LatPA: A Training-Free Latent Proactive Adapter for Streaming Video-LLMs

Anonymous authors
Paper under double-blind review

Abstract

Streaming video assistants must decide when to act, not just what to answer. Existing proactive triggers typically obtain this decision from an additional trained module or a yes/no query to the backbone. This leaves open how to obtain a temporal decision directly from the frozen model's representations. To this end, we introduce **LatPA** (**Lat**ent **P**roactive **A**dapter), a training-free alternative that extracts latent signals from a frozen Video-LLM's intermediate hidden states. The key idea is to use the video's own history as the reference for the current observation. LatPA compares the question alignment of the current frame with that of an exponential memory of preceding frames and acts when the resulting score is positive. This self-referenced rule requires neither an additional learned scoring module nor yes/no generation, and uses a shared configuration across backbones and benchmarks. Experiments on StreamingBench-PO show that LatPA improves average timing-and-answer accuracy over the evaluated learned and queried triggers with Qwen2-VL backbones. Across the evaluated model families and scales, LatPA consistently outperforms the evaluated queried triggers on the average scores of StreamingBench-PO and IPIBench. These results demonstrate that a video-specific historical reference enables effective proactive timing directly from frozen latent representations. Code is available at `https://anonymous.4open.science/r/LatPA-D405`.

1 Introduction

Streaming video assistants play an increasingly important role in our daily lives. In scenarios such as embodied robotics, the model must do more than understand a video stream. It also needs to respond to human instructions in time and interact with people. This requires the model to watch what is in front of it in real time and to answer at the most suitable moment. In other words, a streaming video assistant needs to recognize the right moment to act. Multimodal large language models (MLLMs) already provide increasingly capable video understanding (Wang et al., 2024; Li et al., 2024; Zhu et al., 2025; Wang et al., 2025b; Chen et al., 2024b; Bai et al., 2025), supporting a growing range of streaming applications (Qian et al., 2024; Zhang et al., 2024; Liu et al., 2024; Chen et al., 2025; Yao et al., 2025). A proactive assistant takes this interaction further: a user can leave a question pending, and the assistant responds when the relevant event occurs. The assistant must therefore decide whether the current observation warrants a response or whether it should keep watching.

The task works as follows. Before the video starts, the model receives a question. The question may ask about something that only the video can reveal, or about a decision that becomes possible once certain content appears. The model then watches the video with this question in mind. As soon as the video has shown enough to answer, the model should respond and give its answer. We call the component that decides when to act a proactive trigger. It makes this decision while the backbone itself stays frozen.

Existing approaches obtain the trigger through two main routes. **Learned triggers** acquire a decision criterion from proactive training data, using an activation head (Wang et al., 2025a), a specialized token (Chen et al., 2024a; Wang et al., 2025d), or a dedicated decision module (Qian et al., 2025; Fu et al., 2025). These methods depend on high-quality training data, and such data is costly to collect.

Queried triggers elicit a decision from the backbone, for example through a yes/no prompt at each step (Shi et al., 2026; Wang et al., 2025c). These methods make each decision by asking the model

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: **Deployable proactive timing with LatPA. Left:** a frozen Video-LLM decides every second whether to wait or act. LatPA decides by comparing the current frame with the video's own history in the model's latent space. **Right:** a real StreamingBench-PO example. The LatPA score stays below zero before the event and crosses zero at 14 s, where LatPA acts.

directly. Each query also focuses on the current moment, so it cannot make full use of what the model has already seen in a long video.

Both routes described above place the final criterion in a learned component or a token-level response. However, we hope to make use of the MLLM's own strong video understanding. While keeping its generalization ability, we use the more informative internal signals of the model and solve this problem in the latent space. We expect such an approach to capture finer-grained signals.

To this end, we propose **LatPA** (**Lat**ent **P**roactive **A**dapter), a training-free adapter for frozen Video-LLMs. At each poll, LatPA performs one prefill and reads question and current-frame representations from an intermediate layer. It maintains an exponential moving average of preceding frame representations and compares the question's cosine similarity to the current frame with its similarity to this memory (Section 3). The difference defines a self-referenced score with a natural zero threshold. LatPA adds no trainable parameters, generates no yes/no decision, and applies the same read-out depth, memory decay, and threshold across the evaluated backbones and benchmarks. The adapter thus separates proactive triggering from both additional model training and token generation.

Experiments establish the value of this historical reference for training-free proactive timing. On StreamingBench-PO, LatPA improves average timing-and-answer accuracy over the strongest evaluated learned or queried trigger by 18.8 and 8.8 percentage points on Qwen2-VL-2B and 7B, respectively. With one shared operating point, LatPA exceeds both evaluated queried triggers on the average scores of StreamingBench-PO and IPIBench across all four Qwen2-VL and InternVL3 backbones (Table 1). On IPIBench, it also ranks first or second among the evaluated learned and queried triggers. Analysis on a constructed validation set further shows that the latent read-out concentrates responses at target onsets, whereas the yes/no gate acts more often during subsequent target presence (Section 4.3). Memory and decision-rule ablations isolate the benefits of the historical reference and its zero point. OVO-Bench FAR and ProactiveVideoQA characterize how this shared rule interacts with long-horizon response selection and multi-turn answering.

In summary, our main contributions are as follows:

- We identify a signal in the latent space of frozen Video-LLMs. This signal carries a high-dimensional projection of the model's understanding of the scene. Operations such as aggregation extract the model's internal representation of its understanding of a subset of tokens.
- We propose **LatPA**, a training-free adapter that converts this signal into a score for decision making in existing proactive settings. This score captures finer changes in the video frames.
- Further experiments show that our signal generalizes well across backbones of different families and sizes. In-depth analysis also shows that, compared with token-level scores, the LatPA score more easily distinguishes frames that require an action from those that do not.

2 Related work

Proactive assistants. A proactive assistant (Lu et al., 2025) differs from a reactive one in that it must decide when to act, not only what to say, while an input stream continues to arrive; streaming Video-LLMs (Qian et al., 2024; Zhang et al., 2024; Liu et al., 2024; Chen et al., 2025) are its most studied instance. Existing assistants obtain this decision in two ways. Learned triggers train the timing either through streaming interaction training or through an additional decision component (Chen et al., 2024a; Wang et al., 2025d; Fu et al., 2025; Qian et al., 2025; Wang et al., 2025a). Queried triggers instead obtain a decision from an existing model at inference time: some ask the backbone whether the current observations justify a response, in the manner of self-evaluation prompts (Kadavath et al., 2022; Shi et al., 2026; Wang et al., 2025c), while matching-based methods translate the query into visual proposals and act on lightweight embedding matches with an external encoder (Zheng et al., 2026). LatPA differs from both categories: it extracts latent signals directly from the frozen backbone's hidden states, trains no scoring component, and uses the video's own history as the reference for the decision.

Latent communication between LLMs. Latent communication enables LLM agents to exchange information through continuous internal representations, extending latent-space reasoning within one model (Hao et al., 2024) to several. One line of work passes internal states between models during collaboration, through activation-level communication, key–value transfer, shared latent working memory, or compressed hidden-state messages (Ramesh & Li, 2025; Fu et al., 2026; Shkolnikov, 2026; Zou et al., 2026; Du et al., 2026). Another line studies what can be read from hidden representations, using probing classifiers, lens-style read-outs, layer-wise analyses, internal-state probes, representation engineering, and vision-language representation diagnostics (Alain & Bengio, 2016; Belrose et al., 2025; Tenney et al., 2019; Herrmann et al., 2025; Azaria & Mitchell, 2023; Zou et al., 2025; Neo et al., 2025). Together, these works motivate using hidden states as an interface for downstream computation. LatPA applies this idea to a different setting: a parameter-free trigger reads representations from a single frozen backbone to decide when to act. It compares the question relevance of the current frame with that of historical representations, without exchanging messages between agents or fitting a decision threshold.

3 Method

LatPA turns a frozen Video-LLM's intermediate representations into a training-free rule for deciding when to act. The central idea is to evaluate the current observation against the video's own history. As shown in Figure 2, the adapter has three components: a hidden-state read-out that represents the question and current frame, an exponential memory that supplies a temporal reference, and a zero-threshold rule that detects an increase in question alignment. Together, these components provide a content-driven trigger with no additional trainable parameters or decision-token generation.

This design is inspired by Bayesian surprise (Itti & Baldi, 2005), an information-theoretic account of what attracts human attention in videos. Bayesian surprise measures how much a new observation changes what an observer expected from the past. We follow the same idea, but measure the change with respect to the pending question. LatPA measures the increase in the current frame's relevance to the question relative to the content already observed in the video. A positive gap means that the frame brings new question-related content, and LatPA should act. Unlike Bayesian surprise, LatPA keeps no explicit probability model and computes no KL divergence. Instead, the gap is measured between cosine similarities in the model's latent space.

3.1 Problem setup and hidden-state read-out

A video stream arrives at 1 fps as frames $f_1, f_2, \ldots$, and a question q arrives at time t_q and stays pending. At every subsequent second $t > t_q$, called a *poll*, the system decides whether to keep watching or to act now. Two constraints define the setting: the backbone $\mathcal{A}$ stays frozen, and the decision at poll t may use only the frames observed up to t. Benchmarks score first-response timing, repeated responses, or timing together with answer correctness (Section 4).

To make this decision, LatPA reads intermediate states that provide visual and question representations before they are mapped to a next-token distribution. At poll t, LatPA prefills $\mathcal{A}$ once with up

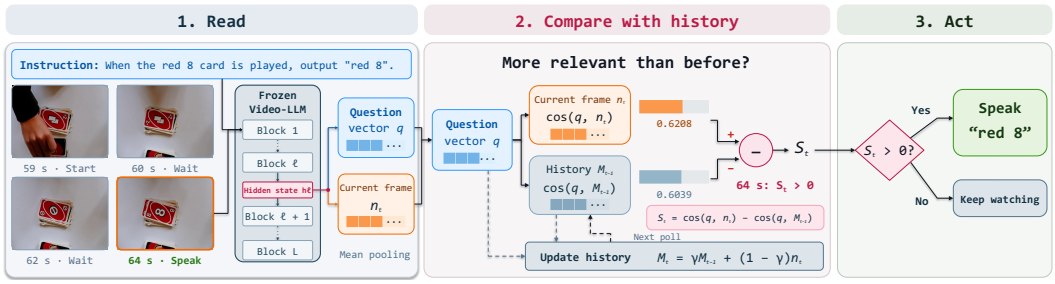

Figure 2: Overview of LatPA. LatPA reads intermediate representations from a frozen Video-LLM and acts when current visual evidence exceeds the running historical reference.

to W recent frames followed by the question and generates no token. Let $H_\ell \in \mathbb{R}^{N\times d}$ denote the hidden states at layer ℓ, where N is the sequence length and d the hidden size, and let L be the number of transformer layers. We mean-pool H_ℓ into a question-side representation q_t and a latest-visual representation n_t:

$$q_t = \mathrm{Pool}(H_\ell, \mathcal{I}_q), \qquad n_t = \mathrm{Pool}(H_\ell, \mathcal{I}_{\mathrm{new}}), \tag{1}$$

Here $\mathcal{I}_q$ contains the tokens after the last visual token, including the question and template suffix; $\mathcal{I}_{\mathrm{new}}$ selects the trailing visual tokens used to represent the latest frame. Appendix A specifies the token selection; $q_t, n_t \in \mathbb{R}^d$. We fix the read-out layer at $\ell = 0.5L$. At this depth the representations are expected to have completed cross-modal fusion while not yet being shaped by the next-token geometry of the output layers (Belrose et al., 2025; Tenney et al., 2019; Voita et al., 2019; Skean et al., 2025; Zhang et al., 2025; Neo et al., 2025); Appendix D.2 reports the validation criterion and benchmark depth sweep. These latent signals support temporal comparison across backbones; Appendix D.1 compares their temporal behavior with token-level read-outs.

3.2 Exponential memory of past frames

Relevant content can remain visible after an event begins, so current-frame alignment alone does not distinguish its arrival from its continued presence. LatPA makes this distinction explicit by comparing the current frame with a running summary of the visual history. The summary is an exponential moving average (EMA) $M_t \in \mathbb{R}^d$ of the frame vectors, updated after each poll as

$$M_t = \gamma M_{t-1} + (1-\gamma)\, n_t. \tag{2}$$

Because the update is a convex combination, the memory stays bounded for bounded frame vectors and has no explicit dependence on t. The decay γ controls how quickly it tracks recent observations, with an effective memory length of $1/(1-\gamma)$ polls. We fix $\gamma = 0.5$, corresponding to two polls at the 1 s poll interval; after four updates, the latest four frame vectors account for 94% of the weight ($1 - 0.5^4 \approx 0.94$). This running reference expresses each observation relative to its video's recent context, following the use of historical references in online change detection (Kifer et al., 2004; Adams & MacKay, 2007).

To provide a reference at the first poll, we initialize the memory from all visual tokens in that poll's window:

$$M_{\mathrm{init}} = \mathrm{Pool}(H_\ell^{\mathrm{first}}, \mathcal{I}_{\mathrm{visual}}), \tag{3}$$

where $\mathcal{I}_{\mathrm{visual}}$ includes the current frame and any earlier frames in the window. The window contains at most $W = 8$ frames; pre-question history is used when available but is not required. M_{init} serves as M_{t-1} for the first comparison, after which Equation (2) updates the memory at every poll. When earlier frames are available, this seed allows an immediate comparison against a reference beyond the current frame; initializing with n_t alone would make the first score zero. Section 4 tests the contributions of the historical reference and the window seed, and Appendix D.2 examines the memory timescale.

3.3 Zero-threshold trigger

Given the current representations and the memory before its update, LatPA measures the increase in question alignment at poll t:

$$S_t = \cos(q_t, n_t) - \cos(q_t, M_{t-1}), \tag{4}$$

The score compares the current frame's question alignment with that of the recent visual history. With a fixed threshold $\tau = 0$, LatPA *acts* when

$$\text{Act} \iff S_t > 0. \tag{5}$$

The zero point has a fixed meaning: the current frame is exactly as question-aligned as the recent past. This gives one decision boundary across videos and backbones without fitting a threshold to benchmark labels. Appendix D.2 measures how task-specific thresholds change response behavior.

S_t depends only on the question vector and on the frame vectors observed up to t; it contains no term in the wall-clock time t or in the elapsed delay $t - t_q$. In particular, if the visual representations remain constant and equal to the memory seed, then $n_t = M_{t-1}$ at every poll, so $S_t = 0$ and no response occurs at $\tau = 0$. The content-blind Random baseline in Table 1 tests this rule against a timing prior (Otani et al., 2020).

Deployment controls. Optional controls require m consecutive positive polls, impose a refractory interval, or cap the response count; they operate after the rule and leave S_t unchanged. These controls can only delay or suppress a response, so they cannot trigger one solely because time has elapsed. All main results use $m = 1$. StreamingBench-PO, IPIBench, and FAR impose no refractory interval or cap. ProactiveVideoQA applies a 10 s refractory interval and an eight-response cap to every trigger. Appendix B specifies which responses each benchmark scores.

Cost and operating point. Each poll costs one prefill of the frame window and no decoding; the pooling, the memory update, and the two cosines add $\mathcal{O}(Nd)$ operations, negligible next to the prefill. Table 6 in Appendix D.4 reports the per-decision cost. One declared operating point, $\ell = 0.5L$, $\gamma = 0.5$, $\tau = 0$, the window seed, and $m = 1$, is shared by the four backbones and all benchmarks: the read-out depth was chosen on the constructed validation set of Appendix C, which is disjoint from every benchmark, and the remaining values were fixed before the test evaluations. The frame window and pooling conventions are specified in Appendix A.

4 Experiments

4.1 Experimental setup

Proactive benchmarks. We evaluate on the Proactive Output (PO) task of StreamingBench (Lin et al., 2026) (250 questions, timing and answer accuracy), IPIBench (Li et al., 2026) (Proactive Timing, Repeated Proactiveness, and Proactive Understanding), the Forward Active Responding (FAR) tasks of OVO-Bench (Niu et al., 2025) (172 videos), and ProactiveVideoQA (Wang et al., 2025c) (PAUC on WEB, VAD, and EGO). StreamingBench-PO and IPIBench assess short-lead responses. FAR evaluates responses throughout longer streams. For OVO-Bench FAR, we use the online timing protocol of Em-Garde (Zheng et al., 2026): response-time F1 within ± 2 s of annotated events, macro-averaged over REC, SSR, and CRR, without answer generation. This differs from the official OVO-Bench metric, which grades answers at probe times. ProactiveVideoQA results use GPT-5.6-luna judgments (Zheng et al., 2023) and are reported in Appendix D.3.

Baseline methods and backbones. We compare LatPA with two families of proactive triggers (Table 1). **(1) Learned triggers.** StreamBridge (Wang et al., 2025a) uses a trained activation head to decide when to respond. We train this head on the WebVid subset of its public Stream-IT data, excluding evaluation data from StreamingBench-PO, IPIBench, and FAR. **(2) Queried triggers.** RIVER (Shi et al., 2026) and the ProactiveVideoQA Gate (Wang et al., 2025c) elicit decisions from the frozen backbone through self-querying and yes/no gating, respectively.

The comparison spans four frozen backbones from two model families at different scales: Qwen2-VL-2B/7B (Wang et al., 2024) and InternVL3-2B/8B (Zhu et al., 2025). Each block fixes the answer

Table 1: Trigger comparison with four frozen backbones (scores $\times 100$). Avg: mean within each benchmark. PVQA Gate: ProactiveVideoQA gate. **Bold**/underline: best/second per backbone and column.

	StreamingBench-PO			IPIBench				OVO-Bench FAR (F1@2)			
Method	Time	Ans	Avg	Timing	Repeat.	Underst.	Avg	REC	SSR	CRR	Avg
Qwen2-VL-2B											
RIVER	14.8	12.4	13.6	15.05	0.00	2.89	5.98	10.05	3.95	1.00	5.00
PVQA Gate	25.2	23.2	24.2	11.34	0.47	3.20	5.00	21.30	**8.68**	**4.21**	**11.40**
StreamBridge	29.6	20.4	25.0	**42.98**	**25.47**	19.37	**29.27**	10.77	4.04	0.26	5.02
LatPA (ours)	**48.0**	**39.6**	**43.8**	42.12	22.64	**22.91**	29.22	**21.95**	8.50	0.61	10.35
Qwen2-VL-7B											
RIVER	33.6	32.4	33.0	27.93	0.00	6.19	11.37	10.53	4.09	1.05	5.22
PVQA Gate	35.6	32.4	34.0	17.98	1.42	4.66	8.02	**26.86**	**9.12**	**6.68**	**14.22**
StreamBridge	29.6	25.6	27.6	**43.27**	**25.47**	**14.35**	**27.70**	10.77	4.04	0.26	5.02
LatPA (ours)	**46.4**	**39.2**	**42.8**	41.35	23.58	10.83	25.25	22.77	8.74	0.60	10.70
InternVL3-2B											
RIVER	7.6	7.2	7.4	12.14	0.47	3.41	5.34	11.33	4.10	0.59	5.34
PVQA Gate	22.4	21.6	22.0	8.48	1.42	0.72	3.54	**41.56**	**8.69**	**3.83**	**18.03**
StreamBridge	**30.0**	**29.2**	**29.6**	43.27	**25.47**	**6.37**	25.04	10.77	4.04	0.26	5.02
LatPA (ours)	28.8	28.0	28.4	**51.81**	20.75	4.72	**25.76**	22.92	**8.69**	0.66	10.76
InternVL3-8B											
RIVER	16.4	15.6	16.0	27.88	0.47	4.25	10.87	11.32	4.06	1.17	5.52
PVQA Gate	25.6	23.2	24.4	19.51	1.89	2.77	8.06	**31.94**	**8.95**	**4.49**	**15.13**
StreamBridge	**30.0**	**26.8**	**28.4**	43.27	**25.47**	7.90	25.55	10.77	4.04	0.26	5.02
LatPA (ours)	29.2	26.4	27.8	**49.63**	21.70	**8.21**	**26.51**	22.71	8.13	0.62	10.49

backbone and varies the trigger. Random provides a timing-prior reference (Otani et al., 2020): one uniformly sampled poll per question on StreamingBench-PO and IPIBench, or an independent probability of $1/5$ per poll on FAR, with a fixed seed and no rate tuning. Native proactive systems are compared separately under their released inference settings (Table 2).

Implementation details. All evaluations use bfloat16 (BF16) precision on NVIDIA GPUs. We use a one-second polling grid, with the ProactiveVideoQA Gate following its original two-second decision interval. For LatPA, we fix $\ell{=}0.5L$, $\gamma{=}0.5$, $\tau{=}0$, $W{=}8$, and $m{=}1$ across backbones and benchmarks. StreamingBench-PO supplies 12 s of pre-question context and caps backbone input at 64 frames. IPIBench supplies the observed stream up to the poll. FAR applies no refractory interval or response cap. ProactiveVideoQA applies the same response limits to every trigger. Detailed configurations are provided in Appendix A.

4.2 Main results

Comparison across triggers and backbones. Table 1 compares trigger strategies with each answer backbone held fixed. We highlight two findings. **(1) Strong performance across frozen backbones and benchmarks.** LatPA achieves the highest StreamingBench-PO average on both Qwen2-VL backbones, scoring 43.8 and 42.8 on the 2B and 7B models and surpassing the strongest competing trigger by 18.8 and 8.8 points, respectively. These gains span both response timing and answer accuracy. Across all four backbones, LatPA outperforms RIVER and the ProactiveVideoQA Gate in average scores on both StreamingBench-PO and IPIBench. On IPIBench, it further achieves the highest average with InternVL3-2B and InternVL3-8B, reaching 25.76 and 26.51. **(2) Transfer across model families and evaluation settings.** LatPA applies the same latent decision rule and shared hyperparameters across backbones, without retraining the trigger or fitting benchmark-specific thresholds. The results reveal complementary strengths across model families: Qwen2-VL achieves the leading StreamingBench-PO timing and answer scores, while InternVL3 achieves the leading IPIBench Proactive Timing scores of 51.81 and 49.63, compared with 43.27 for StreamBridge. Across these two benchmarks, LatPA ranks among the top two triggers in average score for every backbone. By reading each frozen model's internal representations and comparing them against its own evolving history, LatPA provides a common mechanism for proactive decisions across architectures and evaluation settings.

Table 2: Native proactive systems versus LatPA (scores $\times 100$). SB-PO denotes StreamingBench-PO and is evaluated on a test subset. Inference settings differ across systems. **Bold**/underline: best/second per column. Protocol details are described in Appendix A.2.

			SB-PO		IPIBench				OVO-Bench FAR			
Method	Params	Frames	Time	Ans	Timing	Repeat.	Underst.	Avg	REC	SSR	CRR	Avg
Flash-VStream	7B	–	2.0	–	–	–	–	–	–	–	–	–
VideoLLM-online	8B	2 fps	0.0	0.0	0.00	0.00	0.00	0.00	9.92	2.70	0.00	4.21
Dispider	7B	4 fps	5.9	2.0	9.10	0.00	0.22	3.11	**27.72**	**13.27**	0.00	**13.66**
MMDuet	7B	1/turn	23.5	23.5	31.56	8.96	6.90	15.81	13.37	7.03	0.74	7.05
MMDuet2	3B	1/turn	19.6	15.7	22.65	1.89	6.22	10.25	24.17	7.36	**3.19**	11.57
LatPA	2B	1 fps	**45.1**	**37.3**	**42.12**	**22.64**	**22.91**	**29.22**	21.95	8.50	0.61	10.35

Table 3: Ablation on StreamingBench-PO (timing accuracy $\times 100$). All rows read S_t at $0.5L$ from the first poll, as in Table 1. Each block changes one component of LatPA (EMA memory, $\tau = 0$) and keeps the other. Rank and Record need three polls of history before acting.

	Qwen2-VL		InternVL3		InternVL3.5	Gemma 3
Variant	2B	7B	2B	8B	4B	4B
Decision rule						
Rank	26.0	28.8	30.0	26.0	30.4	30.8
Record	24.4	27.2	24.8	24.0	27.6	29.6
Memory setting						
Previous frame ($\gamma = 0$)	16.0	18.4	16.4	11.6	17.6	20.8
Cumulative mean	9.6	13.6	31.2	35.2	33.2	37.2
Sliding mean, $k = 8$	5.2	7.2	28.8	30.0	32.4	31.6
First-frame seed	24.4	25.2	24.4	20.8	24.0	29.6
LatPA (ours)	48.0	46.4	28.8	29.2	27.2	36.8

Comparison with native proactive models. Table 2 compares complete proactive systems, pairing LatPA with a frozen Qwen2-VL-2B backbone and evaluating native proactive models with their released checkpoints and model-specific inference configurations (Chen et al., 2024a; Qian et al., 2025; Wang et al., 2025d). LatPA offers two advantages. **(1) Strong response timing with a compact backbone.** With a 2B backbone, LatPA surpasses the compared 3B–8B native proactive models on both StreamingBench-PO Time and IPIBench Proactive Timing. It achieves 45.1 and 42.12, respectively, improving on the strongest native comparator, MMDuet-7B, by 21.6 and 10.56 points. These results demonstrate that effective response timing can be obtained directly from a compact frozen model through latent read-outs. **(2) Strong answer performance without updating the backbone.** LatPA also leads the compared native models on StreamingBench-PO Ans and IPIBench Proactive Understanding, scoring 37.3 and 22.91 versus 23.5 and 6.90 for the strongest native comparator. Its trigger reads internal latent signals while leaving the pretrained backbone parameters unchanged, allowing proactive response control to be added without further fine-tuning the multimodal answer model. The combination of stronger answer-related scores and an unchanged backbone highlights LatPA's ability to support proactive interaction while retaining the original model parameters for multimodal understanding and generation. Appendix A.2 provides the evaluation protocol and inference details.

4.3 In-depth analysis

We examine the contributions of the decision rule and temporal memory, then compare the temporal behavior of latent and token read-outs.

Ablation study. Table 3 ablates the decision rule and temporal memory on StreamingBench-PO across six frozen backbones at $0.5L$. LatPA uses a zero threshold and EMA memory ($\gamma = 0.5$), initialized from the first eight-frame window mean. **(1) Decision rule.** Keeping the latent scores and EMA memory fixed, Rank replaces the zero threshold with the rolling 85th percentile over up to 20 scores, while Record triggers when the current score exceeds its previous running maximum;

both require three polls of history before acting. LatPA outperforms both alternatives on four of the six backbones, with gains of 17.6–23.6 timing points on Qwen2-VL. This supports direct historical-contrast decisions without score ranking or a history-collection delay. **(2) Memory setting.** Keeping the zero-threshold rule fixed, Previous frame ($\gamma = 0$) replaces accumulated history with the immediately preceding frame; Cumulative mean averages the available frame history without forgetting; Sliding mean ($k = 8$) uses the preceding poll's eight-frame window mean; and First-frame seed retains the EMA update but initializes it from the first post-question frame instead of the input-window mean. The full LatPA configuration improves on Previous frame by 9.6–32.0 points across all six backbones, demonstrating the value of accumulated temporal context. Its window-based initialization also improves on First-frame seed by 3.2–23.6 points across all six backbones, establishing the benefit of incorporating available context from the first decision. On Qwen2-VL, the EMA further exceeds both cumulative and sliding averages, supporting recency-weighted history for these backbones. Together, these results support direct historical-contrast decisions with context-initialized memory. Appendix D.2 analyzes the memory timescale.

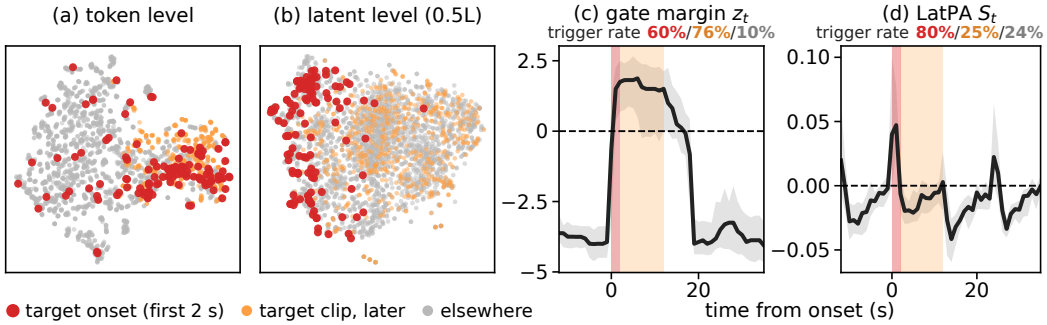

Figure 3: Token and latent read-outs on the constructed validation set (Qwen2-VL-7B, $0.5L$). (a, b) t-SNE embeddings. (c, d) Trigger scores: median and interquartile range. Colored percentages give per-poll trigger rates at $\tau = 0$ for the three legend categories.

Advantages of latent trigger signals. LatPA distinguishes the arrival of target content from its continued presence, concentrating trigger responses around the target onset. To evaluate this behavior, we construct 60 validation streams, each containing five unrelated 12 s clips, and instruct the model to report the appearance of a designated target. Figure 3 compares latent and token-level read-outs from the same frozen Qwen2-VL-7B backbone. The t-SNE views in (a, b) visualize this distinction: target-onset points overlap substantially with later target points at the token level, while the latent features exhibit a more distinct concentration of onset points. The temporal profiles in (c, d) show the corresponding response pattern. The yes/no gate remains positive throughout much of the target clip, whereas LatPA produces a brief peak near the onset and then declines during continued target presence. At zero threshold, LatPA increases the per-poll onset trigger rate from 60% to 80% and reduces the later-target trigger rate from 76% to 25%, improving the onset-to-presence contrast from -16 to 55 percentage points. Appendix D.1 provides cross-backbone comparisons and the token-level memory control. Together, these results highlight LatPA's ability to emphasize newly arriving target content rather than repeatedly respond to its continued visibility.

4.4 Error analysis and limitations

The onset sensitivity in Section 4.3 does not by itself guarantee correct event selection or sparse responses. Figure 4 examines three failure cases involving event identity, spatial aggregation, and response density. Appendix D.5 reports aggregate frequencies and separates trigger decisions from answer correctness.

(a) **Temporal contrast and event identity.** In Figure 4(a), the number 20 is requested and appears at 62 s, but LatPA acts at 32 s when the number changes from 2 to 5. The positive score detects a transition, but this transition does not match the requested number. This illustrates why a response to newly arriving content need not identify the event named in the question.

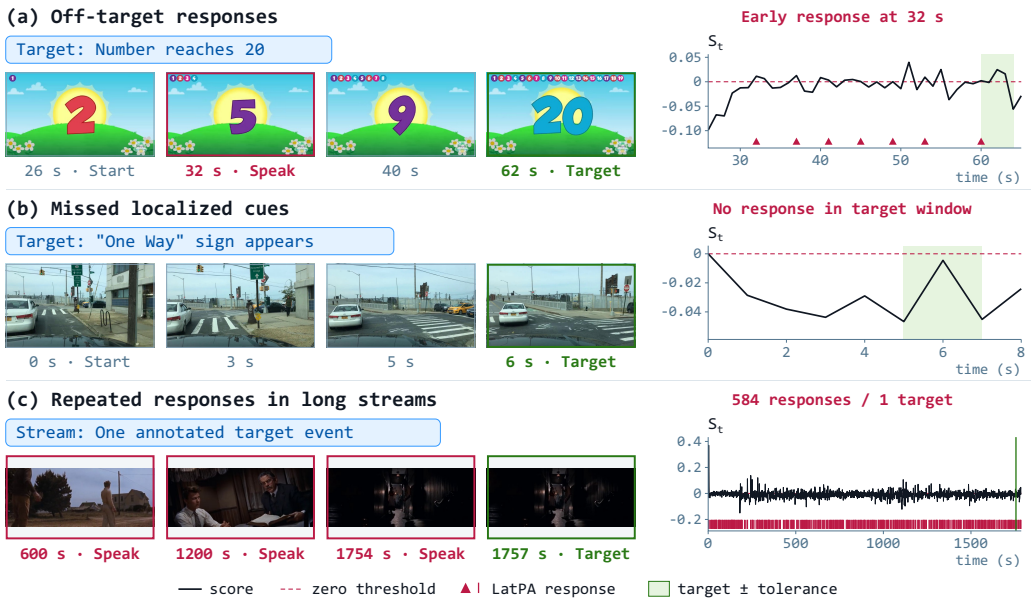

Figure 4: Representative trigger-level error cases (Qwen2-VL-2B, shared operating point): (a) off-target responses, (b) missed localized cues, and (c) repeated responses in a sparse-event stream. Frames are paired with the corresponding score traces.

(b) **Frame-level reference and spatial evidence.** In Figure 4(b), the *One Way* sign appears at 6 s as a small roadside region. The score approaches zero but never crosses it. This is consistent with mean pooling combining a small target region with the surrounding background. The trace identifies spatial aggregation as a factor distinct from the historical comparison.

(c) **Recent change and response density.** In Figure 4(c), a 30-minute CRR video has one annotated clue at 1757 s. LatPA emits 584 responses under the uncapped protocol, two within 2 s of the clue. The EMA tracks recent visual change without encoding an expected event count. This trace illustrates how detecting a sparse target event can coexist with excessive responses elsewhere, as suggested by the FAR results.

Limitations and future directions. The error analysis identifies two primary limitations: **(1) Imprecise trigger thresholds.** Ordinary visual changes can trigger unnecessary responses, while subtle changes may go undetected when a response is required. **(2) Memory context in long videos.** The model may forget visual content seen much earlier, leading to false triggers.

These limitations motivate two directions for future research: **(1) Dynamic action thresholds.** Finer changes in the current video scene could be used to dynamically set the threshold for deciding whether to act, allowing the method to accommodate more diverse video scenarios. **(2) Efficient memory storage and retrieval.** A dedicated memory module could store previously observed visual content and efficiently retrieve relevant memories as new frames arrive.

5 Conclusion

This paper introduces LatPA, a training-free adapter that enables proactive timing directly from frozen Video-LLM representations. LatPA compares the question alignment of the current frame with that of an exponential memory, turning the video's own history into a reference for deciding when to act. The rule requires no additional training or yes/no generation and shares its configuration across backbones and benchmarks. Experiments show higher average scores than queried baselines on StreamingBench-PO and IPIBench, while ablations and controlled analyses support its concentration of responses at target onsets. These findings establish historical comparison in latent space as an effective approach to proactive video interaction.

AI use statement. We used generative AI tools for literature retrieval and summarization, research ideation and execution (including method design and implementation, data cleaning, code debugging, and interpretation of experimental results), figure preparation, and drafting and polishing the manuscript. A privately deployed AI service was used solely to screen the manuscript for inadvertent disclosure of internal information. We did not use generative AI to generate synthetic datasets or prove mathematical claims. The authors reviewed all AI-assisted material, verified the references, and checked and tested the code. We take responsibility for the final text, claims, and artifacts.

Ethics statement. We study response timing for frozen Video-LLMs using public benchmarks (StreamingBench-PO, IPIBench, OVO-Bench FAR, and ProactiveVideoQA) and a validation set constructed from existing WebVid clips (Appendix C). The study does not recruit human participants or collect new personal data. In line with the ICLR Code of Ethics, we consider privacy, bias, and misuse: existing videos may depict identifiable people, and proactive systems could enable unsolicited monitoring or produce biased, inappropriate, or excessive responses. Public availability does not remove the obligation to respect privacy and the original datasets' and models' usage terms. To support responsible assessment, Section 4.4 and Appendix D.5 analyze off-target, missed, and repeated responses, while Appendix D.4 reports decision latency and memory use. Any deployment should obtain appropriate consent, give users control over monitoring and notifications, and retain human oversight in safety-critical settings.

Reproducibility statement. We document the method and evaluation protocols to support independent implementation and verification. Section 3 defines the problem setup, hidden-state read-out, memory initialization and updates, and decision rule. Section 4 and Appendices A–B specify the frozen backbones, shared operating point, input context, response controls, and benchmark-specific metrics. Appendix C describes validation-set construction and its separation from the evaluation benchmarks; Appendix D.4 details the latency and memory measurements. An anonymized reference implementation of LatPA's read-out, memory update, and trigger is available at `https://anonymous.4open.science/r/LatPA-D405`. Only the core method logic is released; the paper and appendices document the experimental protocols and evaluation settings.

References

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it ' s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL `https://arxiv.org/abs/2303.08112`.

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.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning video llm with streaming speech transcription at scale. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29083–29095. IEEE, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao, Xiaoyong Zhu, Yu Cheng, Bo Zheng, Wei Chen, and Haochao Ying. Enabling agents to communicate entirely in latent space. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27106–27129, 2026.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788. IEEE, 2025.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations*, volume 2026, pp. 43130–43158, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *International Conference on Learning Representations*, volume 2025, pp. 41665–41677, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Vincent Herrmann, Róbert Csordás, and Jürgen Schmidhuber. Measuring in-context computation complexity via hidden state prediction. *arXiv preprint arXiv:2503.13431*, 2025.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Advances in neural information processing systems*, 18, 2005.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pp. 180–191. Toronto, Canada, 2004.

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.

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.

Jinzhao Li, Yinuo Chen, Wenxuan Song, Yijia Lei, Yichi Zhang, Honglei Yan, Panwang Pan, and Miao Liu. Ipibench: Evaluating interactive proactive intelligence of mllms under continuous streams. *arXiv preprint arXiv:2605.27074*, 2026.

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.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M Alvare. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*, 2024.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. Proactive agent: Shifting llm agents from reactive responses to active assistance. In *International Conference on Learning Representations*, volume 2025, pp. 47431–47457, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 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.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*, 2020.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Vignav Ramesh and Kenneth Li. Communicating activations between language model agents. *arXiv preprint arXiv:2501.14082*, 2025.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. River: A real-time interaction benchmark for video llms. *arXiv preprint arXiv:2603.03985*, 2026.

Yakov Pyotr Shkolnikov. Agent memory below the prompt: Persistent q4 kv cache for multi-agent llm inference on edge devices. *arXiv preprint arXiv:2603.04428*, 2026.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4396–4406, 2019.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=DTvviEnW2A.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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*, 2025c.

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, 2025d.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yikai Zheng, Xin Ding, Yifan Yang, Shiqi Jiang, Hao Wu, Qianxi Zhang, Weijun Wang, Ting Cao, and Yunxin Liu. Em-garde: A propose-match framework for proactive streaming video understanding. *arXiv preprint arXiv:2603.19054*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL `https://arxiv.org/abs/2310.01405`.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *Forty-third International Conference on Machine Learning*, 2026.

A Implementation details

A.1 LatPA details

LatPA uses one shared trigger configuration across the four main backbones and four benchmarks. Table 4 specifies the read-out depth, memory decay, threshold, pooling, initialization, and confirmation rule. Section 4 and Appendix B specify each benchmark's polling schedule, visual context, and response limits.

Trigger implementations and evaluation context. The ProactiveVideoQA Gate queries the frozen backbone every 2 s and generates an answer after a yes decision. StreamBridge uses a LLaVA-OneVision-0.5B activation model (Li et al., 2024) with a SigLIP encoder (Zhai et al., 2023). Each block of Table 1 uses the same frozen answer backbone; trigger architectures and input windows follow their respective implementations. StreamingBench-PO supplies 12 s of pre-question context and caps the backbone's input at 64 frames. IPIBench provides the observed stream up to the poll, and timing-only evaluations generate no answer. The latency comparison excludes answer generation and is reported separately in Appendix D.4.

Pooling and initialization. We mean-pool all tokens after the last visual token, including the question and template suffix, to obtain q_t. For a window of $F \leq W$ frames and V visual tokens, n_t is the mean of the last $\max(1, \lfloor V/F \rfloor)$ visual tokens. This approximates the latest frame's token group and is exact when each frame has an equal, separate group of visual tokens. The memory is initialized by mean-pooling all visual tokens in the first available window, including the current frame and any available pre-question context.

A.2 Existing proactive MLLM details

Table 2 compares LatPA with native proactive video MLLMs whose trigger and backbone are released as coupled systems. The LatPA row uses the frozen Qwen2-VL-2B backbone and the protocol of Table 1, with StreamingBench-PO rescored on the same 51-question subset used for the native-system comparison. The existing systems retain their released inference settings, with frame budgets adapted to the evaluation harness. Reported averages cover successfully completed samples; failed samples are excluded. A dash denotes an unreported or unmeasured result.

Flash-VStream reports a published StreamingBench-PO score of 1.96, which we round to 2.0 in the table. VideoLLM-online uses immediate-answer inference for StreamingBench-PO, so its proactive timing score is 0.0 under our scoring protocol. Dispider uses 4 s visual windows. MMDuet and MMDuet2 consume one frame per dialogue turn. MMDuet2 is the RL variant of MMDuet; its SSR score is computed on 41 of 42 videos, and its CRR result uses a bounded 64-frame dialogue because the unbounded released recipe exceeds the context limit on long CRR videos. FAR evaluation uses no refractory interval or response cap, so MMDuet2's FAR average combines the unbounded REC/SSR protocol with the bounded-dialogue CRR protocol.

B Benchmark and metric details

StreamingBench-PO (Lin et al., 2026) is the Proactive Output task of StreamingBench. Each of its 250 questions names a trigger condition and a target output, and the model must produce that output as soon as the condition is met. Polling runs at 1 Hz from the question time to 4 s after the ground-truth time; the first response determines the score. Time is correct within ± 2 s of the ground truth, Ans further requires the target output in the response text, and Avg is their mean.

OVO-Bench FAR (Niu et al., 2025) is the Forward Active Responding family of OVO-Bench: REC counts repetitions of an action, SSR tracks tutorial steps, and CRR waits for the clue to a question. We use all 172 videos (82 REC, 42 SSR, 48 CRR). Following the online protocol of Em-Garde (Zheng et al., 2026) (Section 4), we evaluate event timing by F1 of response times within ± 2 s of the events, macro-averaged over sub-tasks. The instruction is given up-front, and no answer is generated. Event times are reconstructed from the official probe lists: for REC, the first probe at which the cumulative count reaches a new value; for SSR, a transition between steps; for CRR, the first probe after the

Table 4: Declared LatPA operating point for all backbones and benchmarks. ℓ: read-out layer; q_t, n_t: question/visual features; W: window; γ: EMA decay; τ: threshold; m: confirmation polls.

ℓ	q_t, n_t	W	γ	M_{init}	τ	m	Poll interval
$0.5L$ (L14 of 28)	mean over tokens	8 frames	0.5	first-window mean	0	1	1 s

clue is revealed. This online metric measures response timing throughout the stream; the official probe-based metric measures answer content at annotated times.

IPIBench (Li et al., 2026) evaluates proactive intelligence under continuous streams. In its Proactive Monitoring setting, Proactive Timing (972 items) asks when a described target first appears, Proactive Understanding (562) also requires the correct content, and Repeated Proactiveness (212) must catch every recurrence of an event. Polling runs at 1 Hz inside the official ± 4 s window around each ground-truth time; a response counts within ± 1 s, scored by the official evaluators.

ProactiveVideoQA (Wang et al., 2025c) evaluates multi-turn proactive answering, where each ground-truth turn has a span in which a reply is credited. As in Em-Garde (Zheng et al., 2026), we use the WEB, VAD and EGO subsets (500, 101 and 326 questions). PAUC has a judge (GPT-5.6-luna) rate the cumulative replies in each span on a three-level rubric (Zheng et al., 2023) and takes the normalized area under the resulting step function; reply times are rescaled by $1 - \omega$, so $\omega{=}1$ ignores timing. Every trigger runs under the same two limits on responses on this benchmark: a 10 s refractory interval after each response and at most eight responses per video.

Diagnostic splits. For the half-based diagnostics and memory measurements in Appendix D, StreamingBench-PO is split into its two official annotation files of 125 questions each, and OVO-Bench FAR is split by video, alternating over the sorted video ids within each sub-task. The FAR event-discrimination diagnostic uses the test half of this split. Table 1 reports full-benchmark results; Table 2 uses the stated 51-question StreamingBench-PO subset.

C Constructed validation set

The constructed set provides controlled onset and target-presence labels for read-out selection, mechanism analysis, and the sensitivity curves in Appendix D.2. Benchmark performance is evaluated separately on the four datasets in Appendix B. The set is built with a fixed seed from WebVid (Bain et al., 2021), a corpus of captioned stock clips. Each stream concatenates five 12 s segments, each cut from the start of a different WebVid clip and re-encoded at 5 fps and 360 px, in random order, so that a stream is 60 s of five unrelated scenes with a hard cut every 12 s; the five captions share no content word, and no clip is used in more than one stream. One segment, in the third, fourth or fifth position with equal counts, is the target. The instruction "Tell me immediately when you see this: *caption*." names the target's cleaned caption and is posed at 12 s, so the first segment is pre-question context. Clip boundaries define the labels: the target onset is at 24, 36 or 48 s, every poll inside the target segment is a target poll, and every other poll is not. The set has 60 streams, 300 clips, and 2,880 polls on the 1 Hz grid from 12 s to 59 s.

The validation set is disjoint from the benchmarks at the stream, clip, and caption levels. WebVid consists of Shutterstock stock footage, whereas StreamingBench, OVO-Bench, IPIBench and ProactiveVideoQA draw on YouTube and web videos (Han et al., 2025), television (Lei et al., 2018), egocentric (Grauman et al., 2022) and surveillance (Sultani et al., 2018) videos.

D Additional results

D.1 Latent versus token-level read-outs

Onset concentration across backbones. LatPA strengthens the onset-to-presence contrast across all four main backbones at the shared $0.5L$ read-out. Beyond Qwen2-VL-7B (Section 4.3), the onset-to-presence contrast improves over the yes/no gate by 41–86 percentage points on the other three backbones. LatPA acts on 78% of onset polls and 24% of later target polls on Qwen2-VL-2B, compared with 55% and 87% for the yes/no gate. The corresponding onset/later-target rates are

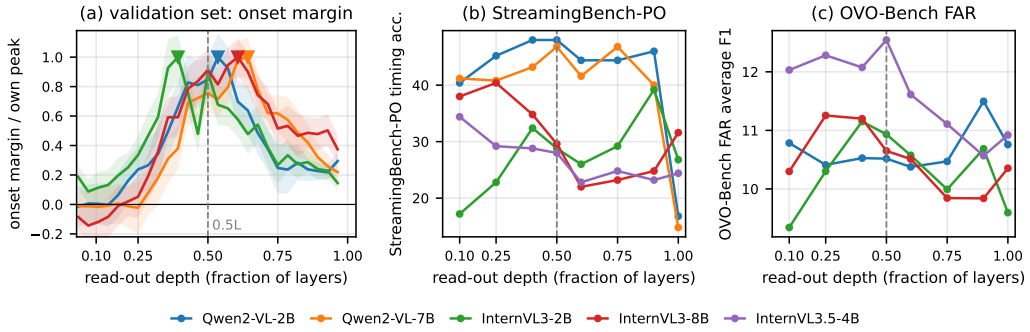

Figure 5: Read-out depth sweep (dashed line: $0.5L$). (a) Validation onset margin per backbone (triangles: peaks, bands: 95% bootstrap intervals). (b, c) StreamingBench-PO and FAR with only the depth changed. InternVL3.5-4B has no validation recording, Qwen2-VL-7B no FAR recording.

59%/29% versus 54%/80% on InternVL3-2B, and 62%/31% versus 75%/85% on InternVL3-8B. LatPA's outside-target trigger rates are 22%, 31%, and 33%, respectively. All rates are measured per poll on the constructed set. Figure 3 visualizes latent features $[q_t \odot n_t,\ n_t - \hat{M}_{t-1}]$ and the gate's next-token distribution with t-SNE (Van der Maaten & Hinton, 2008); $\hat{M}_{t-1}$ has unit norm.

Token score with temporal memory. LatPA also suppresses continued activation more strongly than a token score equipped with temporal memory. On Qwen2-VL-7B, the control subtracts an EMA from the gate's yes/no logit margin z_t. At zero threshold, LatPA acts on 25% of later target polls and 24% of outside-target polls, compared with 61% and 34% for this control: reductions of 36 and 10 percentage points. The per-stream onset AUROCs are 0.86 for LatPA and 0.92 for the token control. Together, these measurements distinguish threshold-free onset ranking from response concentration at the declared zero-threshold operating point.

D.2 Shared operating point and sensitivity

One operating point is shared across backbones and benchmarks: read-out depth is chosen on the independent constructed set, with $\gamma = 0.5$ and $\tau = 0$ fixed before test evaluation. Figures 5 and 6 show its depth and parameter sensitivity.

Read-out depth. The shared $0.5L$ read-out retains 76–91% of each recorded backbone's peak validation onset margin. This margin is the mean S_t over the first two target polls minus its mean over the remaining polls, with peaks between L11 and L18. On StreamingBench-PO, Qwen2-VL-2B/7B achieve 48.0/46.8 timing accuracy at $0.5L$, compared with 16.8/14.8 at the final layer. The benchmark peaks for InternVL3-2B, InternVL3-8B, and InternVL3.5-4B occur at $0.90L$, $0.25L$, and $0.10L$, respectively; FAR varies by less than 2 F1 across the measured depths. The benchmark sweep reuses recorded hidden states with $\gamma = 0.5$ and $\tau = 0$; its $0.5L$ StreamingBench-PO scores agree with the live runs to within one question (0.4 points).

Memory timescale. The EMA decay controls the temporal scale of the historical reference. At $\tau = 0$, the curves peak at $\gamma = 0.5$ on FAR and on StreamingBench-PO for Qwen2-VL. InternVL3 favors 0.7 on StreamingBench-PO, while Qwen2-VL favors 0.3 on IPIBench. For the constructed set's regular 12 s scenes, the onset margin peaks at 0.7 for Qwen2-VL and 0.9 for InternVL3. These profiles characterize the memory timescale alongside the memory-form ablation in Table 3.

Threshold and response allocation. The threshold provides a direct control over responsiveness and selectivity while preserving the latent score. Figure 6 varies τ at $\gamma = 0.5$ and the shared $0.5L$ read-out. Small positive thresholds improve StreamingBench-PO by less than half a point for Qwen2-VL and by eight to ten points for InternVL3. On FAR, they add up to 2.3 and four to five F1, respectively. On IPIBench, negative thresholds add five to thirteen points for Qwen2-VL. On the constructed set, onset recall minus false-response rate improves up to $\tau \approx +0.02$ to $+0.05$. The

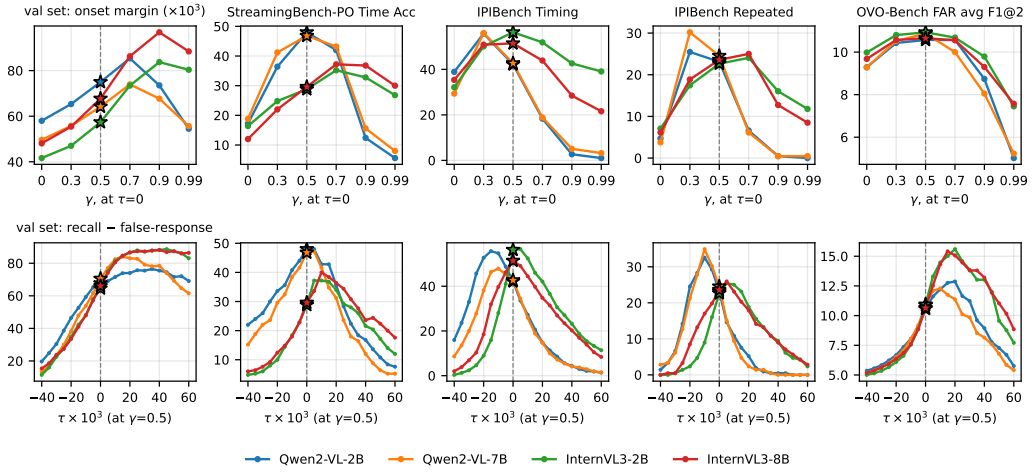

Figure 6: One-dimensional sweeps through the operating point (γ=0.50, τ=0, stars) at $0.5L$: top row varies γ, bottom row varies τ (units of 10^{-3}). The left column shows the constructed validation set, and the other columns show the four timing metrics.

sweeps quantify task-dependent response allocation; all main comparisons use the same predeclared zero threshold.

D.3 ProactiveVideoQA: timing and answer content

LatPA's training-free trigger achieves an EGO PAUC of 31.40, compared with 31.72 for the ProactiveVideoQA Gate, ranking second at every timing weight and remaining within 0.5 points of the Gate. It also ranks second on VAD at $\omega = 0$. Table 5 reports all subsets on Qwen2-VL-2B under the common response limits in Appendix B, with an overall PAUC of 26.19 for LatPA. PAUC complements onset timing by evaluating cumulative answer content and response time across multiple turns; $\omega = 1$ measures content alone.

Table 5: ProactiveVideoQA PAUC by subset and timing weight ω (higher is better; Qwen2-VL-2B, GPT-5.6-luna judge). Avg: mean over $\omega \in \{0, 0.5, 1\}$, equal to the omitted ω=0.5 column (PAUC is linear in ω). All: mean of the subset Avgs. PVQA Gate: ProactiveVideoQA's yes/no gate.

	WEB			VAD			EGO			All
Method	ω=0	1	Avg	ω=0	1	Avg	ω=0	1	Avg	Mean
RIVER	31.40	34.13	32.77	23.79	24.07	23.93	26.04	26.52	26.28	27.66
StreamBridge	34.93	42.55	38.74	22.11	23.36	22.74	26.31	26.68	26.49	29.32
PVQA Gate	32.09	39.53	35.81	21.86	19.63	20.74	30.46	32.98	31.72	29.42
LatPA	23.53	27.97	25.75	22.50	20.33	21.41	30.17	32.63	31.40	26.19

D.4 Decision latency and memory

LatPA is 2.4–2.6× faster per decision than RIVER and uses 0.7–1.5 GB less peak GPU memory than StreamBridge across the four backbones (Table 6). Its single-prefill implementation takes 166–310 ms per decision on InternVL3, within the one-second polling interval, and 1.7–2.4 s on Qwen2-VL. The complete comparison includes each trigger's visual input, prefill, and decision workload, with answer generation excluded. Latency is measured on one dedicated GPU over every decision in the polling windows of all 250 StreamingBench-PO questions. Peak allocated GPU memory is measured separately on six questions, three per benchmark half.

Table 6: Cost of one proactive decision on StreamingBench-PO: mean latency in ms (250 questions) and peak allocated GPU memory in GB (six questions), both lower is better. Measurements include visual input and decision computation; answer generation is excluded. PVQA Gate: the yes/no gate of ProactiveVideoQA, one decision per 2 s. The others decide every second.

	Qwen2-VL-2B		Qwen2-VL-7B		InternVL3-2B		InternVL3-8B	
Method	ms	GB	ms	GB	ms	GB	ms	GB
RIVER	4267	5.9	6150	17.4	417	4.8	739	16.6
PVQA Gate	455	4.3	619	15.7	84	4.0	127	15.0
StreamBridge	25	6.7	25	18.0	25	6.4	25	17.4
LatPA	1693	5.6	2411	17.3	166	4.9	310	16.1

Table 7: Benchmark averages under two read-out prompt orders (scores $\times 100$, higher is better). Frames first: $[\text{frames}, q]$; Question-cond.: $[q, \text{frames}, q]$. Both use $0.5L$, $\gamma = 0.5$, $\tau = 0$, the same polling protocol, and the same frozen backbone. FAR reports average F1@2.

Backbone	Order	StreamingBench-PO	IPIBench	FAR
Qwen2-VL-2B	Frames first	43.8	29.22	10.35
	Question-cond.	41.2	28.90	10.54
Qwen2-VL-7B	Frames first	42.8	25.25	10.70
	Question-cond.	43.0	28.93	10.49
InternVL3-2B	Frames first	28.4	25.76	10.76
	Question-cond.	27.8	23.88	10.56
InternVL3-8B	Frames first	27.8	26.51	10.49
	Question-cond.	28.2	23.74	10.26

D.5 Question dependence and response decomposition

The question controls and response statistics distinguish historical contrast, event selection, and answer generation under the shared visual polling protocol.

Historical contrast and question controls. LatPA's historical comparison outperforms adjacent-frame novelty on StreamingBench-PO under both read-out orders: 28.4–48.0 timing accuracy versus 20.0–22.0 (Table 8). The novelty score $1 - \cos(n_t, n_{t-1})$ uses a threshold selected on the other data half. At $0.5L$, pooled question vectors have pairwise cosines of 0.88–0.93 across the four main backbones; replacing them with their shared direction reproduces 94–95% of score variance. On StreamingBench-PO, the matched-minus-permuted question margin averages -0.2 points for $[\text{frames}, q]$ and -1.5 points for $[q, \text{frames}, q]$. On FAR, the permuted-question scores remain within 0.35 F1 of LatPA under both orders.

Read-out prompt order. The default $[\text{frames}, q]$ order conditions q_t on the question; $[q, \text{frames}, q]$ additionally lets visual states attend to the question while retaining the pooling of q_t after the last frame. With all other settings fixed, moving to the question-conditioned order changes benchmark averages by -2.6 to $+0.4$ points on StreamingBench-PO, -2.8 to $+3.7$ on IPIBench, and at most 0.23 F1 on FAR (Table 7). All main results use $[\text{frames}, q]$.

Response timing and spatial evidence. Table 9 quantifies early, on-time, late, and missing responses for the main results, complementing the traces in Section 4.4. On the 156 road-sign items in IPIBench's Text subset, Qwen2-VL-2B/7B have 7.7%/10.9% on-time and 59.6%/58.3% missing responses, providing a localized-cue diagnostic alongside Figure 4(b).

Event selection over long streams. On FAR, LatPA produces 18.8–21.8 responses per minute, with precision of approximately 13% on REC, 4–5% on SSR, and 0.3% on CRR (Table 9). The

Table 8: Question and temporal controls under both read-out orders ($0.5L$, scores $\times 100$, higher is better). Matched and Permuted use $\tau = 0$; Permuted replaces the question vector with another sample's vector. Frame novelty is $1-\cos(n_t, n_{t-1})$, with its threshold selected on the other data half. Scores are computed offline from recorded states; LatPA's timing scores agree with live evaluation to within 0.8 points.

Backbone	Order	Matched	Permuted	Frame novelty
StreamingBench-PO, timing accuracy				
Qwen2-VL-2B	Frames first	48.0	46.0	20.0
	Question-cond.	44.8	46.4	20.0
Qwen2-VL-7B	Frames first	46.8	46.8	20.0
	Question-cond.	46.0	45.2	20.0
InternVL3-2B	Frames first	28.8	32.0	22.0
	Question-cond.	28.4	31.2	22.0
InternVL3-8B	Frames first	29.6	29.2	20.4
	Question-cond.	30.4	32.8	20.8
OVO-Bench FAR, average F1@2				
Qwen2-VL-2B	Frames first	10.5	10.5	12.0
	Question-cond.	10.7	10.8	11.9
Qwen2-VL-7B	Frames first	10.9	10.9	13.4
	Question-cond.	10.6	10.5	13.2
InternVL3-2B	Frames first	10.9	11.1	13.3
	Question-cond.	10.7	10.9	13.3
InternVL3-8B	Frames first	10.7	10.9	13.7
	Question-cond.	10.4	10.7	13.8

Table 9: Response timing and event precision at the shared operating point. Timing distributions and precision are percentages; response density is measured in responses per minute. IPIBench on-time values are pooled per-item proportions; Table 1 reports macro-averaged Timing scores across four types.

Benchmark	Outcome	Qwen2-VL-2B	Qwen2-VL-7B	InternVL3-2B	InternVL3-8B
StreamingBench-PO	Early	45.2	47.2	70.4	69.6
	On time	48.0	46.4	28.8	29.2
	Late	2.4	2.0	0.4	0.0
	Missing	4.4	4.4	0.4	1.2
IPIBench Timing	Early	12.7	16.2	39.3	33.5
	On time	45.3	44.9	50.7	50.0
	Late	15.0	11.0	5.5	5.7
	Missing	27.1	28.0	4.5	10.8
	Missing, Text subset	59.6	58.3	7.7	30.8
OVO-Bench FAR	Responses per min	18.8	19.0	20.2	21.8
	Precision REC	12.9	13.4	13.3	13.1
	Precision SSR	4.5	4.6	4.6	4.3
	Precision CRR	0.3	0.3	0.3	0.3

Qwen2-VL-2B diagnostic on the FAR test half gives a per-video event AUROC of 0.51 for S_t. These stream-wide measurements complement the onset-to-presence contrast in Appendix D.1.

Answer content at the selected time. On IPIBench Understanding, Qwen2-VL-2B responds on time for 311 of 562 items, with the requested count present in 25 responses; InternVL3-2B produces `<silence>` in 245 of 318 on-time responses. These conditional counts separate trigger timing from generated content within the joint evaluation.