SEEK-VAU: Towards Evidence-Faithful Video Anomaly Understanding via Agentic Search

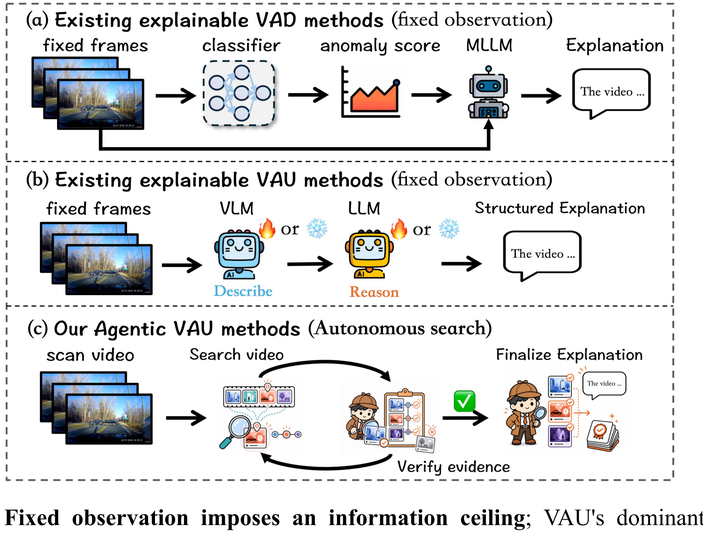
Minghe Wang[1], Weijun Zeng[2], Dapeng Luo[3], Xinyue Wang[3], Nong Sang[1*]

[1]Huazhong University of Science and Technology, [2]Peking University, [3]China University of Geosciences

Code & Paper

https://github.com/MingheWang/SEEK-VAU

Motivation & Contribution



1. Fixed observation imposes an information ceiling; VAU's dominant narrative is "reason better", with stronger MLLMs, longer CoT, and R1-style reinforcement. Yet observation barely changes, as methods still infer feed-forward from fixed frames or events pre-selected by hand-crafted rules.

2. Understanding an anomaly is essentially the ability to recover the anomalous event. An anomaly is rarely a single frame but a causal temporal process: precursors accumulate, a trigger marks the transition, aftermath records the impact. Yet existing VAU benchmarks evaluate only final answers over fixed observations, leaving event-chain recovery itself unmeasured.

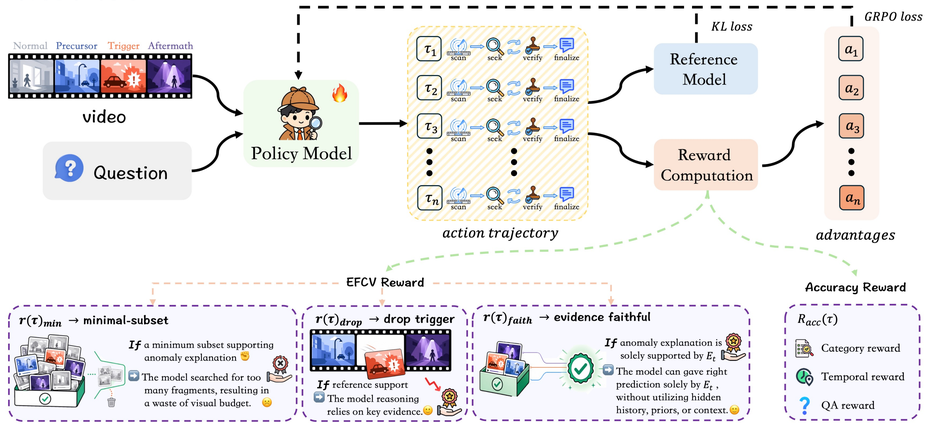
3. **From acquisition to recovery.** We therefore recast VAU as an agentic search-and-verify protocol under a bounded budget: the policy seeks precursors and aftermaths through `{scan, seek, verify, finalize}`; EFCV rewards optimize the search policy, teaching it where to look and when to stop for efficient search; and SEEK-Bench evaluates event-chain recovery itself, making completeness designable, optimizable, and verifiable.

Formulation

Given a video V and a case-level query Q,
find a anomaly case $y = (\text{existence}, \text{category}, \text{grounded interval}, \text{explanation})$ together with the committed evidence set E_T, under a bounded search budget.
Target event chain $C = (C_{\text{pre}}, C_{\text{trg}}, C_{\text{aft}})$ covering Precursor → Trigger → Aftermath, missing stages are allowed;
Recover as much of the event chain as the budget permits.
The policy π acts through `{scan, seek, verify, finalize}`, interleaving search with reasoning:

$$P_\pi(y, E_T \mid V, Q) = \underbrace{P_\pi(E_T \mid V, Q)}_{\text{evidence acquisition}} \cdot \underbrace{P_\pi(y \mid E_T, V, Q)}_{\text{case finalization}}$$

Method



SEEK-VAU: Agentic Search-and-Verify

- Policy interleaves reasoning with evidence acquisition: `scan → seek → verify → finalize`
- verify gates closure: insufficient or redundant support reroutes search to the missing stage
- Bounded budget: 6-frame protected scan, ≤16 active frames, ≤18 frames per episode

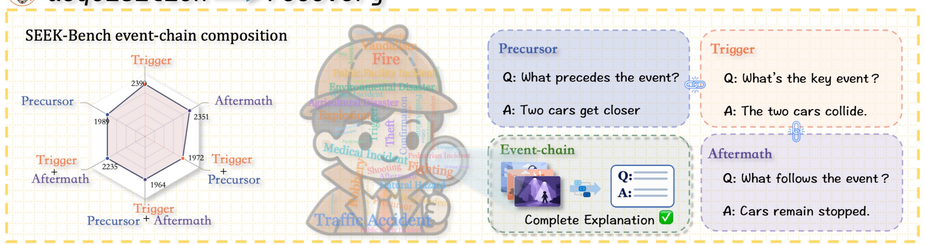
$$R(\tau) = w_{acc}R_{acc} + w_{efcv}(r_{faith} + r_{min} + r_{drop})$$

- EFCV rewards:

$$\begin{cases} r_{faith}(\tau): \text{The selected evidence alone must keep the anomaly confidence above the threshold } \theta. \\ r_{min}(\tau): \text{ Rewarding proximity to the minimal sufficient core, penalizing redundant windows.} \\ r_{drop}(\tau): \text{Removing the trigger evidence must measurably drop the anomaly confidence} \end{cases}$$

SEEK-Bench

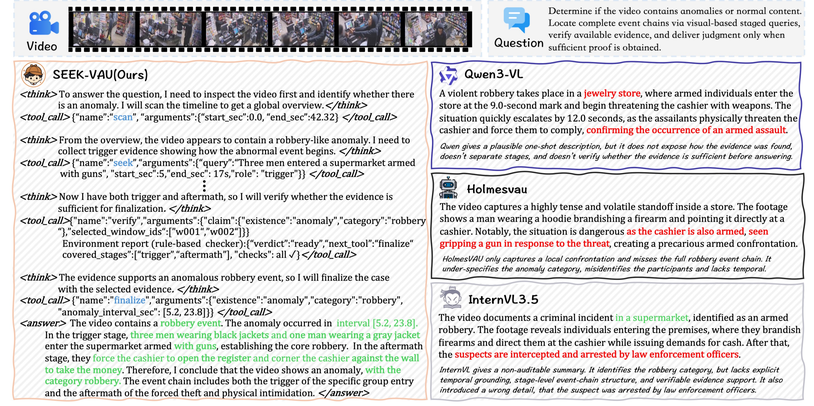
acquisition → recovery



Annotating the Event, Not Just the Answer, From Acquisition to Recovery.

We introduce SEEK-Bench, the first benchmark for interactive search-and-verify VAU, built on the observation that an anomalous event is a temporally extended causal process: precursors accumulate, a trigger marks the transition, and the aftermath records the consequence. SEEK-Bench annotates the event itself: across 2,878 episodes spanning 21 anomaly categories, temporal intervals and semantic QA are linked to explicit precursor-trigger-aftermath stage labels, so that event-chain completeness becomes directly supervisable and measurable.

Experiments



Comparison across model families

Method	Base Model	Temporal mIoU	QA Accuracy	Evidence F1
Video MLLMs				
GLM-4.1v [9]	GLM-4.1v-8B	4.6	16.4	6.5
Kimi-VL [17]	Kimi-VL-A3B	7.3	16.3	0.6
Intern-VL3.5 [18]	InternVL3.5-8B	25.5	19.4	14.3
Qwen3-VL [1]	Qwen3-VL-8B	44.2	37.7	28.1
Video Reasoning MLLMs				
Video-R1 [7]	Qwen2.5-VL-7B	9.8	8.4	0.0
VideoChat-R1 [13]	Qwen2.5-VL-7B	10.0	13.3	7.8
mimo-VL [12]	mimo-VL-8B	39.4	16.5	27.6
Ovis2.5 [14]	Qwen3-8B	41.8	24.7	22.7
MLLM-based VAU Methods				
HAWK [16]	LLaMA-2 7B	3.7	19.2	6.3
Holmes-VAU [28]	InternVL2-2B	7.4	20.7	0.0
VERA [26]	InternVL2-8B	31.5	18.6	6.1
Ours				
SEEK-VAU (ours)	Qwen3-VL-8B	**55.6**	**45.5**	**43.7**

Ablation study

Matched acquisition-mode ablation.

Method	Acquisition	Verify	T. mIoU	QA Acc.	E.F1	Frames
Two-stage matched	One-shot	×	48.2	31.1	32.3	18.0
Closed-loop w/o verify	Closed-loop	×	50.7	33.5	29.1	13.2
SEEK-VAU w/o EFCV	Closed-loop	√	54.4	42.7	41.6	17.1
SEEK-VAU (full)	Closed-loop	√	**55.6**	**45.5**	**43.7**	15.6

Inference efficiency ablation.

Model	Avg. frames	Avg. tokens	Latency (s)	Throughput (ep/min)
Qwen3-VL	18.0	4.33k	8.71	20.22
SEEK-VAU w/o EFCV	17.1	7.17k	10.32	18.45
SEEK-VAU	15.6	5.78k	9.27	19.31

Visual-context cap ablation.

Cap	Temporal mIoU	QA Accuracy	Evidence F1
12	51.6	44.1	38.2
14	53.2	43.8	39.3
16	**55.6**	45.5	**43.7**
18	53.6	45.7	38.2
20	55.3	**45.9**	42.8

Attribution ablation.

Method	T. mIoU	QA Acc.	E.F1
Fixed obs. + basic sup.	44.2	37.7	28.1
Fixed obs. + rich sup.	46.9	38.2	28.3
Active + basic sup.	50.1	40.1	34.3
Active + rich sup. (w/o EFCV)	54.4	42.7	41.6
SEEK-VAU (full)	**55.6**	**45.5**	**43.7**