
A2Seek: Towards Reasoning-Centric Benchmark for Aerial Anomaly Understanding

Mengjingcheng Mo^{1,2}, Xinyang Tong¹, Mingpi Tan¹, Jiaxu Leng^{1,2*},
Jiankang Zheng^{1,2}, Yiran Liu¹, Haosheng Chen¹, Ji Gan^{1,2}, Weisheng Li¹, Xinbo Gao^{1,2*}

¹School of Computer Science and Technology,

Chongqing University of Posts and Telecommunications, Chongqing, China

²Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory, Chongqing, China
{lengjx, gaoxb}@cqupt.edu.cn

Abstract

While unmanned aerial vehicles (UAVs) offer wide-area, high-altitude coverage for anomaly detection, they face challenges such as dynamic viewpoints, scale variations, and complex scenes. Existing datasets and methods, mainly designed for fixed ground-level views, struggle to adapt to these conditions, leading to significant performance drops in drone-view scenarios. To bridge this gap, we introduce A2Seek (Aerial Anomaly Seek), a large-scale, reasoning-centric benchmark dataset for aerial anomaly understanding. This dataset covers various scenarios and environmental conditions, providing high-resolution real-world aerial videos with detailed annotations, including anomaly categories, frame-level timestamps, region-level bounding boxes, and natural language explanations for causal reasoning. Building on this dataset, we propose A2Seek-R1, a novel reasoning framework that generalizes R1-style strategies to aerial anomaly understanding, enabling a deeper understanding of “Where” anomalies occur and “Why” they happen in aerial frames. To this end, A2Seek-R1 first employs a graph-of-thought (GoT)-guided supervised fine-tuning approach to activate the model’s latent reasoning capabilities on A2Seek. Then, we introduce Aerial Group Relative Policy Optimization (A-GRPO) to design rule-based reward functions tailored to aerial scenarios. Furthermore, we propose a novel “seeking” mechanism that simulates UAV flight behavior by directing the model’s attention to informative regions. Extensive experiments demonstrate that A2Seek-R1 achieves up to a 22.04% improvement in AP for prediction accuracy and a 13.9% gain in mIoU for anomaly localization, exhibiting strong generalization across complex environments and out-of-distribution scenarios. Our dataset and code are released at <https://2-mo.github.io/A2Seek/>.

1 Introduction

Traditional anomaly detection [42, 38, 9] relies on fixed-view cameras and primarily focuses on anomaly classification, offering limited semantic interpretation. Their static perspectives and narrow fields of view significantly limit their effectiveness in monitoring large and dynamic environments [78]. With the rapid advancement of unmanned aerial vehicle (UAV) technology, aerial surveillance has emerged as a powerful paradigm for wide-area anomaly detection. Droneview footage introduces frequent viewpoint shifts, scale changes, complex backgrounds and occlusions, as well as environmental disturbances (lighting, weather, moving shadows) [10, 16]. Crucially, anomalous regions in aerial scenes are often subtle, spatially sparse, and occupy only a small por-

*Corresponding authors

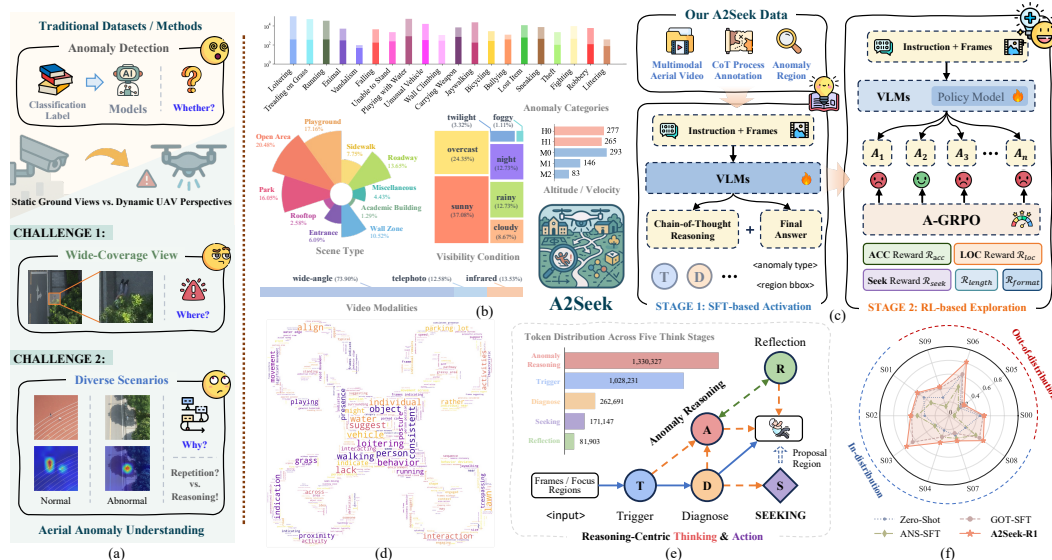


Figure 1: Overview of the A2Seek Benchmark. (a) Challenges in aerial anomaly detection. Traditional methods rely on static surveillance views and focus mainly on classification, making it difficult to answer “Where” and “Why” anomalies occur under dynamic UAV perspectives. (b) Dataset statistics on multiple dimensions. (c) Reasoning pipeline. The method consists of two stages: SFT (supervised fine-tuning) for reasoning activation, and RL (reinforcement learning) for dynamic reasoning. (d) High-frequency word of dataset. (e) Reasoning process. The framework integrates multiple reasoning stages (Trigger, Diagnose, Reasoning, Reflection and Seeking), emphasizing reasoning-driven anomaly understanding. (f) Performance comparison.

tion of the field of view, making them difficult to perceive. Even when alarms are triggered, human observers often struggle to localize these subtle, spatially sparse events. Moreover, traditional models rely on stable spatial priors invalidated when a drone’s pose constantly changes. Consequently, groundview approaches fail to generalize to aerial data, exhibiting poor robustness to dynamic scenes and variable spatial distributions. As Figure 1 illustrates, anomaly detection in drone-view footage requires precise spatial localization (“Where is the anomaly?”) and semantic interpretation (“Why is it an anomaly?”). Drone-captured videos often span large, dynamic environments, where subtle anomalies can be easily overlooked or misinterpreted without robust contextual reasoning. These factors demand models capable of adaptively focusing on critical regions and abstracting high-level reasoning to explain anomalies. In the following, we distill these into two core problems critical for robust aerial anomaly understanding.

CHALLENGE 1: Spatial Localization in Wide-Area, Dynamic Scenes. Traditional ground-based anomaly detection typically relies on fixed-view cameras, where anomalies are often prominent and relatively easy to identify. However, aerial anomaly detection is fundamentally different. Drone-view videos cover expansive and dynamic scenes with complex backgrounds and frequent motion. Anomalies are often small, sparse, and easily obscured by cluttered surroundings, making them difficult to detect even when anomalies are flagged (e.g., a person falling in Figure 1a). Therefore, aerial anomaly detection requires not only determining whether an event is anomalous but also providing accurate localization of the anomalous objects to support anomaly judgment.

CHALLENGE 2: Semantic Generalization Across Diverse Aerial Contexts. Aerial anomaly detection faces significant scene dependency and distributional shifts, as the same behavior (e.g. running) may exhibit vastly different visual characteristics and interactions across environments (e.g. playgrounds or academic buildings). Relying solely on feature classification or reconstruction errors fails to capture this diversity and hinders reliable generalization to unseen scenarios. To address this, models must deeply understand scene context and behavioral semantics to effectively tackle the generalization challenge.

Recent advances in multimodal large language models (MLLMs) [1, 14, 4, 30] have demonstrated significant potential in cross-modal alignment and semantic understanding, particularly in complex reasoning scenarios [45, 65, 74]. However, their application to aerial anomaly understanding re-

mains underexplored due to two major limitations: (1) the lack of large-scale multimodal aerial datasets with temporal annotations, precise spatial localization, and semantic reasoning explanations, which hinders effective training and evaluation; and (2) the absence of structured reasoning frameworks and adaptive strategies, making it difficult to address diverse anomalies in complex aerial perspectives. To address the above challenges, we present A2Seek, a reasoning-centric benchmark specifically designed for aerial anomaly understanding. Collected across 10 campus scenes over one year, the dataset spans 23 hours of UAV footage with diverse flight altitudes, speeds, and trajectories, including 3.79 hours of complex anomalies and the rest normal behaviors. It features 542 untrimmed 4K drone videos and over 32k curated keyframes, annotated with fine-grained anomaly labels, spatiotemporal bounding boxes, and structured reasoning graphs. These annotations enable comprehensive evaluation of detection accuracy and reasoning interpretability. To tackle practical challenges such as occlusion and low-light conditions, A2Seek incorporates telephoto footage for high-altitude scenes and infrared modalities for nighttime scenarios, facilitating the detection of subtle or visually ambiguous anomalies.

Building on this benchmark, we propose A2Seek-R1, a novel reinforcement fine-tuning framework designed to enhance the reasoning capabilities of models for aerial anomaly understanding. A2Seek-R1 first employs a graph-of-thought (GoT)-guided supervised fine-tuning (SFT) approach, which activates the model’s latent reasoning capabilities by leveraging structured reasoning annotations in the A2Seek dataset. These annotations consist of optional stages, including trigger, diagnosis, reasoning, reflection, and seeking, effectively guiding the model to handle anomalies of varying complexity in a progressive manner. Among them, seeking is set as a potential region of interest for the model in video frames with insufficient information, such as blurry or occluded images, thus achieving a new type of seeking mechanism that simulates the flight behavior of unmanned aerial vehicles, enabling the model to dynamically focus on specific regions of interest. Second, it introduces a tailored extension of Group Relative Policy Optimization (GRPO), termed A-GRPO, specifically designed for aerial anomaly understanding. A-GRPO extends the original accuracy and format function rewards by incorporating localization and seeking rewards. Localization rewards enhance the model’s spatial understanding of anomaly regions, while seeking rewards focus on aligning the model’s predictions with human annotations of anomaly candidate areas, ensuring the extraction of valuable spatial information for better understanding. Additionally, to address the diverse perspectives of drones, a length reward function is introduced to encourage concise responses in simple scenarios and allocate more computational effort to complex situations. By combining these components, A2Seek-R1 achieves precise spatial localization and robust reasoning for aerial anomalies, setting a new benchmark for anomaly understanding in complex, real-world environments.

Contributions: (1) We present A2Seek, a large-scale, reasoning-centric benchmark specifically designed for multi-scenario anomaly understanding from aerial perspectives. (2) We propose A2Seek-R1, a novel multi-stage reinforcement fine-tuning framework that significantly enhances the aerial anomaly understanding capabilities of multimodal foundation models. (3) This work is the first to simulate UAV motion characteristics in the context of anomaly understanding, enabling models to actively acquire detailed regional information in challenging scenarios. (4) Extensive experiments across multiple scenarios validate the superiority of A2Seek-R1. Compared to models trained solely with GoT-SFT, A2Seek-R1 achieves an improvement of 6.72% in prediction accuracy.

2 Related Work

Video Anomaly Detection. Early efforts focused on single-scene datasets [34, 42] using fixed-view RGB cameras for pedestrian anomaly detection. Later datasets [38, 52, 15, 47] that towards real-world introduced more complex scenes with crowded traffic, yet remained unimodal, fixed-view, and emphasized foreground representation. Larger-scale datasets [57, 66, 64, 79] improved diversity and duration, but still relied on ground-view perspectives and coarse anomaly labels, lacking spatial localization or causal reasoning. Methodologically, the field evolved from handcrafted features to learned representations [3, 75, 34] and behavior modeling [48, 35, 27]. Recent approaches span reconstruction/prediction-based [20, 46, 19, 67, 70, 39], object-centric [40, 33, 44, 60, 18, 76], distribution-aware [21, 62, 49, 29], and llm-driven paradigms [17, 72, 45, 74], with growing emphasis on generalization and scene dependency. Recent methods [77, 25] incorporate chain-of-reasoning mechanisms to enhance anomaly understanding, enabling more interpretable and goal-directed decision-making. However, existing benchmarks and methods largely overlook aerial-

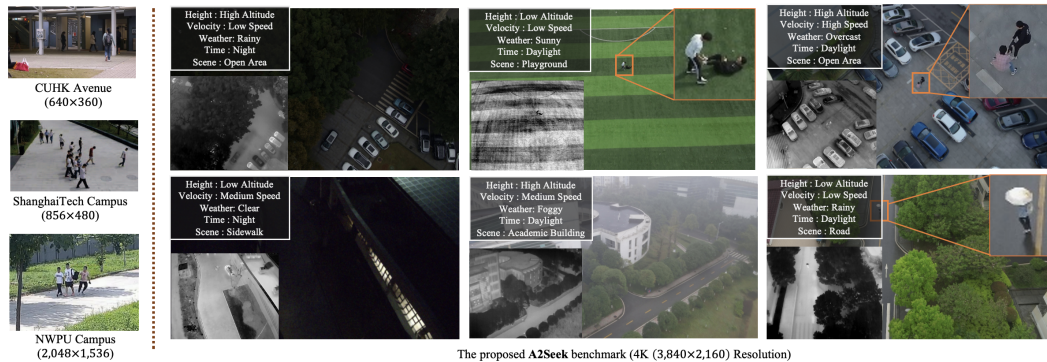


Figure 2: Comparison of scene diversity and complexity. Left: fixed-view surveillance datasets. Right: diverse aerial views in A2Seek.

specific challenges such as extreme viewpoint shifts and scale variation. Moreover, the absence of multimodal and reasoning-oriented annotations limits fine-grained analysis.

Aerial Anomaly Understanding. Anomaly detection in aerial videos remains underexplored. Early pioneer datasets [7, 8, 26, 61] adopt aerial perspectives but provide only frame-level or coarse labels, limiting fine-grained analysis. Existing methods focus on motion cues, including optical flow [8], reconstruction-based schemes[28], or spatiotemporal modeling with 3D CNNs [61] and Transformer [26], but rarely support precise region-level reasoning. While multimodal large language models (MLLMs) [1, 23, 14, 13, 30, 68] have advanced semantic understanding in ground-view tasks, their application to aerial scenarios is limited. Current approaches [6, 59, 74] often lack explicit, grounded reasoning and rely on post hoc explanations. To bridge this gap, we introduce a reasoning-centric aerial anomaly dataset with fine-grained spatial-temporal annotations and dynamic reasoning trajectories.

3 The A2Seek Dataset

Existing video anomaly detection methods primarily rely on fixed-view ground-based cameras, which are designed for limited fields of view and relatively static backgrounds. These methods face significant limitations when applied to drone-view videos, which involve frequent viewpoint changes, scale variations, dynamic occlusions, and complex environmental disturbances (*e.g.*, lighting changes, weather variations). These factors significantly increase the challenges of spatial localization and semantic generalization of anomaly detection. To address these challenges, we introduce A2Seek, a reasoning-centric aerial anomaly understanding benchmark. It supports precise spatial localization of anomalies (“Where is the anomaly?”) and in-depth semantic reasoning explanations (“Why is it anomalous?”). The dataset spans diverse real-world scenarios and anomaly types, providing high-resolution RGB and infrared video data with detailed frame-level labels, region-level annotations, and structured natural language reasoning explanations. Figure 2 illustrates the dataset’s diversity, showcasing various scenes, altitudes, speeds, weather conditions, and times of day. Unlike traditional ground-based datasets, which focus on static viewpoints and limited environmental variations, A2Seek leverages the dynamic and expansive nature of UAV perspectives, making it a more challenging and realistic benchmark. This benchmark aims to advance research on generalization, robustness, and interpretability in aerial anomaly understanding.

3.1 Data Collection and Annotation

The A2Seek dataset was collected using a DJI M30T drone equipped with wide-angle, telephoto, and infrared cameras. Flights were conducted at varying altitudes (10 to 60 meters) and speeds (0 to 20 m/s) to capture diverse aerial perspectives. Trajectory patterns included hovering, linear cruising, curved circling, and area scanning, enabling dynamic viewpoint shifts across scenes. In total, A2Seek comprises 542 untrimmed 4K videos (over 23 hours), recorded across 10 campus environments, covering more than 20 types of anomalous events (*e.g.*, falling, fighting, jaywalking) under varying conditions such as day/night, clear/foggy weather, and so on.

Table 1: Comparison of A2Seek with existing video anomaly detection datasets ([†] denotes web-sourced datasets; [‡] denotes simulated or virtual datasets).

Perspective	Dataset	Frames			Scene Count	Anomaly Types	Resolution	Scene Dependency	Scale Variation	Reasoning Annotation	Multi-modal
		Total	Normal	Abnormal							
Surveillance	CUHK Avenue [42]	30,652	26,832	3,820	1	5	640×360	×	×	×	×
	ShanghaiTech [43]	317,398	300,308	17,090	13	11	856×480	×	×	×	×
	Street Scene [52]	203,257	159,341	43,916	1	17	1,280×720	×	×	×	×
	Subway [3]	209,151	192,548	16,603	2	8	512×384	×	×	×	×
	UBI-Fight [†] [15]	8,530,080	8,287,381	242,699	-	1	1,280×720	×	×	×	×
	LAD [†] [64]	3,625,237	3,016,213	609,024	-	14	320×240	×	×	×	×
	IITB Corridor [11]	483,566	301,999	181,567	1	10	1,920×1,080	×	×	×	×
	UCSD Ped2 [34]	4,560	2,924	1,636	1	5	360×240	×	×	×	×
	UMN [51]	7,741	6,165	1,576	3	1	320×240	×	×	×	×
	XD-Violence [†] [66]	18,697,729	-	-	-	6	640×360	×	×	×	✓
	UCF-Crime [†] [71]	13,741,393	-	-	-	13	320×240	×	×	×	×
	TAD [47]	540,272	462,578	77,694	-	7	1,280×720	×	×	×	×
	UBNormal [‡] [2]	236,902	147,887	89,015	29	22	1,080×720	×	×	×	×
	NWPU Campus [9]	1,466,073	1,400,807	65,266	43	28	2,048×1,536	✓	×	×	×
Drone-view	MSAD [‡] [78]	447,236	308,792	138,444	14	55	1,920×1,080	×	×	×	×
	Minidrone [7]	23,295	14,821	8,474	1	10	1,280×720	×	×	×	×
	AU-AIR-Anomaly [8]	32,823	30,000	2,823	1	8	1,920×1,080	×	×	×	✓
	Drone-Anomaly [26]	87,488	51,635	35,853	7	1	640×640	×	×	×	×
	UIT-ADrone [61]	206,194	142,709	63,485	3	10	1,920×1,080	×	×	×	×
	A2Seek (Ours)	2,485,859	2,087,160	398,699	27	20	3,840×2,160	✓	✓	✓	✓

The anomaly categories in A2Seek were carefully curated based on the principle of “potential disruption to campus public safety or order,” rather than broadly labeling daily activities as abnormal. For instance, running on a playground is not considered anomalous, whereas running in an academic building area is. Similarly, riding a bicycle on the road is normal, but riding on sidewalks is considered anomalous. Drawing from classic VAD datasets [34, 42, 43, 9], we focus on behavior categories with strong relevance to real-world campus safety. Anomalies are categorized into three risk levels. High-risk events such as carrying weapons, fighting, and robbery require immediate intervention. Medium-risk events such as running, falling, animal intrusions, and illegal cycling may escalate but do not cause immediate harm. Low-risk events such as loitering, littering, and walking on grass pertain to order maintenance or environmental management.

To ensure high-quality and comprehensive labeling, A2Seek employs a rigorous multi-level annotation framework. This process includes spatiotemporal bounding boxes, fine-grained categories, and structured reasoning paths. A multi-stage annotation process was designed, encompassing trigger identification, anomaly diagnosis, causal reasoning, reflection, and seeking phases (as shown in Figure 1e). Specific formatting rules were enforced to align model-generated responses with human annotations. During annotation, the model produced multiple candidate explanations, from which professional annotators selected and refined the most appropriate. The final annotations are integrated into a reasoning-centric framework, enabling both precise anomaly localization and high-level semantic understanding. For privacy, all facial regions and identity-related visual attributes are blurred. Further procedural details are provided in Appendix D.

3.2 Dataset Characteristics and Comparative Analysis

Unlike conventional surveillance datasets, where anomalies often occupy the most salient portions of the frame, aerial-captured anomalies are typically small, scattered, and heavily context-dependent due to increased altitude, oblique angles, and limited focal length. As a result, the decisive information for anomaly detection lies not in the entire frame but in semantically rich yet spatially tiny regions, referred to as region-aware anomalies. The A2Seek dataset is specifically curated to address this challenging setting, providing a benchmark that closely mirrors real-world aerial inspection demands. Table 1 compares A2Seek with existing aerial and ground-based anomaly detection datasets. As the largest aerial anomaly detection dataset, A2Seek comprises over 2.4 million frames, including 398k frames containing anomalies, significantly surpassing other drone-view datasets in both scale and diversity. The dataset covers 10 major scene categories, such as roadways, sidewalks, playgrounds, and industrial zones. These categories are further divided into 27 specific sub-scenes, providing comprehensive coverage of complex real-world environments.

Beyond its scale and diversity, A2Seek is the first aerial anomaly detection dataset to provide natural language reasoning annotations with precise spatial localization labels. These annotations enable models to perform deep anomaly understanding and reasoning, bridging the gap between detection and interpretability. Furthermore, A2Seek explicitly addresses challenges unique to aerial perspectives, such as scene dependency, subtle anomalies, complex backgrounds, and scale variations, while also integrating multimodal data (RGB and infrared) to handle diverse environmental conditions. In

summary, A2Seek excels in data modalities, scene richness, and fine-grained annotations, directly addressing the two core challenges of aerial anomaly detection: spatial localization and semantic generalization. By establishing a robust data foundation and offering clear research directions, A2Seek sets a new benchmark for advancing aerial video anomaly detection and promoting deeper exploration in this domain. More details about our A2Seek data are reported in Appendix A.

4 Methodology

Anomaly detection from aerial perspectives poses unique challenges, requiring both precise spatial localization (“Where is the anomaly?”) and comprehensive semantic reasoning (“Why is it an anomaly?”). To address these challenges, we propose A2Seek-R1, a reasoning-centric framework specifically designed for aerial anomaly understanding. The framework integrates two key stages: (1) Supervised Fine-Tuning (SFT), which leverages Graph-of-Thought (GoT) annotations to activate the model’s initial reasoning capabilities by structuring reasoning into explicit semantic stages, and (2) Reinforcement Fine-Tuning (RFT), which employs Aerial Group Relative Policy Optimization (A-GRPO) to refine the model’s reasoning and localization strategies through task-specific reward functions. Among them, the seeking reward enables the model to dynamically identify regions requiring additional high-resolution context for further analysis, while the length reward encourages concise answers in simple scenarios and allocates more reasoning effort to complex and challenging scenes. Through A-GRPO-driven reinforcement fine-tuning, the model not only achieves accurate anomaly detection but also generates interpretable and verifiable reasoning traces, making it effective for real-world aerial applications.

STAGE 1: SFT-based Reasoning Activation. The first stage aims to activate the model’s initial reasoning capabilities through supervised fine-tuning (SFT) on the A2Seek dataset, specifically designed for aerial anomaly understanding. Each sample consists of an n -frame clip, an anomaly behavior label, and optional annotations such as different steps of Graph-of-Thought (GoT) reasoning or candidate bounding boxes. Missing fields are represented by a placeholder token `<NULL>`. To accommodate diverse supervision, we apply a binary mask over the target sequence, activating gradients only on annotated tokens. The supervised loss is defined as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,r,b,a) \sim \mathcal{D}_{\text{A2Seek}}} \sum_{t=1}^T m_t \log \pi_{\theta}(y_t \mid x, y_{<t}), \quad (1)$$

where $\mathcal{D}_{\text{A2Seek}}$ is the data distribution, x represents the input frames, r denotes the reasoning sequence, b refers to the bounding boxes, and a is the anomaly label. The binary mask m_t indicates whether the t -th token in the target sequence y is supervised ($m_t = 1$) or not ($m_t = 0$). The model’s output distribution $\pi_{\theta}(y_t \mid x, y_{<t})$ represents the conditional probability of generating the t -th token y_t , given the input x and the preceding tokens $y_{<t}$. The loss is computed over all supervised tokens in the sequence.

To enhance interpretability, we introduce a Graph-of-Thought (GoT) annotation (as shown in Figure 1e) that structures reasoning into explicit semantic stages, including visual signals (`<|Trigger|>`), anomaly diagnosis (`<|Diagnose|>`), causal explanation (`<|Reasoning|>`), confidence reflection (`<|Reflection|>`), and follow-up action (`<|Seeking|>`). Among these, the `<|Seeking|>` tag introduces a novel mechanism that enables the model to actively identify regions requiring additional high-resolution context, rather than merely localizing anomalies. Specifically, the model predicts a bounding box representing a potential region of interest, rather than a specific object, to be cropped and analyzed further. This mechanism mimics the behavior of UAVs actively adjusting their viewpoints to gather more detailed information. Unlike localization, which aims to pinpoint the exact position of an anomaly, seeking emphasizes identifying broader regions that warrant further inspection, enabling the model to effectively handle complex or ambiguous scenarios. These tags are injected into the target sequence and jointly trained under the same objective, guiding the model to generate coherent and verifiable reasoning traces. This structured supervision enables the model to generalize across varying annotation levels, from label-only samples to full reasoning and localization instances, providing a foundation for interpretable downstream adaptation.

STAGE 2: RL-based Reasoning Exploration. The second stage refines the model’s reasoning and localization strategies through reinforcement fine-tuning (RFT). This stage aims at addressing the dynamic complexity of aerial anomaly detection, where environmental conditions and scene

semantics vary significantly. Anomaly understanding could be formulated as a sequential decision-making problem and optimize the model using reinforcement learning. Following [58], we define the model’s output as a policy $\pi_\theta(y|x)$, which represents the model’s output distribution parameterized by θ . The objective is to maximize the expected reward:

$$J(\theta) = \mathbb{E}_{y \sim \pi_\theta(y|x)} [R(x, y)], \quad (2)$$

where $R(x, y)$ measures the quality of the output y for the given input x . The reward $R(x, y)$ is composed of multiple components, including format, accuracy, localization, seeking, and length rewards. Format and accuracy rewards constrain the model to produce outputs in the correct format and focus on prediction accuracy, while the localization reward evaluates the model’s judgment of anomaly evidence. The seeking and length rewards hierarchically guide the model to achieve dynamic reasoning paths. More detailed reward function designs are provided in Appendix B.

Building on this, we formulate reinforcement fine-tuning as a reward-weighted log-likelihood maximization problem with prior regularization. Specifically, let $\mathcal{X}$ denote the input space, which consists of sequences of aerial video frames, and $\mathcal{Y}$ denote the output space, which includes structured reasoning traces and anomaly prediction. A policy $\pi \in \Pi$ maps inputs $x \in \mathcal{X}$ to a distribution over outputs $y \in \mathcal{Y}$. The reinforcement fine-tuning objective is defined as:

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi} \mathbb{E}_{x \sim \mathcal{D}_{\text{A2Seek}}} \mathbb{E}_{y \sim \pi(y|x)} [R(x, y)] - \beta \cdot D_{\text{KL}}(\pi(y|x) \parallel \pi_{\text{ref}}(y|x)), \quad (3)$$

where π_{ref} is a reference policy. The KL divergence regularizes π to stay close to π_{ref} , with β controlling the trade-off between reward maximization and regularization. To implement this efficiently, we adopt Aerial Group Relative Policy Optimization (A-GRPO). Unlike existing reinforcement learning approaches [12, 41], our A-GRPO algorithm explicitly incorporates anomaly location and region seeking stages, making it particularly effective for aerial anomaly understanding. For each input x , K candidates $\{y^{(1)}, \dots, y^{(K)}\}$ are sampled from a policy group, and reward rankings are computed. The best-performing candidate is used to update the main policy via gradient ascent:

$$\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(y^{(k^*)}|x) \cdot \left(R(x, y^{(k^*)}) - b(x) \right), \quad (4)$$

where $k^* = \operatorname{argmax}_k R(x, y^{(k)})$, and $b(x)$ is a baseline, defined as the mean reward of the group, to reduce variance during optimization. This reflection-aware RFT stage leverages A2Seek’s comprehensive annotations to refine reasoning behaviors, enabling the model to generalize to unseen environments and adaptively revise predictions. More theoretical analysis on the effectiveness of our GoT data and seeking mechanism is discussed in the Appendix C.

5 Experiment

5.1 Implementation Details

We employ Qwen2.5-VL-3B-Instruct [4] as the base model due to its strong performance in vision-language understanding. We employed LoRA [22] to adjust all fully connected layers within the model, and set a learning rate of $1e-5$, using 4 epochs for SFT and 1 epoch for RL. Input resolution is set to 896×448 , with 4 frames per sequence. Batch size is 1, and gradient accumulation is performed over 16 steps. A cosine learning rate scheduler with a warm-up phase (5%) is used. All experiments are conducted in PyTorch on a platform with an Intel Xeon Platinum 8350 CPU, four NVIDIA A100 GPUs, and 1,024 GB of memory.

5.2 Overall Performance

To highlight the contributions of our dataset and method, we design the following experimental settings: (1) Zero-Shot, serving as a baseline without any fine-tuning; (2) RL-Zero, which applies GRPO with basic format and accuracy rewards; (3) ANS-SFT, utilizing anomaly labels for supervised fine-tuning; (4) CoT-SFT, incorporating chain-of-thought to guide inference; and (5) GoT-SFT, leveraging our graph-of-thought reasoning data to enable structured anomaly understanding. We evaluate all methods using category-level average precision (AP_c) for anomaly detection and mean Intersection over Union ($mIoU$) for localization, providing a balanced assessment of classification and grounding performance.

Table 2: Performance comparison across different scenarios on the A2Seek benchmark.

Method	Scene Indices										Average
	S00	S01	S02	S03	S04	S05	S06	S07	S08	S09	
SSRL [32]	11.50	10.20	8.30	9.40	20.10	12.70	5.80	14.20	16.80	11.00	12.00
HSTforU [28]	47.41	50.18	43.52	30.03	29.68	25.01	9.70	37.28	48.72	33.69	47.66
ANDT [26]	10.89	39.00	48.82	25.21	20.20	19.94	30.18	37.68	49.99	28.63	40.42
AnomalyRuler [69]	17.40	15.00	9.80	22.50	25.60	14.90	6.00	19.30	21.70	18.80	17.10
LAVAD [72]	12.80	15.20	11.00	19.50	24.70	20.50	4.30	19.10	22.10	14.90	16.41
Holmes-VAU [74]	17.00	18.50	16.00	23.40	25.00	20.50	13.34	21.12	27.20	17.60	19.97
LLavaVideo [37]	12.00	10.00	16.00	15.00	24.00	13.00	4.00	23.00	12.00	22.00	15.10
InternVL-3.0 [14]	35.62	49.75	48.21	47.08	27.43	23.94	72.40	29.64	36.99	42.33	41.34
Zero-shot [4]	44.83	44.92	35.15	23.88	22.42	19.16	62.80	30.05	29.44	27.62	34.03
RL-Zero	0.56	14.32	7.59	2.06	29.10	4.94	1.60	16.07	18.84	29.10	12.42
ANS-SFT	7.87	32.10	45.89	25.12	32.63	16.79	44.00	34.29	47.95	14.29	30.09
CoT-SFT	42.81	51.36	42.41	62.20	23.47	25.02	73.60	25.36	24.66	39.68	41.06
GoT-SFT	49.44	54.32	54.02	64.96	32.34	23.20	63.20	26.43	54.45	41.80	46.42
A2Seek-R1	51.12	56.54	55.36	75.60	40.84	26.86	81.60	38.21	61.30	43.92	53.14

Scene-wise Performance across Environments. As shown in Table 2, the proposed A2Seek-R1 consistently outperforms other methods in most cases, demonstrating its robustness and adaptability. Notably, in the *Playground* (S03) and *Rooftop* (S06) scenes, our method achieves the highest scores of 75.6% and 81.6%, respectively, significantly surpassing the second-best method. On average, A2Seek-R1 achieves a score of 53.14%, which is 19.11% higher than the baseline and 6.52% higher than GoT-SFT. These results highlight the effectiveness of our proposed method in aerial anomaly understanding and its strong generalization capability across diverse environments.

Table 3: Evaluation of anomaly detection and localization.

(a) Average precision and language metrics							(b) Localization performance						
Method	AP_c	BLEU	METEOR	ROUGE	CIDEr	Samples/s	Method	mIoU	$AP_{0.00}$	$AP_{0.25}$	$AP_{0.50}$	$AP_{0.75}$	$AP_{0.90}$
Zero-Shot	34.03	0.2835	0.2145	0.3263	0.8147	0.6097	Zero-Shot	3.50	72.40	3.94	1.81	0.63	0.00
ANS-SFT	31.10	0.3821	0.3452	0.4379	1.0462	0.9174	ANS-SFT	17.05	50.70	40.01	31.94	15.74	0.70
GoT-SFT	46.42	0.4478	0.3325	0.4439	1.1528	0.2890	GoT-SFT	20.81	43.85	35.88	27.60	16.01	1.03
A2Seek-R1	53.14	0.4564	0.3543	0.4882	1.2989	0.3267	A2Seek-R1	26.03	53.31	45.43	35.33	20.11	4.34

Language Semantic Evaluation. To evaluate the ability of models to handle semantic ambiguity and similarity in aerial anomaly understanding, we adopt language-based metrics such as BLEU [50], METEOR [5], ROUGE [36], and CIDEr [63]. These metrics capture both lexical and structural alignment between predicted and ground-truth descriptions. As shown in Table 3a, A2Seek-R1 achieves the highest scores across all metrics, with BLEU of 0.4564, METEOR of 0.3543, ROUGE of 0.4882, and CIDEr of 1.2989. The throughput is 0.3267 samples per second, which is lower than ANS-SFT (0.9174) and Zero-Shot (0.6097) because our method allocates a larger token budget at test time for reasoning, leading to better anomaly understanding. These results demonstrate its ability to generate accurate and semantically meaningful descriptions even under ambiguous scenarios. By leveraging visual evidence during reasoning, A2Seek-R1 ensures that its language-based predictions are contextually grounded, effectively bridging the gap between semantic understanding and visual perception.

Supporting Region Grounding Accuracy. To further evaluate whether the model relies on visual information to understand anomalies, we assess its localization performance using both mIoU and AP under varying IoU thresholds. As shown in Table 3b, A2Seek-R1 achieves the highest mIoU score of 26.03%, outperforming Zero-Shot and GoT-SFT by 22.53% and 5.22%, respectively. Notably, although Zero-Shot and ANS-SFT achieve relatively high AP values in $AP_{0.00}$, this is mainly attributed to their tendency to predict overly broad and vague regions, rather than accurately indicating the anomalous areas. This limitation is further reflected in their significantly degraded performance (nearly zero in $AP_{0.90}$) at higher IoU thresholds, indicating a tendency toward random guessing. These results confirm that the spatial reasoning mechanism introduced in our approach enhances the model’s ability to precisely localize subtle and ambiguous anomalies, thereby improving both detection accuracy and interpretability for real-world aerial anomaly understanding.

Data Efficiency and Generalization. The in-domain (ID) data in A2Seek includes scenarios present in the training set, while the out-of-domain (OOD) data comprises unseen scenarios, specifically designed to assess the model’s generalization capability in novel and challenging settings. As shown in Figure 3, A2Seek-R1 demonstrates both high data efficiency on in-domain data and strong generalization to out-of-domain scenarios. On ID data (Figure 3b), A2Seek-R1 achieves consistent performance improvements throughout training, whereas GoT-SFT shows stagnation or minor fluctuations. Notably, on OOD data (Figure 3c), A2Seek-R1 continues to gain accuracy as training

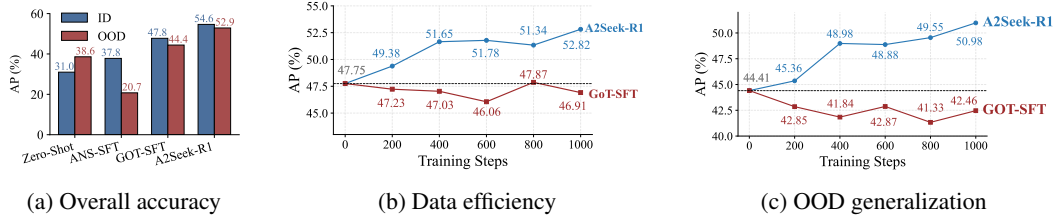


Figure 3: Performance comparison of different settings on A2Seek benchmark.

progresses, reaching 48.98% at 400 steps, a relative improvement of 4.57%. In contrast, GoT-SFT degrades from 44.41% to 41.84% over the same period. This observation suggests that A2Seek-R1 not only leverages training data more effectively but also exhibits better robustness under domain shifts.

Table 4: Ablation studies of the A2Seek-R1 framework.

Idx	Supervised Fine-Tuning			Reinforcement Learning			Metric	
	ANS	CoT	GoT	Accuracy Reward	Location Reward	Seeking Reward	AP_c	$mIoU$
0	✓						31.10	12.13
1	✓	✓					38.24	17.05
2	✓		✓				46.42	20.81
3				✓		✓	12.42	9.66
4	✓		✓	✓			52.82	18.77
5	✓		✓	✓	✓		51.78	24.03
6	✓		✓	✓	✓	✓	53.14	26.03

5.3 Ablation Studies

Table 4 presents the results of ablation studies on the A2Seek dataset, analyzing the contributions of supervised fine-tuning (SFT) strategies and reinforcement learning (RL) reward components.

Impact of Supervised Fine-Tuning on A2Seek. Table 4 (Rows 0-2) evaluate the effect of different SFT strategies: ANS, CoT, and GoT. Using only ANS (Row 0) achieves a baseline AP_c of 31.10% and $mIoU$ of 12.13%. Incorporating CoT (Row 1) improves AP_c to 38.24% and $mIoU$ to 17.05%, demonstrating the benefit of chain-of-thought reasoning. Replacing CoT with GoT (Row 2) further boosts AP_c to 46.42% and $mIoU$ to 20.81%, highlighting the effectiveness of goal-oriented reasoning in establishing a stronger foundation for anomaly detection and understanding.

Impact of Reinforcement Learning Rewards. Rows 3-6 progressively incorporate RL reward components, including accuracy, location, and seeking rewards. Using only accuracy and location rewards (Row 5) achieves AP_c of 51.78% and $mIoU$ of 24.03%. Adding the seeking reward (Row 6) further improves AP_c to 53.14% and $mIoU$ to 26.03%, achieving the best overall performance. These results demonstrate that the synergistic integration of all three reward components is critical for enhancing both detection accuracy and localization precision.

Table 5: Effectiveness of A2Seek-R1 on different base models.

Method	AP_c (%)	ΔAP (%)	$mIoU$ (%)	$\Delta mIoU$ (%)
QwenVL-2.5-7B-Instruct	44.65	—	08.24	—
+ A2Seek-R1	57.07	+12.42	29.71	+21.47
InternVL-3.0-2B-Instruct	20.03	—	00.10	—
+ A2Seek-R1	48.76	+28.73	18.89	+18.79

Impact of Foundation Models. As shown in Table 5, we evaluate the generalization of A2Seek-R1 on two additional vision-language models beyond the base model, *i.e.*, QwenVL-2.5-3B-Instruct [4]. On QwenVL-2.5-7B-Instruct [4], it improves AP by 12.42% and mIoU by 21.47%, indicating effective scaling with capacity. On InternVL-3.0-2B-Instruct [14], it increases AP from 20.03% to 48.76% and mIoU from 0.10% to 18.89%. These results demonstrate consistent gains across model sizes and architectures, supporting strong generalization and transferability.

5.4 Qualitative Visualization

As illustrated in Figure 4, traditional deep neural networks (DNNs) provide only binary classifications (normal or abnormal) without explanations or spatial context. Multimodal large language

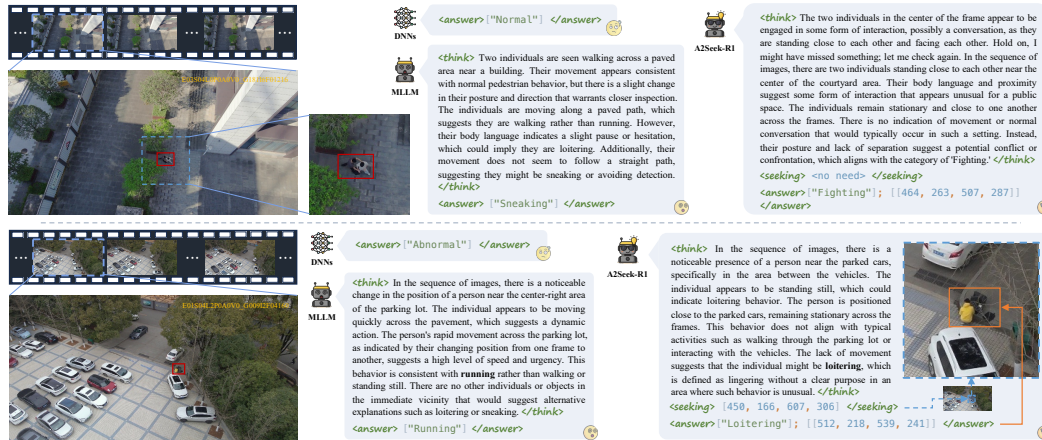


Figure 4: Qualitative results of A2Seek-R1. Beyond predicting anomaly categories, our method provides reasoning traces and accurately localizes the key regions that support its judgment.

models (MLLMs) improve upon this by generating textual descriptions to explain anomalies but lack the ability to localize specific regions, limiting their utility in complex aerial scenarios. In contrast, A2Seek-R1 combines detailed reasoning with precise spatial localization. By reflecting on individuals' body language and proximity, the model iteratively refines its understanding, identifying interactions such as "Fighting" and providing bounding box coordinates to substantiate its reasoning. Additionally, A2Seek-R1 mimics UAV flight behavior by dynamically focusing on regions of interest, enabling adaptive refinement of predictions and enhancing both detection accuracy and localization precision.

6 Conclusions

This work introduces A2Seek, a large-scale, reasoning-centric dataset for aerial anomaly understanding. Built upon diverse aerial scenarios, A2Seek is meticulously annotated with fine-grained anomaly labels, spatiotemporal bounding boxes, and structured reasoning graphs. These annotations enable models to not only detect anomalies but also provide interpretable explanations for "why" and "where" anomalies occur. Building on this benchmark, we introduce A2Seek-R1, a novel multi-stage reinforcement fine-tuning framework that integrates graph-of-thought (GoT)-guided supervised fine-tuning, and an improved A-GRPO algorithm with tailored reward functions, including a seeking mechanism that simulates UAV flight behavior. This framework achieves state-of-the-art performance in both accuracy and interpretability, excelling in complex anomaly scenarios and demonstrating strong cross-domain generalization capabilities.

Limitations and Societal Impact. Despite strong performance, the complexity of reasoning chains and strategy learning limits inference speed and deployment efficiency. The framework mainly focuses on spatial-level anomaly perception, whereas the understanding and modeling of long-duration temporal anomalies are still limited, which we regard as an important avenue for future exploration. Additionally, given its potential applications in sensitive areas like public safety, privacy protection and algorithmic fairness must be prioritized to avoid societal risks from misjudgments.

Acknowledgments and Disclosure of Funding. This work was supported in part by the Science and Technology Innovation Key R&D Program of Chongqing under Grant No. CSTB2023TIAD-STX0016, in part by the National Natural Science Foundation of China under Grants No. 62472060, U23A20318, and 62221005, in part by the Natural Science Foundation of Chongqing under Grants No. CSTB2024NSCQ-QCXM0060 and CSTB2023NSCQ-LZX0061, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJZD-K202300604, in part by the Chongqing Doctoral Student Innovation Talent Project under Grant No. CYB240241, in part by the Chongqing University of Posts and Telecommunications Ph.D. Innovative Talents Project under Grant No. BYJS202404, and in part by the Chongqing Institute for Brain and Intelligence.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022.
- [3] 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.
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [6] Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lms, 2024.
- [7] Margherita Bonetto, Pavel Korshunov, Giovanni Ramponi, and Touradj Ebrahimi. Privacy in mini-drone based video surveillance. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 4, pages 1–6. IEEE, 2015.
- [8] Ilker Bozcan and Erdal Kayacan. Context-dependent anomaly detection for low altitude traffic surveillance. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 224–230. IEEE, 2021.
- [9] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20392–20401, 2023.
- [10] Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, et al. Visdrone-det2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 2847–2854, 2021.
- [11] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Computer Vision-ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pages 136–153. Springer, 2017.
- [12] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.
- [13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pages 24185–24198, 2024.
- [15] Bruno Degardin and Hugo Proença. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters*, 145:50–57, 2021.
- [16] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018.
- [17] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18793–18803, 2024.

- [18] Zhitong Gao, Shipeng Yan, and Xuming He. Atta: anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36:45150–45171, 2023.
- [19] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.
- [20] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.
- [21] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023.
- [22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [23] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- [24] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024.
- [25] Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought, 2025.
- [26] Pu Jin, Lichao Mou, Gui-Song Xia, and Xiao Xiang Zhu. Anomaly detection in aerial videos with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
- [27] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2921–2928. IEEE, 2009.
- [28] Viet-Tuan Le, Hulin Jin, and Yong-Guk Kim. Hstforu: anomaly detection in aerial and ground-based videos with hierarchical spatio-temporal transformer for u-net. *Applied Intelligence*, 55(4):261, 2025.
- [29] Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, and Xinbo Gao. Beyond euclidean: Dual-space representation learning for weakly supervised video violence detection. *arXiv preprint arXiv:2409.19252*, 2024.
- [30] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.
- [31] Feiyang Li, Peng Fang, Zhan Shi, Arijit Khan, Fang Wang, Dan Feng, Weihao Wang, Xin Zhang, and Yongjian Cui. Cot-rag: Integrating chain of thought and retrieval-augmented generation to enhance reasoning in large language models. *arXiv preprint arXiv:2504.13534*, 2025.
- [32] Guoqi Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022.
- [33] Jing Li, Qingwang Huang, Yingjun Du, Xiantong Zhen, Shengyong Chen, and Ling Shao. Variational abnormal behavior detection with motion consistency. *IEEE Transactions on Image Processing*, 31:275–286, 2021.
- [34] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.
- [35] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.
- [36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

- [38] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [39] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.
- [40] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021.
- [41] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [42] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.
- [43] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.
- [44] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15425–15434, 2021.
- [45] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.
- [46] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023.
- [47] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.
- [48] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition*, pages 935–942. IEEE, 2009.
- [49] Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18868–18877, 2024.
- [50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [51] R. Raghavendra, A. D. Bue, and M. Cristani. Unusual crowd activity dataset of university of minnesota. Tech. Rep. / Online Dataset, University of Minnesota, 2006. Available: <http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi>.
- [52] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2569–2578, 2020.
- [53] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.
- [54] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.
- [55] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

- [56] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.
- [57] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.
- [58] Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*, 2025.
- [59] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies, 2024.
- [60] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022.
- [61] Tung Minh Tran, Tu N Vu, Tam V Nguyen, and Khang Nguyen. Uit-drone: A novel drone dataset for traffic anomaly detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:5590–5601, 2023.
- [62] Yuanpeng Tu, Yuxi Li, Boshen Zhang, Liang Liu, Jiangning Zhang, Yabiao Wang, and Cairong Zhao. Self-supervised likelihood estimation with energy guidance for anomaly segmentation in urban scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21637–21645, 2024.
- [63] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [64] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 15(14):3454–3465, 2021.
- [65] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.
- [66] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020.
- [67] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.
- [68] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [69] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- [70] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14592–14601, 2023.
- [71] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22052–22061, 2024.
- [72] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024.
- [73] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

- [74] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024.
- [75] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011.
- [76] Wenjie Zhao, Jia Li, Xin Dong, Yu Xiang, and Yunhui Guo. Segment every out-of-distribution object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3910–3920, 2024.
- [77] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025.
- [78] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *arXiv preprint arXiv:2402.04857*, 2024.
- [79] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 89943–89977. Curran Associates, Inc., 2024.

NeurIPS Paper Checklist

1. Claims

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

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

Justification: The abstract and introduction state that the paper (i) releases A2Seek, a large-scale, multimodal UAVview dataset with fine-grained temporal, spatial and reasoning annotations, and (ii) proposes A2Seek-R1, a multi-stage reinforcement fine-tuning framework with tailored reward functions that markedly boosts detection accuracy and interpretability; they also emphasise the novelty of simulating UAV motion via a seeking mechanism and report consistent performance gains across in-domain and out-of-distribution tests. These claims are precise, limited to the datasets scope and the demonstrated algorithmic improvements, and they align with the detailed contributions and experimental evidence presented later in the paper, so they accurately reflect the works contributions and boundaries.

Guidelines:

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

2. Limitations

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

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

Justification: The section titled "Discussion" and the appendix section titled "Limitations, Open Challenges and Future Directions" explicitly outlines the current shortcomings.

Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

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

Answer: [NA]

Justification: The manuscript's technical content is centred on releasing the reasoning-centric A2Seek dataset and designing an RL-based fine-tuning framework.

Justification: The manuscript's technical content is centred on releasing the A2Seek dataset and designing an RL-fine-tuning framework; although it introduces several scoring functions, reward equations and upper/lower-bound inequalities to motivate the algorithm (*e.g.*, the reflection-reward inequality and the seeking bound analysis), these expressions are illustrative rather than formal theorems none are numbered, no complete set of assumptions is stated, and no rigorous proofs are provided in the main text or appendix. Experimental benchmarking that shows the method works belongs to empirical validation, not to theoretical results. Therefore the checklist item on Theory assumptions and proofs should be marked NA, indicating that the paper does not claim formal theoretical results requiring full proofs.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: Section 5 (*Experiment*) discloses every critical element needed to replicate the study.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (*e.g.*, in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

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

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

Justification: The appendix makes the release plan explicit: all annotations are “released as COCO-VID-style JSON files” and “the final dataset and its metadata are distributed in read-only form under a click-through licence that forbids re-identification”; the same supplemental bundle provides the annotation GUI, data-handling scripts and evaluation code together with step-by-step commands, so any group can download the data, run the scripts and exactly reproduce the results.

Guidelines:

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

6. Experimental setting/details

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

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

Justification: The main papers Implementation Details lists the foundation model, fine tuning method, down-sampling input resolution (896×448), optimiser, learning rate schedule, weight-decay and *etc.*

Guidelines:

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

7. Experiment statistical significance

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

Answer: [No]

Justification: This paper follows standard experimental practices in computer vision, reporting average performance using widely adopted metrics. While explicit error bars or statistical significance tests are not included, the reported results are consistent across benchmarks and reproducible under the same settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [Yes]

Please see Sec. 5 ("Experiment Detail"), where lists the GPU hardware configuration as four NVIDIA A100 40 GB GPUs for running all experiments.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: We have carefully reviewed and adhered to the NeurIPS Code of Ethics: all UAV videos were captured in public spaces with local authority approval, all individuals involved in the data collection, including actors and annotators, received appropriate compensation in accordance with institutional and legal guidelines, personally identifying elements (faces, licence plates) are automatically blurred before release, no sensitive attributes were inferred, and the dataset licence explicitly forbids re-identification or military use; thus the research respects privacy, fairness, safety and legal compliance in every aspect.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The manuscript devotes a titled subsection Limitations and Societal Impact in the discussion part, together with an earlier Privacy Preservation, Licensing and Ethical Compliance appendix, to address both sides of the societal ledger. On the positive side, it argues that releasing a richly annotated 4K UAV-video corpus will advance research on generalisation, robustness and interpretability in aerial anomaly detection and thereby strengthen public-safety applications in complex campus environments. On the risk side, it recognises privacy threats and possible misuse for surveillance; concrete mitigations are spelled out every frame passes an offline sanitisation pipeline that irreversibly scrambles faces and licence plates, and dataset access is gated by a click-through, non-commercial licence that forbids re-identification or military use. By explicitly balancing benefits against hazards and describing safeguards, the paper satisfies the checklist requirement to discuss both potential positive and negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

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

Justification: The paper provides detailed descriptions of the safeguards implemented, including irreversible anonymization of sensitive information, restricted access to the dataset/-model, and the use of a click-through license agreement to prevent unauthorized use or misuse.

Guidelines:

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

12. Licenses for existing assets

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

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

Justification: All third-party assets are properly credited and their terms respected.

Guidelines:

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

13. New assets

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

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

Justification: The new A2Seek dataset is distributed with a structured README, JSON metadata schema, licence, collection protocol and usage examples (see Appendix E), so all new assets are properly documented and released together with their documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

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

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

Justification: All individuals involved in the data collection, including actors and annotators, received appropriate compensation in accordance with institutional and legal guidelines.

Guidelines:

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

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

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

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

Justification: The work does not involve crowdsourcing or research with human subjects; therefore IRB approval is not applicable.

Guidelines:

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

16. Declaration of LLM usage

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

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

Justification: This work employs a multimodal large language model to assist in the data annotation process and also uses a multimodal LLM as a backbone model in the experimental design.

Guidelines:

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

A Data Collection, Annotation and Statistics

A.1 Dataset Acquisition Protocol

To address the challenges of spatial localization and scene generalization, the data collection process was carefully designed to maximize diversity and realism. The A2Seek dataset was collected using a DJI M30T drone equipped with wide-angle, telephoto, and infrared cameras. The drone operated at varying altitudes (10 to 60 meters) and speeds (0 to 20 m/s) to capture a wide range of scale variations for anomalous objects. Flight trajectories included hovering, linear cruising, curved circling, and area scanning, reflecting dynamic viewpoint changes. The dataset comprises 542 untrimmed 4K videos recorded at 30 FPS across 10 campus environments, subdivided into 27 sub-scenes. Each video has an average duration of 153.74 seconds, with a total duration exceeding 23 hours. Among these, 19.3 hours feature normal behaviors, while 3.79 hours capture diverse and complex anomalies, including parallel, sequential, occluded, and scene-dependent events. As shown in Figure 5, our dataset includes over 20 anomaly categories, such as falling, fighting, playing with water, and jaywalking. Spanning nearly a year of recording, it covers various seasons, times of day (daytime, nighttime, twilight), and environmental conditions, including clear, cloudy, foggy, and rainy weather (as shown in Figure 6). This diversity provides a robust foundation for developing and evaluating aerial anomaly detection models. Table 6 presents a unified overview of all categorical codes used in this paper. The *Type* column groups related codes, the *Code* column lists shorthand labels, and the *Definition* column provides detailed descriptions.

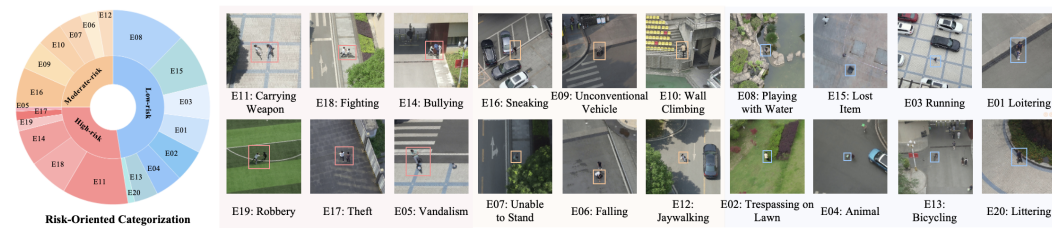


Figure 5: Representative anomaly types in the A2Seek dataset. Our dataset covers a broad spectrum of anomalous behaviors across different risk levels, highlighting the diversity and complexity of aerial anomaly detection.



Figure 6: Examples of scenes from the A2Seek dataset.

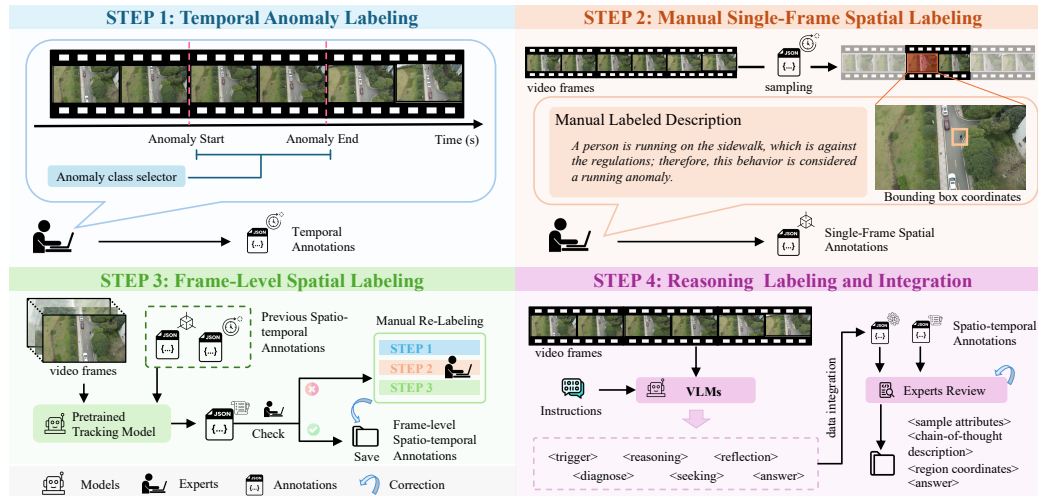


Figure 7: Fourstage annotation workflow. *Step 1* (blue) Temporal annotators use our in-house GUI to mark the start/end frames and class of every anomalous episode, exporting a JSON timeline. *Step 2* (salmon) For the first frame of each event, experts draw a bounding box around the anomalous region and supply a natural-language description, yielding single-frame spatial seeds. *Step 3* (green) A pretrained tracker propagates each seed through the clip to form full-length trajectories; an automated checker screens the results, flags uncertain cases for human correction, and funnels all approved tracks into the spatial-label repository. *Step 4* (violet) Vision-language models (VLMs) ingest the temporal tags, spatial tracks, and human captions; via chain-of-thought reasoning, they merge these cues into consolidated frame-level annotations, producing the final label set.

Our collection workflow for A2Seek streamlines a campus drone survey into a single, reproducible loop while preserving fine-grained control:

- **Scene & Time:** ten functional zones, further divided into 27 sub-scenes, are revisited across four seasons, three day parts (twilight: 06:00 to 08:00 or 16:00 to 19:00, daylight: 10:00 to 16:00, night: 19:00 to 22:00) and five weather classes, producing initial multi-modal videos.
- **Airframe & Optics:** a DJI M30T (1-inch CMOS, 20 MP, 3,840×2,160@30 fps) plus a co-aligned 640×512 thermal sensor for nocturnal sorties; the on-board GNSSIMU logs pose at 50 Hz.
- **Flight Envelope:** grid missions with 80% side and 70% forward overlap (for routine coverage), dynamically switchable to follow or orbit mode for rare anomalies, at altitudes of 10–60 m above ground level (≤ 30 m at night) and ground speeds of 0–15 m/s.
- **Data Handling:** RGB, thermal and telemetry streams are recorded to a 512 GB UHS-II SD card, then mirrored to an offline workstation upon landing; filenames are encoded with UTC timestamps and scene IDs for instant indexing.
- **Quality Gate:** all sorties are checked daily by two senior annotators for focus, exposure and occlusion. Flights flagged for privacy concerns or other disqualifying conditions are scheduled to be re-flown.

This lean yet rigorous protocol provides high-resolution, context-rich footage tailored to drone-view anomaly detection while keeping field operations swift and fully repeatable.

A.2 Annotation Workflow and Quality Assurance

Dataset Annotation. The A2Seek dataset employs a rigorous multi-level annotation framework to ensure high-quality and comprehensive labeling. As illustrated in Figure 7, the annotation process begins with manually labeling anomaly categories and identifying relevant regions in keyframes,

Table 6: Unified code definitions for scenes, actions, parameters, weather conditions, and risk levels.

Type	Code	Definition
Scene	S00	Miscellaneous: other or unspecified scenes
	S01	Roadway: street and vehicular paths
	S02	Sidewalk: pedestrian walkways along roads
	S03	Playground: open recreational areas
	S04	Open Area: unstructured outdoor space
	S05	Park: landscaped public green spaces
	S06	Rooftop: building rooftop areas
	S07	Entrance: school gate or entryway region
	S08	Wall Zone: perimeter walls and boundaries
	S09	Academic Building: building facade and courtyard
Action	E00	Normal: no anomalous activity
	E01	Loitering: prolonged stationary presence near entrances or perimeter areas
	E02	Trespassing on Lawn: entering grass areas illegally
	E03	Running: fast movement on foot in restricted or non-sport areas
	E04	Animal: appearance of animals within the scene
	E05	Vandalism: intentional damage to property
	E06	Falling: loss of balance leading to a fall
	E07	Unable to Stand: person unable to recover after a fall
	E08	Playing with Water: interacting with fountains or water bodies
	E09	Unconventional Vehicle: use of nonstandard or unauthorized vehicles
	E10	Wall Climbing: scaling vertical surfaces
	E11	Carrying Weapon: person holding or drawing a weapon
	E12	Jaywalking: crossing road outside crosswalk
	E13	Bicycling: riding a bicycle in pedestrian or restricted areas
	E14	Bullying: aggressive behavior towards others
	E15	Lost Item: dropping or leaving behind personal belongings
	E16	Sneaking: moving stealthily or furtively
	E17	Theft: unauthorized taking of property
	E18	Fighting: physical aggression between persons
	E19	Robbery: theft involving force or threat
	E20	Littering: discarding waste improperly in public areas
Height	H0	10–20 m (Low altitude flight)
	H1	20–50 m (High altitude flight)
Velocity	M0	0–3 m/s (Low speed)
	M1	3–10 m/s (Moderate speed)
	M2	10–20 m/s (High speed)
Time of Day	L0	10:00–16:00 (Daytime)
	L1	08:00–10:00, 16:00–18:00 (Morning/Evening)
	L2	18:00–22:00 (Evening/Night)
Weather	W0	Clear: no precipitation or clouds
	W1	Cloudy: partial cloud cover
	W2	Overcast: full cloud cover
	W3	Rain: rainfall conditions
	W5	Fog: reduced visibility due to fog
	W8	Night: after sunset until pre-dawn
Risk Level	High-risk	Actions likely to cause serious harm or property damage (E05, E11, E14, E17, E18, E19)
	Moderate-risk	Actions that may result in moderate safety concerns or disruption (E06, E07, E09, E10, E12, E16)
	Low-risk	Actions considered minor with minimal safety risk (E00, E01, E02, E03, E04, E08, E13, E15, E20)

which are then extended to adjacent video frames using Grounded-SAM-2 [53, 54]. These annotations undergo cross-verification by human annotators to ensure consistency and accuracy. Based on this initial information, prompts were designed to guide a vision-language model in generating detailed annotations, including region-level, frame-level, and segment-level information. The model-generated annotations were further reviewed and refined by professional annotators to ensure alignment with human-labeled categories.

To enhance the reliability of the annotations, a multi-stage process was designed, encompassing trigger, diagnose, reasoning, reflection, and seeking phases. Specific formatting rules were employed to align the model outputs with human annotations, ensuring consistency and interpretability. During this process, the model was guided to generate multiple responses, from which professional annotators selected the most appropriate ones for further refinement. The finalized annotations were integrated into a reasoning-centric framework, providing structured explanations that include spatial localization, fine-grained category labels, and causal reasoning paths. This process ensures that the dataset not only supports precise anomaly detection but also facilitates in-depth semantic reasoning. The annotation framework includes precise timestamps for each anomalous event, identifying the specific frame sequences where anomalies occur. Spatial localization is achieved through bound-

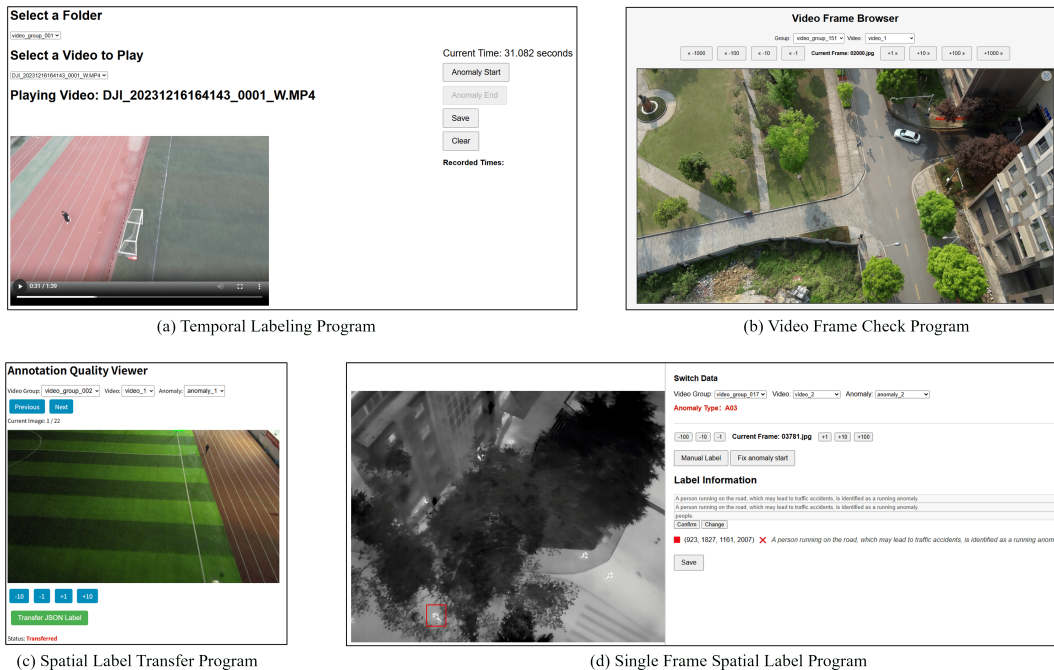


Figure 8: Screenshots of our in-house annotation toolkit. **(a)** Temporal labeling interface: annotators record only the start frame, end frame and anomaly category for each event. **(b)** Frame browser: allows rapid navigation to any frame in a clip so that timestamp alignment between raw videos and labels inside the same *video group* can be visually verified. **(c)** Label-transfer panel: human validators inspect SAM-generated masks and copy only approved instances into the final ground-truth store, filtering out noisy or redundant versions. **(d)** Single-frame spatial labelling: experts draw a bounding box for every anomaly, attach a free-text description and supply key-words that drive the subsequent Grounding-SAM tracking step.

ing boxes that accurately mark the positions of anomalous objects, ensuring high precision and reliability in anomaly detection. Additionally, natural language explanations describe the causes and contextual background of anomalies, forming dynamic reasoning paths based on visual scene content. To maintain data quality, low-quality videos were removed, and privacy-sensitive information, such as faces and license plates, was processed to ensure compliance with ethical standards. This meticulous annotation process ensures that the A2Seek dataset provides a robust foundation for developing and evaluating aerial anomaly detection models in complex scenarios. Fifteen domain specialists spent roughly one month creating the A2Seek labels. The pipeline concentrates on three elements *temporal boundaries*, *spatial trajectories*, and *textual descriptions* implemented with lightweight in-house tools (as shown in Figure 8) and a round-robin verification scheme. All outputs are released as COCO-VID-style JSON files.

Temporal Boundaries. We built a browser-based tool that lets an annotator scrub through each clip and press *Start/End* buttons while selecting one of the 20 anomaly classes. Every video is labelled independently by two experts; a third reviewer resolves mismatches. The final time-stamps are exported to `temporal_labels.json` together with the flight metadata recorded on location.

Spatial Trajectories. For the first frame of each temporally marked event, an expert draws a bounding box and writes a short noun phrase (e.g., “Bicycle”). Two complementary tracking modes are then launched:

- BBox-only: the seed box is propagated frame-by-frame with a pretrained tracker.
- Keyword-guided: Grounding DINO (Grounding DETection with Improved deNoising anchor boxes)[73] detects all boxes matching the phrase; the one with highest IoU to the seed is chosen, and Grounding-SAM (Grounding Segment Anything Model)[55] refines and tracks the mask.

The two results are merged, and a second annotator scrolls through the track to accept or flag it. Flagged cases are re-seeded or re-edited until they pass inspection. Approved trajectories are stored in `spatial_labels.json`.

Textual Descriptions. Grounding-SAM’s <Caption> head produces frame-level and clip-level captions. In addition, annotators provide a multi-steps description for every anomaly to capture intent and context. Captions are lightly proof-read before being written to `text_labels.json`.

Unlike CoT-RAG [31], which models factual relationships between knowledge units, GoT focuses on structuring the *reasoning trajectory*. This design allows adaptive control of reasoning depth in aerial-view anomaly understanding, where scene complexity and anomaly subtlety vary widely.

Graph-of-Thought (GoT) explicitly structures reasoning through **stage-specific tags** inserted into the generated sequence: <|Trigger|>, <|Diagnose|>, <|Reasoning|>, <|Reflection|>, <|Seeking|>.

Each stage represents a distinct cognitive operation:

- **Trigger:** Scene perception and anomaly signal activation.
- **Diagnose:** Hypothesis formation for possible anomaly categories.
- **Reasoning:** Contextual analysis and evidence-driven judgment.
- **Reflection:** Self-evaluation or consistency check of reasoning.
- **Seeking:** Optional localization or reference search for key regions.

Each stage is *skippable*, allowing GoT to shorten reasoning chains for simple scenes and extend them for complex ones. We provide below two representative GoT annotations illustrating both simple and complex reasoning cases.

GoT Annotation Example (a): Simple Scene (Normal)

```
{
  "trigger": "No individuals or objects are present in the scene. The environment consists of a running track, a soccer field, and an adjacent area with no visible activity or anomalies.",
  "diagnose": ["Normal"],
  "reasoning": "<no need>",
  "reflection": "<no need>",
  "seeking": "<no need>"
}
```

GoT Annotation Example (b): Complex Scene (Lost Item)

```
{
  "trigger": "In the sequence of images, there is a noticeable change in the lower left area of the frame. Specifically, a person appears to be walking along a pathway near the edge of the lawn. This individual's presence was not evident in the initial frames but becomes apparent as the sequence progresses.",
  "diagnose": "Possible anomaly categories include 'Trespassing on Lawn' or 'Lost Item'. The appearance of the person near the edge of the lawn could indicate an unusual presence, especially if they are not supposed to be in that area. Alternatively, the person might have dropped something, which would classify as a 'Lost Item'.",
  "reasoning": "The person is seen moving along a designated pathway, which suggests normal pedestrian behavior. However, their proximity to the lawn edge raises suspicion. Upon closer inspection, there is no direct indication of the person interacting with the lawn itself, such as stepping onto it. Instead, the focus shifts to the possibility of a lost item, as the person's movement does not suggest any abnormal interaction with the environment beyond walking.",
  "reflection": "Given the lack of direct interaction with the lawn and the person's adherence to the pathway, the most plausible explanation is that the person has dropped an item. This aligns with the category 'Lost Item', as the person's movement pattern does not suggest trespassing but rather a search for something they may have misplaced.",
  "seeking": "[270, 170, 582, 447]"
}
```

These examples illustrate how GoT represents reasoning as a *semantic graph of thought stages*, bridging perception and interpretation in an interpretable and controllable manner.

A.3 Dataset Features

Anomaly Duration The anomalous events in the A2Seek dataset can be divided into three categories according to their duration: long-term, short-term, and instantaneous anomalies. Long-term anomalies are generally difficult to identify from a single frame and require extended contextual information for accurate judgment; such anomalies are often somewhat predictable, for example, prolonged loitering in restricted areas. Short-term anomalies last for a shorter period and typically require certain prior knowledge for detection, though some can still be recognized from individual frames; for instance, running across a pedestrian walkway may constitute a short-term anomaly. Instantaneous anomalies occur extremely rapidly, usually within only a few frames, making them difficult to detect. For instance, in theft scenarios, the entire action may occur in an instant, thereby complicating the identification process.

	S00	S01	S02	S03	S04	S05	S06	S07	S08	S09	Sum
W0	6	31	23	22	52	32	5	13	17	0	201
W2	12	9	5	35	37	25	2	5	2	0	132
W3	0	18	1	0	8	29	5	8	0	0	69
W1	2	3	3	18	14	1	0	5	0	1	47
W5	0	0	0	0	0	0	0	0	0	6	6
Sum	20	61	32	75	111	87	12	31	19	7	455

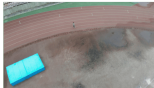
			
W0 Clear	W1 Cloudy	W2 Overcast	W3 Rainy
			
W5 Froggy	L0 Morning	L1 Afternoon	L2 Night

Figure 9: Summary of weather, time, and scene types in our dataset. W0, W1, W2, W3, and W5 represent clear, cloudy, overcast, rainy, and foggy conditions, respectively, while S00 to S09 denote various scenes such as pathways and courtyards. The dataset covers recordings from morning, noon, and afternoon, excluding nighttime due to weather capture limitations. Out of 542 total videos, only 455 are included in the analysis.

Scene-dependent Anomaly Scene-dependent anomalies are a significant feature of the A2Seek dataset, encompassing behavior anomalies that depend on specific scene contexts and occlusion anomalies caused by environmental obstructions. Scene-dependent anomalies can be categorized into two types: cross-video-segment scene-dependent anomalies, where a behavior that appears normal in one segment may be anomalous in another (as shown in Figure 10), and drone-specific scene-dependent anomalies, influenced by dynamic viewpoint shifts within the same segment (as shown in Figure 11). Occlusion anomalies also frequently occur in drone views due to obstructions such as trees or buildings, posing unique challenges that require models to infer anomalies by leveraging temporal cues from preceding and succeeding frames. These complexities make drone-view anomaly detection substantially different from traditional surveillance settings, highlighting the importance of context-aware analysis.

Serial and Parallel Anomalies Our dataset encompasses both serial anomalies and parallel anomalies. Serial anomalies refer to multiple anomalies occurring consecutively within the same time period, often exhibiting causal relationships, *i.e.*, the occurrence of one anomaly triggers another. In contrast, parallel anomalies involve multiple anomalies occurring simultaneously within the same time period. These anomalies are generally independent of one another, lacking any causal connections. Figure 12 demonstrate the serial and parallel anomaly in A2Seek dataset.

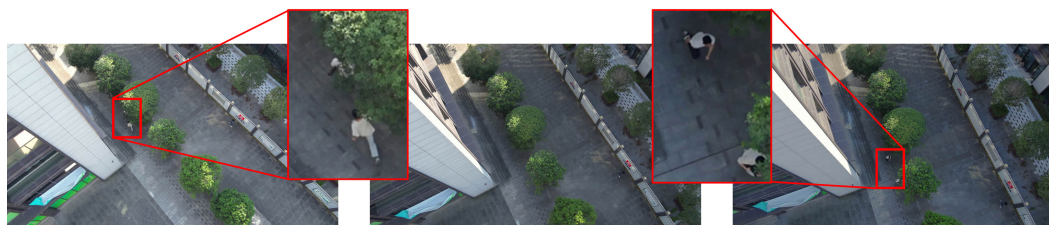


(a) Different Video Clips



(b) Same Video Clips

Figure 10: (a) Scene-dependent anomalies in different environments: An object left on the road is anomalous, while on the playground it is normal. (b) Scene-dependent anomalies in the same environment: Crossing outside a crosswalk is anomalous, while within it is normal.



(a) Before the occlusion

(b) During the occlusion

(c) After the occlusion

Figure 11: The left image shows two individuals chasing each other before entering the shadow of a tree. The middle image depicts the moment of occlusion, where they are fully obscured. The right image shows them emerging from the shadow after occlusion.

B Reward Design for Aerial Anomaly Understanding

Format Reward To encourage structured and interpretable outputs, we introduce a format adherence reward. The model is required to organize its reasoning and final answer within optional intermediate steps, *i.e.*, `<think>`, `<seeking>`, and `<answer>` tags. A reward of 1 is given for strict adherence to the template, while a reward of 0 is assigned if the format deviates.

Accuracy Reward. Let $\hat{c}$ and c^* denote the predicted and reference anomaly classes extracted from the `<answer>` tag. The scalar reward R_{cls} is defined as $R_{\text{cls}} = 1$ when $\hat{c} = c^*$ (exact match); $R_{\text{cls}} = 0.1$ when $\hat{c} \neq c^*$ yet both classes are abnormal (*i.e.*, $\hat{c} \neq \text{Normal}$ and $c^* \neq \text{Normal}$); and $R_{\text{cls}} = 0$ in all other cases, including missing predictions. This scheme awards full credit for correct categorisation, partial credit for correctly flagging abnormality while misidentifying the subtype, and no credit for normal/anomaly confusion or omitted labels.

Localization Reward. The localization reward R_{loc} is defined as the Intersection over Union (IoU) between the predicted bounding boxes $\hat{B}$ and the ground truth boxes B^* . This reward encourages the model to focus on the most relevant regions of the input, thereby improving its ability to identify and understand anomalies in complex aerial scenes.

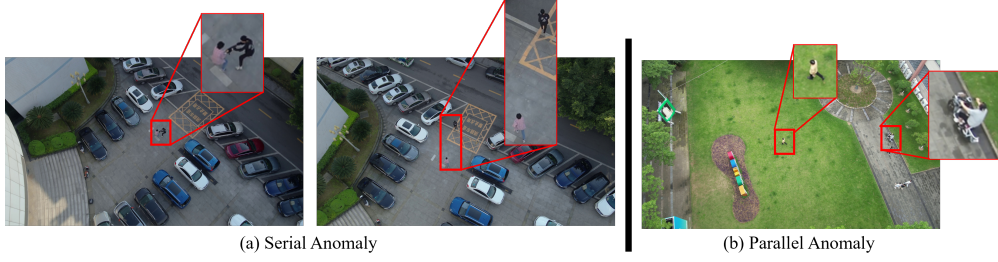


Figure 12: (a) shows a robbery followed by a chase, representing a serial anomaly. (b) depicts simultaneous anomalies at two locations, representing parallel anomalies.

Seeking Reward. The seeking reward R_{seek} is designed to incentivize the model to determine the necessity of additional high-resolution context and, when required, to precisely localize the corresponding regions. Formally, it is expressed as $R_{\text{seek}} = \mathbb{I}[\hat{s} = s^*] \cdot \text{IoU}(\hat{B}, B^*)$, where s^* and $\hat{s}$ are the ground truth and predicted seeking decisions, and $\text{IoU}(\hat{B}, B^*)$ measures the overlap between the predicted bounding box $\hat{B}$ and the ground truth B^* . This formula rewards the model for directly outputting predictions in simple scenarios, while encouraging it to simulate the process of focusing on suspicious areas by cropping and analyzing these regions for further inspection in complex scenarios.

Length Reward. To address the overthinking phenomenon [56] in simple scenarios, we couple answer correctness with reasoning length L (tokens in the <think> segments). If the answer is correct we favour brevity, setting $R_d = \frac{1}{\log(1+L)}$; if it is wrong we encourage elaboration with $R_d = \min(L/L_{\max}, 1)$, where $L_{\max}$ equals the model's maximum output length. The reward is zero whenever either <think> or <answer> is missing.

C Discussed on the effectiveness of A2Seek-R1

C.1 Self-Correction via Reflection with Reasoning Annotations

To further analyze how GoT-guided reasoning annotation in our A2Seek improves reasoning quality, we study model behavior with and without access to reasoning annotations.

Without GoT-guided Reasoning Annotations. For input x , the model predicts y_0 based on a scoring function $s_\theta(x, y)$:

$$P_\theta(y \mid x) = \frac{\exp(s_\theta(x, y))}{\sum_{y'} \exp(s_\theta(x, y'))}. \quad (5)$$

If y_0 is incorrect, there is no mechanism to revise it.

With GoT-guided Reasoning Annotations. Reasoning annotations prompt the model to evaluate its own output and produce a refined prediction $\tilde{y}$:

$$P_\theta(\tilde{y} \mid x, y_0) \propto \exp(s_\theta(x, \tilde{y}) + \lambda \Delta r(x, y_0, \tilde{y})), \quad (6)$$

where $\Delta r(x, y_0, \tilde{y}) = R(x, \tilde{y}) - R(x, y_0)$ is the reflection-induced reward difference.

Assuming auxiliary rewards remain unchanged (e.g., localization), we approximate:

$$\Delta r(x, y_0, \tilde{y}) \approx R_{\text{cls}}(x, \tilde{y}) - R_{\text{cls}}(x, y_0). \quad (7)$$

with

$$R_{\text{cls}}(x, y) = \mathbb{I}[y = y^*], \quad \Delta r(x, y_0, \tilde{y}) = \mathbb{I}[y_0 \neq y^* \wedge \tilde{y} = y^*] - \mathbb{I}[y_0 = y^* \wedge \tilde{y} \neq y^*]. \quad (8)$$

This reward difference yields a reflection-aware update:

$$\Delta \theta = -\eta \nabla_\theta \mathbb{E}[\Delta r(x, y_0, \tilde{y})]. \quad (9)$$

Finally, the improvement in expected classification reward is lower-bounded by:

$$\mathbb{E}[R_{\text{cls}}(\tilde{y})] - \mathbb{E}[R_{\text{cls}}(y_0)] \geq \eta\lambda\mathbb{E}[\Delta r]. \quad (10)$$

This analysis supports our empirical observation that reflection-guided self-correction leads to consistent gains in both accuracy and interpretability, and further highlights the value of structured annotations provided in A2Seek. In summary, the reflection mechanism enables the model to revise suboptimal predictions by leveraging auxiliary hints and structured annotations, proving especially beneficial in scenarios with subtle or ambiguous anomalies. However, when the input information itself is insufficient (*e.g.*, incomplete visual context or occlusions), self-correction alone may fail to imagine the missing clues. This motivates us to introduce a *seeking* mechanism that actively queries for additional data, as detailed in the next section.

Seeking Mechanism. In scenarios where the input information is insufficient for accurate reasoning, the <seeking> mechanism allows the model to actively query for additional context, bridging the gap between the available input and the required information for correct predictions. To formalize this, we leverage the Information Bottleneck (IB) theory [24], which balances the trade-off between the sufficiency of information for the task and the complexity of the representation.

Let I_{input} denote the information provided by the input x , and I_{required} represent the total information needed for accurate reasoning. If $I_{\text{input}} < I_{\text{required}}$, the model's reasoning process is under-constrained, leading to ambiguous or incorrect predictions. The <seeking> mechanism dynamically retrieves additional information I_{seek} , such that the total information available becomes:

$$I_{\text{total}} = I_{\text{input}} + I_{\text{seek}}.$$

The seeking reward R_{seek} is designed to encourage the model to query for I_{seek} only when I_{input} is insufficient. Formally, the reward is defined as:

$$R_{\text{seek}} = \begin{cases} \beta \cdot \frac{I_{\text{seek}}}{I_{\text{required}}}, & \text{if } I_{\text{input}} < I_{\text{required}}, \\ 0, & \text{otherwise,} \end{cases}$$

where β is a scaling factor that controls the weight of the seeking reward. This formulation ensures that the model is incentivized to seek additional information only when it is necessary for accurate reasoning.

Information Bottleneck Perspective. From the perspective of the Information Bottleneck theory, the <seeking> mechanism can be viewed as a way to optimize the mutual information $I(X; Y)$ between the input X and the output Y , while minimizing the complexity of the intermediate representation Z . The objective can be expressed as:

$$\mathcal{L}_{\text{IB}} = I(X; Z) - \lambda I(Z; Y),$$

where Z represents the information retrieved through <seeking>, and λ balances the trade-off between retaining sufficient information for the task and minimizing unnecessary complexity.

By incorporating the seeking reward R_{seek} , the model dynamically adjusts $I(Z; Y)$ based on the complexity of the input. For simple cases where $I_{\text{input}} \approx I_{\text{required}}$, the model minimizes $I(Z; Y)$ by avoiding unnecessary seeking. For complex scenarios where $I_{\text{input}} \ll I_{\text{required}}$, the model increases $I(Z; Y)$ by retrieving additional information, ensuring robust reasoning.

Unified Framework for Seeking and Reasoning. The seeking reward integrates seamlessly with the A-GRPO algorithm, influencing both the reflection reward $\Delta r(x, y_0, \tilde{y})$ and the policy update. Specifically, the total reward $R(x, y)$ now includes R_{seek} as a component:

$$R(x, y) = R_{\text{format}} + R_{\text{acc}} + R_{\text{loc}} + R_{\text{length}} + R_{\text{seek}}.$$

This unified framework ensures that the model balances exploration (querying for additional information) and exploitation (using the retrieved information to refine predictions). By dynamically adjusting the seeking process, the model achieves near-optimal performance across diverse and challenging UAV scenarios.



Figure 13: Examples of privacy protection in aerial video frames. (a) License plates are pixelated to prevent vehicle identification. (b) Facial regions are blurred to anonymize pedestrians. These anonymization procedures ensure compliance with privacy-preserving standards during dataset release.

D Privacy Preservation, Licensing and Ethical Compliance

Because human subjects are typically captured at oblique, metre-scale distances in the drone view, direct identification is already improbable; nevertheless, every frame is passed through an offline sanitisation pipeline. First, a lightweight YOLO-Face detector isolates facial regions, whose pixels are then scrambled with block-wise spectral noise in CIELab space and re-integrated via Poisson blending, leaving no reversible biometric cues while preserving motion and silhouette statistics for anomaly-detection research.

Although licence plates are rarely resolvable at the recorded altitudes, a human-in-the-loop audit team inspected every frame and manually applied the same irreversible scrambling procedure to any plate that remained decipherable. The final dataset and its metadata are distributed in read-only form, and access is gated by a click-through licence that explicitly forbids re-identification attempts and extends identical non-commercial, attribution requirements to all derivative models.

E Limitations, Open Challenges and Future Directions

We introduce A2Seeka largescale, multimodal UAV-view dataset enriched with dense bounding-box tracks and sentence-level descriptions and A2Seek-R1, a multi-stage RL-fine-tuning framework that not only detects anomalies but also explains where and why they occur. To date, we measure performance with frame-level average precision and mean IoU, but our richly annotated data enable much more: future work will add region-aware metrics (*e.g.*, temporal IoU of predicted vs. ground-truth tubes) and spatio-temporal localization scores that jointly reward correct timing and placement of anomaly predictions. Moreover, the moving UAV perspective where objects shrink with altitude, perspective shifts continuously, and the camera itself panscalls for bespoke evaluation criteria that penalize false alarms on distant clutter more heavily and incentivize early anticipation of emerging anomalies.

Methodologically, a UAV frame offers far richer cues than a static CCTV image. Parallax reveals 3D layout, rotor vibration modulates motion patterns, and onboard audio can flag sudden events like collisions or alarms. Building on this, forthcoming models will fuse high-resolution appearance with optical flow, depth priors and language grounding, while dynamically adapting their receptive fields as the drone zooms or sweeps its view. By mining these latent, multimodal signals, we aim to push anomaly detection beyond asking whether something is wrong toward explaining what is wrong, where it happens and why it matters.

Societal Impact

Our proposed dataset focuses on abnormal event detection from UAV (Unmanned Aerial Vehicle) perspectives, with the primary goal of enhancing public safety and promoting robust perception mod-

els for long-range, real-time surveillance in open environments. This effort holds several **positive societal impacts**:

- **Safety and early warning:** UAV-based monitoring allows timely detection of anomalous behaviors in large-scale outdoor areas (*e.g.*, campuses, industrial zones), potentially preventing incidents such as violence or theft.
- **Technological advancement:** The dataset promotes the development of embodied multi-modal AI models capable of spatial-temporal reasoning and active scene understanding.
- **Academic contribution:** It fills a gap in existing datasets by providing controlled abnormal scenarios with diverse contextual challenges, enabling reproducible and fine-grained benchmarking for the research community.

Despite these benefits, we recognize potential **negative societal impacts**, such as:

- **Privacy concerns:** UAVs capture aerial footage, which may inadvertently include individuals or locations without consent.
- **Risk of misuse:** The dataset or associated models could potentially be adapted for surveillance beyond legitimate or ethical use cases.

To **mitigate these risks**, we take several strict measures:

- **Ethical review and anonymization:** All collected data undergoes internal ethical review. No original raw video data is released; only extracted visual features and annotations are provided.
- **Identity protection:** All identifying visual information (*e.g.*, faces, license plates, *etc.*) is blurred or removed. No personal metadata is stored or shared.
- **Controlled data collection:** All actors performing abnormal events are volunteers from our research lab, who participated with full informed consent and received fair compensation.
- **Transparency and access control:** We provide access to visual features and metadata under the CC BY-NC-SA 4.0 license, strictly for academic, non-commercial research. Video URLs are included only to promote reproducibility, and content access remains subject to platform-level permissions.

We believe these efforts ensure that our work advances the field of multimodal abnormal event understanding in a safe, ethical, and socially beneficial manner.