FREEZE THE STUDENT, TRAIN THE RESEARCHER: LEARNING THE VALUE OF EXPERIMENTS FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video systems are typically engineered against a fixed benchmark, yet deployment brings unfamiliar scenes, event statistics, and query types, where the tuning loop must start over. What survives such shifts is the research process itself: hypothesizing about failures, running diagnostic experiments, and revising the design. We propose **StreamRSI**, which keeps the deployed 7B video answerer frozen and instead trains a 27B research policy to conduct this process, learning the value of experiments rather than any single system. The central obstacle is credit: an unsuccessful experiment can rule out a misleading explanation and improve subsequent search, while an immediately successful candidate can teach nothing transferable. StreamRSI therefore assigns controlled evidence credit. Paired continuations restore the same candidate state and differ only in access to a diagnostic result, isolating that result's contribution to later research; a third continuation spends the diagnostic budget on direct search, pricing the experiment's opportunity cost. These comparisons supervise when to investigate, which probe to run, and when to reopen a diagnosis. The discovered inference-time program lifts the frozen answerer to 65.50 OVO-Bench Overall and 80.63 StreamingBench real-time accuracy. Under a matched 20,000 GPU-s budget, full StreamRSI reaches the target quality in all eight seeds and outperforms outcome-only and single-credit training. The resulting research policy generalizes to novel domains, accelerating search on unseen benchmarks even when stripped of its historical archive.

1 INTRODUCTION

Streaming video understanding requires systems to answer questions as visual events unfold, using only the frames available before each query and a bounded model context (Niu et al., 2025; Lin et al., 2026). Progress on this problem is largely driven by benchmark-specific engineering: memory architectures, frame budgets, and prompting schemes are iterated against a fixed evaluation suite until scores saturate. Deployment rarely grants this luxury. A stream in a new domain brings unfamiliar scenes, event statistics, and query types, and the hand-tuning loop starts over. What survives such shifts is not any single tuned system but the process that produced it: hypothesizing about failures, running diagnostic experiments, and revising the design in response. We therefore study a self-improving research agent that learns this process itself, so that given a frozen vision-language model (VLM) and an unfamiliar streaming task, it can autonomously discover the inference-time program that retrieves evidence, arranges it in the input, and elicits an answer.

The design space this agent must navigate is deceptively hard: maintaining a complete high-resolution history is infeasible, and naive strategies such as uniformly increasing frame counts often fail to surpass strong recency-based baselines (Shen et al., 2026). Navigating it requires genuine research moves—generating hypotheses, writing code, and running controlled probes. However, learning to make these research decisions requires evaluating the experiments that inform them, which introduces three obstacles.

❶ **Credit-assignment ambiguity**: the default reward signal, immediate accuracy on a candidate program, does not measure the value of the experiment that produced it. A diagnostic probe that

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.

purposely masks recent frames might drop the overall task score but correctly isolate a recency bias, redirecting the next experiment toward a much better solution. Standard outcome-based training treats such informative failures as uniformly unhelpful.

Existing approaches improve experiment selection, diagnostic recovery, and feedback seeking, while recent intervention-based methods perturb process steps to localize delayed credit (Foster et al., 2026; Wang et al., 2026; Liu et al., 2026; Meng, 2026; Hu, 2026; Guo et al., 2026). The causal target needed by a research agent remains unresolved. ❷ **Artifact–evidence confounding**: candidate ranking and local transition attribution do not identify the downstream value of an executed diagnostic result. To credit the information itself, improving the current artifact must be cleanly separated from acquiring evidence that improves later artifacts, especially when diagnostic code, caches, or side effects carry over into the continuation.

Evaluating experimental evidence also involves strict resource bounds. ❸ **Ignored opportunity cost**: an experiment may yield useful information, but acquiring it consumes training computation, inference tokens, and GPU time that could have been spent on direct search. Evidence that is useful early may arrive too late to exploit near termination, so the net value of an experiment cannot be separated from its cost.

We ask whether a controlled comparison can provide the missing credit signal. We freeze the student and train the researcher on the value of experiments. From the same pre-experiment candidate program and archive, we fork two continuations: one exposes the diagnostic result to the agent, while the other withholds it. A third continuation starts from the same state but spends the experiment's cost on direct search. We call this framework **StreamRSI**. Each component targets one obstacle: downstream utility replaces immediate accuracy ❶, restored-state forks isolate information access ❷, and a direct-search baseline prices the opportunity cost ❸. Because this supervision targets the researcher rather than any benchmark-specific artifact, what is learned is how to investigate: with archives cleared, the trained researcher accelerates fresh searches on held-out categories, an unseen streaming benchmark, and offline long video.

Our study uses streaming video because it offers a rich, executable design space with observable intervention chains from selected frames to task scores. A 27B policy chooses among direct edits, evidence inspection, controlled probes, and diagnosis revision, while the deployed 7B video answerer remains frozen. The researcher checkpoint is fixed within each comparison and updated only after paired outcomes are committed, with task scores supplied by an external evaluator.

Our contributions are:

- We formulate a research learning objective and a controlled attribution mechanism that credit experimental decisions by their downstream utility. By using candidate-state restoration and evidence-exposure interventions, we disentangle artifact quality from net research value, allowing informative negative results to become supervision.
- We validate both the credit terms and the fork protocol through eight-seed, matched-budget ablations: the full objective improves final quality over partial credits by 1.20–2.02 points, and altering the controlled comparisons reduces performance. Additionally, controlled streaming-video interventions and retention–readout probes expose failures of immediate outcome supervision.
- We show that the learned credit transfers: with candidate archives cleared, the trained researcher weights improve fresh searches on held-out task categories, a new streaming benchmark, and an offline long-video benchmark.

2 Related Work

Streaming video memory and readout. Streaming-video systems manage unbounded history via cached states (Di et al., 2025; Ning et al., 2025), hierarchical memories (Kim et al., 2026; Zhang et al., 2026; Xie et al., 2026), persistent event records (Zeng et al., 2026; Wen et al., 2026), or coreset selection (Mahdizadeh et al., 2026). Because a strong recent-window baseline often matches more elaborate memories (Shen et al., 2026), these design choices—including the distinction between internal cache retention (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024) and explicit visual presentation—form a complex optimization space. Rather than proposing a single deployed architecture, we use this space as an executable environment for a research policy.

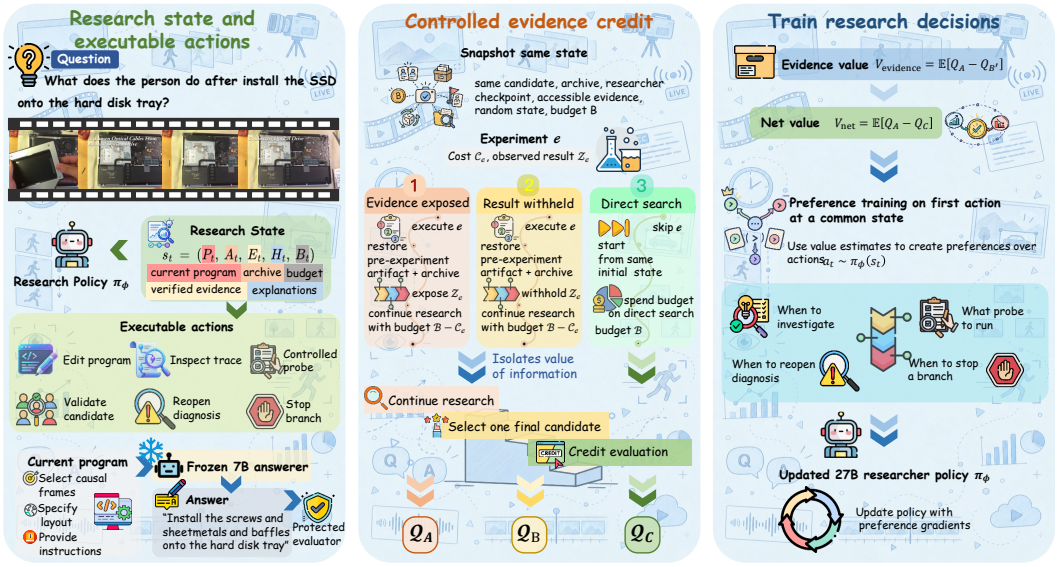


Figure 1: Overview of StreamRSI. **Left:** the research policy edits and diagnoses an inference-time video program while the 7B answerer and protected evaluator remain fixed. **Middle:** from the same snapshot, branch A exposes the observed result, branch B' withholds it under the same spent budget, and branch C skips the experiment and spends the saved budget on direct search; the selected artifacts yield Q_A, $Q_{B'}$, and Q_C. **Right:** evidence value $\mathbb{E}[Q_A - Q_{B'}]$ and net value $\mathbb{E}[Q_A - Q_C]$ create preferences over research actions and update the 27B researcher only.

Automated research and experiment selection. Research agents increasingly automate system design and scientific workflows (Hu et al., 2025; Agrawal et al., 2026; Yamada et al., 2025; Toledo et al., 2026; Cai et al., 2026), shifting evaluation from isolated coding to open-ended search (Lupidi et al., 2026; Chen et al., 2026). Within this broader agenda, active information acquisition chooses costly observations for their downstream utility (He et al., 2016; Wang et al., 2017). Recent agents instantiate related ideas by ranking candidates, running pilot experiments (Foster et al., 2026), and using anomalous observations to revise principles (Pu et al., 2026) or causal models (Cooper & Velasquez, 2026). StreamRSI addresses a specific missing layer in these systems: learning a scalable credit signal for whether a diagnostic experiment causally improved subsequent search.

Intervention-based credit attribution. Recent methods turn failures into localized supervision by freezing policies after diagnosis (Wang et al., 2026), learning when to request feedback (Liu et al., 2026), or attributing long-horizon steps through deletion and substitution interventions (Meng, 2026; Hu, 2026). StreamRSI adapts interventional attribution to the acquisition of research evidence. By restoring the pre-experiment state and adding a direct-search arm, we isolate the utility of the diagnostic information from both artifact changes and opportunity costs.

3 StreamRSI: Learning the Value of Experiments

Figure 1 summarizes StreamRSI as a three-stage research loop. A 27B research policy first operates on an explicit research state (Section 3.1), while the deployed answerer remains fixed. When the researcher proposes a diagnostic experiment, StreamRSI snapshots the pre-experiment state and forks the three controlled continuations (Section 3.2). Their downstream artifact qualities estimate the information value and net value of the experiment. This credit signal updates only the researcher policy, training it to make better decisions about when to investigate, which probe to execute, and when to reopen a diagnosis.

3.1 RESEARCH STATE, ACTIONS, AND OBJECTIVE

Let M_θ be a frozen answerer and P_t the current inference-time program. At a video query, P_t selects and arranges causal evidence under a visual-input cap before calling M_θ. This cap covers the visual input to a single call; frame storage, retrieval, preprocessing, and visual tokens are accounted for separately. The research policy π_ϕ receives

$$s_t = (P_t, \mathcal{A}_t, \mathcal{E}_t, \mathcal{H}_t, B_t), \tag{1}$$

where $\mathcal{A}_t$ is the candidate archive, $\mathcal{E}_t$ the verified evidence, $\mathcal{H}_t$ the current explanations, and B_t the remaining research budget. Actions include editing the program, inspecting an execution trace, proposing a controlled probe, validating a candidate, reopening a diagnosis, or stopping a branch. A probe registers its controlled factors, predicted observations, panel, and cost before execution. Inconclusive measurements may leave several explanations viable.

Let $\widehat{P}_\phi(B)$ denote the program selected when a search driven by π_ϕ exhausts total budget B. With θ fixed, the research objective is

$$\max_\phi \ \mathbb{E}\big[Q\big(\widehat{P}_\phi(B)\big)\big], \tag{2}$$

where Q is measured on data not used to select the program. Every action consumes part of B, so an experiment is worth running only if the information it yields improves this final quality more than spending its cost on direct search would. Sections 3.2 and 3.3 turn this criterion into supervision for π_ϕ.

3.2 CONTROLLED EVIDENCE CREDIT

An experiment can change both the artifact and the information available to the researcher. We isolate the second effect by snapshotting the candidate, archive, researcher checkpoint, accessible evidence, and random-state configuration before the experiment. Let e cost c_e from total budget B, and let z_e denote its observed result. We construct three continuations:

1. A: execute e, restore the pre-experiment artifact and archive, expose z_e, and continue research with budget $B - c_e$.
2. B': execute the same experiment and restoration, withhold z_e, and continue with the same budget $B - c_e$.
3. C: skip e and continue direct search from the same initial state with budget B.

The first two branches share the experiment specification; only its result is exposed differently. Diagnostic candidate code, side effects, caches, and unapproved result channels are excluded from both restored states. The researcher parameters remain fixed throughout each comparison. Paired continuations control sampling seeds where applicable and are replicated to estimate rollout variation.

Let Q_X be the quality of the final program proposed by branch X. To prevent overfitting, the branch selects this program using a dedicated selection panel, while Q_X is evaluated on a disjoint credit panel. We define

$$V_{\text{evidence}}(e \mid s, B) = \mathbb{E}[Q_A - Q_{B'}], \tag{3}$$

$$V_{\text{net}}(e \mid s, B) = \mathbb{E}[Q_A - Q_C]. \tag{4}$$

Equation 3 measures the effect of access to this result under the restored-state protocol. Equation 4 asks whether acquiring it is preferable to spending the same total budget on direct search. Their finite-sample estimates depend on the continuation policy and horizon. Both diagnostic execution and all forked continuations count toward training-data collection cost.

3.3 TRAINING DECISIONS, INCLUDING DIAGNOSIS REVISION

Credit supervises the first research action at a common decision state. Diagnostic versus direct-edit comparisons train when to investigate; comparisons between probes train what to investigate;

reopening versus continuing a diagnosis trains when to revise it. Outcome differences create preferences only when the predeclared uncertainty rule resolves their ordering. Invalid execution and inconclusive measurements are retained as context and receive no positive credit.

For same-state preferred and dispreferred actions a^+ and a^-, one implementation uses reference-relative preference optimization (Rafailov et al., 2023):

$$\mathcal{L}_{\text{research}} = -\mathbb{E} \log \sigma \left(\beta \left[\log \frac{\pi_\phi(a^+ \mid s)}{\pi_{\text{ref}}(a^+ \mid s)} - \log \frac{\pi_\phi(a^- \mid s)}{\pi_{\text{ref}}(a^- \mid s)} \right] \right). \tag{5}$$

Branch outcomes value the pre-experiment action: comparisons of alternative first actions at the shared decision state produce the training pairs. Credit is assigned to the first action only. Reference-checkpoint schedules, target-token limits, and optimizer-state restoration are fixed as part of training.

This objective can be combined with outcome-based preferences for program edits within a fixed training allocation. A negative immediate task-score change can therefore support a preferred research action if its net downstream value is positive. The proposed recursive loop improves artifacts and then uses measured experimental value to update the proposer for subsequent searches. A held-out development gate may reject researcher updates that regress search performance.

3.4 Budgets and evidence boundaries

The generated artifact contains a visual policy for frame selection, layout, and generic instructions. A preprocessing implementation may also be optimized if replay checks preserve the policy's outputs; such an equivalent implementation is credited for reduced cost at preserved correctness. The external evaluator computes correctness from protected labels, and the frozen student produces answers rather than judging its own correctness. Static restrictions and execution checks reduce shortcut opportunities (Appendix A).

The controller exposes the remaining budget and enforces immutable caps on researcher tokens, answerer queries, visual tokens, training computation, and GPU time. We use a 128-round lifecycle and log full-benchmark scores every eight rounds for monitoring. All arms receive the same resource limits and report actual cost. Frozen-answerer replicas can evaluate independent candidates in parallel.

To enforce strict evidence boundaries, the development environment is partitioned: support data guide local probes, the selection panel ranks candidate programs, and the credit panel supplies the supervision signal for the researcher. Full-benchmark scores are logged solely for external monitoring and are never exposed to the agent. A completely unseen transfer set is reserved for final evaluation.

4 Experiments

We organize the evaluation around four questions. **RQ1**: how does the inference-time program discovered by StreamRSI compare with published and reproduced streaming-video baselines? **RQ2**: do evidence credit and opportunity-cost correction improve research quality and efficiency beyond outcome-only learning and non-learning controls, and does the trained researcher transfer to unexposed tasks? **RQ3**: how do visual-context choices change artifact quality, and do selection-panel improvements survive credit-panel evaluation? **RQ4**: when retained evidence does not yield a correct answer, is the limiting factor retention or readout?

4.1 Experimental setup

The evaluation separates artifact quality from research capability. At the artifact level, we report OVO-Bench and StreamingBench results while keeping the deployed answerer frozen. At the researcher level, we compare complete search trajectories under matched budgets and independent seeds, so a strong final artifact alone does not imply a better research policy.

The deployed answerer is Qwen2.5-VL-7B-Instruct (Bai et al., 2025), kept frozen throughout. Among the published baselines, the Qwen2.5-VL-7B and HERMES-7B rows use the same backbone, so they are the closest reference points for our programs. Table 1 reports the published value

Table 1: OVO-Bench accuracy (%). Columns follow the three-track layout of Wen et al. (2026); Overall is the mean of the three category averages.

		Real-Time Visual Perception							Backward Tracing				Forward Active Responding				
Model	**Frames**	**OCR**	**ACR**	**ATR**	**STU**	**FPD**	**OJR**	**Avg.**	**EPM**	**ASI**	**HLD**	**Avg.**	**REC**	**SSR**	**CRR**	**Avg.**	**Overall**
Human	–	94.0	92.6	94.8	92.7	91.1	94.0	93.2	92.6	93.0	91.4	92.3	95.48	89.67	93.56	92.90	92.81
Open-source offline MLLMs																	
Qwen2.5-VL-7B	1 fps	67.8	55.1	67.2	42.1	66.3	60.9	59.9	51.5	58.8	23.7	44.7	29.94	47.85	49.17	42.32	48.97
LLaVA-OneVision-7B	32	66.4	57.8	73.3	53.4	71.3	62.0	63.26	54.2	55.4	21.5	42.78	25.64	65.18	58.33	49.72	51.92
InternVL2-8B	16	67.1	60.6	63.8	46.1	68.3	56.5	60.4	48.2	57.4	24.7	43.4	27.08	57.87	52.92	45.95	49.92
LLaVA-Video-7B	64	69.1	58.7	68.8	49.4	74.3	59.8	63.5	56.2	57.4	7.5	40.4	34.10	69.95	60.42	54.82	52.91
Qwen2-VL-7B	64	69.1	53.2	63.8	50.6	66.3	60.9	60.7	44.4	66.9	34.4	48.6	29.08	65.02	48.33	47.48	52.26
LongVU-7B	1 fps	55.7	49.5	59.5	48.3	68.3	63.0	57.4	43.1	66.2	9.1	39.5	11.89	69.32	61.25	47.49	48.13
Open-source online MLLMs																	
VideoLLM-online-8B	2 fps	8.1	23.9	12.1	14.0	45.5	21.2	20.18	22.2	18.8	12.2	16.02	9.63	21.44	14.63	15.23	17.14
Flash-VStream-7B	1 fps	24.2	29.4	28.5	33.7	25.7	28.8	28.4	39.1	37.2	5.9	27.4	8.02	67.25	60.00	45.09	33.61
Dispider-7B	1 fps	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1	18.05	37.36	48.75	34.72	41.78
TimeChat-Online-7B	1 fps	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	29.94	45.15	41.67	38.92	47.51
StreamForest-7B	1 fps	68.5	53.2	71.6	47.8	65.4	60.9	61.2	58.9	64.9	32.3	52.0	32.81	70.59	57.08	53.49	55.57
Streamo-7B	1 fps	79.2	57.8	75.0	49.4	64.4	70.1	66.0	54.6	52.0	31.7	46.1	29.96	51.03	83.33	54.77	55.61
HERMES-7B	1 fps	85.2	64.2	71.6	53.4	74.3	65.2	71.62	48.5	62.2	37.6	52.01	28.80	43.24	60.42	44.15	55.93
Recency baseline and ours																	
SimpleStream-4f (7B)	4	94.0	72.5	80.2	68.0	76.2	79.3	78.4	54.5	60.8	40.3	51.9	19.34	63.59	42.08	41.67	57.32
Ours (frozen 7B)	≈9	91.95	78.90	82.76	71.91	79.20	83.70	81.40	59.93	66.89	46.24	57.69	35.10	75.04	62.08	57.41	65.50

when the source paper gives that quantity for the stated configuration, and otherwise our reproduction of the same configuration. Reproduced OVO entries use the full 3,035-query evaluation. Published rows retain their source protocols and differ in backbone, frame budget, and training procedure, so cross-paper differences are descriptive.

We compare direct program search, ordinary reflection, a parameter-frozen researcher with the same active tools, active diagnosis trained only on final outcomes, and variants trained with controlled evidence credit. Candidate expressivity, initial programs, answerer, data access, and resource limits are matched across methods. Each method receives 20,000 GPU-s per seed, and all results use the same eight independent research seeds. We report final artifact quality with a 95% confidence interval, the number of seeds that reach the prespecified quality target $Q = 80.0$, the first cost at which the target is reached, and area under the quality–cost curve (AUC; higher is better).

4.2 Comparison experiments

The RSI OVO program combines a recent visual tail with evenly sampled history in a bounded prompt and uses a benchmark-specific readout configuration. It averages about 8.9 selected frames. The RSI StreamingBench program is a distinct recent4 policy with eight frames for Counting, different from the OVO union program. We report task-conditioned routing as part of the evaluator interface. Neither program updates the deployed answerer.

OVO-Bench. Ours achieves 65.50 Overall (69.55 RT/BT mean). As visualized in Figure 2, our program improves both real-time perception (+3.00 RT Avg.) and historical recall (+5.76 on EPM/ASI) relative to SimpleStream-4f, uniquely occupying the upper-right quadrant. Section 4.3.4 provides within-backbone ablations isolating these gains.

StreamingBench. The frozen-answerer program reaches 80.63 RTVU accuracy, 7.32 points above the Qwen2.5-VL-7B row and 1.91 points above HERMES-7B, which share its backbone. Together, the two benchmark results establish the artifact-level outcome; the following ablations isolate the research-credit signals and analyze the context choices associated with that outcome.

4.3 Ablation studies

4.3.1 Credit objective

We evaluate the full objective alongside its individual components ($A - B'$ and $A - C$), an outcome-only variant, and non-learning controls (frozen tools, direct search, ordinary reflection) to determine whether learned experimental credit is necessary under matched budgets.

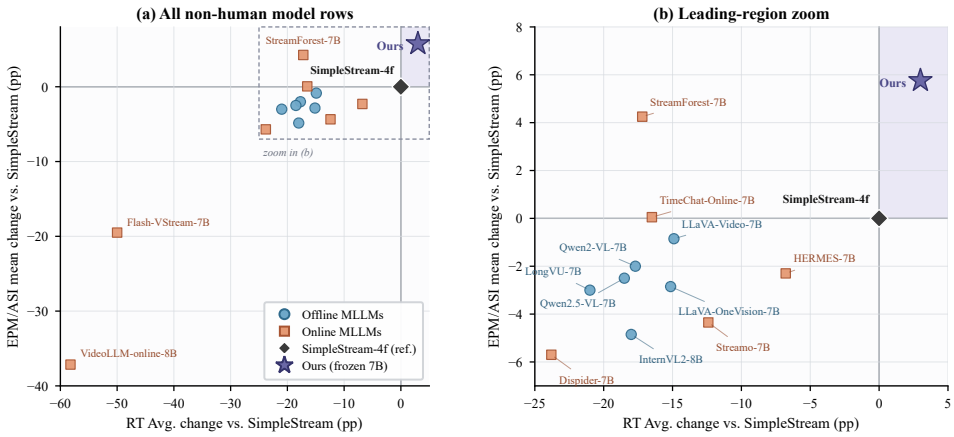


Figure 2: OVO-Bench deltas relative to the printed SimpleStream-4f row (Table 1). Panel (b) enlarges the leading region; the shaded quadrant marks improvement on both axes.

Table 2: StreamingBench real-time visual understanding accuracy (%) by official task. Column abbreviations follow the benchmark taxonomy (Lin et al., 2026). Published rows keep their own protocols and serve as context (Lin et al., 2026; Wen et al., 2026; Zhang et al., 2026). ALL is the question-weighted RTVU accuracy. For Ours, Frames is the per-query frame budget (8 for Counting).

Model	Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	ALL
Human	–	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Proprietary MLLMs												
Gemini 1.5 Pro	1 fps	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
GPT-4o	64	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Claude 3.5 Sonnet	20	80.49	77.34	82.02	81.73	72.33	75.39	61.11	61.79	69.32	43.09	72.44
Open-source offline MLLMs												
InternVL-V2-8B	16	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Qwen2-VL-7B	0.2–1 fps	75.20	82.81	73.19	77.45	68.32	71.03	72.22	61.19	61.47	46.11	69.04
Qwen2.5-VL-7B	1 fps	77.93	76.56	78.55	80.86	76.73	76.95	80.56	65.45	65.72	52.85	73.31
Qwen3-VL-8B	0.2–1 fps	73.57	71.09	76.34	80.39	65.22	74.77	78.70	67.89	61.47	47.67	70.20
Open-source online MLLMs												
StreamForest-7B	1 fps	83.11	82.81	82.65	84.26	77.50	78.19	76.85	69.11	75.64	54.40	77.26
HERMES-7B	1 fps	83.38	78.91	86.12	87.13	78.62	86.60	84.26	74.80	71.39	46.63	78.72
Online video agents												
EventMemAgent	$\leq$32	83.92	76.56	87.70	85.29	77.02	79.44	77.78	68.29	72.24	48.70	77.00
Ours (frozen 7B)	4 / 8	85.83	68.75	91.17	88.89	73.13	90.03	80.56	82.11	77.05	43.52	80.63

Table 3 shows that full StreamRSI consistently outperforms all baselines, reaching $Q = 80.0$ in all eight seeds within 9,400 GPU-s (13–19% faster than single-credit variants). In contrast, no outcome-only or non-learning baseline achieves the target within the 20,000 GPU-s budget.

The paired contrasts in Table 4 confirm the distinct value of each credit component: evidence credit improves upon outcome-only learning by isolating information value, while net credit further accounts for opportunity cost. Combining both achieves the highest final quality, with all pairwise confidence intervals strictly excluding zero.

4.3.2 Fork-protocol ablation

To isolate the contribution of the fork protocol, we vary its components while keeping the training pipeline, budget, and seeds fixed (Figure 3). *Naive fork* exposes the result but continues without

Table 3: Matched-budget researcher ablation over eight independent seeds. All methods receive 20,000 GPU-s per seed. Final quality reports the mean and 95% confidence interval; $>$ 20,000 denotes right-censoring because no seed reaches $Q = 80.0$ within budget. AUC summarizes quality over cumulative research cost; higher is better.

Researcher	Final quality ↑ (95% CI)	Seeds at $Q \geq 80$	First target cost ↓ (GPU-s)	AUC ↑
StreamRSI full	**82.70 [82.28, 83.12]**	**8/8**	**9,400**	**0.723**
Net credit ($A - C$)	81.50 [80.95, 82.05]	8/8	10,800	0.680
Evidence credit ($A - B'$)	80.68 [80.10, 81.25]	8/8	11,600	0.672
Outcome-only researcher	79.60 [79.05, 80.15]	0/8	$>$ 20,000	0.509
Frozen researcher + same tools	78.80 [78.21, 79.39]	0/8	$>$ 20,000	0.449
Direct search	78.40 [77.73, 79.07]	0/8	$>$ 20,000	0.420
Ordinary reflection	77.90 [77.17, 78.63]	0/8	$>$ 20,000	0.389

Table 4: Paired final-quality contrasts over the same eight seeds.

Contrast	Δ ↑	95% CI
Full − Net credit	+1.20	[+0.89, +1.51]
Full − Evidence credit	+2.02	[+1.71, +2.34]
Full − Outcome-only	+3.10	[+2.84, +3.36]
Full − Frozen + tools	+3.90	[+3.58, +4.22]
Full − Direct search	+4.30	[+3.92, +4.68]
Net − Evidence	+0.82	[+0.64, +1.01]
Evidence − Outcome-only	+1.08	[+0.91, +1.24]

Table 5: Context-design ablation on frozen Qwen2.5-VL-7B. Frame budgets differ; the local last-4 baseline differs from the published SimpleStream row.

Design	#Fr	RT	Bwd	OVO Avg	Fwd
last-4 (recency)	4	79.5	43.7	61.6	40.9
last-12 (recency)	12	78.1	45.6	61.8	44.1
last-16 (recency)	16	78.0	45.5	61.8	45.8
union-compiled (RSI)	≈9	**81.4**	**57.7**	**69.5**	57.4

restoring the pre-experiment state; *mask only* withholds the result but retains diagnostic side effects; *evidence permutation* and *noise evidence* expose a mismatched or random result; *unstructured MC credit* replaces the $A - B'$ comparison with a three-rollout average.

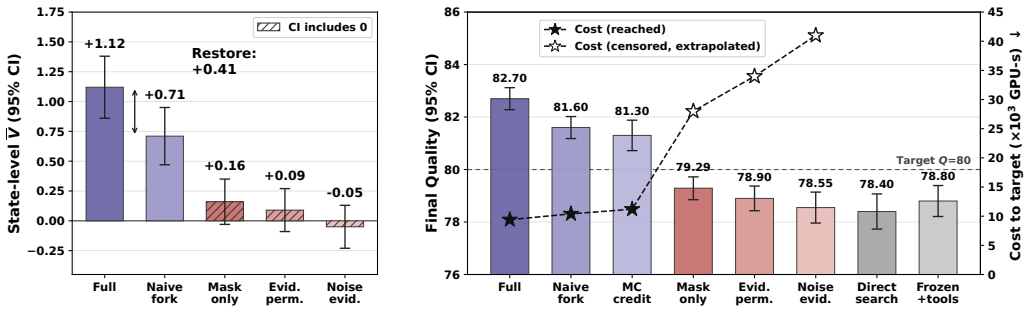
State restoration prevents signal dilution. Without restoration, the naive fork carries the diagnostic artifact into the continuation. This mixes the value of the information with the effect of the code change, dropping final quality by 1.10 points below full StreamRSI and increasing the cost to target. Masking the result without restoring the state is worse still, as retained caches allow the masked branch to partially recover the withheld result. This leakage drives the evidence value close to zero and pushes final quality below even the outcome-only baseline.

The gain requires accurate results and a structured comparison. Replacing the true result with a permuted or random one eliminates the evidence value entirely and drops final quality to the level of direct search, confirming that the researcher relies on the actual diagnostic content. The structure of the comparison also matters: unstructured MC credit uses the same number of continuations but lacks the $A - B'$ contrast. It yields a lower final quality and costs nearly 20% more than full StreamRSI, showing that denser rollouts alone are insufficient.

4.3.3 Archive-free transfer

We run fresh searches on three pre-registered targets of increasing distance, all built from public benchmarks: OVO-Bench and StreamingBench categories held out from research feedback (T1), the streaming benchmark unused during research, StreamingBench for OVO-Bench runs and vice versa (T2), and the offline long-video benchmarks Video-MME (Fu et al., 2025) and MLVU (Zhou et al., 2025) (T3). Starting from the untrained researcher, we transfer the trained weights (with candidate code, diagnosis archives, and task memory cleared), the source archive, or both.

Table 6 shows the same ordering on every target: joint $>$ weights only $>$ archive only $>$ base. Weights carry most of the gain, so the learned credit resides mainly in researcher parameters. Gains shrink with task distance, and archives decay faster: from T1 to T3, the weight-only gain keeps 49% of its value and the archive-only gain 37%. Joint transfer exceeds the sum of the two single-carrier gains, increasingly so on distant targets (by 0.60 points on T3), suggesting that a trained researcher makes better use of an archive that no longer applies directly.



(a) Evidence Value by Fork Protocol **(b) Trained Researcher Quality and Cost**

Figure 3: Fork-protocol ablation over eight seeds. **(a)** State-level evidence value $\overline{V}$ with 95% CIs; hatched bars have CIs containing zero. Unstructured MC credit has no B' branch and therefore no $\overline{V}$. **(b)** Final quality of the trained researcher (bars, left axis) and cost to reach $Q = 80$ (stars, right axis). Hollow stars denote variants for which no seed reaches the target within 20,000 GPU-s; their costs are extrapolated (Appendix D).

Table 6: Transfer to unexposed tasks (mean over three seeds). Parentheses give the gain over the base search.

Transferred	T1	T2	T3
None (base search)	78.83	77.87	76.84
Weights only	80.32 (+1.49)	79.08 (+1.21)	77.57 (+0.73)
Archive only	79.62 (+0.79)	78.37 (+0.50)	77.13 (+0.29)
Weights + archive	**81.23** (+2.40)	**79.98** (+2.11)	**78.46** (+1.62)

Table 7: Readout-asymmetry probe on frozen Qwen2.5-VL-7B. Paired-bootstrap 95% CI on the focus arm; open word uses exact match.

Evidence at question time	4-choice	Open word
long KV only	0.17	–
+ 16 uniform frames (event not among them)	0.17	–
+ all 9.3 retained frames (event included)	0.42	–
+ one focus frame (event)	**0.75** [0.58, 0.92]	**1.00** (24/24)
recency window only	0.17	0.33

4.3.4 Context interventions and failure evidence

Table 5 shows that the union view outperforms denser recent windows on backward tracing and real-time perception, indicating that temporal distribution matters. Gains of this kind, however, do not always survive held-out evaluation: on OVO backward queries split by video into selection and credit panels, similarity-based historical reselection improves accuracy on selection but drops on credit, and energy-based reselection shows a similar reversal. Immediate selection gains are therefore unreliable research supervision. Whether retained evidence is actually used is a separate question (Table 7). In synthetic segments where an early event is followed by a question, the compressor successfully retains the event states in all cases. Yet the cache-only arm answers below the chance rate. Presenting all retained frames improves accuracy, and supplying a single focus frame at the known event location nearly solves the task (perfect accuracy in the open-word arm). Retention therefore does not guarantee readout, so StreamRSI treats retention and explicit presentation as separate probe types. Appendix B gives a separate cache-compression guarantee.

5 Discussion and Conclusion

StreamRSI trains a research policy to assess experiments by their downstream utility rather than their immediate task score. The resulting objective outperforms outcome-based training and either credit term alone under a matched compute budget. Our ablations show two requirements for this gain: the opportunity cost of an experiment must be priced against a direct-search baseline, and the causal contribution of its result must be isolated by restoring the pre-experiment state. Without this state restoration, the credit signal is diluted by artifact changes and loses its advantage.

The learned credit transfers to new tasks. A trained researcher with cleared archives still accelerates search on held-out categories, new streaming benchmarks, and offline long video, though the gain decays with distance from the source environment. Isolating the contribution of specific research actions, such as diagnosis revision, remains a direction for future work.

AI USE STATEMENT

Large language models (LLMs) were used solely for text polishing, grammar checking, and minor language revisions. They were not used to generate experimental designs, scientific claims, analyses, or conclusions. All scientific content was authored and verified by the authors.

REPRODUCIBILITY STATEMENT

The core methodology, action spaces, and credit objectives are defined in Section 3. Experimental protocols, compute budgets, frame allocations, and seed counts are reported in Section 4. Implementation details, execution sandbox constraints, budget tracking, and data splits are provided in Appendix A and Appendix E. Source code, discovered programs, prompt configurations, and evaluation manifests will be open-sourced upon publication.

ETHICS STATEMENT

All experiments in this paper are conducted on publicly available benchmarks (OVO-Bench, StreamingBench, Video-MME, and MLVU). The study does not involve human subjects, personally identifiable information, or sensitive surveillance material. To mitigate risks associated with autonomous code generation, all researcher actions execute in a sandboxed environment with strict boundaries on dynamic code execution, filesystem I/O, and external network access (Appendix A).

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 671–680, 2014.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Ziyang Cai, Amir Saeidi, and Harkirat Behl. Ml-autoresearch: Training machine learning research agents with automatically generated environments, 2026. URL https://arxiv.org/abs/2603.17216.

Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pp. 626–635, 1997.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Advances in Neural Information Processing Systems*, 38, 2026.

Patrick Cooper and Alvaro Velasquez. Active causal experimentalist (ace): Learning intervention strategies via direct preference optimization. *arXiv preprint arXiv:2602.02451*, 2026.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Duc Duong, Hoang Anh Duy Le, Jianwen Xie, Anshumali Shrivastava, and Zhaozhuo Xu. Forget without compromise: Nexus sampling for streaming kv-cache eviction under fixed budgets. *arXiv preprint arXiv:2606.23961*, 2026.

Thomas Simon Foster, Bassel Al Omari, Tingchen Fu, Thomas Mann, Carl Domond, Lucia Cipolina-Kun, Bhavul Gauri, Muna Aghamelu, Alexander D Goldie, Eryk Helenowski, et al. Ai research preference models. *arXiv preprint arXiv:2608.13940*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, et al. Memeye: A visual-centric evaluation framework for multimodal agent memory. *arXiv preprint arXiv:2605.15128*, 2026.

He He, Paul Mineiro, and Nikos Karampatziakis. Active information acquisition, 2016. URL https://arxiv.org/abs/1602.02181.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, volume 2025, pp. 21344–21377, 2025.

Xuchun Hu. Credo: Variance-guided rubric evolution for replay-corrected credit assignment, 2026. URL https://arxiv.org/abs/2609.24174.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Ehsan Kazemi, Marko Mitrovic, Morteza Zadimoghaddam, Silvio Lattanzi, and Amin Karbasi. Submodular streaming in all its glory: Tight approximation, minimum memory and low adaptive complexity. In *International Conference on Machine Learning*, pp. 3311–3320. PMLR, 2019.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *Advances in Neural Information Processing Systems*, 38: 138983–139013, 2026.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Boyang Liu, Senjie Jin, Peixin Wang, Zhangyue Yin, Yibo Wang, Yuhao Zhou, Xinbing Liang, Shizheng Zhu, Yuhui Wang, Jingqi Tong, et al. Cafe: Self-improving search agents need co-evolving feedback. *arXiv preprint arXiv:2608.24794*, 2026.

Alisia Lupidi, Bhavul Gauri, Thomas Simon Foster, Bassel Al Omari, Despoina Magka, Alberto Pepe, Alexis Audran-Reiss, Muna Aghamelu, Nicolas Baldwin, Lucia Cipolina-Kun, et al. Airs-bench: a suite of tasks for frontier ai research science agents. *arXiv preprint arXiv:2602.06855*, 2026.

Ailar Mahdizadeh, Puria Azadi, Muchen Li, Xiangteng He, and Leonid Sigal. Cords: Coreset-based representative and diverse selection for streaming video understanding. *arXiv preprint arXiv:2605.14310*, 2026.

Richard Matthew McCutchen and Samir Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pp. 165–178. Springer, 2008.

Renwei Meng. Policy-conditioned counterfactual credit for verifiable reinforcement learning of long-horizon language agents. *arXiv preprint arXiv:2606.05263*, 2026.

Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Yingming Pu, Tao Lin, and Hongyu Chen. Principle-evolvable scientific discovery via uncertainty minimization. *arXiv preprint arXiv:2602.06448*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, et al. Ai research agents for machine learning: Search, exploration, and generalization in mle-bench. *Advances in Neural Information Processing Systems*, 38:35309–35348, 2026.

Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11 (1):37–57, 1985.

Pan Wang, Yihao Hu, Hang Wang, Zirui Lv, Xin Zhang, Jianshe Li, Jiang-Ming Yang, Wei Wu, and Yongqi Tong. Diagnosis before recovery: Turning agent failures into selective self-correction. *arXiv preprint arXiv:2608.11772*, 2026.

Yingfei Wang, Chu Wang, and Warren Powell. Optimal learning for sequential decision making for expensive cost functions with stochastic binary feedbacks. *arXiv preprint arXiv:1709.05216*, 2017.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use. *arXiv preprint arXiv:2602.15329*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. *arXiv preprint arXiv:2603.02096*, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Haowei Zhang, Shudong Yang, Jinlan Fu, See Kiong Ng, and Xipeng Qiu. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8411–8430, 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

A RESEARCH-LOOP CONTRACTS

Executable scope and validity. Generated policy code selects causal frame indices, chooses a permitted layout, and supplies generic instructions. Protected labels are available only to the evaluator. I/O, dynamic execution, literal answer tables, and identity-based routing are restricted; task-specific routing in the harness is disclosed and shared across controls. Equivalent preprocessing changes require replay checks. Action, execution, measurement, improvement, and learning eligibility are recorded separately; failed code stays available as context.

Fork identity, cost, and stopping. Each fork records candidate and archive hashes, researcher checkpoint, evaluator identity, data manifests, random-state configuration, and resource reservations. The pre-experiment artifact is restored before comparison, and the masked branch cannot access the result through caches, paths, or logs. The same checkpoint generates all paired continuations, and parameter updates occur after their credit is committed. We track tokens, answerer queries, visual tokens, optimizer steps, and GPU time, including diagnostic execution and all forked continuations. Exhausting the budget stops new actions without deleting evidence.

B STREAMING QUERY-AGNOSTIC KV EVICTION (SQA-KVC)

This section analyzes attention-output fidelity under cache compression; the visual-program search does not depend on it.

B.1 Definition. A stream produces KV pairs $(k_i, v_i) \in \mathbb{R}^d \times \mathbb{R}^{d_v}$, $i = 1 \dots T$, T unbounded. For a query q, full attention is $o(q) = \sum_i a_i(q) v_i$ with $a_i(q) = \exp(\langle q, k_i\rangle/\sqrt{d}) / \sum_j \exp(\langle q, k_j\rangle/\sqrt{d})$. An online policy keeps $S_t \subseteq S_{t-1} \cup N_t$ with $|S_t| \le B$ (irreversible, one-pass). The compressed output $\hat{o}(q)$ renormalizes over S; $\mathrm{err}(S, q) = \|o(q) - \hat{o}(q)\|_2$. **SQA-KVC:** minimize, over one-pass policies, the worst case of err over streams and q in a bounded query set $\mathcal{Q} = \{q : \|q\| \le \rho\}$. Assumptions: **(A1)** $\|v_i\| \le C$; **(A3)** $\|k_i\| \le \kappa$; **(A2′)** transport-local consistency: for the transport φ assigning each evicted token to its retained representative, $\widehat{L}_\varphi := \max_i \frac{\|v_i - v_{\varphi(i)}\|}{\|k_i - k_{\varphi(i)}\|}$—measurable per instance, per layer, per head.

B.2 Impossibility (Proposition 1). For $B < D \le \min(d, d_v)$, there is a $(B+1)$-token stream over D orthogonal key/value directions on which every deterministic online policy has $\mathrm{err} \ge C\big(1 - (B+1)e^{-\rho\kappa/\sqrt{d}}\big)$ for some q with $\|q\| \le \rho$, while the clairvoyant budget-B evictor achieves $\mathrm{err} \le \frac{2CB}{B + e^{\rho\kappa/\sqrt{d}}}$. The competitive ratio is $\Omega(e^{\rho\kappa/\sqrt{d}}/B)$ in this adversarial construction. This is a worst-case limitation on deterministic selectors and does not imply failure on natural streams.

B.3 Characterization and certificate. Theorem A (upper). With mass-conserving attention (each retained token carries its absorbed count n_j as a $\log n_j$ logit bias) and (A1)(A2′)(A3): $\mathrm{err}(S, q) \le \widehat{L}_\varphi\, \delta(S) + C\big(e^{2\rho\delta(S)/\sqrt{d}} - 1\big)$ for all $\|q\| \le \rho$, where $\delta(S) = \max_i \min_{j \in S} \|k_i - k_j\|$. **Proposition 2 (lower).** For every retained set S there exist admissible instances with $\mathrm{err} \ge \min\{L\,\delta(S), 2C\}(1 - o(1))$ as $\rho \to \infty$: in the unsaturated regime the covering radius is a *necessary* scale—no other statistic can uniformly upper-bound the error. **Theorem A′ (certificate).** $\mathrm{err}(S, q) \le \sum_i a_i(q)\, \|v_i - v_{\varphi(i)}\| + C \sum_{j \in S} |w_j(q) - \hat{a}_j(q)|$, where w_j sums the true attention transported onto j and $\hat{a}_j$ is the biased compressed attention on j. Both terms are computable in a post-hoc audit with full attention weights, original values, compressed output, and transport; the retained cache alone is insufficient after irreversible deletion.

B.4 Algorithm and schedules. Theorem B (StreamKC). A one-pass, only-discard streaming k-center (arrival absorbed within threshold τ; on overflow τ doubles and a maximal τ-separated net is rebuilt, with token-level mass additivity) maintains $\delta(S_t) \le 8\, \delta^*_B(t)$ in $O(B)$ memory, combining with Theorem A into an always-valid bound. Sourcing: doubling-tree streaming k-center (Charikar et al., 1997); the best known streaming k-center guarantee is $(2{+}\varepsilon)$ (Matthew McCutchen & Khuller, 2008); the 8-approximation is the price of the one-pass, only-discard constraint. **Theorem C (recursive compression).** Over R compression rounds with per-round merge radii τ_r, the composed

transport radius is $\leq \sum_r \tau_r$ with exact mass additivity: a fixed-radius schedule gives $R\tau$ (linear growth), while a geometric schedule $\tau_r = 2^r\tau_0$ gives $\leq 2\tau_R$—compression triggers should be geometric in radius, not periodic in time. (Reservoir-style randomization, in the spirit of (Vitter, 1985; Duong et al., 2026), improves expected survival but does not recover a worst-case bound; submodular-streaming guarantees (Badanidiyuru et al., 2014; Kazemi et al., 2019) do not transfer because facility-location objectives are not one-pass evaluable.)

B.5 SurpCache (infeasible regime). When the covering number exceeds the budget, no subset selector retains a worst-case guarantee; natural streams may still expose a relevance prior. Define an entry's causal surprisal $u(e) = |e|^{-1} \sum_{t \in e} -\log p_\theta(x_t \mid x_{<t})$—already available from the frozen model's forward pass. SurpCache keeps the exact top-B entries by u: scores are fixed at arrival and the top-B threshold only rises, so retention is one-pass, irreversible, query-agnostic, and $O(B)$. Under **(A4)**—the target of a sparse-recall query lies in the top-B of surprisal with probability $\geq 1 - \beta$—the target is retained with probability $\geq 1 - \beta$ (Theorem S). (A4) is an added, measurable assumption that does not remove Proposition 1.

B.6 Validation. We check Theorem A′ per query position, layer, and head on real attention (correlated checks of the implemented inequality, above floating-point noise tolerance):

Model (modality)	**Checks**	**Violations**	**Median tightness**
Qwen2.5-7B (text) (Hui et al., 2024)	30,240	0	7.7×
Mistral-7B (text) (Jiang et al., 2023)	34,560	0	8.2×
Qwen2.5-VL-7B (video) (Bai et al., 2025)	8,960 + 6,720	0	7.6×
Total (4 runs)	**80,480**	**0**	–

Score-based vs coverage-based retention (needle recall): 240-event streams, 20 seeds, 2,020 model calls—window 0.00, attention-based 0.00, reservoir 0.07, coverage-based 1.00, full 1.00; white-box 7B KV (10 seeds, physical eviction)—score-based 0.00, coverage-based 0.90 (B=16) / 1.00 (B=32), full 1.00. (A2′) profile (Qwen2.5-7B, p50/p90 of $\widehat{L}_\varphi$): L0 0.14/0.35, L2 0.30/0.48, L7 0.74/1.09, L14 0.58/0.80, L21 0.80/1.15, L27 3.54/7.50—the layer dependence motivates instance-specific diagnostics.

C READOUT PROBE DETAILS

The probe uses 24 segments with a full-screen pure-color event at approximately $t = 8$–10 seconds. All arms start from copies of the same compressed stream; the later question is absent from cache construction. Cache-only answers use the long KV context. Uniform replay adds 16 frames that omit the event; all-retained replay presents approximately 9.3 frames including it; focus replay presents one known event frame; recency-only uses the latest window. Four-choice scoring extracts the option letter. The focus-arm interval in Table 7 is the reported bootstrap interval over segments. The open-word arm uses exact color-name matching. Event-aware focus is an oracle diagnostic.

D EXTRAPOLATED COST TO TARGET

Variants that do not reach $Q = 80.0$ within the budget $B = 20{,}000$ GPU-s are right-censored. Figure 3 shows extrapolated costs; all statistical comparisons of censored variants use final quality.

Model. Under a fixed resource allocation, research quality typically saturates as cost increases. We therefore model each quality–cost curve as

$$Q(c) = Q_0 + w\left(1 - e^{-c/\tau}\right), \tag{6}$$

where $Q_0 = 70.1$ is the measured quality of the initial program, w is the asymptotic gain so that $Q_\infty = Q_0 + w$, and τ is a pace constant in GPU-s. The target requires a gain of $\Delta = 80.0 - Q_0 = 9.9$.

Fitting. For variants that reach the target, we use two observations, the first target cost c^* and the final quality $Q(B)$:

$$w\left(1 - e^{-c^*/\tau}\right) = \Delta, \qquad w\left(1 - e^{-B/\tau}\right) = Q(B) - Q_0. \tag{7}$$

Dividing the two equations eliminates w; we solve the resulting one-dimensional equation for τ by root finding and then substitute back to obtain w. For censored variants, only $Q(B)$ is observed, so the model is underdetermined. We fix the asymptote slightly above the target, $Q_\infty \in [80.2, 80.9]$, which encodes the assumption that the variant would eventually pass the target with a small margin. Given $w = Q_\infty - Q_0$, the endpoint yields

$$\tau = -\frac{B}{\ln\left(1 - \left(Q(B) - Q_0\right)/w\right)}. \tag{8}$$

Extrapolation and uncertainty. Solving $Q(c^*) = 80.0$ gives

$$c^* = \tau \ln \frac{w}{w - \Delta}, \tag{9}$$

which is finite only when $Q_\infty > 80.0$ and diverges as $Q_\infty \to 80.0^+$. If the true asymptote of a censored variant is at most 80.0, the variant never reaches the target, and the extrapolated cost is therefore optimistic. We report uncertainty by perturbing τ by $\pm 20\%$ with w fixed. Because c^* is linear in τ, the resulting interval is $[0.8c^*, 1.2c^*]$. For censored variants, w is assumed rather than fitted, so this τ sensitivity and the Q_∞ assumption jointly constitute the uncertainty statement. Table 8 lists the parameters.

E REPRODUCIBILITY AND REMAINING VALIDATION

Limitations. Forked continuations add training cost and estimate value relative to a particular continuation policy and horizon. Repeated reuse of credit panels risks overfitting. Existing benchmark results include differing budgets, task-conditioned configurations, and coverage differences. The readout study is limited to 24 synthetic segments on one frozen 7B backbone.

The benchmark tables distinguish published source values from forward scores obtained in our 3,035-query reproduction, OVO RT/BT from three-category Overall, and question-weighted StreamingBench RTVU accuracy. Reproducing each result requires raw media, frame indices, prompts, predictions, errors, and code versions.

Table 8: Parameters of the saturating quality–cost model ($Q_0 = 70.1$, $B = 20{,}000$ GPU-s). For reached variants, c^* is observed and (w, τ) are fitted exactly. For censored variants, Q_∞ is assumed; the last column shows the range of c^* over $Q_\infty \in [80.2, 80.9]$.

Variant	$Q(B)$	Q_∞	τ (GPU-s)	c^* (GPU-s)	$\pm 20\%\,\tau$	c^* over Q_∞ range
Full	82.70	83.45	6,950	9,400 (obs.)	–	–
Naive fork	81.60	81.98	5,800	10,400 (obs.)	–	–
Unstructured MC credit	81.30	81.66	5,770	11,200 (obs.)	–	–
Mask only	79.29	80.53	9,380	≈28,000	[22,400, 33,600]	[26,100, 32,600]
Evidence permutation	78.90	80.38	10,330	≈34,000	[27,200, 40,800]	[29,500, 38,300]
Noise evidence	78.55	80.27	11,240	≈41,000	[32,800, 49,200]	[32,600, 43,300]

Evidence-credit training freezes support, selection, and credit manifests by video/source, records all exposures, and reserves new transfer tasks before optimization. Archive-free transfer resets candidate code and task-specific memory. We record gradients, adapter and optimizer hashes, reload checks, paired continuation costs, and all negative or inconclusive outcomes.