bear, raccoon, deer) in the area.
- vehicle incident: Vehicle collision or a vehicle striking a person or object.
- normal activity: Ordinary daily activity with no safety or security risk.

Output Format (strictly follow):
Video category: <exact category name>
Video description: <event summary>

Prompt for Dynamic Viewpoint Evaluation on CCTV-Anomaly

You are an expert in video tracking and security video analysis.

In a continuous video, each second you can only see a cropped local window from the current video frame, not the full frame. Based on the visible content inside the window, decide whether to move or zoom the view to keep the people involved in the primary anomalous event, the objects they interact with, and their actions and interactions clearly visible across time. After the video ends, complete the final task using the window views observed throughout the process.

During each second, output only the action JSON. Use this format: {"action": ...}

Available actions:
{"action": None} (no movement);
{"action": left/right/up/down, "offset": 0-1} (move the view left/right/up/down, range 0-1);
{"action": zoom_in, "scale": s}, where 0<s<1 (shrink the view by the direct scale factor);
{"action": zoom_out, "scale": s}, where s>1 (enlarge the view by the direct scale factor).
To apply multiple actions, separate them with semicolons, for example:
{"action": zoom_out, "scale": 1.5}; {"action": down, "offset": 0.12}.

When the user later indicates that the video has ended, stop tracking and follow the final task instructions.

The video has ended. Stop tracking.
+
<Prompt for training and passive viewpoint evaluation on CCTV-Anomaly>

Figure 7: Prompt templates for training and passive and dynamic viewpoint evaluation on CCTV-Anomaly.

Prompt for LLM-Based Semantic Evaluation of Video Captioning on CCTV-Anomaly

```
You are a lenient semantic evaluator for security-video descriptions.
Judge whether the model description covers every GT item in the structured fields.

General rules:
- Synonyms, paraphrases, broader terms, and common variants count when they express the same visible meaning.
- Do not require exact wording, direction, position, order, or object subtype unless a subject attribute is required.
- Do not reward information absent from the description. Contradictions are not hits.

Field rules:
- location: a compatible scene type counts; driveway and street may represent the same relevant outdoor access area.
- time_of_day: compatible cues such as night/dark/evening or day/daylight count.
- subjects: identify the same entity. If a GT subject contains attributes such as color, clothing, count, role, age,
  species, or vehicle type, the description must contain the entity plus at least one compatible GT attribute.
  Saying only "a man" does not hit "a man wearing a dark blue top and black pants"; "a person in blue" does.
  If the GT item is generic and has no attribute, such as "a person", "a man", "people", or "vehicles",
  matching the entity alone is sufficient.
- activities: the core action is enough. walk/run/move/approach/head/go and similar movement expressions may match.
- objects: the object meaning is enough and attributes are ignored. car/pickup/sedan/vehicle and similar variants may match.

Model description:
{description}

GT structured fields:
{structured JSON}

Return strict JSON only. Preserve each GT list item exactly and return one result per item:
{
  "location": {"hit": true, "reason": "..."},
  "time_of_day": {"hit": true, "reason": "..."},
  "subjects": [{"item": "...", "hit": true, "reason": "..."}],
  "activities": [{"item": "...", "hit": true, "reason": "..."}],
  "objects": [{"item": "...", "hit": true, "reason": "..."}]
}
```

Figure 8: Prompt template for LLM-based semantic evaluation of video captioning on CCTV-Anomaly.

Prompt Template for Training on CamTrack-53K and Dynamic Viewpoint Evaluation on UDVideoQA*

You are an expert in video tracking and video analysis.

In a continuous video, each second you can only see a cropped local window from the current video frame, not the full frame. Based on the visible content inside the window, you need to decide each second whether to move or zoom the window to track all objects referred to by the question, ensuring they remain clearly visible inside the window with an appropriate size. After the video ends, answer the question using the window views observed throughout the process.

Question:
{question}

During each second, output only the action JSON. Use this format: {"action": ...}

Available actions:
{"action": None} (no movement);
{"action": left/right/up/down, "offset": 0-1} (move the view left/right/up/down, range 0-1);
{"action": zoom_in, "scale": s}, where 0<s<1 (shrink the view by the direct scale factor);
{"action": zoom_out, "scale": s}, where s>1 (enlarge the view by the direct scale factor).
To apply multiple actions, separate them with semicolons, for example:
{"action": zoom_out, "scale": 1.5}; {"action": down, "offset": 0.12}.

When the user later indicates that the video has ended, stop window tracking and answer directly with <answer>your answer</answer>.

The video has ended. Stop window tracking and answer the question directly.
Output only the answer using this format: <answer>your answer</answer>.

Prompt Template for Passive Viewpoint Evaluation on UDVideoQA*

You are a traffic video analyst. Watch the video carefully before answering.

The video has ended.

Question:
{question}

Output the answer using this format: <answer>your answer</answer>.

Figure 9: Prompt templates for training on CamTrack-53K and dynamic and passive viewpoint evaluation on UDVideoQA*.

Prompt for LLM-Based Semantic Evaluation on on UDVideoQA*

You are an impartial, strict grader for VideoQA semantic correctness.

Question: {question}
Ground truth answer: {ground_truth}
Model answer: {model_answer}

Assign correct as 1 or 0 only; no partial credit.
Accept exact or unambiguous semantic equivalence for binary or categorical answers.
Accept a short textual paraphrase only when it has the same meaning. Missing key information or contradicting the ground truth is 0.
A blank answer, NA, I don't know, unsure, or a contradiction is 0.
Do not infer beyond the ground truth.

Return JSON only: {"correct": 0|1}

Figure 10: Prompt template for LLM-based semantic evaluation on UDVideoQA*.