PROV-MEM: PROACTIVE MEMORY FOR STREAMING VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Proactive streaming video question answering requires a model to continuously monitor an incoming stream and autonomously decide when to speak, what to say, and whether to emit a response. Existing streaming systems face a fundamental difficulty: retaining long visual histories is costly, whereas discarding them can lead to missed events and repeated responses. We propose **ProV-Mem**, a training-free proactive memory framework built around a frozen VLM backbone. ProV-Mem decomposes the proactive decision into three lightweight memory tiers: a temporal-trigger memory controls when to query the backbone and recovers from prolonged silence, an instructional memory guides what to say through a fixed prompting prior for concise and visually grounded generation, and an episodic memory determines whether to emit by filtering near-duplicate responses. Under an idealized conditional-sufficiency assumption, we derive a stagewise approximation bound relating the decomposed process to a full-state joint policy. We further introduce Cost-Aware Area Under the Curve (**CAUC**), which extends PAUC by explicitly discounting redundant, overly frequent, and unnecessarily long emissions. Experiments on the four scenarios of ProactiveVideoQA show that ProV-Mem achieves the best CAUC across all scenarios and the best PAUC on WEB, EGO, and TV, while maintaining low per-response inference latency.

1 INTRODUCTION

Recent advances in video-language models (VLMs) (Lin et al., 2024; Maaz et al., 2024; Li et al., 2024) have driven a paradigm shift from retrospective video understanding (Xu et al., 2016; Lei et al., 2018) to real-time streaming applications, such as live moderation (Xu et al., 2026) and embodied agents (Driess et al., 2023; O'Neill et al., 2024). Unlike traditional passive streaming benchmarks where explicit user queries drive every turn, Proactive Video Question Answering (Proactive-VQA) (Wang et al., 2025a) requires an agent to continuously monitor an unbounded video stream without explicit prompts. Under strict real-time constraints, the agent must autonomously determine the timing (when to speak), content (what to say), and necessity (whether to emit) of each response, integrating trigger control directly into the task formulation.

A dual predicament: passive memory and a leaky metric. To sustain real-time throughput on unbounded streams, existing architectures typically discard raw video frames, which deprives the model of explicit temporal visual history. This lack of historical awareness causes trained-head baselines to exhibit severe templated repetitions and coverage gaps, frequently missing critical events across target windows, as shown in Figure 1(a). Latest pipelines attempt to mitigate this by relying on lightweight feature caches or hidden states to implicitly represent past context. However, these implicit representations remain fundamentally passive: they store compressed context but play no active role in governing when, what, or whether to respond, so any degradation inevitably causes all three controls to collapse simultaneously. Crucially, this systemic failure is further masked by a structural loophole in the standard evaluation protocol. The primary metric, Proactive Area Under the Curve (PAUC) (Wang et al., 2025a), integrates per-fire quality scores over ground-truth windows, which allows degenerate, over-firing policies to artificially inflate their scores by accumulating window duration rather than increasing information density.

Our solution: from passive caching to proactive memory. The root cause of implicit context decay is that a passive buffer plays no structural role in decision-making, forcing a monolithic head to

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.

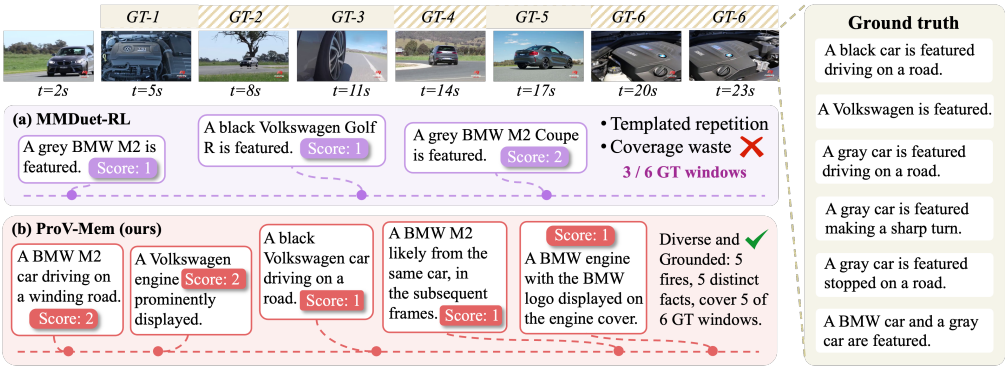

Figure 1: Comparison of (a) existing methods and (b) our ProV-Mem method.

recover timing, content, and redundancy judgements from compressed history alone. We propose a fundamentally different paradigm: *proactive memory*, in which each memory tier does not merely store context but actively governs a distinct axis of the streaming decision. We instantiate this paradigm as **ProV-Mem**, a training-free framework comprising three constant-size, non-parametric proactive memory tiers sharing a frozen VLM backbone. A temporal-trigger memory actively controls *when* to speak via deterministic gating over a compact timestamp state, with a silence-recovery for re-engagement; an instructional memory actively shapes *what* to say by imposing a canonical surface-form prior that maximises visual grounding and aligns generations with downstream lexical filtering; and an episodic memory actively decides *whether* to emit by caching the latest response and filtering redundancy via character-level similarity. Because each tier maintains constant-size state rather than raw-frame buffers, ProV-Mem eliminates coherence degradation and reduces cross-scenario adaptation to tuning a few interpretable scalars. We further show that these three proactive memory tiers can be optimized independently, with an approximation gap characterized by three stagewise discrepancy terms. Finally, we introduce Cost-Aware Area Under the Curve (CAUC), a metric that penalises redundancy, excessive frequency, and token inefficiency to close the evaluation loophole.

Experimental success and contributions. Experiments on the ProactiveVideoQA benchmark (Wang et al., 2025a) show that ProV-Mem consistently outperforms existing baselines. Without scenario-specific fine-tuning, the framework achieves state-of-the-art proactive performance while maintaining low response latency. In summary, our main contributions are threefold:

- We propose ProV-Mem, a proactive memory framework factorising the proactive decision function into three independent memory stages to enable seamless cross-scenario adaptation.
- We expose a structural reward-hacking loophole in existing evaluation metrics and propose CAUC, a rigorous cost-aware metric that penalises redundancy, chatter frequency, and token inefficiency at each emission.
- We provide comprehensive empirical evidence showing that ProV-Mem achieves excellent performance on ProactiveVideoQA under both standard and cost-aware evaluation.

2 Related Work

Proactive-VQA and Streaming Video-Language Models. Passive VQA s, in which the model consumes a full video before answering, has been extensively studied (Lei et al., 2020; Zhang et al., 2023). Wang et al. (2025a) depart from this paradigm by requiring frame-by-frame streaming with autonomous decisions on timing, content, and emission necessity. Broader benchmarks such as StreamingBench (Lin et al., 2026) and OVOBench (Niu et al., 2025) advance online video understanding but remain question-driven, leaving autonomous firing outside their protocol. Existing methods split into trained trigger-head systems such as MMDuet (Wang et al., 2025b) and VideoLLM-Online (Chen et al., 2024a), and training-free adaptations via chunk-based (Zhang et al., 2024b; Chen et al., 2025) or prompt-based (Han et al., 2024; Wang et al., 2024) reuse of offline VLMs. A parallel line on long-context streaming, including StreamingVLM (Xu et al., 2026) and StreamingLLM (Xiao et al., 2024), scales to unbounded inputs but assumes passive interaction. The trained-head

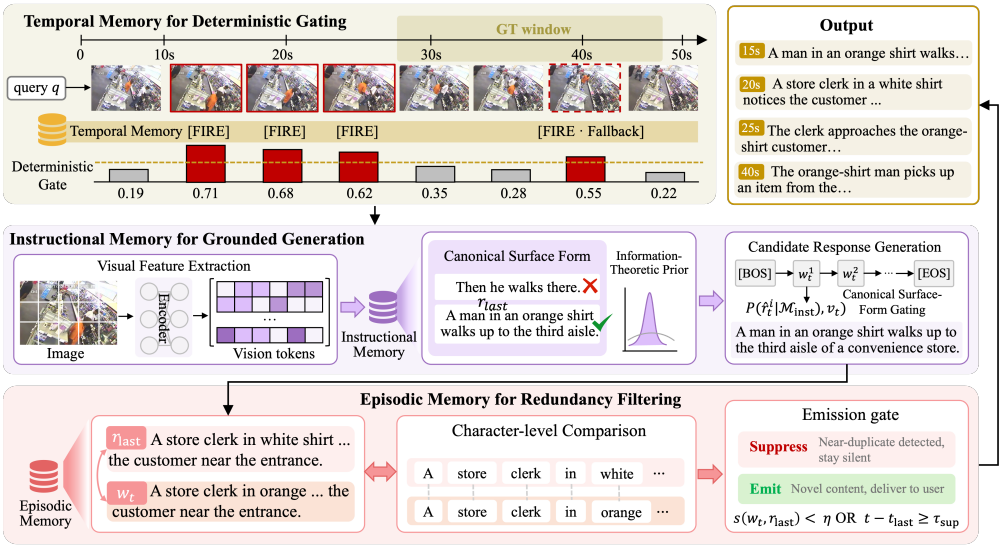

Figure 2: Framework of our ProV-Mem method.

family degrades outside its training distribution, while training-free and long-context families lack autonomous firing, leaving a structurally complete, proactive controller as the gap our work targets.

Prompt-Based Control and Memory-Augmented Models Since CLIP (Radford et al., 2021) and Flamingo (Alayrac et al., 2022), a substantial line of work shapes frozen VLM behaviour at inference time via system prompts and in-context examples, where grounded-detail and persona prompts (You et al., 2024; Maaz et al., 2024) elicit specificity or stylistic surface form. In parallel, a memory-augmented lineage attaches explicit memory modules read at inference to modulate generation, including MemoryBank (Zhong et al., 2024) which stores summarised interactions in a vector database, and Memorizing Transformers (Wu et al., 2022) and RETRO (Borgeaud et al., 2022) which retrieve nearest-neighbour examples during generation. These two lineages are typically pursued in isolation, and the memory-augmented one is moreover parametric, requiring target-domain training and retrieval-time operations that block the streaming forward pass; the absence of a single non-parametric formulation that unifies inference-time prompting and explicit memory while staying backbone-agnostic and streaming-compatible is precisely the gap our work targets.

3 Method

This section presents the mathematical and structural design of ProV-Mem shown in Figure 2. We first formalize Proactive-VQA under streaming causality constraints. We then show that the proactive decision can be decomposed into three coordinated stages, with an explicit suboptimality bound governed by measurable stagewise violation terms. Next, we describe the three lightweight memory tiers that implement this decomposition in practice. Finally, we introduce CAUC, a cost-aware refinement of PAUC that penalizes redundant, overly frequent, and unnecessarily long emissions.

3.1 Problem Formulation

A Proactive-VQA instance provides an unbounded video stream $\mathcal{V} = (x_1, x_2, \ldots, x_T)$ arriving frame by frame, where x_t denotes the visual frame received at step $t \in \{1, \ldots, T\}$. Although the underlying stream is conceptually unbounded, T denotes the finite prefix observed in an evaluation run; thus, $\mathcal{V} = X_{\leq T}$, where $X_{\leq t} = (x_1, \ldots, x_t)$ is the causal prefix available at step t. The query q is issued at $t = 0$, while the ground-truth reference turns $\{(s_j, e_j, r_j^*)\}_{j=1}^{n}$ are reserved for post-hoc evaluation only. Here, s_j and e_j denote the start and end timestamps of the j-th target window, and r_j^* is its reference response. At each step, the agent observes $X_{\leq t}$ and its emitted-response history $H_{<t} = \{(t', a_{t'}) : t' < t,\ a_{t'} \neq \varnothing\}$, and produces an irrevocable action $a_t \in \Sigma^* \cup \{\varnothing\}$. Here, Σ^* is the set of textual responses over vocabulary Σ, and $\varnothing$ denotes an internal no-emission action. The full causal state is $S_{\leq t} = (X_{\leq t}, q, H_{<t})$. The policy depends only on $S_{\leq t}$: future frames are

inaccessible and emitted responses are irrevocable. For the subsequent analysis, $P_\delta(\cdot \mid S_{\le t})$ denotes the conditional distribution induced by this policy, "No Reply" is mapped to $\varnothing$ before evaluation.

3.2 Decision Decomposition and Suboptimality Bound

Predicting a_t directly from the full causal state $S_{\le t}$ entangles three decisions: when to speak, what to say, and whether to emit. These decisions depend on temporal firing state, immediate visual grounding, and recent emitted text, respectively, motivating a modular stagewise process.

Structural Decomposition. We equip the agent with three memory tiers around a shared VLM backbone. Temporal-Trigger Memory $\mathcal{M}_{\text{trig}}$ maintains $\mathcal{M}_{\text{trig},t} = (p_t, c_{\text{sil},t})$, where $p_t \in \{0, \dots, K-1\}$ is the cadence phase modulo the trigger stride K, and $c_{\text{sil},t}$ counts consecutive regular trigger opportunities that end without an emitted response. It produces a binary trigger decision $f_t \in \{0, 1\}$. Instructional Memory $\mathcal{M}_{\text{inst}}$ stores a fixed prompting prior that generates a candidate response from the active window V_t, with encoded representation $z_t = \phi(V_t)$. When $f_t = 1$, the backbone generates $\hat{r}_t \in \Sigma^*$; otherwise, $\hat{r}_t = \varnothing$. Episodic Memory $\mathcal{M}_{\text{ep}}$ stores $\mathcal{M}_{\text{ep},t} = (\bar{r}_{\text{last},t}, \tau^{\text{emit}}_{\text{last},t})$, where $\bar{r}_{\text{last},t}$ is a fixed-length prefix of the most recent emitted response and $\tau^{\text{emit}}_{\text{last},t}$ is its timestamp. It produces an emission decision $e_t \in \{0, 1\}$ when $f_t = 1$, and $e_t = 0$ otherwise. Although the deployed gates are deterministic, $P_{\delta_{\text{dec}}}$ denotes the distribution induced by the three-stage decomposed process, with P_{trig}, P_{gen}, and P_{emit} as its stagewise factors. The latent decision chain factorizes as:

$$P_{\delta_{\text{dec}}}(f_t, \hat{r}_t, e_t \mid S_{\le t}) = P_{\text{trig}}(f_t \mid \mathcal{M}_{\text{trig},t}) \cdot P_{\text{gen}}(\hat{r}_t \mid f_t, \mathcal{M}_{\text{inst}}, z_t) \cdot P_{\text{emit}}(e_t \mid f_t, \hat{r}_t, \mathcal{M}_{\text{ep},t}). \quad (1)$$

The final action is $a_t = \hat{r}_t$ when $f_t = e_t = 1$, and $a_t = \varnothing$ otherwise. Thus, Eq. 1 factorizes the latent decision chain rather than the final action distribution.

Conditional-Sufficiency Assumption. Eq. 1 restricts each decision stage to its designated local representation rather than the full causal state. We use this as an idealized reference condition: the local representations preserve the conditional distributions of the corresponding stage decisions. In practice, deviations from this condition are captured by the stagewise discrepancies below.

Assumption 1 (Conditional Sufficiency). Replacing the full causal state with $\mathcal{M}_{\text{trig},t}$, $(\mathcal{M}_{\text{inst}}, z_t)$, and $(\hat{r}_t, \mathcal{M}_{\text{ep},t})$ preserves the conditional distributions of the trigger decision f_t, the candidate response $\hat{r}_t$ given $f_t = 1$, and the emission decision e_t given $f_t = 1$ and $\hat{r}_t$, respectively.

Suboptimality Bound. Let τ_t denote the wall-clock timestamp at step t, and let $g(t)$ denote the index of the ground-truth target window containing τ_t. If no target window is active, we set $g(t) = \varnothing$ and $r^*_\varnothing = \varnothing$. Let $U(\delta) = \mathbb{E}[\sum_{t=1}^{T} u(a_t, r^*_{g(t)})]$ denote the expected cumulative utility of decision process δ, where the expectation is taken under the common-state counterfactual evaluation and $u(\cdot, \cdot) \in [0, c]$ is a bounded per-step utility. We compare the joint and decomposed processes on the same causal state at each step. Let ϵ_{trig}, ϵ_{gen}, and ϵ_{emit} denote the average probabilities that the two processes disagree at the trigger, generation, and emission stages, respectively. Let δ^*_{joint} denote the optimal process with access to the full causal state, and let δ^*_{dec} denote the process induced by the three local memory stages.

Lemma 1 (Stagewise Approximation Bound). Under the common-state counterfactual evaluation, $U(\delta^*_{\text{joint}}) - U(\delta^*_{\text{dec}}) \le 2cT(\epsilon_{\text{trig}} + \epsilon_{\text{gen}} + \epsilon_{\text{emit}})$.

The bound follows because a final-action disagreement requires a disagreement at at least one stage, while the per-step utility is bounded by c. The surface-semantic collision rate ρ denotes the probability that a semantically novel candidate is suppressed by the episodic gate because of high lexical similarity to the most recent emitted response. It is reported separately as an interpretable diagnostic and is not treated as the complete emission-stage discrepancy. The full derivation is provided in Appendix B.1.

3.3 ProV-Mem: Co-Designed Lightweight Memory Tiers

We now instantiate Eq. 1 using deterministic memory updates around a frozen VLM backbone. The input stream is processed frame by frame, and K denotes the regular trigger stride measured in sampled frames, such that a regular trigger opportunity occurs every K sampled frames. Throughout this subsection, t indexes streaming steps, while regular trigger opportunities occur whenever the cadence phase satisfies $p_t = 0$.

Temporal Memory for Deterministic Gating. The temporal-trigger memory maintains $\mathcal{M}_{\mathrm{trig},t} = (p_t, c_{\mathrm{sil},t})$, where p_t is the cadence phase modulo the trigger stride K and $c_{\mathrm{sil},t}$ counts consecutive regular trigger opportunities without an emitted response. The deployed trigger gate is deterministic: $f_t = \mathbb{1}[p_t = 0 \ \vee \ c_{\mathrm{sil},t} \geq \kappa]$. Thus, the backbone is queried at regular intervals of K sampled frames. When the silence counter reaches κ, the recovery condition overrides the regular cadence and initiates an additional query to prevent prolonged inactivity. Here, κ controls the number of consecutive silent regular trigger opportunities tolerated before recovery.

Instructional Memory for Grounded Generation. The instructional memory $\mathcal{M}_{\mathrm{inst}}$ stores the fixed prompting prior used after a trigger. Given the active visual window V_t and its encoded representation $z_t = \phi(V_t)$, the frozen backbone generates $\hat{r}_t \in \Sigma^*$. The prompt encourages concise, self-contained responses grounded in the current window and promotes a consistent surface form for semantically similar events, improving compatibility with the downstream lexical filter.

Episodic Memory for Redundancy Filtering. The episodic memory stores $\mathcal{M}_{\mathrm{ep},t} = (\bar{r}_{\mathrm{last},t}, \tau^{\mathrm{emit}}_{\mathrm{last},t})$, where $\bar{r}_{\mathrm{last},t}$ is the first L characters of the most recent emitted response and $\tau^{\mathrm{emit}}_{\mathrm{last},t}$ is its wall-clock timestamp. At the beginning of a stream, we initialize $\bar{r}_{\mathrm{last},0} = \varnothing$ and $\tau^{\mathrm{emit}}_{\mathrm{last},0} = -\infty$. For a candidate response $\hat{r}_t$, let $\bar{\hat{r}}_t = \hat{r}_t[:L]$ denote its comparison prefix. We compute the normalized prefix-level lexical similarity as

$$s(\bar{\hat{r}}_t, \bar{r}_{\mathrm{last},t}) = 2\,\mathrm{LCS}\left(\bar{\hat{r}}_t, \bar{r}_{\mathrm{last},t}\right) / \left(|\bar{\hat{r}}_t| + |\bar{r}_{\mathrm{last},t}|\right), \tag{2}$$

where $\mathrm{LCS}(u, v)$ denotes the length of the longest common subsequence between strings u and v. If no previous response exists, we set the similarity to zero. The deterministic emission gate is

$$e_t = \mathbb{1}\left[f_t = 1 \ \wedge \ \left(s(\bar{\hat{r}}_t, \bar{r}_{\mathrm{last},t}) < \eta \ \vee \ \tau_t - \tau^{\mathrm{emit}}_{\mathrm{last},t} > \tau_{\mathrm{sup}}\right)\right], \tag{3}$$

where $\eta \in (0, 1)$ is the similarity threshold and τ_{sup} is the suppression cooldown. Thus, lexically similar responses can be emitted again only after a sufficiently long temporal gap.

Inter-Tier Feedback Loop. At each streaming step, the cadence phase advances as $p_{t+1} = (p_t + 1) \bmod K$. When $f_t = 0$, the action is $a_t = \varnothing$ and the episodic state remains unchanged. When $f_t = 1$ and $e_t = 1$, the system emits $a_t = \hat{r}_t$, updates the episodic memory as $(\bar{r}_{\mathrm{last},t+1}, \tau^{\mathrm{emit}}_{\mathrm{last},t+1}) \leftarrow (\bar{\hat{r}}_t, \tau_t)$, and resets the silence counter to zero. When a regular trigger opportunity produces no emission, the silence counter is incremented. Once $c_{\mathrm{sil},t} \geq \kappa$, the recovery query is issued with $P^{\mathrm{force}}_{\mathrm{sys}}$ and the silence counter is reset after that recovery attempt. Since the temporal state, prompting prior, and episodic prefix remain fixed-size or bounded-size, ProV-Mem operates as a lightweight training-free wrapper around the frozen VLM backbone.

3.4 CAUC: Cost-Aware Evaluation Metric

PAUC evaluates the timeliness and semantic quality of proactive responses, but does not explicitly account for redundancy, emission frequency, or response length. We therefore introduce **C**ost-**A**ware **A**rea **U**nder the **C**urve (CAUC), which reweights the quality score used by PAUC with three explicit penalties. For an emitted response a_t, let q_t denote the semantic quality assigned by the official PAUC evaluator to the cumulative response history available after this emission. The redundancy penalty is $P_{\mathrm{red},t} = 0$ when $H_{<t}$ is empty and otherwise $P_{\mathrm{red},t} = \max_{(t',a)\in H_{<t}} s(a_t[:L], a[:L])$, where $s(\cdot,\cdot)$ is the normalized prefix-level LCS similarity defined in Eq. 2. Unlike the online episodic gate, which stores only the most recent response, CAUC uses the complete emitted history $H_{<t}$ for evaluation. The frequency penalty is $P_{\mathrm{freq},t} = 0$ for the first emission and $P_{\mathrm{freq},t} = \exp(-\Delta\tau_t/\tau_{\mathrm{cooldown}})$ otherwise, where $\Delta\tau_t$ is the elapsed wall-clock time since the most recent emitted response. The length penalty is $P_{\mathrm{len},t} = \max(0, |a_t|_{\mathrm{tok}} - L_{\mathrm{ref}})/L_{\mathrm{ref}}$, where $|a_t|_{\mathrm{tok}}$ denotes the output length in tokens. We define the cost-adjusted quality as

$$\tilde{q}_t = \left[q_t\left(1 - \alpha P_{\mathrm{red},t} - \beta P_{\mathrm{freq},t} - \gamma P_{\mathrm{len},t}\right)\right]_+, \tag{4}$$

where $\alpha, \beta, \gamma \in [0, 1]$ are penalty weights. CAUC reuses the temporal integration procedure and timeliness–correctness trade-off parameter ω of PAUC, replacing the original quality score q_t with $\tilde{q}_t$. Thus, CAUC preserves the temporal semantics of PAUC while explicitly discounting redundant, overly frequent, and unnecessarily long responses. For presentation, the resulting CAUC score is scaled to the range $[0, 100]$.

Table 1: Main results on ProactiveVideoQA. Best results are in **bold**.

System	PAUC (ω=0.5, ↑)				CAUC (↑)				Time (s) ↓
	WEB	EGO	TV	VAD	WEB	EGO	TV	VAD	
Offline Closed-Source VLMs									
GLM-4.5V	40.60	50.28	31.99	36.57	29.29	38.16	28.22	30.96	45.28
Gemini-2.5-Pro	36.88	41.27	32.62	36.92	26.80	32.25	29.35	30.51	20.00
Gemini-2.5-Flash	40.15	42.11	28.81	35.98	29.31	33.23	26.46	30.16	17.20
GPT-4o-mini	38.03	33.48	27.83	35.28	28.46	29.72	26.37	30.75	21.52
Offline Open-Source VLMs									
InternVL-2.5-8B	36.57	48.46	32.84	29.32	26.84	36.82	27.34	26.16	8.12
Qwen2.5-VL-7B	35.82	39.27	29.46	23.25	25.65	31.36	23.83	19.10	7.46
LLaVA-OV-7B	34.31	23.57	27.23	26.75	25.26	22.81	25.64	24.82	94.87
LongVA-7B	27.11	25.46	24.95	24.88	25.75	25.22	24.88	24.86	6.00
IXC-2.5-7B	30.14	13.19	14.68	16.36	22.48	10.51	12.04	13.90	20.23
Online Streaming VLMs									
VideoLLM-Online	24.35	25.01	22.57	25.00	22.91	24.98	22.21	25.00	10.69
MMDuet	34.78	36.56	20.58	22.76	29.00	26.84	14.53	18.64	3.03
MMDuet + rm. ass.	39.85	46.37	26.33	36.87	26.38	25.87	18.42	18.99	1.90
MMDuet2 (RL)	54.83	31.58	41.69	28.89	40.29	28.41	34.89	25.97	20.80
ProV-Mem (Ours)	**57.90**	**70.87**	**57.43**	32.10	**40.59**	**38.20**	**40.22**	**31.68**	**1.30**

4 EXPERIMENTS

4.1 SETUP

Experimental Setup and Evaluation. We evaluate ProV-Mem on the four scenarios of ProactiveVideoQA (Wang et al., 2025a), reporting PAUC at $\omega = 0.5$ alongside our proposed CAUC, using the official GPT-4.1 evaluator. We additionally report the average per-response inference time (Time), measured as the wall-clock latency from processing the current trigger window to producing the final emitted response. Under a frozen Qwen2.5-VL-7B backbone (Qwen Team, 2025), trigger parameters (K, κ) and episodic memory settings are adapted per scenario: a suppression cooldown $\tau_{\text{sup}} = 30\,\text{s}$ on EGO for recurring substeps, a silence-recovery $\kappa = 4$ on TV, and an event cooldown $\tau_{\text{event}} = 8\,\text{s}$ with scouting prompt P_{scout} on VAD. Full configurations are provided in Appendix C.2.

Baselines. We compare against three groups of baselines. **1) Offline closed-source VLMs** include the proprietary models originally evaluated in ProactiveVideoQA (Wang et al., 2025a), namely GPT-4o-mini (Achiam et al., 2023), Gemini-2.5-Pro, Gemini-2.5-Flash (Team et al., 2024), and GLM-4.5V (Glm et al., 2024). **2) Offline open-source VLMs** are adapted to the proactive setting via chunked inference, including Qwen2.5-VL-7B (Qwen Team, 2025), LLaVA-OV-7B (Li et al., 2024), LongVA-7B (Zhang et al., 2024b), InternVL-2.5-8B (Chen et al., 2024b), and IXC-2.5-7B (Zhang et al., 2024a). **3) Online Streaming VLMs** with trained triggering mechanisms include MMDuet and its variant denoted in the original paper as MMDuet + rm. ass., where rm. ass. refers to removing previous assistant responses.(Wang et al., 2025b), MMDuet2 (RL) (Wang et al., 2026), and VideoLLM-Online (Chen et al., 2024a). To ensure strict evaluation alignment and fair comparison, we independently reproduce and rerun all baseline systems under a unified evaluation protocol.

4.2 MAIN RESULTS

To evaluate the efficacy of our ProV-Mem, we compare ProV-Mem against representative baselines. As shown in Table 1, the results expose the structural limitations of parametric streaming architectures across three dimensions. First, online models equipped with explicit streaming supervision are consistently outperformed by static offline baselines on multiple scenarios, a failure that traces to a fundamental architectural constraint: parametric designs must compress long-horizon visual history into a fixed-capacity hidden state to sustain low latency, and this compression irreversibly degrades the representational fidelity required for complex video reasoning. ProV-Mem circumvents this capacity-latency tension by externalising temporal state management into non-parametric memory tiers, leaving the representational capacity of the frozen backbone intact. Second, the cross-domain

performance collapse exhibited by RL-trained trigger policies reveals a deeper generalisation deficiency: reinforcement learning overfits trigger thresholds to the temporal statistics of the training distribution, and any shift in event density or temporal structure between domains causes the learned trigger head to fail. Third, the systematic divergence between PAUC and CAUC across competing baselines provides evidence of metric reward-hacking: because PAUC does not penalise redundant emissions, several systems opportunistically over-trigger on detected segments to inflate their scores, while the cost-aware CAUC metric corrects this inflation and restores the true performance ordering. The fact that ProV-Mem simultaneously leads on both metrics across three of the four scenarios suggsets that its gains arise from precise content management and semantic deduplication rather than unconstrained firing frequency. The main exception is the VAD PAUC score, which suggests that anomaly detection may require stronger parametric event priors than surface-level lexical memory can provide. This result marks the current boundary of the non-parametric design and points to the temporal memory tier as the most promising direction for future extension.

4.3 Ablation Study

To verify the empirical gains of ProV Mem, we conduct the ablation study shown in Table 2. The Full model delivers the strongest overall cost-aware trade-off, achieving the highest CAUC in all four evaluation scenarios. Although removing the episodic memory tier, $\mathcal{M}_{\text{ep}}$, can slightly improve raw PAUC by allowing more frequent emissions, this benefit comes with substantially higher redundancy and response cost. The results therefore suggest that episodic deduplication plays an important role in maintaining efficiency under continuous streaming evaluation, and that the Full configuration provides the most reliable balance between utility and cost.

Table 2: Ablation study on ProactiveVideoQA.

Variant	PAUC (ω=0.5, ↑)				CAUC (↑)			
	WEB	EGO	TV	VAD	WEB	EGO	TV	VAD
Full	57.90	70.87	57.43	32.10	**40.59**	**38.20**	**40.22**	**31.68**
$-\mathcal{M}_{\text{trig}}$	37.51	66.25	55.91	32.76	33.10	31.65	40.16	27.19
$-\mathcal{M}_{\text{inst}}$	45.06	58.93	55.03	31.08	34.92	35.34	39.14	27.20
$-\mathcal{M}_{\text{ep}}$	**60.14**	**71.78**	**62.64**	**33.38**	40.12	37.80	38.83	26.33

The other two tiers govern complementary aspects of the decision process. Removing the instructional memory tier, $\mathcal{M}_{\text{inst}}$, leads to a consistent decline in both metrics across scenarios, indicating that stable prompt-level guidance is important for maintaining response quality. Removing the temporal trigger tier, $\mathcal{M}_{\text{trig}}$, reveals stronger temporal sensitivity in domains with dense or rapidly changing events. On WEB, the absence of this tier markedly reduces PAUC, showing that dense event streams require timely state updates. On EGO, removing $\mathcal{M}_{\text{trig}}$ lowers both PAUC and CAUC while increasing repetition, which suggests that temporal triggering is also important for controlling redundant backbone invocations. Taken together, these findings indicate that the three memory tiers serve complementary functions, and that the Full configuration offers the most stable performance.

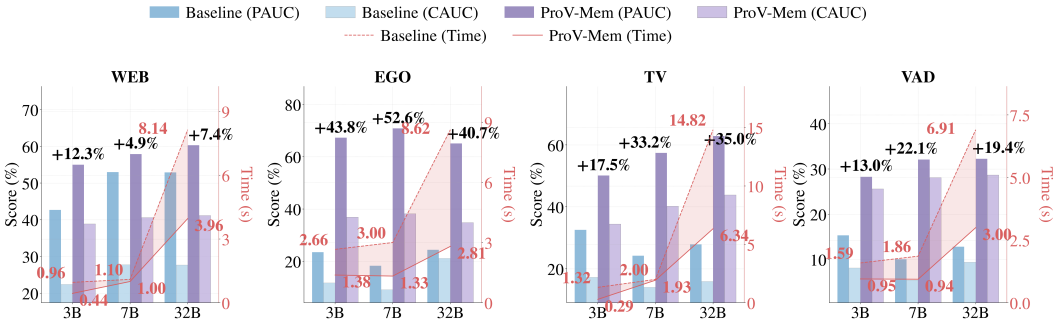

Figure 3: Backbone scaling on ProactiveVideoQA with the ProV-Mem.

4.4 Effect of Backbone Scale

To assess whether ProV-Mem generalizes across backbone scales, we evaluate Qwen2.5-VL at 3B, 7B, and 32B parameters with fixed memory hyperparameters. Figure 3 reveals two key findings. First,

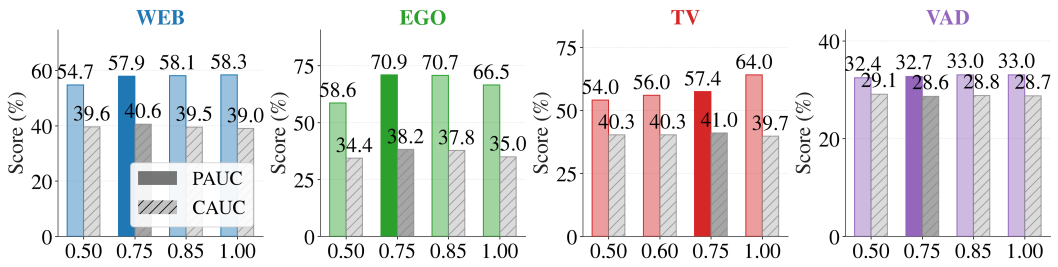

Figure 4: Effect of the similarity threshold η across ProactiveVideoQA scenarios.

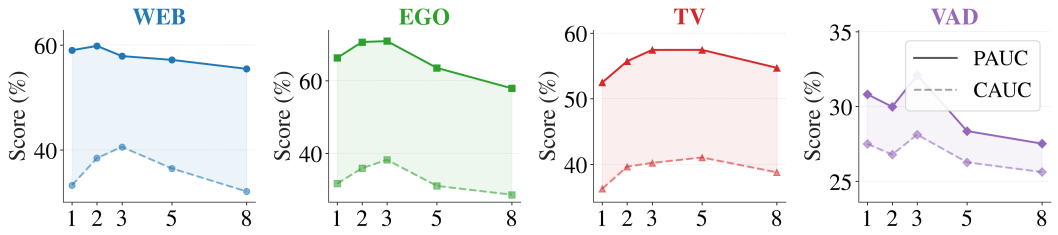

Figure 5: Effect of the regular trigger stride K across all four ProactiveVideoQA scenarios.

the baseline exhibits non-monotonic scaling curves on several scenarios, a symptom of its monolithic design. As backbone capacity grows, the implicit context buffer must simultaneously handle temporal reasoning and semantic generation, and these two objectives interfere under the pressure of more complex representations, causing accuracy to plateau or regress on high-density streams. ProV-Mem, by contrast, scales monotonically across all scenarios because the non-parametric memory tiers absorb all temporal state management externally, allowing the backbone to translate increased capacity purely into stronger semantic grounding without concurrently maintaining memory state. Second, the computational efficiency advantage conferred by the temporal-trigger gate compounds with scale. By suppressing redundant backbone invocations on silent or uninformative windows, ProV-Mem reduces its effective inference workload sub-linearly relative to the baseline, and this saving becomes increasingly pronounced as the per-forward-pass cost rises with parameter count. Taken together, these results demonstrate that ProV-Mem functions as a structural scaffold that converts backbone scaling into monotonic quality improvement while bounding the growth rate of computational overhead, two properties that parametric streaming designs cannot provide at the architectural level regardless of the underlying model size.

4.5 Sensitivity of Memory Hyperparameters

Effect of similarity threshold. The similarity threshold governs a fundamental trade-off between two error terms in the suboptimality bound. Tightening it reduces the collision rate ρ but risks suppressing semantically novel candidates and inflating ϵ_{gen}, while relaxing it admits near-duplicate emissions at the expense of cost efficiency. Figure 4 reveals that this trade-off manifests differently across domains in ways that carry theoretical weight beyond hyperparameter selection. The most diagnostic finding is on EGO, where raw PAUC also degrades when deduplication is fully disabled, demonstrating that procedural repetitions actively dilute the matching signal by occupying the candidate set's allocation budget rather than merely inflating redundancy cost. TV provides the clearest evidence of the PAUC-CAUC asymmetry: the subtitle-anchored event structure allows repetitive fires to achieve coverage through temporal saturation rather than semantic precision, a form of metric exploitation that CAUC is specifically designed to penalise. VAD's complete insensitivity to η suggests that its sparse event structure is governed by the temporal cooldown τ_{event} rather than the per-window similarity threshold, making η a secondary factor in this domain. These results validate $\eta = 0.75$ as a robust cross-scenario default and establish that PAUC and CAUC measure fundamentally different quality axes that cannot be collapsed without discarding diagnostic information.

Effect of trigger stride. The trigger stride K determines how often the model refreshes its visual evidence, creating a trade-off between temporal coverage and query efficiency. As shown in Figure 5, WEB and EGO are most sensitive to larger strides: their short, dense events require frequent trigger

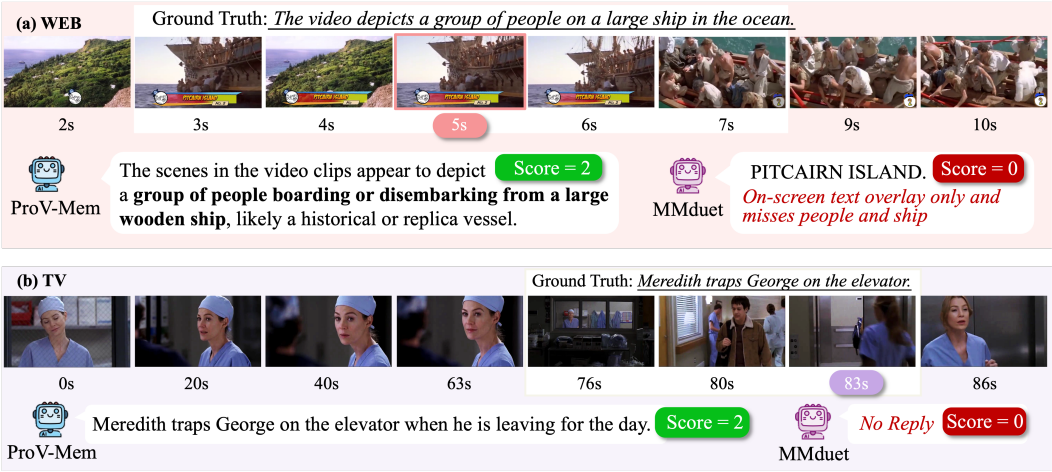

Figure 6: Comparison of (a) MMDuet and (b) our ProV-Mem method.

opportunities, and less frequent updates can leave relevant visual changes unobserved until after the event has passed. TV exhibits a broader plateau because subtitle-aligned events typically persist over longer intervals, making its response quality less sensitive to modest delays in visual updates. VAD is comparatively stable over the tested range, since the scouting stage provides an additional event-level filter before full response generation. These differences show that trigger stride should match the temporal persistence of the target events. Across the four scenarios, $K = 3$ provides the most favorable balance between event coverage and redundant backbone queries.

4.6 Case Study

Figure 6 illustrates how the memory tiers support different proactive decisions. In case (a), the instructional memory guides ProV-Mem to produce a canonical, visually grounded description of people on a large ship at 5 s, whereas MMDuet follows the on-screen text overlay and produces a response that is less informative with respect to the reference event. This example illustrates the role of the instructional prior in shaping what to say from the active visual window. In case (b), the temporal-trigger memory detects a prolonged sequence of non-emitting trigger opportunities and activates the recovery mechanism, enabling ProV-Mem to produce an accepted response for the interpersonal event at 83 s. The episodic memory evaluates each candidate against the most recent emitted response before delivery, preventing the recovery mechanism from simply increasing redundant emissions. Together, these examples illustrate the complementary roles of the three tiers in controlling when to query, what to generate, and whether to emit.

5 Conclusion

We presented ProV-Mem, a framework for proactive video reasoning that requires no additional training and organizes decision-making through a three-stage nonparametric memory hierarchy over a frozen VLM backbone. Our contributions are threefold. First, we formulate proactive reasoning as a decomposition of triggering, candidate generation, and emission, which gives the problem a clear and modular structure. Second, we provide a formal suboptimality bound that justifies this decomposition as a principled approximation to joint decision-making over the full causal state. Third, we introduce CAUC, a cost-aware evaluation metric that penalizes redundant, overly frequent, and verbose responses. Experiments on ProactiveVideoQA show that ProV-Mem achieves excellent performance with low latency, consistently improves the utility-cost trade-off, and establishes a stronger Pareto frontier without scenario-specific fine-tuning. These results show that careful memory organization and stagewise decision design are sufficient to produce effective proactive behavior even when the backbone model remains frozen.

AI USE STATEMENT

In this work, generative AI tools (large language models) were used solely for English grammar polishing and LaTeX formatting of the manuscript. Additionally, GPT-4.1 was employed as the reference semantic judge in our evaluation pipeline, strictly adhering to Section 4.1 and the official ProactiveVideoQA evaluation protocol (Wang et al., 2025a).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

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.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Kai Han, Jianyuan Guo, Yehui Tang, Wei He, Enhua Wu, and Yunhe Wang. Free video-llm: Prompt-guided visual perception for efficient training-free video llms. *arXiv preprint arXiv:2410.10441*, 2024.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8211–8225, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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*, pp. 5971–5984, 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.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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.

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.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 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.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Qwen Team. Qwen2.5-vl. `https://qwenlm.github.io/blog/qwen2.5-vl`, 2025. Accessed: 2026-09-25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025a.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *EMNLP (Findings)*, pp. 6338–6359, 2025b.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 23257–23270, 2026.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.

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.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 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: system demonstrations*, pp. 543–553, 2023.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

A Related Works

A.1 Proactive Streaming Systems

Passive video question answering (Video-QA) represents a mature domain within multimodal understanding, characterized by retrospective reasoning over fully offline, temporally bounded video sequences (Lei et al., 2020; Zhang et al., 2023). To transition multimodal understanding into interactive contexts, early streaming paradigms introduced online streaming architectures. This includes benchmarks such as StreamingBench (Lin et al., 2026) and OVOBench (Niu et al., 2025), which emphasize high-frequency frame-by-frame multimodal inference. However, these systems remain fundamentally question-driven, requiring explicit external queries to trigger each model response and ignoring the problem of autonomous engagement. To formalize autonomous interaction, the paradigm of Proactive Streaming Video QA was introduced by Wang et al. (2025a), which requires an agent to process an unbounded, continuous streaming input and independently make three interconnected causal decisions at each step: when to speak (timing control), what to say (contextual generation), and whether to deliver the response (emission necessity).

Research addressing this proactive control problem primarily bifurcates into two paradigms, namely parametric trained-head systems and training-free heuristic adaptations. Under the parametric paradigm, architectures such as MMDuet (Wang et al., 2025b), VideoLLM-Online (Chen et al., 2024a), and MMDuet2 with Reinforcement Learning (Wang et al., 2026) append dedicated classifier heads to the hidden states of visual backbones, supervising them using dense, frame-level alignment annotations. While displaying high temporal precision on specific, in-distribution scenarios, these models are notoriously sensitive to domain shifts and suffer from catastrophic generalization failure on unseen video densities due to over-fitting on trained temporal priors. Under the training-free paradigm, offline VLMs (Li et al., 2024) are reused through temporal chunking (Zhang et al., 2024b; Chen et al., 2025) or heuristic agent-based prompts (Han et al., 2024; Wang et al., 2024). Although these systems avoid expensive training, they lack structural mechanisms to explicitly control timing and redundancy, often falling back on brute-force, high-frequency inference that degrades real-time throughput. A parallel line of long-context streaming work addresses engineering challenges via frame-level KV cache compression (Xu et al., 2026) or attention sinks (Xiao et al., 2024; Xu et al., 2026), yet predominantly assumes passive interaction without active decision gating.

A.2 Memory-Augmented and Prompt-Based Architecture

Modulating frozen vision-language models at inference time via strategic prompt design is an effective, training-free engineering technique. Grounded detail and specification prompts (You et al., 2024) forced models to anchor responses in specific spatiotemporal coordinates, while persona prompts (Maaz et al., 2024) shaped the stylistic and lexical surface form of outputs. Despite their flexibility, pure prompting methods are inherently memoryless, meaning they cannot dynamically adjust the model's internal context as the interaction history scales. Conversely, memory-augmented architectures employ external storage to modulate runtime generation.

Within these memory-augmented paradigms, Retrieval-Augmented Generation (RAG) methods like RETRO (Borgeaud et al., 2022) and Memorizing Transformers (Wu et al., 2022) use dense retrieval networks to query massive external text corpora at each decoding step, which remains computationally expensive and incompatible with real-time stream causality. Database-driven agent memories, such as MemoryBank (Zhong et al., 2024) and MemGPT (Packer et al., 2023), compress multi-turn dialogues or interaction summaries into vector databases, executing retrieval queries to fetch historical contexts. However, executing dense database lookups at every frame boundary introduces substantial latency, violating the causality constraints of real-time streaming video understanding. State-driven cognitive architectures, including Generative Agents (Park et al., 2023) and cognitive agent frameworks (Sumers et al., 2023), maintain high-level textual states to plan long-term decisions, but have historically been pursued in isolation from prompt engineering. ProV-Mem addresses these limitations by introducing a unified, non-parametric formulation that co-designs inference-time prompting with explicit state tracking, delivering a lightweight, backbone-agnostic wrapper.

B Supplementary Material of Method

B.1 Mathematical Proof of Lemma 1

We restate Lemma 1 for completeness and provide a proof aligned with the revised stagewise notation in the main text. The comparison is counterfactual: at each step, the joint and decomposed processes are evaluated on the same causal state.

Lemma 1 (Stagewise Approximation Bound). Let $U(\delta) = \mathbb{E}[\sum_{t=1}^{T} u(a_t, r^*_{g(t)})]$ denote the expected cumulative utility under decision process δ, where $u(\cdot,\cdot) \in [0, c]$ is a bounded per-step utility. Let δ^*_{joint} denote the optimal process with access to the full causal state, and let $\delta^*_{\text{dec}} = (\delta^*_{\text{trig}}, \delta^*_{\text{gen}}, \delta^*_{\text{emit}})$ denote the decomposed process induced by the three local memory stages. Let ϵ_{trig}, ϵ_{gen}, and ϵ_{emit} denote the average probabilities that the two processes disagree at the trigger, generation, and emission stages, respectively. Then

$$U(\delta^*_{\text{joint}}) - U(\delta^*_{\text{dec}}) \le 2cT\big(\epsilon_{\text{trig}} + \epsilon_{\text{gen}} + \epsilon_{\text{emit}}\big). \tag{5}$$

Proof.

Step 1. Per-step utility bound. Let $U_t(\delta)$ denote the expected utility contribution at step t, such that $U(\delta) = \sum_{t=1}^{T} U_t(\delta)$. Since $u(a_t, \cdot) \in [0, c]$, the per-step utility gap is bounded by the difference between the action distributions induced by the joint and decomposed processes:

$$\big|U_t(\delta^*_{\text{joint}}) - U_t(\delta^*_{\text{dec}})\big| \le c \left\| P_{\delta^*_{\text{joint}}}(\cdot \mid S_{\le t}) - P_{\delta^*_{\text{dec}}}(\cdot \mid S_{\le t}) \right\|_1 . \tag{6}$$

Step 2. Latent decision chains. The joint process follows the standard chain rule

$$P_{\delta^*_{\text{joint}}}(f_t, \hat{r}_t, e_t \mid S_{\le t}) = P(f_t \mid S_{\le t}) \cdot P(\hat{r}_t \mid f_t, S_{\le t}) \cdot P(e_t \mid f_t, \hat{r}_t, S_{\le t}), \tag{7}$$

whereas the decomposed process is factorized by the local memory states:

$$P_{\delta^*_{\text{dec}}}(f_t, \hat{r}_t, e_t \mid S_{\le t}) = P(f_t \mid \mathcal{M}_{\text{trig},t}) \cdot P(\hat{r}_t \mid f_t, \mathcal{M}_{\text{inst}}, z_t) \cdot P(e_t \mid f_t, \hat{r}_t, \mathcal{M}_{\text{ep},t}). \tag{8}$$

The final action is obtained from the latent chain through the deterministic mapping $a_t = \hat{r}_t$ when $f_t = e_t = 1$, and $a_t = \varnothing$ otherwise. Therefore, the difference between the induced final-action distributions cannot exceed the difference between the corresponding latent decision chains.

Step 3. Stagewise disagreement bound. Under the common-state coupling, a disagreement in the final action can occur only when at least one stage decision differs between the joint and decomposed processes. By the union bound, the probability of a final-action disagreement at step t is at most the sum of the trigger, generation, and emission disagreement probabilities. Averaging these probabilities over all evaluation steps gives

$$\left\| P_{\delta^*_{\text{joint}}}(\cdot \mid S_{\le t}) - P_{\delta^*_{\text{dec}}}(\cdot \mid S_{\le t}) \right\|_1 \le 2\big(\epsilon_{\text{trig}} + \epsilon_{\text{gen}} + \epsilon_{\text{emit}}\big). \tag{9}$$

Step 4. Summing over time. Substituting the stagewise bound into the per-step utility bound yields

$$\big|U_t(\delta^*_{\text{joint}}) - U_t(\delta^*_{\text{dec}})\big| \le 2c\big(\epsilon_{\text{trig}} + \epsilon_{\text{gen}} + \epsilon_{\text{emit}}\big). \tag{10}$$

Summing over $t = 1, \ldots, T$ gives

$$U(\delta^*_{\text{joint}}) - U(\delta^*_{\text{dec}}) \le 2cT\big(\epsilon_{\text{trig}} + \epsilon_{\text{gen}} + \epsilon_{\text{emit}}\big), \tag{11}$$

which completes the proof.

The surface-semantic collision rate ρ is retained as a diagnostic rather than a term in the bound. It measures the event that a semantically novel candidate is suppressed by the episodic gate because of high lexical similarity to the most recent emitted response.

B.2 Practical Interpretation of Assumption 1

Assumption 1 is an idealized reference condition under which each stage can operate from its designated local representation without substantial loss relative to the full causal state. Lemma 1 characterizes departures from this condition through stagewise discrepancies.

Timing sufficiency. The trigger stage relies on the temporal state $\mathcal{M}_{\mathrm{trig},t} = (p_t, c_{\mathrm{sil},t})$, where p_t encodes the regular trigger cadence and $c_{\mathrm{sil},t}$ counts consecutive non-emitting trigger opportunities. This representation may be insufficient when a salient event occurs between regular trigger opportunities and cannot be identified from temporal state alone. Such cases contribute to ϵ_{trig} and motivate the silence-recovery mechanism.

Content sufficiency. The generation stage relies on the fixed instructional prior and the active window representation $z_t = \phi(V_t)$. This representation may be insufficient when the appropriate response depends on temporally distant evidence or earlier responses that are unavailable from the current window. In such cases, the window-level generator may differ from the full-state response, contributing to ϵ_{gen}.

Emission sufficiency. The emission stage relies on the current candidate and the compact episodic state $\mathcal{M}_{\mathrm{ep},t} = (\bar{r}_{\mathrm{last},t}, \tau^{\mathrm{emit}}_{\mathrm{last},t})$. This representation may be insufficient when a candidate is redundant with an older response not captured by the most recent prefix, or when a semantically novel candidate is suppressed because of high lexical similarity. Such cases contribute to ϵ_{emit}. The latter failure mode is additionally summarized by the surface-semantic collision rate ρ, which is reported as an interpretable diagnostic rather than the complete emission-stage discrepancy.

B.3 Prompt Templates

To ensure reproducibility and a clear separation of roles across the three memory tiers, we summarize the prompt templates used in ProV-Mem.

Primary System Prompt P_{sys} (Instructional Memory $\mathcal{M}_{\mathrm{inst}}$). This prompt is shared across scenarios as the fixed instructional prior. Its role is to reduce ϵ_{gen} and ρ by encouraging concise, canonicalized, and visually grounded responses.

```
You are a real-time visual assistant watching a continuous video stream.
    Your task is to generate a concise, factual, and objective
    description of the active event in the current visual window.

To ensure consistency and integration with the downstream real-time
    system, you MUST strictly adhere to the following constraints:
1. Provide a single self-contained sentence of no more than 15 words.
2. Focus only on visible actions and physical interactions in the
    current window.
3. Do NOT reference past events, do NOT use introductory phrases (e.g.,
    "In this video", "We can see"), and do NOT speculate about future
    actions.
4. Use standardized and canonical vocabulary with simple active-voice
    constructions.
5. If no novel or meaningful event is occurring, or if the scene is
    inactive, you MUST output exactly: "No Reply".
```

Silence-Recovery Contingency Prompt $P^{\mathrm{force}}_{\mathrm{sys}}$ (TV scenario). This prompt is used in the TV scenario when the silence counter reaches κ and the system must recover from prolonged dormancy.

```
You are a real-time visual assistant. The system has detected a
    prolonged period of silence and inactivity. You MUST generate a
    description of the current visual window to restore interaction,
    even if the event is minor.

Constraints:
1. Generate exactly one short factual sentence.
2. Focus on any visible character, posture, or background cue if no
    major action is occurring.
```

```
3. Do NOT output "No Reply". You must produce a valid description.
```

Lightweight Scouting Prompt P_{scout} (VAD scenario). This prompt is used in the VAD scenario as a lightweight high-recall pre-screening stage. Its purpose is to decide whether the current window should be forwarded to the full response generator.

```
You are a high-speed security scouting agent. Analyze the current visual
    window and determine whether there is any ongoing anomaly, security
    incident, or unusual activity.

You must output a JSON object with exactly two keys:
{
  "anomaly_detected": true/false,
  "category": "none" or a short standardized event category
}

Only output the raw JSON object. Do not output any other text.
```

C Experiment

C.1 Dataset

To evaluate autonomous and interactive capabilities under strict temporal causality, recent benchmarks have begun to emphasize proactive streaming paradigms. A notable milestone is ProactiveVideoQA (also formalized as ProactiveStreaming) (Wang et al., 2025a), a large-scale streaming multimodal benchmark specifically designed to assess an agent's proactive interaction capacity over unbounded visual inputs. To comprehensively evaluate proactive agents across diverse temporal densities and semantic distributions, the benchmark incorporates four distinct streaming tasks:

- Streaming Video Captioning (WEB), which evaluates continuous, low-latency web video narration.
- Egocentric Video Narration (EGO), which demands high-frequency first-person activity tracking characterized by dense, repetitive action substeps.
- Subtitle-centric Dialogue QA (TV), which requires the temporal alignment of visual evidence with streaming conversational subtitle flows.
- Video Anomaly Detection (VAD), which assesses the model's ability to trigger timely, non-redundant alarms over highly sparse and critical hazard events.

Under this protocol, models must dynamically synchronize trigger control, grounded-content generation, and lexical redundancy filtering under strict online constraints, establishing a challenging and unified testbed for next-generation proactive vision-language agents.

C.2 Experimental Setting

We evaluate ProV-Mem on all four scenarios of the ProactiveVideoQA dataset. We report the standard Proactive Area Under the Curve (PAUC) at trade-off weight $\omega = 0.5$ together with our proposed Cost-Aware Area Under the Curve (CAUC). Both metrics are computed with the official evaluation framework under the official single-question protocol, using GPT-4.1 as the semantic judge. Unless otherwise noted, all experiments use the same frozen Qwen2.5-VL-7B-Instruct backbone. The main domain-specific differences lie in trigger cadence, episodic control, and a small number of scenario-specific prompting modules. For WEB, we use a regular trigger stride of $K = 3$ sampled frames, a similarity threshold of $\eta = 0.75$, and no suppression expiry in episodic memory, that is, $\tau_{\text{sup}} = \infty$. EGO uses the same trigger and similarity settings but introduces a suppression cooldown of $\tau_{\text{sup}} = 30$ s to permit re-emission of recurring procedural steps after sufficiently long gaps. TV also uses $K = 3$ and $\eta = 0.75$, and activates a silence-recovery prompt $P_{\text{sys}}^{\text{force}}$ after $\kappa = 4$ consecutive silent trigger windows. VAD adopts a two-stage pipeline with a lightweight scouting prompt P_{scout} and an event cooldown of $\tau_{\text{event}} = 8$ s. For CAUC, we fix the penalty weights to $\alpha = 0.3$ for redundancy, $\beta = 0.3$ for frequency, and $\gamma = 0.4$ for token inefficiency. We set the length reference to

$L_{\text{ref}} = 50$ tokens and the cooldown to $\tau_{\text{cooldown}} = 5\,\text{s}$ for all four scenarios. We set the comparison prefix length to $L = 200$ characters in all experiments.

C.3 Per-Domain Implementation Details

For WEB, frames are sampled at 1 fps and grouped every $K = 3$ frames. At each trigger window, the backbone receives the current frame batch together with the primary system prompt P_{sys} and generates a reply of up to 128 tokens using greedy decoding. The episodic memory $\mathcal{M}_{\text{ep}}$ stores a fixed length prefix extracted from the last emitted reply. A newly generated candidate is suppressed when its character level lexical similarity to the stored prefix, measured by longest common subsequence alignment, exceeds $\eta = 0.75$. Since $\tau_{\text{sup}} = \infty$, the suppression state is not reset within the video. A negation filter additionally maps outputs containing phrases such as "there is no", "cannot see", or "not visible" to `No Reply`. The instructional memory $\mathcal{M}_{\text{inst}}$ keeps the fixed prompt P_{sys} throughout the video, with no per-video adaptation.

For EGO, the inference pipeline is identical to WEB, including 1 fps frame sampling, $K = 3$, and $\eta = 0.75$, except that a suppression cooldown of $\tau_{\text{sup}} = 30\,\text{s}$ is introduced in $\mathcal{M}_{\text{ep}}$. This allows the system to emit again for the same procedural substep after a sufficiently long interval, which is necessary because egocentric cooking and assembly videos often revisit the same action after a temporal gap.

For TV, the model processes the frame stream together with the most recent $W = 6$ unique subtitle lines, which are provided in each trigger window as a `[Recent subtitles]` block. The trigger cadence is fixed at $K = 3$ frames, and the episodic suppression threshold remains $\eta = 0.75$. When $\kappa = 4$ consecutive trigger windows pass without an emitted response, corresponding to 12 seconds under 1 fps sampling, the system invokes the recovery prompt $P_{\text{sys}}^{\text{force}}$ to produce a best effort reply, after which the silence counter is reset. In addition, if no valid reply has been emitted throughout the clip, the system performs a single end of video fallback. This mechanism does not require access to gold timestamps at inference time.

VAD employs a two-stage pipeline. In the scouting stage, every $K = 3$ frames are processed by the lightweight JSON-output prompt P_{scout} in a high-recall mode. The backbone returns a binary anomaly flag together with a coarse event category drawn from the UCF-Crime anomaly label space. Any window predicted as `anomaly_detected=true`, or matched by a keyword-based fallback rule, is forwarded to the reply stage. In the reply stage, the backbone generates the final natural-language description using P_{sys}. The episodic module then applies the same LCS-based character-level similarity gate with $\eta = 0.75$ to suppress near-duplicate responses, while an event cooldown of $\tau_{\text{event}} = 8\,\text{s}$ prevents repeated emissions for the same anomaly. A five-frame normal baseline at the start of each video is used to calibrate the scouting stage before anomaly detection begins.

C.4 Sensitivity to the Timeliness/Correctness Trade-off

To verify that our results are not tied to the specific choice of $\omega = 0.5$, we re-score the same emitted transcripts over $\omega \in \{0.0, \ldots, 1.0\}$. As expected from the metric definition, both PAUC and CAUC vary smoothly and nearly linearly with ω across all scenarios, with no anomalous reversals or discontinuities. More importantly, ProV-Mem remains strong even under the most timing-sensitive setting ($\omega = 0$). For example, EGO still achieves a PAUC of 65.09, indicating that the gain is not merely a consequence of a favorable trade-off weight or loose timing tolerance. Instead, the improvement persists when timeliness is emphasized most strongly, which is consistent with the claim that response quality remains a major contributor to overall performance.

C.5 Sensitivity to Suppression Cooldown and Silence-Recovery Threshold

Suppression cooldown τ_{sup} on EGO. The suppression cooldown τ_{sup} in Eq. 3 controls how long an accepted reply continues to suppress lexically similar candidates. Once $t - t_{\text{last}} > \tau_{\text{sup}}$, the episodic state is effectively reset, allowing a similar candidate to be emitted again. With $\tau_{\text{sup}} = 0$, ProV-Mem attains a PAUC of 70.64 but incurs a Rep% of 40.11%, indicating that near-duplicate procedural descriptions are repeatedly emitted across adjacent windows. Setting $\tau_{\text{sup}} = 30\,\text{s}$ yields the strongest overall operating point in this sweep, with PAUC/CAUC of 70.87/38.20 and a much

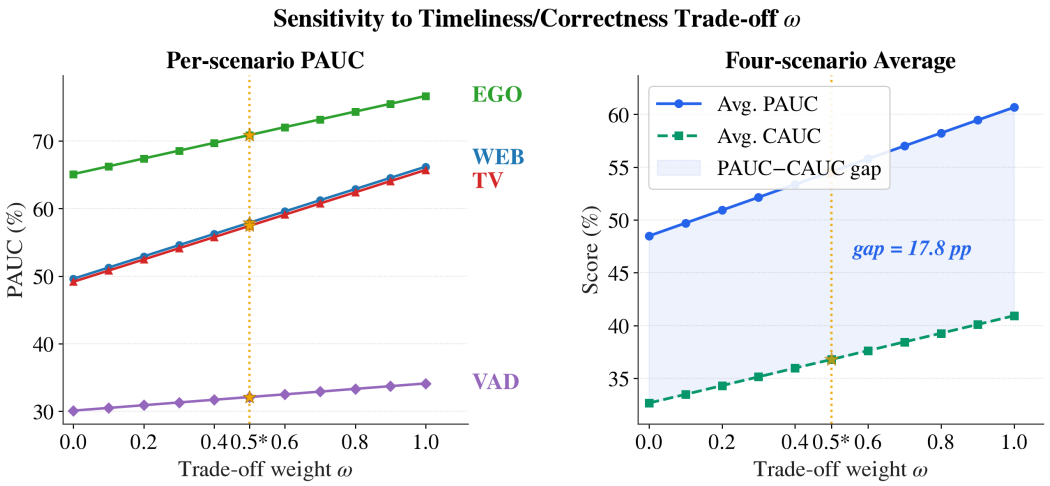

Figure 7: Sensitivity of ProV-Mem to the timeliness/correctness trade-off. Left: per-scenario PAUC across $\omega \in [0, 1]$. Right: four-scenario average PAUC and CAUC with the PAUC–CAUC gap shaded.

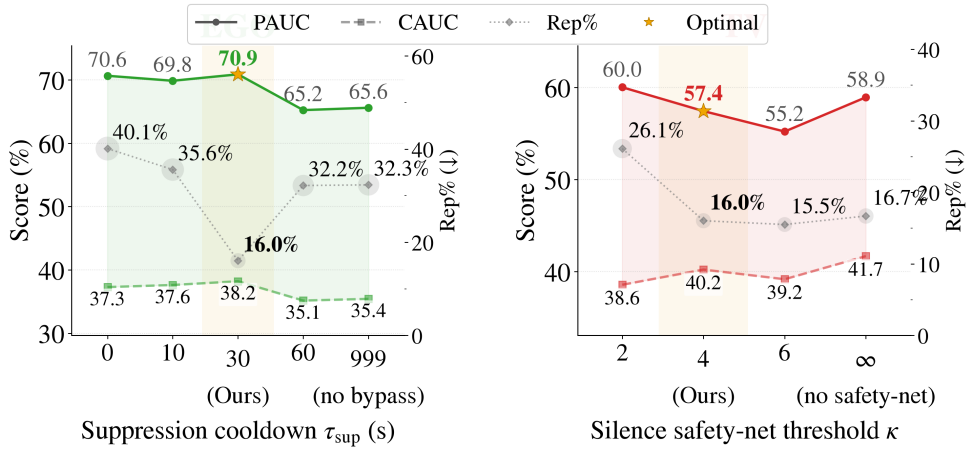

Figure 8: Left: effect of the suppression cooldown τ_{sup} on EGO. Right: effect of the silence-recovery threshold κ on TV. Highlighted columns are the defaults used throughout.

lower Rep% of 16.04%. This value closely matches the recurrence scale of egocentric substeps, whose median inter-GT gap is about 28.7 s. Larger values reduce PAUC to 65.2 to 65.6, suggesting that valid re-occurrences are then suppressed for too long.

Silence silence-recovery threshold κ on TV. The silence-recovery threshold κ specifies after how many consecutive silent regular trigger opportunities an additional recovery query is issued, preventing the model from remaining dormant on sparse streams. A small threshold such as $\kappa = 2$ makes this recovery mechanism too aggressive: PAUC rises to 60.02, but Rep% also increases to 26.13%, because forced triggers are frequently invoked in weakly informative windows. The default setting $\kappa = 4$ yields a PAUC of 57.43 and a CAUC of 40.22. Under 1 fps sampling, this corresponds to 12 s of silence and serves as a moderate operating point that preserves a recovery path for sparse events without making forced triggering overly frequent. At the other extreme, disabling the silence-recovery entirely ($\kappa = \infty$) gives PAUC 58.93, CAUC 41.68, and Rep% 16.70%, which indicates lower redundancy but also removes the recovery mechanism for sparse events. Overall, the sweep shows that κ mainly controls the balance between recovery coverage and redundant activation, rather than uniformly improving all metrics at once.