ANOMALYMINER: CLOSED-LOOP EVIDENCE MINING FOR AERIAL ANOMALY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial anomaly understanding requires models to identify, localize, and explain abnormal events in complex aerial videos. Existing methods based on fixed perception, visual enhancement, or global fusion often assume that providing more visual information yields more useful anomaly evidence. In practice, additional information may emphasize background structures, irrelevant thermal responses, or misleading regions rather than support the current anomaly judgment. This information–evidence gap is especially pronounced in low-light RGB–infrared scenes: RGB may miss subtle anomalous cues, whereas infrared may introduce salient but anomaly-irrelevant responses. The key challenge is therefore not simply to enhance, fuse, or acquire more visual inputs, but to continually admit, validate, and repair anomaly-relevant evidence throughout reasoning. Accordingly, we propose AnomalyMiner, a closed-loop evidence-mining framework for aerial anomaly understanding, which organizes perception and reasoning into a stateful evidence trajectory: perception supports the current anomaly hypothesis, while reasoning specifies what evidence should be sought next. Evidence-Gain Interleave first controls the forward flow of evidence by admitting localized infrared only when it yields counterfactual gain for downstream anomaly reasoning, rather than indiscriminately expanding multimodal context. However, admitted evidence can still update the reasoning state in a way that redirects subsequent perception toward misleading cues. Evidence-State Backtrack therefore monitors whether the current state regresses from the preceding state and repairs a misdirected search from that reliable preceding state. To make these admission and repair decisions directly learnable, Evidential Preference Policy Optimization routes process feedback to the corresponding key evidence-bearing tokens and action tokens for acquisition, interaction, and repair. AnomalyMiner achieves 74.75% AP on complete A2Seek and 45.18% on its nighttime subset, while showing consistent cross-dataset generalization on M3OT and VTUAV.

1 INTRODUCTION

Recent advances in multimodal large language models are moving video anomaly analysis from anomaly scoring toward anomaly understanding, where a model must identify an abnormal event and explain why it is abnormal (Yang et al., 2024; Zanella et al., 2024; Du et al., 2024; Ye et al., 2025; Zhang et al., 2025a; Huang et al., 2025b; Zhu et al., 2025b). Aerial video is widely used in public safety, traffic management, and emergency response, yet dynamic viewpoints, severe scale variation, complex backgrounds, and small targets make aerial anomaly understanding particularly challenging (Mo et al., 2025). Across such wide spatial and temporal contexts, anomaly understanding depends on identifying the sparse, task-relevant evidence that supports an anomaly judgment.

Despite different visual processing strategies, existing methods commonly follow an open-loop assumption: stronger or more visual information supplied before reasoning should improve anomaly understanding. Fixed perception compresses a video into a representation that remains unchanged thereafter; visual enhancement and global multimodal fusion increase visible content or feature capacity. Even when a system can request an additional observation, without evidence-level admission it may simply pass more information into reasoning. Information, however, is not equivalent to anomaly evidence: new content may amplify background structure, ordinary objects, or salient but irrelevant regions. Evidence demand also evolves with reasoning; only after a preliminary hy-

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

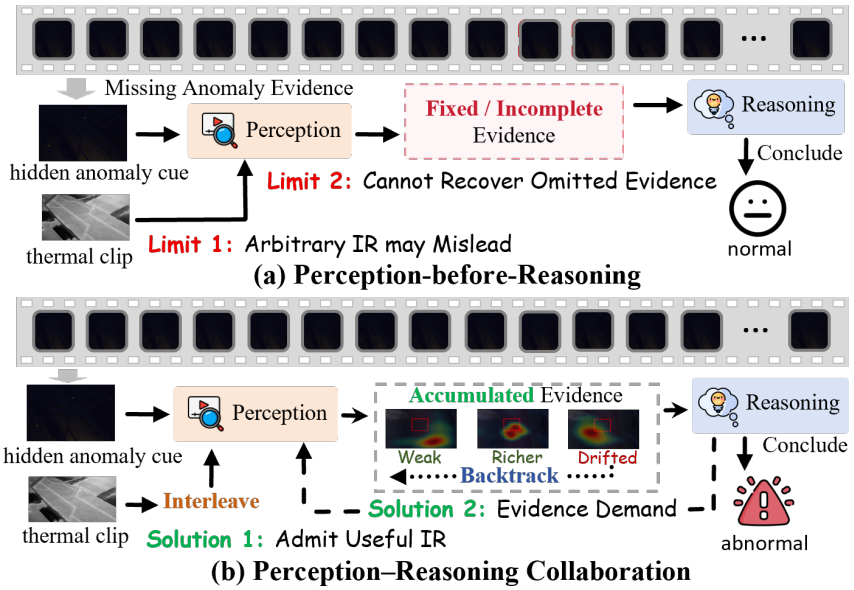


Figure 1: **From open-loop visual processing to closed-loop evidence mining.** Top: Fixed perception, enhancement, or fusion supplies visual information only once before reasoning. Anomaly cues omitted at this stage cannot be recovered, while arbitrary infrared may mislead the decision. Bottom: AnomalyMiner lets reasoning specify evidence demand, admits localized infrared only when beneficial, and learns when to backtrack to repair a drifting evidence path. The low-light RGB–infrared example is a representative stress case in which the gap between visual information and anomaly evidence is especially clear.

pothesis is formed may the model know which person, action, or contextual cue is missing. Aerial anomaly understanding therefore requires reasoning demand to drive perception and newly acquired information to be validated against the current anomaly hypothesis.

Low-light RGB–infrared aerial scenes provide a representative stress case for this general information–evidence gap. Low illumination degrades RGB observations and makes aerial perception challenging (Zhu et al., 2024; Ye et al., 2022). Infrared provides complementary visual information in multimodal aerial perception (Chen et al., 2024; Iaboni & Abichandani, 2026). It may also introduce broad background responses and unrelated objects. Regions emphasized by generic low-light enhancement likewise need not be anomaly relevant. Rather than proposing another low-light enhancement or fixed RGB–infrared fusion method, we use this demanding setting to study a more general question: how can additional information be converted into anomaly evidence, and how should a misdirected evidence-search path be repaired?

We therefore propose **AnomalyMiner**, a closed-loop evidence-mining paradigm for aerial anomaly understanding. Instead of fixing perception before reasoning, AnomalyMiner organizes both processes into an evolving evidence trajectory: at each node, perception supports the current anomaly hypothesis, while reasoning updates the event interpretation and issues the next evidence demand. Figure 1 contrasts the two paradigms. An open-loop method cannot recover anomaly cues omitted by its initial perception and may be distracted by irrelevant additions; AnomalyMiner continually admits, validates, and repairs evidence along the same trajectory.

The loop first controls what evidence enters the trajectory. Evidence derived solely from RGB observations may be insufficient, particularly under low illumination, while indiscriminately incorporating infrared observations may introduce anomaly-irrelevant information that biases downstream anomaly reasoning. We therefore propose **Evidence-Gain Interleave**, which treats localized infrared retrieval as a forward admission action triggered by current reasoning demand and validates it through counterfactual gain to downstream anomaly reasoning; only beneficial evidence is admitted. Yet the resulting reasoning may still direct the next perception toward a wrong region. Backtracking based only on answer correctness cannot reveal whether the evidence path itself is regressing (Yao et al., 2023; Koh et al., 2024). We therefore propose **Evidence-State Backtracking**, which com-

pares the current and preceding states with an anomaly-grounding proxy. Cross-state changes in this proxy provide feedback for learning when to trigger backtracking, enabling the model to resume perception from a reliable preceding state and correct the evidence-search direction. The two mechanisms thus form forward admission and path repair within one loop. To make these loop behaviors directly learnable, we propose **Evidential Preference Policy Optimization (EvPO)**, which routes evidence-mining, modality-interaction, backtracking-repair, and final-task feedback to the corresponding key evidence-bearing tokens and action tokens. Our contributions are threefold: (1) we reformulate aerial anomaly understanding as closed-loop evidence mining, distinguish visual information from anomaly evidence, and use low-light RGB–infrared scenes as a representative stress test; (2) we develop two coupled controls over one evidence trajectory—Evidence-Gain Interleave admits evidence according to its gain for anomaly reasoning, while Evidence-State Backtracking repairs misdirected perception according to cross-state changes in the anomaly-grounding proxy; and (3) we introduce EvPO to assign each process signal to its corresponding key evidence-bearing tokens and action tokens, and validate AnomalyMiner on complete A2Seek, its nighttime subset, and external aerial datasets.

2 Related Work

With the development of multimodal large language models (MLLMs), recent studies have begun to leverage their open-world semantic understanding and reasoning capabilities to advance anomaly analysis from clip-level discrimination to more fine-grained anomaly understanding. For example, training-free methods exploit pretrained large models for anomaly analysis (Zanella et al., 2024; Shao et al., 2025), while HiProbe-VAD trains a lightweight anomaly scorer on hidden states of frozen MLLMs (Cai et al., 2025). Other methods associate anomalous cues with interpretable semantic spaces through rule-based reasoning (Yang et al., 2024), verbalized learning (Ye et al., 2025), or vision-language alignment (Chen et al., 2025; Huang et al., 2025a). Related research also explores causation understanding (Du et al., 2024), long-term temporal modeling (Zhang et al., 2025a), and reinforcement fine-tuning (Huang et al., 2025b; Zhu et al., 2025b) to support reasoning about complex anomalous events. Aerial anomaly understanding has also attracted increasing attention. Dynamic viewpoints, scale variations, complex backgrounds, and small targets in UAV videos introduce distinctive visual perception challenges. Prior studies combine object detections with spatial and temporal context for traffic anomaly detection (Bozcan & Kayacan, 2021), learn spatio-temporal representations through Transformer-based future-frame prediction (Le et al., 2025; Jin et al., 2022), and establish aerial traffic anomaly benchmarks (Tran et al., 2023). These efforts advance aerial anomaly detection, with evaluation primarily focused on anomaly scores and spatio-temporal localization. Recently, A2Seek has advanced aerial anomaly analysis toward reasoning-centric understanding through fine-grained temporal and spatial reference labels, natural-language explanations, and an active seeking mechanism (Mo et al., 2025). Nevertheless, existing methods generally retain a fixed-perception paradigm: evidence gaps emerging during anomaly reasoning do not trigger a return to perception to acquire additional evidence. This limitation motivates closed-loop evidence mining that acquires and validates evidence according to evolving reasoning demands throughout the perception–reasoning process. Low-light RGB–infrared scenes expose this general information–evidence gap particularly clearly: RGB may omit needed cues, while complementary information may also introduce irrelevant responses.

3 Method

AnomalyMiner formulates aerial anomaly understanding as closed-loop evidence mining, coupling perception and reasoning along an evolving evidence trajectory (Figure 2). Perception first extracts visual evidence from RGB observations to support the current anomaly hypothesis. To supplement insufficient RGB evidence, we introduce Evidence-Gain Interleave, which couples anomaly reasoning with the on-demand acquisition of localized infrared observations. Reasoning then integrates the available evidence to update the anomaly interpretation and assess whether a conclusion is supported. The model resumes perception if it deems the evidence insufficient. Even with informative evidence, reasoning may still drift and misdirect further evidence mining (Wang et al., 2026). We therefore introduce Evidence-State Backtrack, which revisits a reliable preceding state and uses its reasoning context to guide subsequent perception and correct the evidence-search direction. Finally,

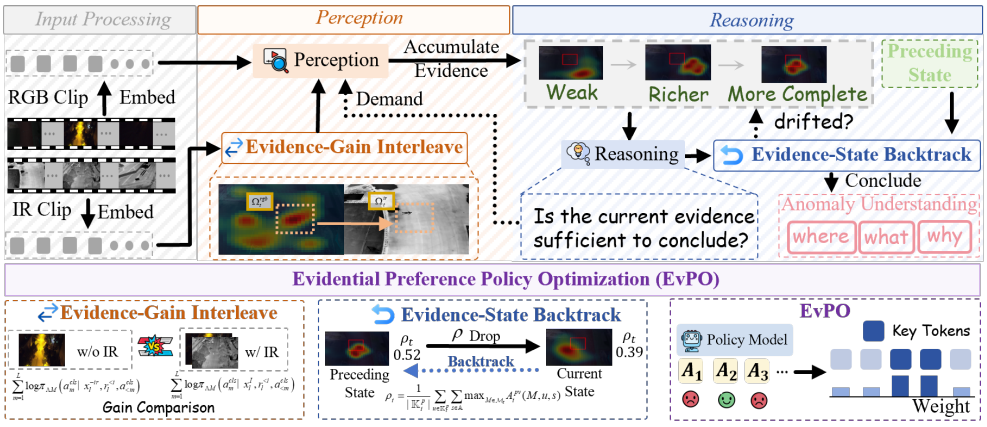


Figure 2: **Overview of AnomalyMiner.** Perception provides evidence for anomaly reasoning, which assesses whether the available evidence supports a conclusion or further perception is required. Evidence-Gain Interleave couples anomaly reasoning with the on-demand acquisition of localized infrared observations to supplement RGB evidence. Evidence-State Backtrack resumes perception from a reliable preceding state to correct deviations in evidence search. During reinforcement fine-tuning, counterfactual reasoning gains and cross-state changes in the anomaly-grounding proxy provide feedback for interleaving and backtracking, respectively. Evidential Preference Policy Optimization routes process feedback to the corresponding key evidence-bearing tokens and action tokens, placing greater optimization emphasis on these positions.

we propose Evidential Preference Policy Optimization, which assigns process feedback to the corresponding key evidence-bearing tokens and action tokens, making evidence acquisition, interleaving, and backtracking directly optimizable.

3.1 Evidence-Gain Interleave

The model first acquires perceptual evidence from RGB observations to support anomaly reasoning. Let Z^{rgb} and Z^{ir} denote the visual-token sequences of paired RGB and infrared clips, respectively. At node t, Z_t^m denotes the visual observation accessed in modality m, and r_{t-1} denotes the preceding reasoning state. The policy π_θ generates perceptual evidence e_t^m as

$$e_t^m \sim \pi_\theta(\cdot \mid Z_t^m, r_{t-1}), \qquad m \in \{rgb, ir\}. \tag{1}$$

Evidence obtained from RGB alone may be insufficient to support an anomaly judgment, particularly under low illumination. Infrared observations are commonly used to complement RGB information in multimodal visual perception (Chen et al., 2024; Zhang et al., 2024). However, acquiring infrared observations at every reasoning step or using the full infrared clip increases the available information but may also introduce anomaly-irrelevant evidence.

To address this information–evidence gap, we propose Evidence-Gain Interleave, which couples anomaly reasoning with the on-demand acquisition of localized infrared observations. Conditioned on the current RGB perceptual evidence, the model determines whether complementary infrared observations are needed. Infrared acquisition is restricted to a local region Ω_t. A small set of carefully selected attention heads maps the region most attended to by the current RGB perception onto the paired infrared observation.

The resulting perceptual evidence e_t is then supplied to the model to update its reasoning, where $e_t = e_t^{rgb}$ without interleaving and $e_t = [e_t^{rgb}; e_t^{ir}]$ after interleaving:

$$r_t \sim \pi_\theta(\cdot \mid e_t, r_{t-1}). \tag{2}$$

The updated reasoning refines the anomaly interpretation and assesses evidence sufficiency. The model resumes perception if evidence is insufficient and reaches a final conclusion when it deems the anomaly evidence sufficient.

To learn this interleaving behavior, we measure how perceptual evidence supports the correct anomaly category throughout reasoning. Let y^* denote the correct anomaly category and $r_t^{<i}$ the reasoning prefix before position i. We score each position by the autoregressive log-likelihood of y^* conditioned on evidence e and that prefix. The top-K^r positions are selected as reasoning-side key evidence-bearing tokens, and their average score serves as a proxy for the support that the current reasoning provides for the correct category:

$$q_t^i(e) = \log \pi_\theta\big(y^* \mid e, r_t^{<i}\big), \qquad \mathbb{K}_t^r = \mathrm{IdxTopK}_i\big(q_t^i(e_t), K^r\big), \\ s_t(e) = \frac{1}{|\mathbb{K}_t^r|} \sum_{i \in \mathbb{K}_t^r} q_t^i(e). \tag{3}$$

We evaluate each interleaving decision against its alternative while holding the generated reasoning and selected positions fixed. The paired conditions differ only in whether infrared perceptual text is included. Let $a_t^{\mathrm{int}} = 1$ denote requesting infrared and $a_t^{\mathrm{int}} = 0$ denote skipping it. The counterfactual reward is

$$\mathcal{R}_{\mathrm{int}}^t = (2a_t^{\mathrm{int}} - 1)\left[s_t\Big([e_t^{rgb}; e_t^{ir}]\Big) - s_t\Big(e_t^{rgb}\Big)\right]. \tag{4}$$

The interleaving decision is rewarded when it yields stronger support for the correct category than its counterfactual alternative and penalized when it yields weaker support. This signal optimizes both decisions according to their relative value for reasoning, encouraging complementary-information acquisition when it advances anomaly inference. Appendix B.3 describes how the attention heads are selected; Appendix B.3.4 specifies how Ω_t is constructed and accessed in the paired infrared observation.

3.2 Evidence-State Backtrack

After acquiring perceptual evidence, the model updates its anomaly interpretation and directs subsequent evidence mining. Even when the perceptual evidence is informative, reasoning may still deviate from the correct anomaly interpretation (Wang et al., 2026), redirecting subsequent perception toward irrelevant regions. Search-based reasoning methods commonly revisit preceding states or explore alternative branches according to task progress or state value (Yao et al., 2023; Koh et al., 2024). We therefore propose Evidence-State Backtrack, which corrects a misdirected evidence search by resuming perception from a reliable preceding state.

Attention weights alone do not establish the causal contribution of an input to the final prediction (Jain & Wallace, 2019; Kobayashi et al., 2020). However, attention assigned to task-relevant image regions has been used to measure visual grounding (Hudson & Manning, 2019), and selected localization heads provide spatial signals aligned with textual semantics (Kang et al., 2025). Following these observations, we define an anomaly-grounding proxy ρ_t to measure how strongly the perceptual evidence associated with the current reasoning attends to the anomaly region. Let $\mathbb{K}_t^p$ denote the perception-side key evidence-bearing tokens, $\mathcal{H}^{\mathrm{loc}}$ the fixed localization heads, and $\mathbb{A}$ the anomaly region:

$$\rho_t = \frac{1}{|\mathbb{K}_t^p|} \sum_{u \in \mathbb{K}_t^p} \mathrm{Attn}_{\mathcal{H}^{\mathrm{loc}}}(u, \mathbb{A}). \tag{5}$$

Here, $\mathrm{Attn}_{\mathcal{H}^{\mathrm{loc}}}(u, \mathbb{A})$ aggregates attention within $\mathbb{A}$, taking the maximum across modalities at corresponding visual positions and averaging over localization heads.

When the current evidence state allocates substantially less attention to the anomaly region than the preceding state, we treat this decline as a signal of possible drift in evidence search and encourage resuming exploration from that preceding state. Comparing the two states provides directional information: a low grounding score may be improving from an earlier state or may result from a decline, calling for different responses.

Let $\bar{\rho}_{t-1}$ and $\tilde{\rho}_t$ denote the proxies of the retained preceding state and the current candidate state, respectively, and let $b_t = 1$ indicate backtracking and $b_t = 0$ otherwise. We define their relative grounding decline and the corresponding reward as

$$\gamma_t = \frac{\bar{\rho}_{t-1} - \tilde{\rho}_t}{\bar{\rho}_{t-1} + \epsilon_\rho}, \qquad \mathcal{R}_{\mathrm{bt}}^t = (2b_t - 1)\gamma_t\, \mathbb{I}[|\gamma_t| > \tau_{\mathrm{bt}}]. \tag{6}$$

where ϵ_ρ ensures numerical stability and τ_{bt} specifies a tolerance within which the reward is zero. Outside this tolerance, the reward favors backtracking when anomaly grounding deteriorates and retaining the current state when it improves, while penalizing the opposite decisions. This feedback encourages the model to repair evidence-search directions when anomaly grounding deteriorates while preserving trajectories that continue to improve. Appendix B.3 details the selection of key evidence-bearing tokens and the aggregation of visual attention; Appendix B.5 specifies the complete backtracking procedure.

3.3 Evidential Preference Policy Optimization

We adopt reinforcement fine-tuning to optimize the closed-loop evidence-mining policy. GRPO provides group-relative training signals across multiple trajectories sampled for the same input (Shao et al., 2024). However, assigning a shared trajectory-level advantage to all generated tokens does not distinguish their different roles in evidence acquisition and reasoning. We therefore propose Evidential Preference Policy Optimization (EvPO), which preserves baseline credit across generated tokens while emphasizing process-specific feedback at the corresponding key evidence-bearing tokens and action tokens.

For rollout g, let $\widehat{A}_g^c$ denote the group-relative advantage of process feedback c, and let $\mathcal{K}_g^c$ contain the key evidence-bearing tokens and action tokens associated with that process. EvPO defines the policy advantage at position i as

$$A_{g,i}^{\mathrm{EvPO}} = \sum_c \left(1 + \lambda_c \mathbb{I}[i \in \mathcal{K}_g^c]\right) \widehat{A}_g^c, \tag{7}$$

where $\lambda_c > 0$ controls the additional weight of process signal c at its key positions. Other positions retain unit weight. The training procedure, including supervised fine-tuning and policy optimization, is detailed in Appendices B.2 and B.6.

4 Experiments

4.1 Setting

We evaluate AnomalyMiner on the full A2Seek benchmark and its nighttime subset, and assess cross-dataset generalization on VTUAV (Zhang et al., 2022) and M3OT (Nie et al., 2025) without target-domain fine-tuning. Stage-wise SFT initializes the executable loop interface, after which EvPO trains the model to make interleaving and backtracking decisions while optimizing the overall closed-loop policy.

4.2 Main Results

Complete A2Seek benchmark. We first report scene-wise and overall results under the standard complete benchmark protocol in Table 1. AnomalyMiner trained on the complete split obtains 74.75% overall AP, exceeding A2Seek-R1 by 21.61 percentage points. On the more complex Open Area and Gate scenes, AnomalyMiner exceeds A2Seek-R1 by 41.84 and 37.14 percentage points, respectively. The model trained only on nighttime data already reaches 59.97% on the complete test set; expanding training coverage further adds 14.78 percentage points. The performance of the nighttime-trained model beyond its training conditions suggests that the learned interleaving and backtracking strategies are transferable, supporting the effectiveness of closed-loop evidence mining across illumination conditions.

A2Seek nighttime subset. Nighttime scenes form the primary stress test because anomaly semantics are most likely to be omitted by initial perception. As shown in Table 2, AnomalyMiner achieves 45.18% overall AP, outperforming the strongest infrared-enabled baseline, A2Seek-R1 w/ IR, by 10.06 percentage points. Improvements occur in all five nighttime scenes, including gains of 12.04 and 15.13 percentage points on Playground and Wall Zone, respectively.

Cross-dataset generalization. We further evaluate cross-dataset generalization on VTUAV and M3OT without target-domain fine-tuning. As shown in Table 2, AnomalyMiner achieves 37.54% AP and 68.58% AUC on VTUAV, and 34.19% AP and 67.66% AUC on M3OT, outperforming all

Table 1: Performance comparison on the complete A2Seek benchmark. We report scene-wise AP (%) and the overall AP under the standard full-dataset protocol. For AnomalyMiner, * denotes training on the nighttime subset only, whereas † denotes training on the complete training split.

Method	Misc.	Road.	Sidew.	Playgr.	Open	Park	Gate	Wall	Acad.	Avg.
SSRL Li et al.	11.50	10.20	8.30	9.40	20.10	12.70	14.20	16.80	11.00	12.00
HSTforU Le et al.	47.41	50.18	43.52	30.03	29.68	25.01	37.28	48.72	33.69	47.66
ANDT Jin et al.	10.89	39.00	48.82	25.21	20.20	19.94	37.68	49.99	28.63	40.42
AnomalyRuler Yang et al.	17.40	15.00	9.80	22.50	25.60	14.90	19.30	21.70	18.80	17.10
LAVAD Zanella et al.	12.80	15.20	11.00	19.50	24.70	20.50	19.10	22.10	14.90	16.41
Holmes-VAU Zhang et al.	17.00	18.50	16.00	23.40	25.00	20.50	21.12	27.20	17.60	19.97
LLaVA-Video Zhang et al.	12.00	10.00	16.00	15.00	24.00	13.00	23.00	12.00	22.00	15.10
InternVL-3.0 Zhu et al.	35.62	49.75	48.21	47.08	27.43	23.94	29.64	36.99	42.33	41.34
Zero-shot Bai et al.	44.83	44.92	35.15	23.88	22.42	19.16	30.05	29.44	27.62	34.03
VAD-R1 Huang et al.	38.49	22.71	33.97	40.78	11.81	12.14	18.95	53.87	30.44	33.07
Hawk Tang et al.	28.92	13.03	46.52	21.73	18.59	8.56	8.37	40.42	31.71	23.79
VERA Ye et al.	17.92	31.63	28.68	55.44	13.32	30.59	27.49	39.01	18.86	27.51
VADTree Li et al.	22.93	14.68	15.33	33.10	21.08	30.11	24.42	38.35	12.55	22.68
EventVAD Shao et al.	13.93	15.20	21.00	33.14	19.85	15.71	22.44	13.08	25.60	18.57
A2Seek-R1 Mo et al.	51.12	56.54	55.36	75.60	40.84	26.86	38.21	61.30	43.92	53.14
AnomalyMiner*	44.89	41.45	40.82	28.29	73.28	74.32	67.98	50.90	51.67	59.97
AnomalyMiner†	71.92	64.70	61.27	43.75	82.68	84.44	75.35	64.92	68.44	**74.75**

Table 2: Performance comparison on the A2Seek nighttime subset, VTUAV, and M3OT. VTUAV and M3OT are unseen during training and are evaluated by direct transfer of the model trained on A2Seek.

	A2Seek Nighttime (AP)						VTUAV		M3OT	
Method	Road.	Sidew.	Playgr.	Roof	Wall	Avg.	AP	AUC	AP	AUC
SSRL Li et al.	7.82	6.14	9.35	4.26	12.48	10.73	13.80	52.41	8.45	48.75
HSTforU Le et al.	31.45	21.76	14.82	7.93	36.18	19.84	6.22	48.75	4.85	47.15
ANDT Jin et al.	27.63	19.84	11.26	18.71	34.52	18.27	14.76	53.12	9.80	49.21
AnomalyRuler Yang et al.	13.28	8.47	16.92	5.61	20.35	13.62	11.07	51.68	6.55	48.33
LAVAD Zanella et al.	14.06	9.72	18.45	4.38	21.57	14.08	18.86	55.30	11.50	59.85
Holmes-VAU Zhang et al.	24.83	17.61	21.54	13.76	30.48	20.46	25.12	60.85	21.75	52.18
LLaVA-Video Zhang et al.	10.72	12.38	13.54	3.95	14.16	10.71	13.33	57.20	7.90	50.40
InternVL-3.0 Zhu et al.	35.74	25.63	29.18	31.45	32.86	23.74	21.40	58.94	23.40	52.64
Zero-shot Bai et al.	28.51	22.46	17.35	30.82	25.74	21.36	16.49	54.22	14.15	51.09
VAD-R1 Huang et al.	26.94	24.83	18.76	29.57	31.62	22.48	28.11	63.55	25.18	53.79
Hawk Tang et al.	23.67	21.54	12.83	19.46	29.71	20.15	15.89	53.90	10.25	49.47
VERA Ye et al.	25.38	20.17	14.52	12.84	30.46	19.63	20.05	56.45	16.30	50.88
VADTree Li et al.	18.74	15.62	11.35	13.47	27.83	17.92	22.86	59.10	12.80	55.12
EventVAD Shao et al.	16.53	18.76	7.42	6.18	19.84	16.58	8.53	49.30	5.20	57.90
A2Seek-R1 Mo et al.	40.31	28.54	19.76	33.82	37.65	27.84	29.96	62.18	18.55	61.24
A2Seek-R1 w/ IR Mo et al.	43.86	30.72	27.94	40.65	45.38	35.12	31.12	65.88	20.01	63.79
AnomalyMiner (Ours)	48.62	33.75	39.98	47.76	60.51	**45.18**	**37.54**	**68.58**	**34.19**	**67.66**

competing methods across both metrics on both benchmarks. These results demonstrate that the learned closed-loop evidence-mining capability generalizes to unseen aerial scenarios.

4.3 Component Ablation

Table 3(a) introduces the closed-loop components in their functional order. Building on evidence mining, Evidence-Gain Interleave raises AP to 37.14%, showing that admitting complementary evidence according to reasoning demand is effective. Under the same EvPO setting, adding Evidence-State Backtrack raises AP to 45.18%, indicating that returning to a reliable preceding state can repair subsequent perception when the evidence-mining direction drifts. With the same complete rollout, uniform-credit GRPO obtains 41.76%, whereas EvPO adds a further 3.42 percentage points.

The fixed-fusion baselines in Table 3(b) use pretrained DDBF (Zhang et al., 2024), DCEvo (Liu et al., 2025), and SAGE (Wu et al., 2025) models to replace the RGB input with fused RGB–infrared frames for the same anomaly reasoner. They obtain 32.11%, 29.50%, and 33.91% AP, respectively. The configuration with interleaving achieves 37.14% AP and exceeds all three fixed-fusion baselines

even before backtracking is introduced. AnomalyMiner exceeds the strongest fixed-fusion baseline by 11.27 percentage points. These results highlight the limitations of conventional enhancement and fusion in bridging the gap between additional visual information and anomaly-relevant evidence.

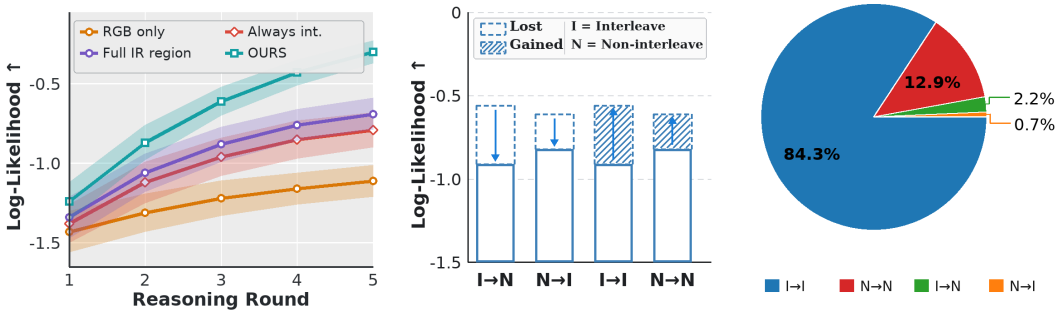
Table 3: Comparisons on A2Seek nighttime. (a) Cumulative component ablation (left); GRPO and EvPO use the same complete loop. (b) Fixed RGB–infrared fusion (right).

Configuration	Mining	Interleave	Backtrack	Optimizer	AP (%)
SFT	–	–	–	–	28.33
+ Evidence mining	✓	–	–	EvPO	30.46
+ Evidence-Gain Interleave	✓	✓	–	EvPO	37.14
Complete loop	✓	✓	✓	GRPO	41.76
AnomalyMiner	✓	✓	✓	EvPO	**45.18**

Visual input	AP (%)
DDBF Zhang et al.	32.11
DCEvo Liu et al.	29.50
SAGE Wu et al.	33.91
AnomalyMiner	**45.18**

4.4 Mechanism Analysis

4.4.1 Effectiveness of Interleave



(a) Comparison of infrared-use strategies.

(b) Forced decision reversal.

(c) Decision-reversal statistics.

Figure 3: Quantitative analysis of interleaving. (a) Comparison of infrared-use strategies. Correct-category log-likelihood is averaged across trajectories at each reasoning round; shaded bands indicate one standard deviation. (b) Forced decision reversal. Bars compare mean correct-category log-likelihood under the original and forcibly reversed interleaving decisions. Empty dashed and hatched regions indicate decreases and increases, respectively. (c) Decision-reversal statistics. Percentages report cases in which retaining (I $\rightarrow$ I, N $\rightarrow$ N) or forcibly reversing (I $\rightarrow$ N, N $\rightarrow$ I) the original decision yields higher correct-category log-likelihood. I and N denote interleaving and non-interleaving, respectively.

Figure 3a shows that our method consistently achieves higher correct-category log-likelihood than RGB-only reasoning, unconditional interleaving, and full-region infrared access, with the differences becoming more pronounced over successive rounds. The improvement over RGB-only reasoning suggests that complementary infrared observations help address evidence missing from RGB. The advantage over unconditional interleaving and full-region infrared access further suggests that indiscriminately introducing infrared information may bring redundant or anomaly-irrelevant cues that interfere with the current judgment.

In Figure 3b, suppressing requested interleaving reduces mean correct-category log-likelihood relative to retaining the original decision. Forcing infrared access when the original decision is non-interleaving likewise yields lower likelihood. These counterfactual comparisons support the effectiveness of the learned policy in adapting infrared acquisition to the current evidence state. Retaining the original interleaving decision yields higher correct-category log-likelihood than forced reversal in 97.2% of paired comparisons (Figure 3c). This statistical trend corroborates the average advantage in Figure 3b and supports the effectiveness of the learned policy in adapting infrared acquisition to the current evidence state.

4.4.2 BACKTRACKING RECOVERS ANOMALY-RELEVANT EVIDENCE

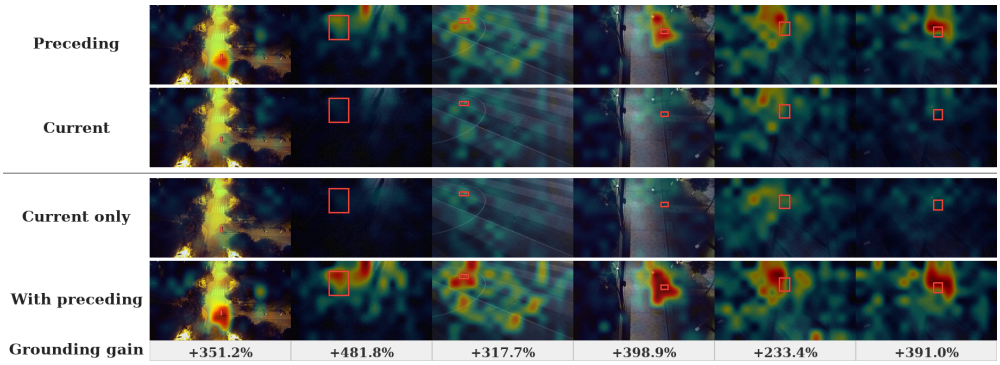


Figure 4: Comparison of backtracking strategies. Heatmaps visualize attention from perception-side key evidence-bearing tokens, contrasting perception based only on the current state with perception resumed from the preceding state. The bounding boxes indicate the ground-truth anomaly regions. "Preceding" and "Current" denote the preceding and current states, respectively; the current state exhibits attention drifting away from anomaly evidence relative to the preceding state. "Current only" and "With preceding" show the results of backtracking based solely on the current state and resuming perception from the preceding state, respectively. Grounding gain measures the recovery of attention to anomaly evidence. Percentages denote the relative increase in the anomaly-grounding proxy after backtracking, measured against the deviated current state.

To further evaluate the effectiveness of preceding-state backtracking for aerial anomaly understanding, we quantitatively compare it with backtracking based solely on the current state. For the deviated evidence-search trajectories shown in Figure 4, using only the current state produces minor changes in the anomaly-grounding proxy ρ relative to the deviated state. This limited recovery suggests that the current state alone provides insufficient reference for assessing whether the search has deteriorated and determining how its direction should be corrected. In contrast, the preceding state provides both a reference for assessing the deviation and a basis for corrective perception. Evaluating the current state can support backtracking when its correctness is directly assessable. For evidence-search drift in aerial anomaly understanding, however, comparison with the preceding state provides a more effective basis for directional repair.

At the dataset level, backtracking based solely on the current state achieves 38.91 AP, compared with 45.18 AP when the preceding state is incorporated. This improvement further supports the effectiveness of preceding-state backtracking for aerial anomaly understanding. Details of this comparison are provided in Appendix D.1.

5 CONCLUSION

We presented AnomalyMiner, a closed-loop evidence-mining framework for addressing the information–evidence gap in aerial anomaly understanding. Within the loop, perception provides evidence for anomaly reasoning, while reasoning guides further perception when the evidence is insufficient. Evidence-Gain Interleave couples anomaly reasoning with the on-demand acquisition of localized infrared observations to supplement RGB evidence. Evidence-State Backtrack corrects deviations in evidence search by resuming perception from a reliable preceding state. Evidential Preference Policy Optimization places greater emphasis on key evidence-bearing tokens and action tokens during reinforcement fine-tuning to optimize the closed-loop policy. Comparisons and ablations on A2Seek show that reasoning-guided evidence acquisition and preceding-state backtracking improve aerial anomaly understanding. Results on VTUAV and M3OT further support the generalization of the closed-loop policy across datasets.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Ilker Bozcan and Erdal Kayacan. Context-dependent anomaly detection for low altitude traffic surveillance. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 224–230. IEEE, 2021.

Zhaolin Cai, Fan Li, Ziwei Zheng, and Yanjun Qin. HiProbe-VAD: Video anomaly detection via hidden states probing in tuning-free multimodal LLMs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 592–601, 2025.

Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26836–26845, 2024.

Yingxian Chen, Jiahui Liu, Ruidi Fan, Yanwei Li, Chirui Chang, Shizhen Zhao, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu. Aligning effective tokens with video anomaly in large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22695–22706, 2025.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 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*, pp. 18793–18803, 2024.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25750–25761. PMLR, 2025a. URL https://proceedings.mlr.press/v267/huang25ad.html.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-R1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 38, pp. 118486–118518, 2025b. doi: 10.52202/085713-3952. URL https://papers.nips.cc/paper_files/paper/2025/hash/abccc325c84dedf23dbe8de3f686c733-Abstract-Conference.html.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Craig Iaboni and Pramod Abichandani. Tri-modal fusion transformers for uav-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4373–4382, 2026.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357.

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.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9339–9350, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7057–7075. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.574.

Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*, 2024.

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.

Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022.

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *arXiv preprint arXiv:2510.22693*, 2025.

Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2226–2235, 2025.

Mengjingcheng Mo, Xinyang Tong, Mingpi Tan, Jiaxu Leng, Jiankang Zheng, Yiran Liu, Haosheng Chen, Ji Gan, Weisheng Li, and Xinbo Gao. A2Seek: Towards reasoning-centric benchmark for aerial anomaly understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5057. URL https://papers.nips.cc/paper_files/paper/2025/hash/de02de513503962e1d21035ab50ce661-Abstract-Datasets_and_Benchmarks_Track.html.

Zhihao Nie, Luyi Xue, Zhenyu Fang, Jinchang Ren, Yufeng Wei, and Jiangbin Zheng. M3ot: A multi-drone multi-modality dataset for multi-object tracking. *Scientific Data*, 12(1):1927, 2025.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. EventVAD: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025.

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

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. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024.

Tung Minh Tran, Tu N Vu, Tam V Nguyen, and Khang Nguyen. Uit-adrone: A novel drone dataset for traffic anomaly detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:5590–5601, 2023.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? rewarding perception for multimodal reasoning. *arXiv preprint arXiv:2605.14054*, 2026.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17882–17891, 2025.

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 *European Conference on Computer Vision*, pp. 304–322. Springer, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8896–8905, 2022.

Muchao Ye, Weiyang Liu, and Pan He. VERA: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8679–8688, 2025.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

Hao Zhang, Linfeng Tang, Xinyu Xiang, Xuhui Zuo, and Jiayi Ma. Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26487–26496, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13843–13853, 2025a.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8886–8895, 2022.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=EElFGvt39K.

Jiawen Zhu, Huayi Tang, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Shihao Qiu, Shengming Li, and Huchuan Lu. Dcpt: Darkness clue-prompted tracking in nighttime uavs. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7381–7388. IEEE, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a. URL https://arxiv.org/abs/2504.10479.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. VAU-R1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025b. URL https://arxiv.org/abs/2505.23504.

A AI USE STATEMENT

During manuscript preparation, we used large language models to assist with grammar and syntax checking, language polishing, and improvements to clarity and readability. This assistance included identifying awkward expressions, refining sentence structure, and checking the consistency of terminology. The authors reviewed and revised the suggested edits to ensure that they accurately conveyed the intended technical meaning and remained consistent with the paper's arguments. The authors take full responsibility for the final manuscript, including its scientific claims, experimental results, and conclusions.

B FURTHER FORMULATION OF CLOSED-LOOP EVIDENCE MINING

This section connects four definitions summarized in the main paper: how the closed-loop evidence state progresses, how carefully selected functional heads provide localization and token-ranking signals, how the anomaly-grounding proxy supplies cross-state reward feedback for learning backtracking decisions, and how EvPO assigns process feedback to the corresponding key evidence-bearing tokens and action tokens. The expanded architecture below exposes relations omitted from the compact overview.

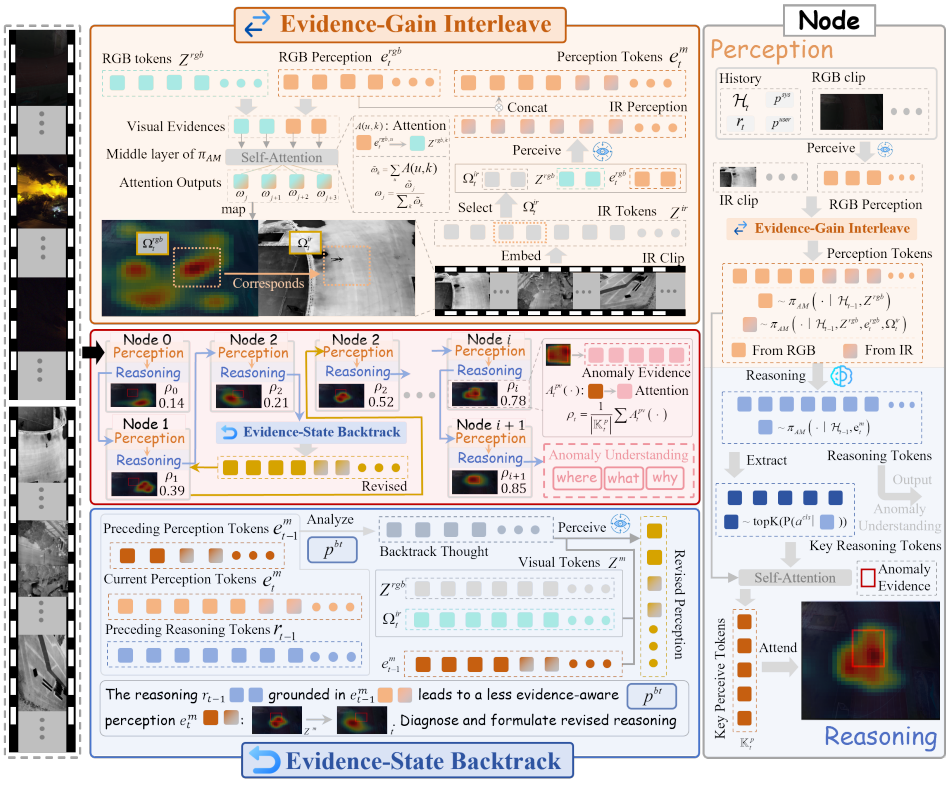


Figure 5: **Detailed realization of the closed-loop evidence trajectory.** At each node, RGB perception and, when admitted, localized infrared perception form the current evidence. Functional attention heads connect perceptual responses to visual tokens to localize the Ω region and compute the anomaly-grounding proxy. Reasoning both updates the anomaly hypothesis and specifies the next perceptual demand. Cross-state changes in the anomaly-grounding proxy provide a reward signal for learning when to reject a candidate and backtrack; upon backtracking, the model's diagnosis redirects subsequent perception. Key evidence-bearing tokens and action tokens provide the mechanism-aligned credit targets optimized by EvPO.

B.1 Closed-Loop Evidence Trajectory

AnomalyMiner treats each perception–reasoning update as a node on an evidence trajectory. Node t starts from the preceding committed state $\overline{\mathcal{S}}_{t-1}$, mines RGB evidence, optionally acquires a localized infrared observation, and then updates anomaly reasoning. Before this update is accepted, the new quantities form the candidate state $\widetilde{\mathcal{S}}_t$:

$$\widetilde{\mathcal{S}}_t = \big(\overline{\mathcal{S}}_{t-1}, e_t^{rgb}, e_t^{ir}, r_t\big), \tag{8}$$

Reasoning assesses whether the candidate supports a conclusion. If the model considers further evidence necessary, it specifies the next perceptual demand, and the candidate, once accepted, becomes the preceding state of the next node. If the model considers the available evidence sufficient, it produces the final anomaly judgment and explanation. Upon backtracking, the model rejects the candidate, retains the preceding committed state, and produces a diagnosis $\mathcal{B}_t$ that redirects subsequent perception.

B.2 Initializing an Executable Closed-Loop Policy

We initialize AnomalyMiner from Qwen2.5-VL-7B (Bai et al., 2025). A base vision–language model must first learn to generate according to the perception, reasoning, and state-transition semantics of the loop. We therefore use stage-level supervised fine-tuning (SFT) to establish an executable initial policy, select functional heads on that model, and then use EvPO to learn evidence admission and path repair.

SFT uses standard assistant-only cross-entropy to learn perceptual evidence generation, reasoning updates, trajectory decisions, and perception redirection after backtracking. It initializes the closed-loop behavior; later mechanism feedback optimizes how the policy responds to reasoning-gain and cross-state backtracking signals.

B.3 Functional Heads and Key Evidence-Bearing Tokens

B.3.1 Post-SFT Functional-Head Discovery

Attention heads in a multimodal model are functionally heterogeneous, and averaging all heads can obscure a small set of stable localization responses. Prior work shows that selected heads can localize image regions aligned with text semantics (Kang et al., 2025). We borrow only this localization-readout observation, not the claim that attention causally explains the final prediction (Jain & Wallace, 2019; Kobayashi et al., 2020). After SFT, we use forced-decoding forward passes over training responses to select RGB-localization, IR-localization, and evidence-to-reasoning families, with the selected head sets kept fixed thereafter.

For stage response i, let $A_i^{l,h}(q,k)$ be attention from response position q to context position k in layer l, head h. After masking control and EOS positions, we average over response queries $\mathcal{Q}_i$ to obtain the mean attention $\overline{A}_i^{l,h}(k)$:

$$\overline{A}_i^{l,h}(k) = \frac{1}{|\mathcal{Q}_i|} \sum_{q \in \mathcal{Q}_i} A_i^{l,h}(q,k). \tag{9}$$

Here, $|\mathcal{Q}_i|$ is the number of retained query positions; vertical bars around a set denote its cardinality throughout.

Each family g has function-aligned target positions $\mathcal{T}_i^g$: RGB and IR localization use the corresponding visual patches, while evidence-to-reasoning uses current perceptual-text spans. We define target mass $M_i^{l,h,g}$ as the fraction of the pre-response attention assigned to those positions:

$$M_i^{l,h,g} = \frac{\sum_{k \in \mathcal{T}_i^g} \overline{A}_i^{l,h}(k)}{\sum_{k \in \mathcal{U}_i} \overline{A}_i^{l,h}(k) + \epsilon}, \tag{10}$$

Here, $\mathcal{U}_i$ contains all context positions before the response, and $\epsilon > 0$ is a numerical stabilizer that prevents a zero denominator.

High mass can still arise from a response dispersed across an image or text. We therefore measure cohesion by thresholding target attention at its mean, forming 8-connected visual components or one-dimensional textual components with sizes $\{c_j\}$. We compute each component's share p_j of all active positions and the entropy $E_i^{l,h,g}$ of these shares:

$$p_j = \frac{c_j}{\sum_r c_r}, \qquad E_i^{l,h,g} = -\sum_j p_j \log p_j. \tag{11}$$

Here, j and the denominator index r enumerate connected components, and $\log$ denotes the natural logarithm. Lower entropy indicates concentration in fewer coherent regions.

We first retain heads with above-average target mass, then retain cohesive responses with below-average component entropy, and rank the remainder by low entropy, high mass, and high dominant-component ratio. The same fixed budget is used for each family, and the selected heads remain fixed thereafter.

B.3.2 Key Evidence-Bearing Tokens

Key evidence-bearing tokens comprise reasoning-side and perception-side subsets. Key always refers to the ranking rules below; it does not claim that an individual token has a proven causal effect on the final answer. The two subsets connect target-category support at the task end to perceptual-text positions.

Let the complete target-category token sequence be $y^* = (y_1^*, \ldots, y_J^*)$, and let x_t^r be the realized context of the current reasoning update. For every candidate reasoning position i, we compute the category-support score q_t^i as the autoregressive log-likelihood of the complete category sequence after the prefix ending before that position, and use these scores to select the reasoning-side key positions $\mathbb{K}_t^r$:

$$q_t^i = \sum_{j=1}^{J} \log \pi_\theta\big(y_j^* \mid x_t^r, r_t^{<i}, y_{<j}^*\big), \qquad \mathbb{K}_t^r = \mathrm{IdxTopK}_i(q_t^i, K^r). \tag{12}$$

Here, J is the target-category sequence length, j indexes its tokens, and $y_{<j}^*$ contains the preceding $j-1$ category tokens. The symbol θ denotes the policy parameters, and $\mathrm{IdxTopK}$ returns the indices of the highest-scoring candidates up to the specified budget.

The up-to-K^r highest-ranked positions are reasoning-side key evidence-bearing tokens: their preceding prefixes provide the strongest support for the target category. This is not a leave-one-token-out attribution.

Next, in the fixed evidence-to-reasoning heads $\mathcal{H}^{rp}$, we average attention from the reasoning-side key evidence-bearing positions i to each perceptual-text position u to obtain scores $s_t^{M,u}$, then select the highest-scoring positions within each available modality to form the perception-side key sets $\mathbb{K}_t^{p,M}$:

$$\begin{aligned} s_t^{M,u} &= \frac{1}{|\mathcal{H}^{rp}|\,|\mathbb{K}_t^r|} \sum_{(l,h)\in\mathcal{H}^{rp}} \sum_{i\in\mathbb{K}_t^r} A_t^{l,h}(i,u), \\ \mathbb{K}_t^{p,M} &= \mathrm{IdxTopK}_{u\in e_t^M}\big(s_t^{M,u}, K^p\big). \end{aligned} \tag{13}$$

Here, $M \in \{rgb, ir\}$ denotes the perceptual modality, K^p is the per-modality selection budget, and the subscript condition $u \in e_t^M$ restricts the candidates to that modality's perceptual-text positions. The attention $A_t^{l,h}(i,u)$ is from reasoning position i to perceptual position u at update t; the superscript p marks perception-side quantities, and rp marks the reasoning-to-perceptual-text attention direction.

A perception-side key evidence-bearing token therefore means a perceptual-text token receiving high mean attention under this fixed head/query rule. It is not a feature-importance or causal-explanation claim. Available modality-specific sets are united as $\mathbb{K}_t^p$.

B.3.3 Anomaly-Grounding Proxy

Attention alone does not establish a causal explanation of the final prediction (Jain & Wallace, 2019; Kobayashi et al., 2020). However, prior visual-grounding work uses attention mass within task-

relevant regions to measure spatial alignment (Hudson & Manning, 2019). Following this measurement, we define ρ_t as the anomaly-grounding proxy: it measures how much attention perception-side key evidence-bearing tokens allocate, through fixed visual-localization heads, to the anomaly region, providing a signal for process supervision and cross-state diagnosis.

Localized infrared attention, when available, is mapped back to Ω_t on the RGB spatial grid. For each query and head, we form the cross-modal attention map $U_t^{l,h}(u,s)$ by taking the maximum across modalities at corresponding visual positions. We compute ρ_t by summing this map within the dataset-provided reference anomaly region $\mathbb{A}$, then averaging over localization heads and perception-side key evidence-bearing positions:

$$U_t^{l,h}(u,s) = \max_{M \in \mathcal{M}_t} A_t^{M,l,h}(u,s),$$
$$\rho_t = \frac{1}{|\mathbb{K}_t^p|} \sum_{u \in \mathbb{K}_t^p} \frac{1}{|\mathcal{H}_u^{\mathrm{loc}}|} \sum_{(l,h) \in \mathcal{H}_u^{\mathrm{loc}}} \sum_{s \in \mathbb{A}} U_t^{l,h}(u,s) \le 1. \tag{14}$$

Here, $\mathcal{M}_t$ is the set of available modalities, and $\mathcal{H}_u^{\mathrm{loc}}$ is the fixed localization family associated with perceptual position u. The index s identifies a visual position on the shared RGB grid, and $A_t^{M,l,h}(u,s)$ is modality M's visual attention for that query and head, mapped onto the grid. A modality absent from the query's context contributes zero attention. The proxy is disabled for normal samples and anomalous samples without a valid box.

The spatial mining reward $\mathcal{R}_{\mathrm{mine}}^t$ is defined by the proxy change from the preceding committed state:

$$\mathcal{R}_{\mathrm{mine}}^t = \rho_t - \bar{\rho}_{t-1}. \tag{15}$$

The initial round receives zero mining reward because no preceding evidence state is available; the reward is also zero when either proxy cannot be evaluated. A rejected candidate does not update $\bar{\rho}_{t-1}$. Thus, ρ is used only for relative grounding changes under the same sample, model, and fixed-head rule; it does not quantify how much anomaly evidence the model causally used.

B.3.4 Localization of the Ω Region

Selected localization heads can use text-to-image attention to identify spatial locations aligned with text semantics (Kang et al., 2025). Following this observation, we locate the visual token receiving the strongest aggregated attention from the current RGB perceptual response and expand its corresponding spatial region to retrieve complementary infrared information.

Let $\mathcal{H}^{rgb}$ denote the fixed RGB-localization heads, $\mathcal{E}_t^{rgb}$ the RGB perceptual-text positions, and $\mathcal{Z}^{rgb}$ the RGB visual-token positions. We aggregate attention over the localization heads and perceptual-text positions as

$$\alpha_t(k) = \frac{1}{|\mathcal{H}^{rgb}|\,|\mathcal{E}_t^{rgb}|} \sum_{(l,h) \in \mathcal{H}^{rgb}} \sum_{q \in \mathcal{E}_t^{rgb}} A_t^{l,h}(q,k), \qquad k \in \mathcal{Z}^{rgb}. \tag{16}$$

The highest-response visual token determines the region to be expanded:

$$k_t^\star = \arg\max_{k \in \mathcal{Z}^{rgb}} \alpha_t(k), \qquad \Omega_t = \mathrm{Expand}_{\gamma_\Omega}\big(\mathcal{P}(k_t^\star)\big). \tag{17}$$

Here, $\mathcal{P}(k_t^\star)$ denotes the image region represented by the selected visual token. The operator $\mathrm{Expand}_{\gamma_\Omega}$ enlarges this region about its center by a factor of γ_Ω in both width and height, with the resulting boundaries constrained to the image. We set $\gamma_\Omega = 1.5$. The resulting region Ω_t is mapped to the paired infrared observation to obtain the localized input for Evidence-Gain Interleave.

B.4 Evidence-Gain Interleave: Admitting Additional Evidence for Reasoning

Additional visual information can replenish cues omitted by fixed RGB perception, but uncontrolled infrared input can also introduce salient content unrelated to the anomaly. Evidence-Gain Interleave therefore learns whether current reasoning needs a localized infrared read. The interleaving action a_t^{int} is

$$a_t^{int} \sim \pi_\theta\Big(\cdot \mid Z^{rgb}, e_t^{rgb}\Big), \qquad a_t^{int} \in \{0,1\}, \tag{18}$$

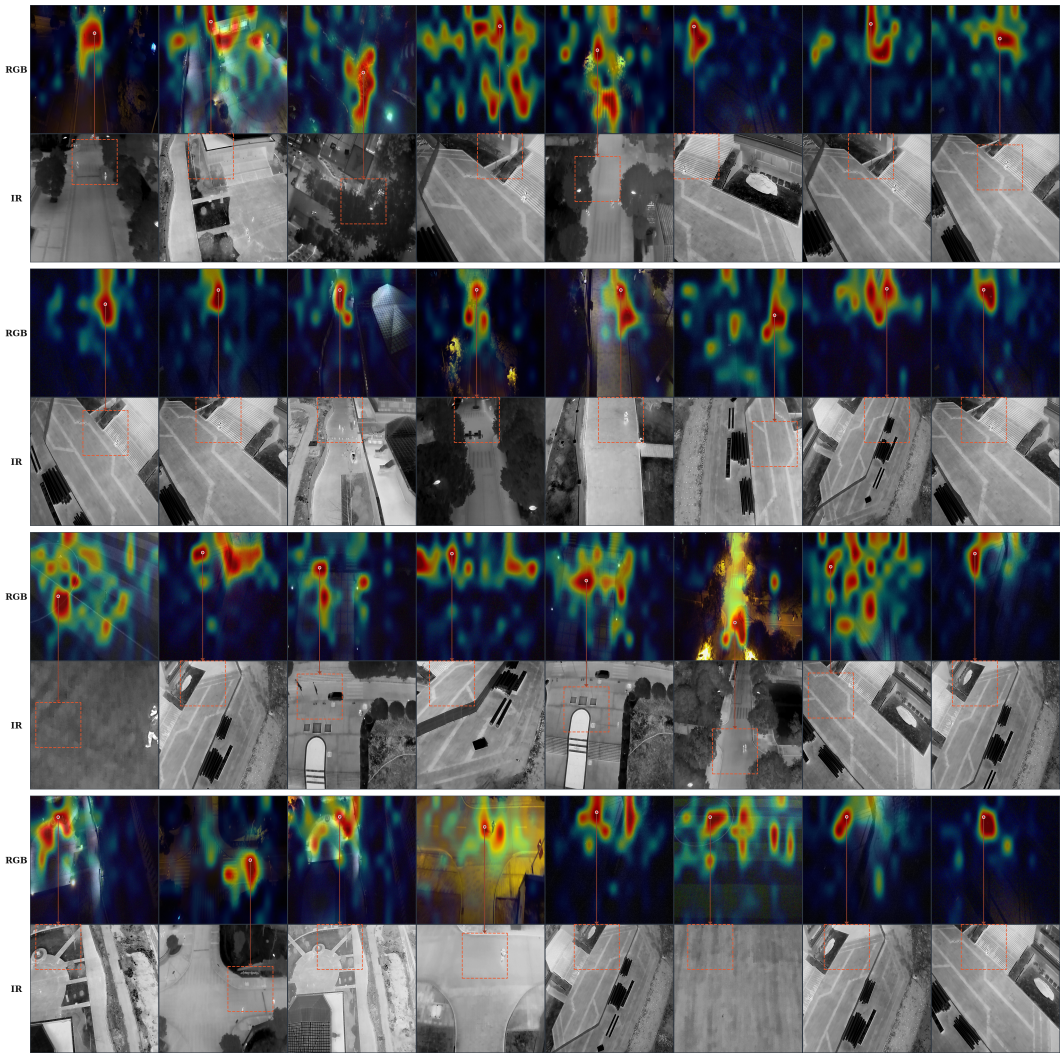


Figure 6: RGB-guided infrared region retrieval during interleaving. For each pair of rows, the upper row visualizes RGB attention, and the lower row shows the aligned infrared view. The marked point denotes the visual token receiving the strongest RGB attention. Its normalized location is transferred to the infrared view and enlarged into the dashed region supplied to infrared perception.

During interleaving, the model reads the local infrared region from Section B.3.4 and generates complementary perceptual text e_t^{ir},

$$e_t^{ir} \sim \pi_\theta\Big(\cdot \mid Z^{rgb}, Z^{ir}_{\Omega_t}, e_t^{rgb}\Big). \tag{19}$$

where $Z^{ir}_{\Omega_t}$ denotes the infrared visual tokens from region Ω_t. Without interleaving, the node updates reasoning from RGB perception alone.

Figure 6 illustrates how RGB attention guides the retrieval of localized infrared observations during interleaving.

B.4.1 Validating Admission by Reasoning Gain

A matched counterfactual compares the model's actual decision with its alternative. We fix the realized reasoning and the reasoning-side key evidence-bearing tokens selected from it, and compare contexts x_t^I and x_t^{-ir} with and without the infrared perceptual text, respectively, keeping the remaining context unchanged. When the model skips infrared, an auxiliary infrared perception is generated

from the same RGB perception and local-region construction solely for counterfactual scoring. This auxiliary perception is excluded from the realized reasoning trajectory, grounding measurements, and subsequent state updates.

For reasoning-side key evidence-bearing position i, we define category support $q_t^{i,\zeta}$ as the autoregressive log-likelihood of the target-category sequence under that reasoning prefix. The interleaving gain Δ_t^{int} averages the with-IR minus without-IR difference over these positions:

$$\begin{aligned} q_t^{i,\zeta} &= \sum_{j=1}^{J} \log \pi_\theta\Big(y_j^* \mid x_t^\zeta, r_t^{<i}, y_{<j}^*\Big), \quad \zeta \in \{I, -ir\}, \\ \Delta_t^{\text{int}} &= \frac{1}{|\mathbb{K}_t^r|} \sum_{i \in \mathbb{K}_t^r} \Big(q_t^{i,I} - q_t^{i,-ir}\Big). \end{aligned} \tag{20}$$

where ζ indexes the contexts with (I) and without $(-ir)$ infrared perceptual text. A positive gain means that the infrared content increases support for the target category at the selected reasoning prefixes; a negative gain means that it reduces this support. The interleaving reward measures the advantage of the model's actual decision over its alternative, rewarding the decision when it provides stronger category support and penalizing it otherwise:

$$\mathcal{R}_{\text{int}}^t = (2a_t^{int} - 1)\, \Delta_t^{\text{int}}. \tag{21}$$

Evidential Preference Policy Optimization uses the counterfactual reward to optimize the interleaving decision, retaining baseline credit across generated tokens while assigning additional weight to feedback at the same-round reasoning-side key evidence-bearing tokens and the action tokens for requesting or skipping infrared. Forced skips caused by unavailable infrared observations have no valid paired counterfactual and receive no interleaving-process credit.

B.5 Evidence-State Backtrack: Correcting Direction with a Preceding State

Interleaving can improve current reasoning without guaranteeing that the updated reasoning will direct the next perception toward a better region. Search-based backtracking commonly revisits an earlier branch when an intermediate state is unpromising (Yao et al., 2023; Koh et al., 2024). Evidence-State Backtrack instead uses changes in anomaly grounding from the preceding committed state to the current candidate as reward feedback for learning when to backtrack. This feedback encourages the policy to repair a drifting evidence-search direction even when the current interpretation remains locally plausible.

Let $\overline{\mathcal{S}}_{t-1}$ be the preceding committed state and $\widetilde{\mathcal{S}}_t$ the current uncommitted candidate, with anomaly-grounding proxies $\bar{\rho}_{t-1}$ and $\tilde{\rho}_t$. We define their relative grounding decline γ_t as

$$\gamma_t = \frac{\bar{\rho}_{t-1} - \tilde{\rho}_t}{\bar{\rho}_{t-1} + \epsilon_\rho}. \tag{22}$$

A positive value means only that anomaly grounding is lower for the candidate than for the preceding state under the same fixed protocol. It neither asserts that the current answer is already wrong nor provides a causal explanation. This relative decrease supplies the backtracking reward when the required grounding-proxy values are available.

Conditioned on both the preceding committed state and the current candidate, the policy determines whether reasoning should return to the preceding state, continue from the current state, or conclude:

$$a_t^{bt} \sim \pi_\theta\Big(\cdot \mid \overline{\mathcal{S}}_{t-1}, \widetilde{\mathcal{S}}_t\Big), \quad a_t^{bt} \in \{\texttt{BACKTRACK}, \texttt{CONTINUE}, \texttt{CONCLUDE}\}. \tag{23}$$

Let $b_t = \mathbb{I}[a_t^{bt} = \texttt{BACKTRACK}]$. The reward couples the action to cross-state direction: a significant proxy decrease rewards backtracking, whereas an increase rewards continuing or concluding,

$$\mathcal{R}_{\text{bt}}^t = \mathbb{I}[|\gamma_t| > \tau_{\text{bt}}](2b_t - 1)\gamma_t. \tag{24}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function, equal to one when its condition holds and zero otherwise.

B.5.1 Rejecting the Candidate and Re-entering the Loop

A valid backtrack does not merely delete a reasoning sentence. The model compares the preceding state with the deviated candidate and generates a diagnosis $\mathcal{B}_t$ that analyzes the lost or displaced anomaly cue and formulates what the next perception should inspect. The rejected candidate does not overwrite the committed state.

After backtracking or continuing, the committed state $\overline{\mathcal{S}}_t$ is updated according to

$$\overline{\mathcal{S}}_t = \begin{cases} \overline{\mathcal{S}}_{t-1}, & a_t^{bt} = \texttt{BACKTRACK}, \\ \widetilde{\mathcal{S}}_t, & a_t^{bt} = \texttt{CONTINUE}. \end{cases} \tag{25}$$

The conclude action instead enters the final-answer stage from the candidate.

Let $\bar{r}_{t-1}$ denote the reasoning retained in the preceding committed state. After a backtrack, the model generates the next RGB perceptual text $\widehat{e}_{t+1}^{rgb}$ from this reasoning and the diagnostic goal, before proceeding with the same interleaving and reasoning updates:

$$\widehat{e}_{t+1}^{rgb} \sim \pi_\theta\big(\cdot \mid Z^{rgb}, \bar{r}_{t-1}, \mathcal{B}_t\big)\,. \tag{26}$$

Backtracking therefore repairs the evidence-search direction without removing the preceding committed state.

B.6 Evidential Preference Policy Optimization

Training proceeds from the base model to stage-level SFT, functional-head discovery, and EvPO. SFT establishes an executable loop; functional heads are selected on the SFT model and then frozen; EvPO learns evidence mining, admission, and path repair.

Group-relative optimization compares multiple trajectories sampled for the same input (Shao et al., 2024). Applying one trajectory signal uniformly to every generated token, however, dilutes feedback at the few positions that implement closed-loop decisions. EvPO retains the within-group comparison and baseline credit across generated tokens, while assigning additional weight to each process signal at the corresponding key evidence-bearing tokens and action tokens.

B.6.1 Mechanism-Aligned Credit

For each input, a group of trajectories is sampled. Each feedback unit c identifies a process signal together with its update round, or a trajectory-level outcome component. We normalize each unit independently within the group to obtain its group-relative advantage $\widehat{A}_g^c$. For interleaving feedback, the group statistics use only trajectories with a valid paired counterfactual. Thus, a signal for one interleaving or backtracking decision is not overwritten by the scale of another round or the final outcome:

$$\widehat{A}_g^c = \frac{R_g^c - \mu_G^c}{\sigma_G^c + \epsilon_A}, \tag{27}$$

where G denotes the group of trajectories sampled for the same input, and R_g^c is the reward for feedback unit c in trajectory g. The quantities μ_G^c and σ_G^c are the group mean and standard deviation of that unit's rewards, restricted to valid paired trajectories for interleaving feedback as specified above. The constant $\epsilon_A > 0$ is a numerical stabilizer.

Let $\mathcal{K}_g^c$ be the set of token positions receiving additional weight for feedback unit c in trajectory g. The EvPO advantage $A_{g,i}^{\mathrm{EvPO}}$ at position i combines baseline credit with additional process-specific credit:

$$A_{g,i}^{\mathrm{EvPO}} = \sum_c \big(1 + \lambda_c \mathbb{I}[i \in \mathcal{K}_g^c]\big)\, \widehat{A}_g^c, \tag{28}$$

where $\lambda_c > 0$ controls the additional weight for feedback unit c, and $\mathbb{I}[\cdot]$ is the indicator function. Positions outside $\mathcal{K}_g^c$ retain unit weight, while positions within it receive weight $1 + \lambda_c$.

Mining feedback receives additional weight at perception-side and reasoning-side key evidence-bearing tokens in the same update; when a valid paired counterfactual is available, interleaving feedback receives additional weight at reasoning-side key evidence-bearing tokens and the action

tokens for requesting or skipping infrared; and backtracking feedback receives additional weight at the trajectory decision. Outcome feedback receives additional weight at the final-answer fields and the key evidence-bearing tokens selected in the final reasoning update. Other positions retain baseline credit, while reference-policy regularization covers the generated sequence. EvPO is therefore distinguished by the additional weight assigned to mechanism feedback at the corresponding positions.

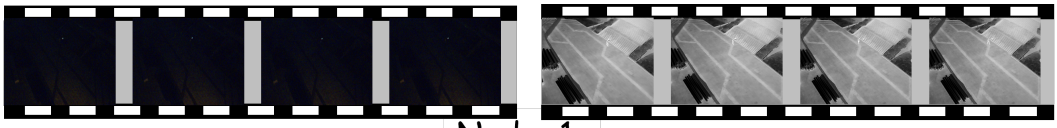
C Complete Rollout Demonstration

Appendix B.1 summarizes the node-level state transitions. To show how they compose an actual inference trajectory, we present a complete rollout on a nighttime loitering example. This visualization illustrates the model's observable trajectory decisions rather than serving as an additional aggregate performance comparison. In particular, it exposes the evidence available at each node, the modality-interleaving decision, the resulting reasoning-state update, and the transition to the next node.

The first part of Figure 7 shows the first node. The RGB clip is severely underexposed and provides insufficient evidence for identifying the small foreground actor. The model therefore emits <INTERLEAVE>, localizes the visually relevant region, and retrieves the corresponding infrared observation. Although the complementary thermal cue reveals a human target, the evidence is not yet sufficient for a reliable anomaly decision, so the model retains a conservative hypothesis and continues perception.

In the second part of Figure 7, the newly acquired evidence resolves the target's temporal persistence and limited motion. The reasoning state is consequently revised from Normal to Loitering, while the remaining uncertainty is expressed as a concrete next-perception goal rather than an immediate terminal prediction. This transition illustrates that modality interleaving is used to update the evidence-grounded hypothesis, instead of merely appending an infrared description to the RGB observation.

In the final part of Figure 7, the accumulated evidence is judged sufficient. The model therefore emits <SKIP_IR> instead of performing another redundant infrared retrieval, preserves the supported loitering hypothesis, and concludes with <CONCLUDE>. It then produces the machine-readable answer shown at the end of Figure 7, including the anomaly status, canonical category, supporting spatial location, and a concise evidence-grounded explanation. Taken together, the rollout demonstrates conditional cross-modal acquisition, iterative hypothesis revision, and structured termination under the shared control-token protocol.



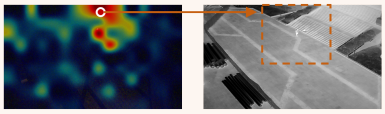
Node_1

Perception

[RGB PERCEPTION]

The video depicts a dimly lit environment, likely captured during nighttime conditions, with minimal visible details due to low lighting. The scene appears to be outdoors, possibly near a structure such as a wall or building, as suggested by faint vertical lines or edges in the background. There is no discernible movement or activity within the frames provided, indicating a static or slow-paced setting. The overall darkness and lack of clear features make it challenging to identify specific objects or individuals, contributing to an ambiguous and somewhat eerie atmosphere. The absence of significant motion or identifiable elements suggests a focus on subtle anomalies or unusual patterns that might emerge over time.<PERCEPTION_END>

[INTERLEAVE DECISION - INTERLEAVE]



[IR PERCEPTION]

The infrared observation reveals a distinct thermal signature corresponding to a human figure moving across the scene. This figure exhibits a consistent heat profile indicative of a person, contrasting sharply with the cooler surroundings. The movement is persistent and linear, suggesting a walking or running action. The thermal data provides additional detail about the figure's shape and trajectory, which aligns with the general direction observed in the RGB footage but offers enhanced clarity due to the thermal contrast. The surrounding area remains relatively uniform in temperature, further emphasizing the figure's presence and movement.<PERCEPTION_END>

Reasoning

[REASONING]

Current Best-Supported Hypothesis: Normal

Perceptual Evidence Supporting or Contradicting Hypothesis:

- The scene is dimly lit with minimal visible details, suggesting a static or slow-paced setting.
- A distinct thermal signature indicates a human figure moving consistently across the scene, exhibiting a walking or running action.

Evidence Sufficiency: Insufficient

Next Perception Goal: Identify the figure's posture, gait, or any associated objects to determine if the movement aligns with typical human behavior.<REASONING_END>

[NEXT ACTION - CONTINUE]

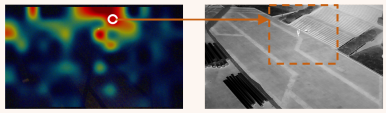
Node_2

Perception

[RGB PERCEPTION]

The scene remains consistently dimly lit throughout the sequence, maintaining a dark ambiance with limited visibility. There is no discernible change in lighting conditions or environmental elements, reinforcing the static nature of the setting. The absence of significant movement or activity suggests a lack of dynamic events occurring within the frame. The overall lack of detail and consistency supports the hypothesis of normalcy, as there are no apparent anomalies or unusual behaviors observed. The scene continues to appear calm and undisturbed, with no notable deviations from expected patterns.<PERCEPTION_END>

[INTERLEAVE DECISION - INTERLEAVE]



[IR PERCEPTION]

The infrared observation reveals a distinct thermal signature corresponding to a human figure located near the center-right of the frame, persistently present across multiple frames. This figure exhibits minimal movement, suggesting a stationary posture consistent with loitering behavior. The thermal contrast between the figure and the surrounding environment highlights its presence more clearly than in the RGB footage, where details are obscured due to low lighting conditions. The thermal data indicates a stable heat emission pattern over time, further supporting the notion of prolonged stay without significant activity. This evidence complements the RGB findings by providing additional confirmation of the figure's presence and lack of dynamic movement, enhancing the perception of potential loitering activity.<PERCEPTION_END>

Reasoning

[REASONING]

Current Best-Supported Hypothesis: Loitering

Perceptual Evidence Supporting or Contradicting Hypothesis:

- The scene is dimly lit with minimal visible details, indicating a static or slow-paced environment.
- A distinct thermal signature shows a human figure near the center-right, persistently present with minimal movement, suggesting a stationary posture.

Evidence Sufficiency: Insufficient

Next Perception Goal: Determine the figure's proximity to structures or objects to assess intent or purpose for remaining stationary.<REASONING_END>

[NEXT ACTION - CONTINUE]

Node_3

Perception

[RGB PERCEPTION]

The scene remains consistently dimly lit across the frames, with very limited visibility due to low ambient lighting conditions. There is no discernible change in the environment or additional objects entering the frame, maintaining a static and unchanging setting. The lack of significant movement or activity suggests a continued focus on the potential loitering behavior observed earlier. The overall darkness and absence of clear features make it challenging to identify specific elements or interactions within the scene.<PERCEPTION_END>

[INTERLEAVE DECISION – SKIP IR]

Reasoning

[REASONING]

Current Best-Supported Hypothesis: Loitering

Perceptual Evidence Supporting or Contradicting Hypothesis:

- The scene is consistently dimly lit with minimal visible details, reinforcing the static nature of the environment.
- A persistent thermal signature indicates a human figure near the center-right, showing minimal movement over time, consistent with a stationary posture.

Evidence Sufficiency: Sufficient

Next Perception Goal: Investigate the spatial relationship between the figure and nearby structures or objects to determine if there is a clear intent or purpose for their stationary position.<REASONING_END>

[NEXT ACTION - CONCLUDE]

Final Answer

<status>abnormal</status>

<what>Loitering</what>

<where>[487, 163, 509, 216]</where>

<why>The scene remains consistently dimly lit with minimal visible details, suggesting a static environment. A persistent thermal signature indicates a human figure near the center-right, showing minimal movement over time, which is consistent with a stationary posture.</why>

<ANSWER_END>

Figure 7: Complete multiround rollout on a nighttime loitering case. The first two nodes interleave infrared observations to resolve evidence that is ambiguous in the underexposed RGB clip. The complementary thermal observation updates the best-supported hypothesis from Normal to Loitering. Once the accumulated evidence is sufficient, the final node skips redundant infrared acquisition, terminates the collaborative analysis, and produces a structured answer containing the anomaly status, canonical category, localization, and evidence-grounded explanation.

D SUPPLEMENTARY EXPERIMENTS

D.1 ADDITIONAL ABLATION STUDIES

D.1.1 TOKEN SELECTION

Table 4: Token selection for process feedback and comparison of backtracking states. The mechanism blocks are cumulative: interleaving retains mining, and backtracking retains both mining and interleaving. Random Tokens uses the same number of tokens as Key Tokens, which denotes key evidence-bearing tokens.

Mechanism	**Token Selection**			**Backtracking State**		**AP (%)**
	All Tokens	**Random Tokens**	**Key Tokens**	**Current State**	**Preceding State**	
Mining	✓	–	–	–	–	25.84
	–	✓	–	–	–	25.17
	–	–	✓	–	–	30.46
Interleave	✓	–	–	–	–	33.26
	–	✓	–	–	–	32.71
	–	–	✓	–	–	37.14
Backtracking	–	–	–	✓	–	38.91
	–	–	–	–	✓	**45.18**

Table 4 examines the effect of token selection for computing process signals on anomaly-understanding performance. Using our key evidence-bearing tokens achieves 30.46% AP for evidence mining and 37.14% after adding interleaving. Averaging over all candidate token positions instead reduces AP to 25.84% and 33.26%, respectively. This decrease suggests that including less informative token positions in the average dilutes the process signals. To test whether the advantage follows simply from reducing the number of contributing tokens, we randomly sample a subset matched in size to the key-token set. This yields 25.17% and 32.71% AP, respectively, remaining close to the all-token results. The advantage of key-token selection under the same token budget supports the importance of which positions are used to compute the process signals.

The token-selection comparisons above exclude backtracking. We then add backtracking to the variant combining mining and interleaving with key evidence-bearing tokens. Backtracking based solely on the current state achieves 38.91% AP, whereas using the preceding state reaches 45.18%. This improvement suggests that reliance on the current state alone may preserve the reasoning deviation underlying a misdirected evidence search, while the preceding state provides a reference for correcting its direction.

D.1.2 EVPO CREDIT ALLOCATION

Table 5: Ablation of credit allocation to evidence-bearing positions. All concentrated variants use the same token budget. "Norm." denotes per-trajectory rescaling of token weights to unit mean.

Allocation target	Percept.	Reason.	Norm.	AP (%)
Uniform (GRPO)	–	–	–	41.76
Random key	–	–	✓	36.74
Non-key	–	–	✓	35.92
Perception key only	✓	–	✓	41.71
Reasoning key only	–	✓	✓	43.26
P.+R. key, unnorm.	✓	✓	–	42.84
OURS	✓	✓	✓	**45.18**

We further test whether EvPO benefits specifically from assigning additional credit to key evidence-bearing tokens. Random-key, non-key, perception-only, and reasoning-only variants use matched

token budgets. In Table 5, "Norm." rescales the token weights after reweighting within each trajectory to have a mean of one, preserving their relative ratios; "unnorm." applies the reweighting without this rescaling. Full EvPO obtains 45.18%, exceeding uniform-credit GRPO at 41.76% and substantially outperforming random or non-key allocation. Joint perceptual and reasoning credit is stronger than either set alone, while removing normalization also degrades performance. These comparisons support assigning greater relative weight to key evidence-bearing tokens from both perception and reasoning.

D.1.3 Effect of Infrared Input on Aerial Anomaly Understanding

We assess the effect of infrared input on aerial anomaly understanding by comparing AnomalyMiner with and without infrared observations on the nighttime scenes of A2Seek. Both variants use the same trained model and evaluation settings, with the latter relying solely on RGB observations.

Removing infrared input reduces AP from 45.18% to 34.22%, with performance declining across all five nighttime scenes. These results show that infrared observations provide useful additional information for aerial anomaly understanding, some of which constitutes anomaly-relevant evidence.

Table 6: Effect of complementary infrared evidence across the nighttime scenes of the A2Seek benchmark. The two variants share the same trained model and rollout settings; only infrared availability is changed.

Variant	**Nighttime Scene**					**AP (%)**
	Roadway	**Sidewalk**	**Playground**	**Rooftop**	**Wall Zone**	**Avg.**
AnomalyMiner w/o IR	33.92	20.26	13.91	19.74	20.08	34.22
AnomalyMiner	48.62	33.75	39.98	47.76	60.51	**45.18**

E Reproducibility Settings

We follow the official A2Seek training and evaluation partitions and preserve their original normal/anomalous distribution. The same training partition is used for closed-loop policy initialization and EvPO, and all ablations share the same visual preprocessing and decoding conditions. VTUAV and M3OT are used solely for testing, without training or fine-tuning on either dataset. Both SFT and EvPO are conducted in BF16 on two NVIDIA H800 GPUs.