MEMVAU: AGENTIC VISUAL EVIDENCE MEMORY FOR VIDEO ANOMALY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Anomaly Understanding (VAU) requires reports of what happened, who was involved, where it occurred, and how it unfolded. Single-pass vision-language models map a video directly to a report without exposing dimension-wise visual support, making weak actors or temporal relations difficult to repair. We introduce **MemVAU**, an agentic framework with an explicit four-dimensional evidence state. An **Observer** proposes localized evidence, a frozen **Verifier** assigns model-based support scores, and a rule-based **Mem-Optimizer** maintains a bounded pool by prioritizing dimensional gap, threshold surplus, and secondary support. Unresolved dimensions guide further acquisition until all thresholds are met or the round budget is exhausted, after which a **Reporter** compiles the pool without acquiring new evidence. We train the Observer with listwise Group Relative Policy Optimization (GRPO) for anomaly-type accuracy and evidence coverage. On a fixed, locally available 2,335-sample HIVAU-70k test subset, the full 8B system ranks first on 9 of 12 metrics across all reported systems, with its largest gains over prior methods at the Event and Video granularities.

1 INTRODUCTION

Video Anomaly Understanding (VAU) extends frame-level anomaly detection (Sultani et al., 2018; Wu et al., 2020) to language-grounded explanation. A VAU system must identify *What* happened, *Who* was involved, *Where* it occurred, and *How* it unfolded. HIVAU-70k (Zhang et al., 2025a) evaluates this four-dimensional report at Clip, Event, and Video granularities. The task therefore requires not only recognizing a salient anomaly, but also collecting sufficient visual support for less explicit actors and temporal relations.

Recent training-free, instruction-tuned, and chain-of-thought systems (Zanella et al., 2024; Huang et al., 2025a; Zhang et al., 2025a; Ma et al., 2025; Huang et al., 2025b) generate the four fields in a single pass. Their output text is the final answer, without an explicit state that measures visual support separately for each dimension. This design is especially limiting when the visually salient *What* field is easy to describe but *Who* or *How* depends on small actors, multi-step actions, or evidence distributed across time.

We term this limitation *evidence insufficiency*: a report may appear complete while its support remains uneven and unobserved. Related studies document unreliable spatial grounding and missed critical visual clues (Agarwal et al., 2026; Cheng et al., 2025). These works address different symptoms, but in the VAU and multimodal video studies considered here, we did not identify an explicit four-dimensional state that checks whether each field meets an operational support threshold. Such a state would allow a system to identify a weak dimension, preserve dimension-covering and secondary support, and direct another acquisition round toward the unresolved gap.

MemVAU addresses this gap both at inference time and during Observer training, as illustrated in Figure 1. During inference, an Observer proposes temporally localized evidence, a frozen Verifier assigns four support scores, and a rule-based Mem-Optimizer maintains a bounded pool under a lexicographic coverage objective. Per-dimension thresholds expose the remaining gap, while a tertiary second-support term distinguishes pools whose dimensional gap and surplus are identical. A structured snapshot then directs the next Observer round, or the Reporter synthesizes the final report when every threshold is met or the round budget is exhausted. During training, listwise GRPO (Shao

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.

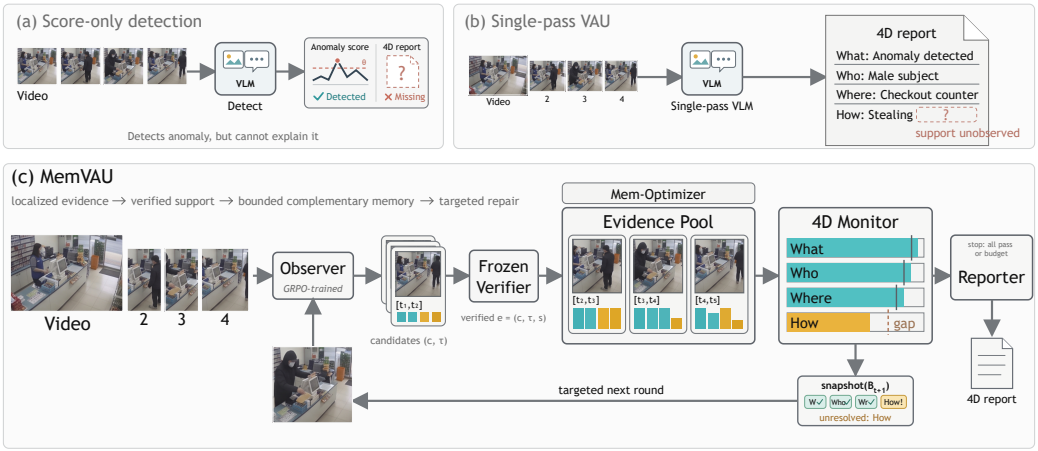

Figure 1: From score-only detection and single-pass VAU to targeted evidence repair. **(a)** Detection flags but does not explain an anomaly. **(b)** Single-pass VAU emits a four-field report without exposing dimension-wise support. **(c)** MemVAU retains localized, Verifier-scored evidence in a bounded pool and targets unresolved dimensions in subsequent rounds; otherwise, the Reporter synthesizes the report. Scores and pass/gap states are model-based operational signals, not calibrated probabilities or grounding certificates.

et al., 2024; Guo et al., 2025) trains the Observer to prioritize anomaly-type accuracy, balanced coverage, and evidence diversity. The Verifier scores are model-based estimates rather than grounding certificates; Section 4.4 evaluates their alignment and residual errors.

Our contributions are threefold:

- We formalize *evidence insufficiency* in single-pass VAU and expose a per-dimension evidence state that provides an operational measure of the remaining support gap under the frozen Verifier's scores and guides targeted evidence acquisition during inference.
- We propose **MemVAU**, combining Verifier-scored evidence memory, threshold-guided competitive pool updates with a secondary-support tie-break, and a listwise-GRPO-trained Observer.
- On the evaluated HIVAU-70k subset, the full 8B MemVAU system ranks first on 9 of 12 metrics across all reported systems. Targeted analyses further characterize Verifier alignment, Reporter evidence utilization, and the contribution of structured memory and GRPO.

2 Related Work

Video Anomaly Understanding. Large VLMs (Radford et al., 2021; Liu et al., 2023; Wang et al., 2024) extend anomaly detection from scores to language-grounded explanations. LAVAD (Zanella et al., 2024) combines frozen captioning with LLM reasoning in a training-free pipeline. Holmes-VAU (Zhang et al., 2025a) introduces HIVAU-70k and four-dimensional instruction tuning, while Sherlock (Ma et al., 2025) extracts global-local multi-scene event quadruples. Vad-R1 (Huang et al., 2025b) trains a perception-to-cognition chain of thought with GRPO (Shao et al., 2024). Within the VAU studies reviewed here, these systems are primarily formulated as generators of descriptions or reasoning traces. We did not identify, in this scoped set, an explicit four-dimensional operational support state that controls subsequent evidence acquisition. MemVAU focuses on this interface between explanation generation and evidence management.

Visual Grounding, Evidence Verification, and Test-Time Self-Correction. Visual grounding methods align language with spatial regions or temporal segments, making the referenced visual content inspectable (Bai et al., 2023; Ma et al., 2025). Evidence-oriented studies in multimodal and anomaly-focused video understanding examine unreliable spatial grounding and missed critical

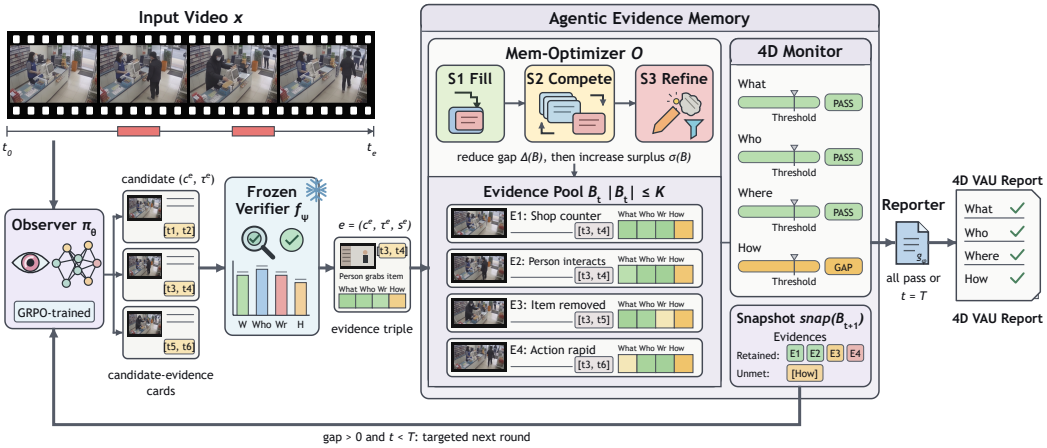

Figure 2: MemVAU evidence-state architecture. At round t, the GRPO-trained Observer proposes temporally localized candidates, and the frozen Verifier maps each candidate to an evidence triple with four-dimensional model-based support scores. The rule-based Mem-Optimizer updates the bounded pool through the applicable S1 Fill, S2 Compete, or S3 Refine stage, lexicographically prioritizing a smaller dimensional gap, a larger threshold surplus, and then stronger secondary support. The Dimension Monitor compares the updated pool with the per-dimension thresholds. If $\Delta(\mathcal{B}_{t+1}) > 0$ and $t < T$, the structured snapshot guides a targeted next round; otherwise, the Reporter synthesizes the four-dimensional VAU report from the retained pool.

visual clues (Agarwal et al., 2026; Cheng et al., 2025). A related line of work improves responses through interaction, critique, reflection, or consensus across repeated inference (Yao et al., 2023; Shinn et al., 2023; Chen et al., 2024a). These lines address related but distinct questions: where language is grounded, whether a claim is visually supported, and how a response can be revised after feedback. In the VAU and video-agent studies considered here, we did not identify a formulation that combines these operations into a four-dimensional operational support state, a bounded evidence pool with explicit secondary-support refinement, and threshold-guided acquisition and stopping. MemVAU makes this combination explicit, while treating its support values as model-based operational signals rather than calibrated probabilities or grounding certificates.

Agentic Memory. Agentic memory builds on retrieval, tool use, and reflection (Lewis et al., 2020; Yao et al., 2023; Shinn et al., 2023; Packer et al., 2023). Long-video systems organize entity-centric, episodic, hierarchical, or raw-frame memories (Long et al., 2025; Yeo et al., 2025; Yin et al., 2025; Li et al., 2026), typically to support free-form question answering across temporal scales. Unlike these general-purpose memory systems, MemVAU organizes retained entries around the four semantic dimensions of VAU. It stores localized evidence with model-based support scores, exposes the remaining dimensional gap, and updates a bounded pool under threshold-preserving max-score and secondary-support objectives. This task-specific state supports threshold-or-budget stopping and keeps each retained entry traceable through its temporal interval and four-dimensional score vector.

3 Method

3.1 Problem Formulation and Framework

For video $x = (x_1, \ldots, x_L)$, VAU produces a report over $\mathcal{D} = \{\text{what}, \text{who}, \text{where}, \text{how}\}$ (Zhang et al., 2025a). MemVAU replaces the single-pass mapping $x \mapsto y$ with the explicit evidence-state loop in Figure 2. The Observer proposes localized descriptions and intervals, the frozen Verifier converts them into scored evidence triples, and the Mem-Optimizer updates a bounded pool. The Dimension Monitor then either returns a structured snapshot for another round or routes the updated pool to the Reporter.

Evidence Triple and Pool. The atomic unit of our process is a *visual evidence triple*

$$e = \big(c_e,\, \tau_e,\, \mathbf{s}_e\big), \tag{1}$$

where c_e is a free-form textual description, $\tau_e = [t_1, t_2] \subseteq [1, L]$ is a temporal interval in x, and $\mathbf{s}_e = (s_{e,d})_{d\in\mathcal{D}} \in [0,1]^4$ is a per-dimension *support-score vector* produced by a frozen Verifier f_ψ on the corresponding sub-segment,

$$\mathbf{s}_e = f_\psi\big(x_{\tau_e},\, c_e\big), \tag{2}$$

Each entry therefore receives model-based support scores from its corresponding video sub-segment rather than relying only on the generator's textual assertion. These scores are operational estimates of visual support, not calibrated probabilities or ground-truth grounding labels. Triples are stored in a bounded pool $\mathcal{B}$ satisfying $|\mathcal{B}| \le K$, with per-dimension max aggregation

$$\bar{s}_d(\mathcal{B}) = \max_{e\in\mathcal{B}} s_{e,d}, \quad \bar{s}_d(\emptyset) \triangleq 0, \tag{3}$$

so one entry can cover a dimension under the pool objective.

Convergence and Closed-Loop Update. Given per-dimension thresholds $\boldsymbol{\theta} = (\theta_d)_{d\in\mathcal{D}} \in [0,1]^4$, the pool's *dimensional gap* and *surplus* are

$$\Delta(\mathcal{B}) = \textstyle\sum_{d\in\mathcal{D}} \big(\theta_d - \bar{s}_d(\mathcal{B})\big)_+, \tag{4}$$

$$\sigma(\mathcal{B}) = \textstyle\sum_{d\in\mathcal{D}} \big(\bar{s}_d(\mathcal{B}) - \theta_d\big)_+, \tag{5}$$

with $(\cdot)_+ = \max(\cdot, 0)$. We define *operational convergence* of the internal evidence state as

$$\Delta(\mathcal{B}_t) = 0 \iff \forall d \in \mathcal{D},\ \bar{s}_d(\mathcal{B}_t) \ge \theta_d, \tag{6}$$

Thus, every dimension must independently meet its Verifier threshold. Unlike a weighted sum, the gap cannot be hidden by a strong score on another dimension. However, coordinate-wise maxima alone cannot value additional non-maximal evidence: selecting one maximizer for each of the four dimensions always gives a subset $\mathcal{B}_{\max} \subseteq \mathcal{B}$ with $|\mathcal{B}_{\max}| \le 4$ and identical $(\bar{s}_d)_{d\in\mathcal{D}}$, Δ, and σ. To make repeated within-dimension support visible to the pool objective, let $s_d^{(2)}(\mathcal{B})$ be the second-largest value in the multiset $\{s_{e,d} : e \in \mathcal{B}\}$, with $s_d^{(2)}(\mathcal{B}) \triangleq 0$ when $|\mathcal{B}| < 2$, and define

$$q_2(\mathcal{B}) = \textstyle\sum_{d\in\mathcal{D}} s_d^{(2)}(\mathcal{B}). \tag{7}$$

The Mem-Optimizer compares trial pools by the lexicographic key

$$\Phi(\mathcal{B}) = \big(-\Delta(\mathcal{B}),\, \sigma(\mathcal{B}),\, q_2(\mathcal{B})\big), \tag{8}$$

so gap reduction remains primary, surplus remains secondary, and q_2 distinguishes pools only when the first two terms are equal. This leaves the convergence criterion unchanged while allowing an additional entry to receive objective value without becoming a coordinate-wise maximum. The term measures secondary score support only; without an explicit redundancy or consistency model, it does not establish that the two supporting descriptions are independent, non-duplicative, or mutually consistent. The thresholds are operational decision points rather than calibrated probabilities, and convergence does not establish factual completeness. We use $(0.70, 0.50, 0.60, 0.75)$ for What/Who/Where/How and assess local threshold sensitivity in Appendix C. One round is

$$\begin{aligned} &\mathcal{C}_t \sim \pi_\theta\big(\cdot \mid x,\, \mathrm{snap}(\mathcal{B}_t)\big), \\ &\mathcal{E}_t = \big\{(c_e, \tau_e, f_\psi(x_{\tau_e}, c_e)) : (c_e, \tau_e) \in \mathcal{C}_t\big\}, \\ &\mathcal{B}_{t+1} = \mathcal{O}\big(\mathcal{B}_t,\, \mathcal{E}_t\big), \end{aligned} \tag{9}$$

where $\mathcal{C}_t$ contains unverified candidates and $\mathcal{E}_t$ their scored triples. The loop continues if $\Delta(\mathcal{B}_{t+1}) > 0$ and $t < T$; otherwise, the Reporter emits $\hat{y} = g_\phi(\mathcal{B}_{t+1})$.

3.2 Observer: Visual Evidence Generation

At round t, the Observer samples $\mathcal{C}_t = \{(c_i, \tau_i)\}_{i=1}^N$ from x and $\mathrm{snap}(\mathcal{B}_t)$. The first round uses an empty snapshot and relies on the video alone. Later rounds receive the retained evidence and explicit pass state of every dimension, turning unconstrained retries into targeted completion. We train the Observer with GRPO (Shao et al., 2024) for anomaly-type accuracy and balanced evidence.

Listwise Reward. For each input, a frozen Qwen3-VL-4B-Instruct (Bai et al., 2025) ranks G candidate sets under three ordered criteria: anomaly-type accuracy, balanced coverage across $\mathcal{D}$, and intra-set diversity. We map rank to

$$r_{\text{rank}}(o_i) = \frac{G - \text{rank}(o_i)}{G - 1}. \tag{10}$$

This bounded signal preserves within-group ordering without assuming cross-video score calibration.

Reward Composition. The total reward

$$R(o_i) = w_{\text{fmt}}\, r_{\text{fmt}}(o_i) + w_{\text{qual}}\, r_{\text{qual}}(o_i) \tag{11}$$

combines a parseability reward with

$$\begin{aligned} r_{\text{qual}}(o_i) = \underbrace{\mathbb{1}\big[\hat{y}^{(i)}_{\text{abn}} = y_{\text{abn}}\big]}_{\text{hard gate}} \cdot \tau^{(i)}_{\text{type}} \\ \cdot\; r_{\text{rank}}(o_i) \cdot \big(1 - p^{(i)}_{\text{rep}}\big). \end{aligned} \tag{12}$$

The hard gate removes quality reward for an incorrect normal/abnormal decision; τ_{type} measures type consistency and p_{rep} penalizes repetition. We use $w_{\text{fmt}} = 0.10$ and $w_{\text{qual}} = 0.90$. Appendix A.3 gives the complete definitions and their limitations.

3.3 Verifier: Per-dimension Visual Evidence Scoring

The frozen Verifier is Qwen3-VL-8B-Instruct (Bai et al., 2025). It scores each x_{τ_e} rather than the full video, making $\mathbf{s}_e$ a localized support estimate. At most N candidates are verified in each of T rounds, yielding $O(NT)$ worst-case calls. Freezing f_ψ prevents Observer optimization from updating the scoring function, but does not make the same-family models independent or error-free.

3.4 Mem-Optimizer: Agentic Memory Coordinator

The Mem-Optimizer $\mathcal{O}$ is a training-free, rule-based scheduler that converts the verified candidate set $\mathcal{E}_t$ into the next-round pool $\mathcal{B}_{t+1}$ together with its snapshot $\text{snap}(\mathcal{B}_{t+1})$.

Per-entry Retention Value. For post-update interpretation, we define a vector-valued per-entry *retention value*:

$$\boldsymbol{\rho}(e; \mathcal{B}) = \Big(\Delta(\mathcal{B} \setminus \{e\}) - \Delta(\mathcal{B}),\ \sigma(\mathcal{B}) - \sigma(\mathcal{B} \setminus \{e\}),\ q_2(\mathcal{B}) - q_2(\mathcal{B} \setminus \{e\})\Big). \tag{13}$$

Read lexicographically, the three coordinates measure the gap increase, surplus loss, and secondary-support loss caused by removing e. The vector is recomputed after each update and serves as an interpretable contribution diagnostic aligned with Eq. (8). It is not by itself a globally optimal eviction rule; candidate–entry swaps are evaluated by the resulting pool key.

Three-stage Competitive Update. **S1 Fill** inserts the candidate giving the greatest resulting lexicographic key Φ while $|\mathcal{B}_t| < K$, but only when the addition strictly improves the current key; hence the bounded pool need not be filled by an objective-neutral entry. Once full, **S2 Compete** searches all current candidate–entry pairs and applies the best swap only when it strictly improves Φ. The search repeats on the updated pool, so each decision uses the current state rather than a fixed initial ranking. After $\Delta = 0$, **S3 Refine** discards swaps that would violate any threshold and accepts the feasible replacement with the greatest resulting pair (σ, q_2) only when that pair improves lexicographically. Thus S1–S2 are monotone in Φ, while S3 preserves $\Delta = 0$ and is monotone in (σ, q_2); Appendix B gives the full algorithm.

Memory Snapshot. After the update, $\mathcal{O}$ serializes $\mathcal{B}_{t+1}$ as

$$\text{snap}(\mathcal{B}_{t+1}) = \big(\text{sum}(\mathcal{B}_{t+1}),\ \text{list}(\mathcal{B}_{t+1})\big), \tag{14}$$

Here $\text{sum}(\cdot)$ records each dimension's maximum score, second-highest score, threshold, and pass state; $\text{list}(\cdot)$ records every retained description, interval, and score vector. This is the only inter-round interface. It exposes both unmet dimensions and existing primary and secondary support to the next Observer round.

3.5 REPORTER: STRUCTURED SYNTHESIS

At convergence or budget exhaustion, the Reporter maps the updated pool to a four-field report, $\hat{y} = g_\phi(\mathcal{B}_{t+1})$. It uses only the serialized pool, including the best available pool after non-convergent exhaustion. Pool conditioning provides an explicit evidence basis but not clause-level factuality; Section 4.4 measures evidence utilization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate all reported systems and ablations on the same fixed, locally available subset of 2,335 samples from the Holmes-VAU (HIVAU-70k) test set (Zhang et al., 2025a). The samples are drawn from UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020), with each sample evaluated at the *Clip* (C), *Event* (E) and *Video* (V) granularity. Following the official protocol we report twelve metrics in total: cumulative BLEU-1 to BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005) and ROUGE-L (Lin, 2004) at each granularity.

Baselines. We compare against two groups: **(i) general-purpose video VLMs** that perform single-pass VAU inference, namely Video-ChatGPT (Maaz et al., 2024), Video-LLaMA (Zhang et al., 2023), Video-LLaVA (Lin et al., 2024), LLaVA-Next-Video (Zhang et al., 2025b), Qwen2-VL (7B) (Wang et al., 2024) and InternVL2 (8B) (Chen et al., 2024b); and **(ii) anomaly-specific video understanding models**, including Hawk (7B) (Tang et al., 2024), which is trained to explain open-world video anomalies, and Holmes-VAU (2B) (Zhang et al., 2025a), the previous state-of-the-art on HIVAU-70k.

Implementation Details. For the full-scale MemVAU configuration, the Observer is fine-tuned with GRPO on Qwen3-VL-8B-Instruct (Bai et al., 2025) (Qwen-VL family (Bai et al., 2023; Wang et al., 2024)) using LoRA (Hu et al., 2022) of rank 8 on the *q/k/v/o* projections, with $G = 4$, $w_{\text{fmt}} = 0.1$, $w_{\text{qual}} = 0.9$, learning rate 2×10^{-6} and 2,204 optimization steps; the listwise reward is issued by a frozen Qwen3-VL-4B-Instruct to keep training cost tractable. At inference time the Verifier is a frozen Qwen3-VL-8B-Instruct (matching the Observer family for distributional consistency), and the Reporter is a text-only Qwen3-8B trained on the official HIVAU-70k instruction data via SFT: it consumes only the structured pool snapshot and therefore does not need a vision encoder. System hyper-parameters: pool capacity $K = 5$, per-dimension thresholds $(\theta_{\text{what}}, \theta_{\text{who}}, \theta_{\text{where}}, \theta_{\text{how}}) = (0.70, 0.50, 0.60, 0.75)$, maximum rounds $T = 3$, and 32 uniformly sampled frames per clip. Because four coordinate-wise maxima can always be preserved by at most four entries, $K = 5$ permits one additional slot beyond this worst-case max-score cover; the q_2 tie-break gives that slot explicit secondary-support value when it improves the second-highest score state. This interpretation does not make five entries necessary, and K remains a tunable capacity hyper-parameter.

Scale-Aligned Visual Agents. To control visual-backbone scale, we additionally evaluate O2/V2/R8, where O/V/R denote Observer/Verifier/Reporter: the Observer and frozen Verifier use Qwen3-VL-2B-Instruct, the Observer follows the same GRPO strategy, and the Reporter remains 8B. Other components are unchanged. This matches Holmes-VAU's per-visual-agent 2B scale, not aggregate parameters or inference calls.

4.2 MAIN RESULTS

Comparison with Prior Methods. Across all rows in Table 1, full 8B MemVAU ranks first on 9 of 12 metrics, second on CIDEr-C (0.478 versus 0.481), and third on BLEU-C and ROUGE-L-C. Among prior systems, Hawk is strongest at Event and Video granularities across all four metric families and has the best prior METEOR scores ($0.205/0.197/0.201$).

Event- and Video-Level Interpretation. MemVAU exceeds reproduced Hawk on all twelve metrics, improving BLEU by $+0.415/+0.746/+0.410$ and METEOR by $+0.074/+0.234/+0.189$

Table 1: Results on the fixed, locally available 2,335-sample subset of the Holmes-VAU (HIVAU-70k) test set. BLEU is cumulative BLEU-1–4; C/E/V denote Clip/Event/Video. Bold/underline mark the best/second-best values across all rows. † is our Hawk reproduction; ‡ is the scale-aligned O2/V2/R8 MemVAU configuration.

Method	Params	BLEU ↑			CIDEr ↑		
		C	E	V	C	E	V
Video-ChatGPT	7B	0.152	0.068	0.066	0.033	0.011	0.013
Video-LLaMA	7B	0.151	0.079	0.104	0.024	0.014	0.017
Video-LLaVA	7B	0.164	0.046	0.055	0.032	0.009	0.013
LLaVA-Next-Video	7B	0.435	0.091	0.120	0.102	0.015	0.031
Qwen2-VL	7B	0.312	0.082	0.155	0.044	0.020	0.044
InternVL2	8B	0.331	0.101	0.145	0.052	0.022	0.035
Hawk†	7B	0.320	0.246	0.338	0.167	0.143	0.144
Holmes-VAU	2B	**0.913**	0.804	0.566	0.467	1.519	1.437
MemVAU‡	2B	0.743	0.559	0.661	**0.481**	0.892	1.036
MemVAU (Ours)	8B	0.735	**0.992**	**0.748**	0.478	**1.616**	**1.795**

Method	Params	METEOR ↑			ROUGE ↑		
		C	E	V	C	E	V
Video-ChatGPT	7B	0.102	0.069	0.044	0.153	0.048	0.079
Video-LLaMA	7B	0.112	0.076	0.057	0.156	0.067	0.090
Video-LLaVA	7B	0.097	0.022	0.014	0.132	0.023	0.045
LLaVA-Next-Video	7B	0.117	0.085	0.096	0.198	0.080	0.106
Qwen2-VL	7B	0.133	0.092	0.112	0.163	0.081	0.137
InternVL2	8B	0.141	0.095	0.101	0.182	0.102	0.122
Hawk†	7B	0.205	0.197	0.201	0.163	0.150	0.166
Holmes-VAU	2B	0.190	0.165	0.121	**0.329**	0.370	0.355
MemVAU‡	2B	0.278	0.375	0.334	0.264	0.354	0.342
MemVAU (Ours)	8B	**0.279**	**0.431**	**0.390**	0.263	**0.474**	**0.444**

at the Clip, Event, and Video granularities, respectively. The largest margins occur at the Event and Video granularities. Against Holmes-VAU, MemVAU improves METEOR at all granularities ($+0.089/+0.266/+0.269$), CIDEr at all granularities ($+0.011/+0.097/+0.358$), and BLEU-E/V by $+0.188/+0.182$. Holmes-VAU retains BLEU-C and ROUGE-L-C. These comparisons favor the full pipeline but do not isolate the causal contributions of evidence scoring, memory accumulation, backbone scale, or inference cost.

Performance under Visual-Backbone Scale Alignment. Excluding our O2/V2/R8 configuration, full 8B MemVAU leads prior methods on 10 of 12 metrics. With both visual agents reduced to 2B, O2/V2/R8 still exceeds reproduced Hawk on all twelve metrics and Holmes-VAU on five: BLEU-V, CIDEr-C, and all three METEOR scores. The gains therefore persist under per-visual-agent scale alignment rather than arising solely from larger visual backbones, particularly for METEOR-measured semantic coverage. This controls the scale of each visual agent, but not aggregate parameters or the number of inference calls.

4.3 Component Analysis

Additional Inference Calls vs. Structured Evidence Memory. Under shared preprocessing, we compare one-call Single-pass and CoT, two-call Self-correction, and three-call Multi-round controls. We also compare raw Observer evidence against the same SFT Observer with verification and structured memory. Table 2 reports the core Event/Video results; Appendix A.4 gives all C/E/V values.

Extra reasoning or correction calls do not consistently improve Event/Video performance over Single-pass. In contrast, verified structured memory substantially improves the matched raw-

Table 2: Component analysis on the same fixed 2,335-sample evaluation subset at Event (E) and Video (V) granularities. Panel (a) compares direct inference controls and matched SFT-Observer memory variants; panel (b) changes only Observer training. B/C/M/R denote BLEU/CIDEr/METEOR/ROUGE-L.

	B		C		M		R	
System	E	V	E	V	E	V	E	V
(a) Additional inference calls vs. structured evidence memory								
Single-pass	0.539	0.335	0.421	0.446	0.337	0.303	0.337	0.322
CoT	0.416	0.224	0.346	0.408	0.258	0.260	0.292	0.299
Self-correction	0.384	0.154	0.350	0.389	0.296	0.279	0.298	0.291
Multi-round	0.358	0.142	0.358	0.397	0.294	0.274	0.296	0.289
Raw evidence	0.471	0.301	0.384	0.419	0.292	0.264	0.305	0.299
MemVAU (SFT Obs.)	**0.999**	**0.715**	**1.628**	**1.760**	**0.428**	**0.383**	**0.470**	**0.434**
(b) Effect of GRPO training on the Observer								
MemVAU (SFT Obs.)	**0.999**	0.715	**1.628**	1.760	0.428	0.383	0.470	0.434
MemVAU (GRPO Obs.)	0.992	**0.748**	1.616	**1.795**	**0.431**	**0.390**	**0.474**	**0.444**

Table 3: Same-model, score-blinded audit of 150 statements from 30 stratified videos. Verifier and Audit are mean production and separate-prompt scores; Agree. is exact $\{0, 0.5, 1\}$ agreement. Results measure cross-prompt alignment, not agreement with independent or human labels.

Dimension	Verifier	Audit	Agree.	MAE ↓
What	0.759	0.920	68.0%	0.235
Who	0.749	0.763	74.0%	0.167
Where	0.939	0.923	83.3%	0.125
How	0.762	0.873	66.7%	0.225

evidence variant. Because that comparison jointly introduces Verifier scoring and pool management, it supports their combined effect but does not assign the gain to either component alone.

Effect of GRPO Training on the Observer. With the Mem-Optimizer and Reporter fixed, GRPO improves all four Video metrics and Event-level METEOR/ROUGE. BLEU-E and CIDEr-E decrease slightly, so the effect is not uniformly positive across metrics or granularities. The comparison isolates the Observer-training change within this pipeline, but the reported single run does not quantify training variance.

4.4 Evidence Reliability and Utilization

Verifier Cross-Prompt Alignment and Local Stability. We diagnose local prompt-level stability using a score-blinded AI-assisted audit of 150 first-round statements from 30 videos, stratified into 10 high-, middle-, and low-scoring videos. The audit uses the frozen production Qwen3-VL-8B weights with a separate prompt and no production scores or decisions. Blinding reduces direct anchoring, but shared weights may retain correlated errors and biases; we therefore measure same-model cross-prompt alignment, not independent visual correctness, calibrated accuracy, or human agreement. Each dimension receives $\{0, 0.5, 1\}$ for unsupported, partial/uncertain, or clear support. Table 3 shows the lowest exact agreement for *How* (66.7%) and largest MAE for *What* (0.235).

Of 83 threshold-accepted statements, 70 receive full audit support on all four dimensions (84.3% accepted-set cross-prompt alignment). The remaining 13 (15.7%) are not fully supported: three of the 83 accepted statements (3.6%) contain at least one unsupported dimension, while the other ten (12.0%) contain a dimension scored 0.5, indicating partial support or uncertainty. The audit therefore supports local cross-prompt stability but also reveals non-trivial residual risk, especially for fine-grained *How*. Because production and audit prompts use shared weights, these figures cannot be interpreted as overall accuracy, independent visual validation, human agreement, or benchmark-wide reliability.

Reporter Evidence Utilization. To examine whether the Reporter incorporates the accumulated evidence, we analyze reports stratified across the Event and Video temporal scopes. Each report is decomposed into atomic factual statements and matched against the exact evidence pool supplied to the Reporter. A report is counted as evidence-utilizing when at least one atomic statement is fully supported by its corresponding pool. Under this report-level criterion, 94% of reports contain at least one directly pool-supported statement, with utilization rates of 96% at Event level and 92% at Video level. This indicates substantial use of the supplied pool, but the criterion does not require every clause to be supported and therefore does not establish universal clause-level grounding or visual factuality.

End-to-End Sensitivity. On a fixed screening set of 120 videos and 317 Event/Video instructions, O8B/V8B/R8B sweeps over T, N, K show non-monotonic quality–cost trade-offs. $T = 2$ gives the highest BLEU, whereas $T = 3$ gives the highest CIDEr, METEOR, and ROUGE-L at higher cost. Increasing N from 3 to 7 raises average calls from 21.05 to 34.42 and frames from 322.05 to 583.39 without improving every metric or the operational full-4D pass rate. $K = 5$ maximizes operational coverage and full-threshold passing, while $K = 3$ yields the strongest text scores and measured pool diversity. These screening results characterize local trade-offs, not globally optimal settings or monotonic scaling. Appendix A.1 reports full diagnostics; Appendices B and C give pseudocode and fixed-input heuristic sensitivity.

5 Conclusion

MemVAU recasts VAU as closed-loop evidence acquisition: a GRPO-trained Observer proposes localized evidence, a frozen Verifier assigns four-dimensional model-based support scores, and a rule-based Mem-Optimizer maintains a bounded pool. The Dimension Monitor requests another round for unresolved dimensions or routes the best available pool to the Reporter at stopping.

On the reported subset, full 8B MemVAU has the highest point estimate on 9 of 12 metrics, with its largest gains at Event and Video. Report-level utilization is 94%; the same-weight, score-blinded audit gives 84.3% accepted-set cross-prompt alignment and weaker fine-grained *How* stability. Jointly shifting all thresholds by ± 0.1 changes one acceptance decision in either direction (accepted-set Jaccard ≈ 0.988), indicating local decision stability rather than calibration or end-to-end robustness.

These results support bounded evidence memory as an operational mechanism, not calibrated probabilities, human-ground-truth visual accuracy, independent Verifier validation, or universal claim-level factuality. Future work should use distinct-model validation, uncertainty intervals, cheaper verification, claim-level error analysis, and cross-domain tests.

Limitations

MemVAU requires up to $O(NT)$ Verifier calls, uses greedy within-round updates, and relies on hand-crafted reward, pool, and threshold settings. Fixed-input sensitivity establishes only local stability: threshold replay fixes evidence and scores and does not rerun acquisition, pool updates, stopping, reporting, or downstream evaluation, so it does not establish global optimality, calibration, or end-to-end robustness.

Clip captioning bypasses the multi-round Verifier, Mem-Optimizer, and Reporter, so Clip metrics do not isolate structured-memory effects. The same-model AI audit covers only 150 statements from 30 stratified videos under one prompt protocol; it provides no human-ground-truth accuracy or independent-model validation. We also report neither multi-seed variation nor bootstrap intervals for endpoint metrics. Finally, the four-dimensional schema and evaluation are limited to HIVAU-70k surveillance videos, and transfer to other anomaly domains remains unverified.

AI use statement

Generative AI assisted migration to the ICLR 2027 LaTeX template and format checks. The authors reviewed the output and take responsibility for all content.

REFERENCES

Sakshi Agarwal, Aishik Konwer, and Ankit Parag Shah. Reasoning-guided grounding: Elevating video anomaly detection through multimodal large language models. *arXiv preprint arXiv:2605.02912*, 2026. URL https://arxiv.org/abs/2605.02912.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

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*, pp. 65–72, 2005.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. RECONCILE: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can MLLM think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. URL https://arxiv.org/abs/2505.21374.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25750–25761. PMLR, 2025a.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026. URL https://arxiv.org/abs/2605.16481.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5971–5984, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025. URL https://arxiv.org/abs/2508.09736.

Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive LLM. In *Proceedings of the ACM Web Conference 2025 (WWW)*, 2025. doi: 10.1145/3696410.3714617.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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 (ACL)*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

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

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 (CVPR)*, pp. 6479–6488, 2018.

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. In *Advances in Neural Information Processing Systems*, 2024.

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 (CVPR)*, pp. 4566–4575, 2015.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–339, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. GCAgent: Long-video understanding via schematic and narrative episodic memory. *arXiv preprint arXiv:2511.12027*, 2025. URL https://arxiv.org/abs/2511.12027.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. *arXiv preprint arXiv:2512.12360*, 2025. URL https://arxiv.org/abs/2512.12360.

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 (CVPR)*, pp. 18527–18536, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2023.

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 (CVPR)*, pp. 13843–13853, 2025a.

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 (TMLR)*, 2025b.

Table 4: Maximum-round sensitivity with $K = N = 5$. Calls and frames are averaged over 124 pipeline executions. B/C/M/R denote BLEU/CIDEr/METEOR/ROUGE-L.

T	B	C	M	R	Calls	Frames
1	0.6153	0.1006	0.3497	0.3472	21.32	326.30
2	**0.6257**	0.0912	0.3512	0.3476	24.46	380.82
3	0.6165	**0.1058**	**0.3601**	**0.3571**	26.58	421.14

Table 5: Candidate-count sensitivity with $K = 5, T = 3$. Full-4D is the fraction of segments whose final pool meets all four Verifier thresholds; it is an operational state measure, not visual-ground-truth accuracy.

N	B	C	M	R	Calls	Frames	Full-4D
3	0.6292	0.0922	**0.3561**	**0.3509**	21.05	322.05	85.37%
5	0.6065	0.0936	0.3439	0.3389	26.82	434.14	**91.60%**
7	**0.6365**	**0.1005**	0.3450	0.3440	34.42	583.39	89.97%

Table 6: Pool-capacity sensitivity with $N = 5, T = 3$. Cov. is mean operational dimension coverage; Full is the four-threshold pass rate; Lex./Temp. are within-pool lexical and temporal diversity.

K	B	C	M	R	Pool	Cov.	Full	Lex./Temp.
3	**0.6289**	**0.1003**	**0.3640**	**0.3552**	3.00	96.48%	89.70%	0.881/0.827
5	0.6065	0.0936	0.3439	0.3389	5.00	**96.95%**	**91.60%**	0.846/0.730
7	0.5787	0.0959	0.3471	0.3422	5.30	96.00%	88.62%	0.854/0.726

A Additional Experimental Analyses

A.1 End-to-End $N/K/T$ Sensitivity

We conduct end-to-end sensitivity screening with the same O8B/V8B/R8B system, thresholds, and fixed manifest of 120 unique videos and 317 Event/Video instructions. Text metrics use all 317 instructions. Trace and pool statistics use 124 pipeline executions because four videos cross shard boundaries and are independently executed in two shards. These experiments characterize the evaluated screening subset rather than benchmark-wide robustness or globally optimal hyperparameters.

The quality metrics vary non-monotonically with T: $T = 2$ gives the highest BLEU, whereas $T = 3$ gives the highest CIDEr, METEOR, and ROUGE-L. The average number of calls and processed frames increases with the round budget, providing an end-to-end cost–quality trade-off rather than evidence that more rounds uniformly improve the report.

Increasing N from 3 to 7 raises calls per video from 21.05 to 34.42 and frames per video from 322.05 to 583.39. $N = 7$ improves BLEU and CIDEr, but not METEOR or ROUGE-L, while $N = 5$ gives the highest operational full-threshold pass rate. Thus, the additional candidates produce a cost–quality trade-off rather than a uniform gain.

$K = 5$ gives the highest operational dimension coverage, whereas $K = 3$ gives the strongest text metrics and the highest lexical and temporal diversity under the reported measures. The mean final pool size at $K = 7$ is only 5.30, showing that adaptive stopping and competitive updates do not necessarily fill the available capacity. Lexical or temporal diversity describes pool variation and does not establish semantic correctness or visual factuality.

The evidence-count trace requires an implementation qualification. The runtime parser deduplicates model outputs and pads underfilled candidate lists to the requested N before recording `generated_count`. Consequently, the recorded 3/5/7 candidates per round are post-parser, post-padding counts rather than raw parse-success rates. Placeholder entries constitute 4.98%, 13.12%, and 7.75% of the final pools for $N = 3, 5, 7$, respectively. All final prediction files contain the expected 317 unique, non-empty outputs. Earlier concurrent launches encountered GPU-memory



Figure 3: Four uniformly sampled frames from the robbery clip used in the qualitative case.

contention and were completed by configuration-validated serial resume; the final telemetry sidecars record no failed model calls.

A.2 Qualitative Case Study

We illustrate the change in generated evidence on a representative robbery clip from the HIVAU-70k test set (Figure 3). The Observer is queried on the same 32-frame input before and after GRPO training while the Mem-Optimizer and Reporter are kept identical.

Before GRPO. The base Observer produces a description that mischaracterizes the abnormal signal:

> **What:** a customer arranging items on the counter. **Who:** a customer. **Where:** inside a convenience store. **How:** reaching across the counter to organize merchandise.

The *What* field is misclassified as a normal activity, and *Who*/*How* reuse the same surface cue without introducing any independently verifiable visual evidence.

After GRPO. The trained Observer produces a more dimensionally specific account associated with the Verifier-scored evidence pool:

> **What:** `theft` occurring at the cashier counter. **Who:** a masked individual standing on the customer side, with a second individual loitering near the entrance. **Where:** the cashier area of a small convenience store, late evening. **How:** the individual leans across the counter, grabs items from the back shelf, conceals them inside a backpack and moves toward the entrance.

Source-video audit. The main qualitative benefit of GRPO is a shift from a generic, single-cue description to a structured account spanning all four VAU dimensions. Before training, the base Observer describes a customer arranging items on the counter and reuses this surface cue for *Who*

and *How*. After training, the Observer frames the salient event as theft, characterizes the actor as a dark-clothed, hooded person, localizes the activity to the cashier counter, and expands *How* into a multi-step action sequence involving leaning across the counter, handling an open paper bag, transferring nearby countertop objects, and moving out of the visible camera field. In this qualitative case, GRPO improves descriptive coverage, action specificity, and cross-dimensional organization, providing a richer candidate report for the subsequent evidence-memory pipeline. The complete-video audit supports this core account, although not every additional clause in the raw output is independently verifiable from the clip.

A.3 Complete Reward Definition

The format component in Eq. (11) is

$$r_{\text{fmt}}(o_i) = 0.3\,\mathbb{1}\big[\texttt{is_abnormal}\ \text{parseable}\big] + 0.7\,\frac{n_{\text{valid}}(o_i)}{N}, \tag{15}$$

where n_{valid} counts successfully parsed fields and the second term is clipped to $[0, 1]$. The soft anomaly-type gate and repetition penalty in Eq. (12) are

$$\tau^{(i)}_{\text{type}} = 0.3 + 0.7\, s^{(i)}_{\text{type}} \in [0.3, 1.0], \tag{16}$$

$$p^{(i)}_{\text{rep}} = \min\big(1,\ (\bar{J}^{(i)} - 0.3)_+ / 0.4\big), \tag{17}$$

where $s^{(i)}_{\text{type}}$ is the keyword-hit rate over the ground-truth category and $\bar{J}^{(i)}$ is the mean pairwise Jaccard similarity among entries. The penalty is inactive below 0.3 and saturates above 0.7. Keyword overlap is an auxiliary consistency prior rather than a semantic classifier, and these boundaries are empirical operating points rather than theoretically optimal settings.

A.4 Full Component Results

Tables 7 and 8 provide the complete Clip, Event, and Video results underlying the compact main-text comparison.

A.5 Sensitivity of Heuristic Components

We conduct two targeted offline analyses for the heuristic components. First, on fixed logged completions, nearby format/quality weights ($0.05/0.95$, $0.10/0.90$, and $0.20/0.80$) preserve the observed candidate ordering under the baseline repetition boundary, while tightening the repetition interval to $[0.2, 0.6]$ changes some rankings and widening it to $[0.4, 0.8]$ preserves the observed top-set ordering. Second, on a fixed 10-video diagnostic subset, jointly shifting all four production thresholds by $\{-0.1, 0, +0.1\}$ leaves the evidence-level acceptance count unchanged at $38/50$, although three Where and one Who dimension decisions flip at the lower and upper settings, respectively. The complete results are in Appendix C. These analyses are fixed-completion and fixed-decision sensitivity studies: they do not rerun the Observer, Judge, pool update, stopping policy, Reporter, or downstream evaluation.

A.6 Training Dynamics

Figure 4 reports the joint evolution of the listwise reward and the `is_abnormal` hard-gate pass rate (200-step bins): both rise together, with reward growing from 0.202 to 0.228 ($+13\%$) and pass rate from 54.4% to 60.8% ($+6.4$ pp). The first ~ 400 steps form a *Rapid Improvement* phase, followed by a slower-improvement phase in which the pass rate continues to increase. Their concurrent rise indicates improved performance under the training reward, but does not by itself establish improved visual grounding or semantic factuality.

B Mem-Optimizer Algorithm

Algorithm 1 summarizes the three-stage, secondary-support-aware competitive update applied by the Mem-Optimizer in each round.

Table 7: Complete inference and memory controls on the same fixed 2,335-sample evaluation subset. The final two rows compare the same SFT Observer with raw evidence or with verification and structured memory.

Metric	System	C	E	V
BLEU	Single-pass	0.677	0.539	0.335
	CoT	0.743	0.416	0.224
	Self-correction	0.677	0.384	0.154
	Multi-round	0.684	0.358	0.142
	Raw evidence	**0.744**	0.471	0.301
	MemVAU (SFT Obs.)	0.733	**0.999**	**0.715**
CIDEr	Single-pass	**0.518**	0.421	0.446
	CoT	0.443	0.346	0.408
	Self-correction	0.507	0.350	0.389
	Multi-round	0.515	0.358	0.397
	Raw evidence	0.488	0.384	0.419
	MemVAU (SFT Obs.)	0.468	**1.628**	**1.760**
METEOR	Single-pass	0.324	0.337	0.303
	CoT	0.272	0.258	0.260
	Self-correction	0.321	0.296	0.279
	Multi-round	**0.325**	0.294	0.274
	Raw evidence	0.282	0.292	0.264
	MemVAU (SFT Obs.)	0.278	**0.428**	**0.383**
ROUGE	Single-pass	0.261	0.337	0.322
	CoT	0.255	0.292	0.299
	Self-correction	0.260	0.298	0.291
	Multi-round	0.261	0.296	0.289
	Raw evidence	**0.264**	0.305	0.299
	MemVAU (SFT Obs.)	0.263	**0.470**	**0.434**

Table 8: Complete effect of GRPO Observer training on the same fixed 2,335-sample evaluation subset; the Mem-Optimizer and Reporter are fixed. The GRPO row is the same full MemVAU configuration reported in Table 1, so its twelve values are identical.

Metric	System	C	E	V
BLEU	MemVAU (SFT Obs.)	0.733	**0.999**	0.715
	MemVAU (GRPO Obs.)	**0.735**	0.992	**0.748**
CIDEr	MemVAU (SFT Obs.)	0.468	**1.628**	1.760
	MemVAU (GRPO Obs.)	**0.478**	1.616	**1.795**
METEOR	MemVAU (SFT Obs.)	0.278	0.428	0.383
	MemVAU (GRPO Obs.)	**0.279**	**0.431**	**0.390**
ROUGE	MemVAU (SFT Obs.)	**0.263**	0.470	0.434
	MemVAU (GRPO Obs.)	**0.263**	**0.474**	**0.444**

Algorithm 1: Secondary-support-aware Mem-Optimizer round update

Input: current pool $\mathcal{B}_t$ (capacity K), verified candidates $\mathcal{E}_t$, thresholds $\boldsymbol{\theta}$
Output: updated pool $\mathcal{B}_{t+1}$
Procedure:

1. While $|\mathcal{B}_t| < K$, select the candidate that gives the lexicographically greatest resulting key $\Phi(\mathcal{B}_t \cup \{e\})$. Add it and remove it from the remaining set only if it strictly improves the current key; otherwise stop filling, so an objective-neutral candidate is not retained merely to reach capacity K.

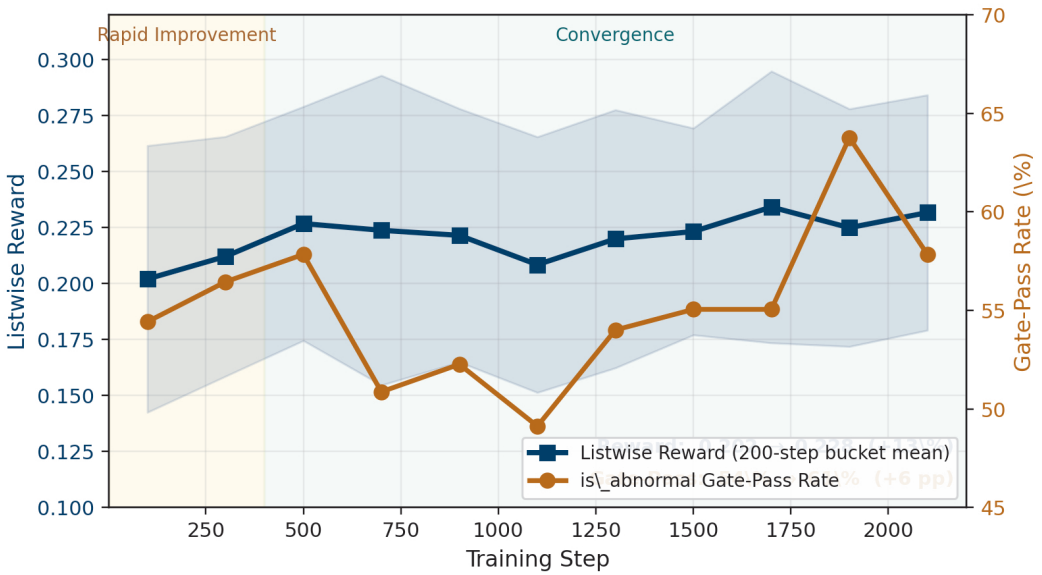

Figure 4: Joint evolution of two metrics during the Observer's GRPO training, aggregated into bins of **200 steps**. **Left axis**: listwise average reward (dark-blue squares for bin means, light-blue shading for $\pm 1\sigma$). **Right axis**: `is_abnormal` hard-gate pass rate (orange dots). Both rise together: reward $0.202 \rightarrow 0.228$ ($+13\%$) and pass rate $54.4\% \rightarrow 60.8\%$ ($+6.4$ pp); the first $\sim$400 steps form a *Rapid Improvement* phase, followed by a slower-improvement phase.

2. While $\Delta(\mathcal{B}_t) > 0$, enumerate all swaps between the current remaining candidates and $\mathcal{B}_t$, choose the trial with maximum Φ, and apply it only when it strictly improves the current pool key. Repeat on the updated pool; if no strict improvement exists, discard the remaining candidates for this stage.
3. Once $\Delta(\mathcal{B}_t) = 0$, process the remaining candidates. For each candidate, enumerate all evictions, retain only trials with $\Delta = 0$, and apply the trial with the lexicographically greatest (σ, q_2) only when it strictly improves the current pair. Thus a swap that preserves both coverage and surplus can still be accepted when it increases secondary support.
4. Refresh $\boldsymbol{\rho}$ for every entry using Eq. (13) and return $\mathcal{B}_{t+1} = \mathcal{B}_t$.

C Heuristic Sensitivity Analysis

Reward Re-scoring Protocol. We perform an offline reward-ranking sensitivity analysis on 136 logged completion records containing 1,088 candidates. These form 272 valid $G = 4$ prompt groups, of which 271 can be re-scored. The analysis fixes the generated completions and does not re-invoke the frozen 4B Judge. Because the historical logs do not retain the raw Judge ranking or all reward components, we reconstruct the historical format and repetition terms and hold the remaining composite quality term fixed while varying the format/quality weights and repetition boundaries. The reconstructed baseline has a maximum absolute error of 3×10^{-8} from the logged reward.

The nearby weight settings do not produce an observed ranking reversal under the baseline repetition boundary, while the stricter $[0.2, 0.6]$ interval changes some deterministic top-1 selections and increases the group reward variance. However, only 6 of 271 groups have a unique baseline top-1 and 79.3% of groups are fully tied at the baseline. Consequently, the agreement values indicate no detected systematic reversal in the available logs, not strong evidence of reward informativeness or globally optimal hyperparameters. We retain $0.10/0.90$ and $[0.3, 0.7]$ as empirical middle operating points.

Table 9: Offline reward sensitivity on fixed logged completions. The weight scan holds the repetition boundary at $[0.3, 0.7]$; the boundary scan holds the weights at $0.10/0.90$. Kendall's τ_b is defined for 56 of 271 groups; all-tied groups are excluded rather than assigned an arbitrary correlation.

Component	Setting	Top-1 agree	Top-set overlap	Mean τ_b	Group variance	Zero-var. groups
Format/quality	$0.05/0.95$	100.0%	100.0%	1.000	0.009211	79.3%
	$0.10/0.90$	100.0%	100.0%	1.000	0.008296	79.3%
	$0.20/0.80$	100.0%	100.0%	1.000	0.006812	79.3%
Rep. boundary	$[0.2, 0.6]$	91.9%	99.6%	0.861	0.009011	69.4%
	$[0.3, 0.7]$	100.0%	100.0%	1.000	0.008296	79.3%
	$[0.4, 0.8]$	100.0%	100.0%	0.993	0.008039	79.3%

Table 10: Local threshold-decision stability on the same 150-statement audit set as Table 3. Accepted requires all four recorded Verifier scores to meet their corresponding thresholds. Audit- relative precision, recall, and F1 use full four-dimensional support under the same-model, score-blinded audit rubric as a diagnostic reference; they are not independent visual-accuracy estimates.

Shift	What/Who/ Where/How	Accepted	Audit-rel. P	Audit-rel. R	Audit-rel. F1	Decision changes
-0.1	$0.60/0.40/0.50/0.65$	84/150	84.5%	73.2%	78.5%	Evidence $+1$; Where: 6; How: 1
0	$0.70/0.50/0.60/0.75$	83/150	84.3%	72.2%	77.8%	None
$+0.1$	$0.80/0.60/0.70/0.85$	82/150	84.1%	71.1%	77.1%	Evidence -1; Who: 6

Verifier-threshold Protocol. We additionally evaluate local threshold-decision stability on the same 150 round-0 evidence statements and score-blinded audit labels used in Table 3, comprising 600 dimension judgments. The audit labels come from the same-model, separate-prompt diagnostic audit and are not human annotations. We hold the recorded evidence and Verifier scores fixed, start from $\boldsymbol{\theta} = (0.70, 0.50, 0.60, 0.75)$, and jointly shift every dimension threshold by -0.1 and $+0.1$.

The evidence-level accepted set remains nearly unchanged under the joint threshold perturbations. Lowering all four thresholds by 0.1 admits one additional statement, whereas increasing them by 0.1 rejects one previously accepted statement; the accepted-set Jaccard similarity with the baseline is approximately 0.988 in both cases. This provides evidence of local threshold-decision stability on the fixed audit set. The stability is not uniform at the dimension level: the lower setting changes six Where and one How decisions, while the higher setting changes six Who decisions. Thus, the conjunctive evidence-level acceptance decision is locally stable even though some individual dimensions remain threshold-sensitive. Because this replay holds evidence and Verifier scores fixed and does not rerun Observer generation, pool updates, stopping, reporting, or downstream metrics, it neither calibrates the thresholds nor establishes rerun or end-to-end robustness.