INTERACTIONBENCH: A REAL-TIME INTERACTION BENCHMARK FOR STREAMING VIDEO SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A video assistant must speak when its instruction warrants a response and stay silent otherwise. We introduce a benchmark that evaluates this decision for the complete system of model, memory, and response controller. INTERACTIONBENCH covers query responses, event triggers, and ongoing updates in 1,060 interactions over 812 videos, with 69 negative streams and 53 suites that pair counted events with look-alike near misses. It scores content accuracy, timing accuracy, and silence compliance on the video clock and reports wall-clock latency separately. Timely speech costs silence across systems. Polled Qwen3-VL-8B (Bai et al., 2025) reaches 77.8 timing accuracy but 10.9 silence compliance. A native real-time interaction system reaches 29.2 silence compliance at 66.8 timing accuracy, yet emits on 89.9% of negative streams. No open-weight system clears a third of the near-miss suites. Fewer replies help only when chosen, as random deletion merely trades timing for silence. Offline scores miss these failures and mispredict online behavior. Adding restraint is costly, as the native system's controller adds little by itself and agentic systems add it only at about 30 s per poll.

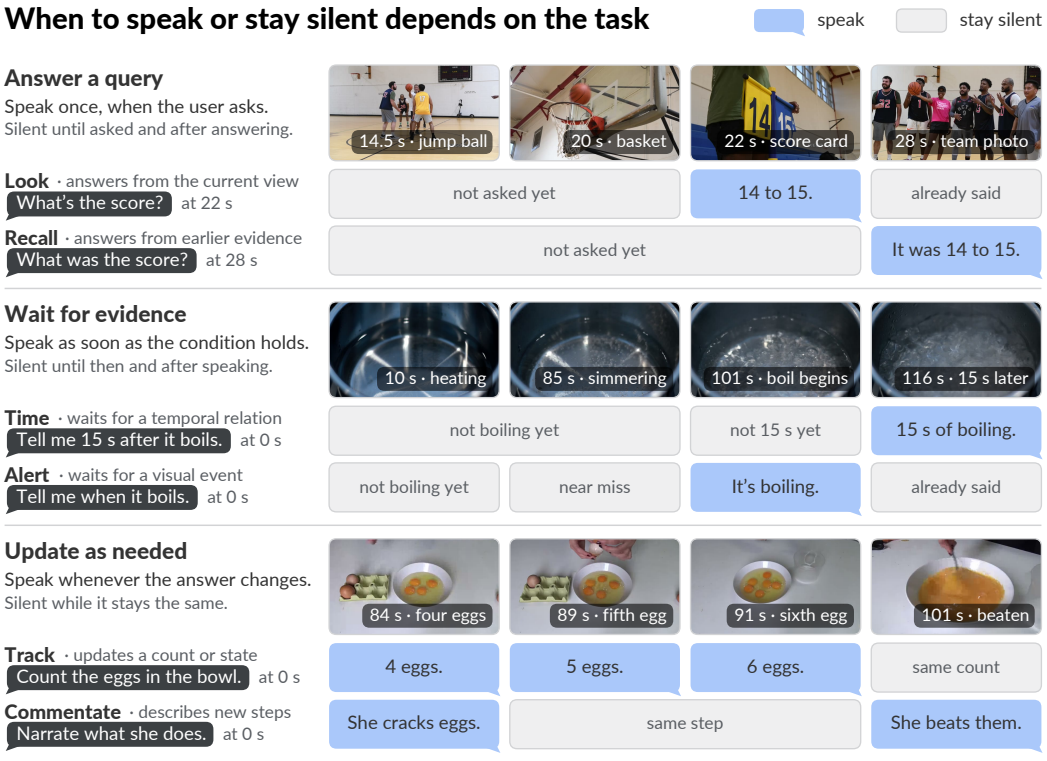


Figure 1: **A frame can require speech in one task and silence in its pair.** Each task pair shares one video, and each cell shows the reply or silence required at that frame.

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.

1 INTRODUCTION

"Tell me when the water boils" asks an assistant to make two decisions. It must respond when boiling becomes visible and remain silent while the pot merely simmers. After reporting the event, it should keep watching without repeating the same message. Answering a question about a sampled frame tests only part of this behavior. The complete interaction also depends on when the system speaks, when it waits, and when it resumes monitoring after it has responded.

Streaming video models process frames as they arrive and respond during the stream (Chen et al., 2024; Zhang et al., 2024; Wang et al., 2025b). Interaction systems add memory and turn management around a visual backbone (Yao et al., 2026). Their observable behavior therefore belongs to the complete system. A capable model can still speak at every poll, repeat an answer, or stop attending after its first response. Conversely, a system can avoid false alarms by missing every event. Useful interaction requires both timely speech and disciplined silence, and an evaluation must score both.

Existing benchmarks study proactive responses and streaming interaction (Wang et al., 2025a; Shi et al., 2026; Zhao et al., 2026a; Lu et al., 2026). SPOT-Bench (Chatterjee et al., 2026) and LiveProBench (Du et al., 2026) penalize premature and repeated replies but include no negative streams or near misses. INTERACTIONBENCH focuses on the paired obligation created by a standing instruction, to respond to sufficient evidence and withhold speech when that evidence is absent. We annotate the earliest answerable moments, score repeated emissions, include streams where the requested event never occurs, and pair real events with visually similar near misses. These tests separate detecting a scene from judging whether its instruction is met.

INTERACTIONBENCH contains 812 videos and 1,060 interactions across six tasks. LOOK and RECALL answer a user query. TIME and ALERT wait for temporal evidence or an event. TRACK and COMMENTATE update repeatedly as the stream changes. Tasks that wait or update hold 69.6% of interactions. The same visual task can require speech at one moment and silence at another. Our protocol scores content accuracy, timing accuracy, and silence compliance, and combines them into one Overall score. It reports wall-clock latency separately, as a property of the serving stack.

The evaluated systems show different imbalances between these obligations. Qwen3-VL-8B (Bai et al., 2025) under a polling adapter reaches 34.8 Overall, with timing accuracy of 77.8 but silence compliance of 10.9. Only 5 of 18 timed configurations finish 90% of one-second polls in time, and agents take about half a minute per poll. Negative streams and matched near misses reveal additional false alarms, even among systems that speak less often. Controlled adapter changes and offline comparisons on identical items show that offline grounding and online decisions rank systems differently, and that denser frame sampling buys little silence. We make three contributions:

- We introduce INTERACTIONBENCH, which tests when a video assistant should speak or stay silent across query responses, event triggers, and ongoing updates in 1,060 interactions.
- Our stream-time protocol separates answer content, response timing, and unnecessary speech, and tests silence directly with negative streams and matched near-miss suites.
- We evaluate native streaming systems, polled turn-based models, and agents on INTERACTIONBENCH, separate repeated speech, premature responses, false alarms, and missed follow-up events, and find that timely speech costs silence across systems.

2 EVALUATING SPEECH AND SILENCE

2.1 THE SYSTEM DECISION

A system receives a video stream v and a question or standing instruction q. At stream time t, it can access only the video prefix $v_{\leq t}$, its instruction, and its own interaction history. It chooses to remain silent or emit text y. The output is a time-ordered sequence $\hat{\mathcal{E}} = ((t_1, y_1), \ldots, (t_M, y_M))$ of M non-empty emissions, where emission i carries text y_i at stream time t_i on the video clock. The evaluated system includes its visual model, memory, sampling policy, and response controller. A polling adapter is one such controller for turn-based models. It asks the model at every stream second whether to speak at that moment, and silence is a valid answer to any poll.

Each positive item has N reference events. Event n has reference content c_n and an earliest answerable moment r_n^*, defined by the video prefix that first supplies sufficient evidence. A negative item has no event, and silence is correct throughout. As shown in Figure 1, the required response pattern depends on the task. Accurate content alone cannot establish a correct interaction, because the same answer can be premature, timely, or redundant. Case Studies in the appendix add late replies, missed responses, wrong content, and false alarms on a negative stream, all on one railway video.

2.2 Three Complementary Scores

For event-based items, event n owns the response window $W_n = [r_n^* - \epsilon, r_{n+1}^* - \epsilon)$. The last window extends to the end of the stream. We use $\epsilon = 1\,\text{s}$, so a reply in the same second as the evidence is on time. The first emission in a window is its matched response, regardless of content. A matched response at time t_i has delay $d_n = \max(0, t_i - r_n^*)$. Later emissions in the same window are redundant, and emissions before the first window are premature. Redundant and premature emissions form the violation set V. With $H(a, b) = 2ab/(a + b)$ for $a + b > 0$ and $H(0, 0) = 0$, the item scores are

$$\begin{aligned} \text{TA} &= \frac{100}{N} \sum_{n=1}^{N} \left(1 - \frac{d_n}{\Delta}\right)_+, \quad \text{SC} = 100 \left(1 - \frac{|V|}{N}\right)_+, \\ \text{Overall} &= \frac{\text{Acc} + H(\text{TA}, \text{SC})}{2}, \qquad \Delta = 5\,\text{s}. \end{aligned} \tag{1}$$

Here $(x)_+ = \max(x, 0)$, an unmatched event adds nothing to TA, and timing credit decays linearly to zero once lateness reaches the delay scale Δ. Content accuracy Acc grades what the matched response says, timing accuracy TA measures when it arrives, and silence compliance SC penalizes unnecessary emissions. The harmonic mean balances the two behavioral components, and averaging it with Acc keeps credit for accurate content when timing or silence fails. Dataset Overall averages item scores rather than combining the averaged Acc, TA, and SC.

Scoring follows the response format. Query-answer items use the first post-query response, with the query as the timing anchor. Trigger items average content scores over matched responses, while counting items use exact final-count accuracy. Free-form updates concatenate speech within each reference segment, and VDCSCORE (Chai et al., 2025) checks whether that text answers questions derived from the reference. Their Overall equals segment accuracy, including zeros for empty segments. Commentary may need several utterances per segment, and their timing and silence scores therefore serve as diagnostics. A negative item with M emissions scores $\text{Overall} = \text{SC} = 100 \max(0, 1 - M)$, and a single emission forfeits all of its credit. See Scoring Specification in the appendix for edge cases, a uniform variant, and fallback matching.

Constant speech and blanket silence both score poorly. A fully silent positive item receives zero Overall. Its silence component is undefined and drops out of the silence aggregate, and blanket silence therefore earns no positive-item credit. Silence compliance averages over negative streams and the positive items a system speaks on. Among six scripted policies, one that always speaks reaches 98.8 timing accuracy but 0.0 Overall. One that never speaks earns only the 6.5 Overall of the negative streams, and no policy exceeds 15.6 Overall, see Protocol Audit in the appendix.

2.3 Three Response Obligations

As shown in Figure 1, the six tasks differ in the evidence they need and in what each new response must contain. Look answers a query from the current view, whereas Recall draws on earlier visual evidence for a later query. Time waits for a temporal relation such as an order, duration, or scheduled condition, whereas Alert watches for a specified visual event. Track maintains an accumulated count or state across events, whereas Commentate describes the scene as it unfolds, as in sports play-by-play. Each pair shares one response obligation. Silence is required before evidence arrives, between unchanged states, and after a reported event, and negative streams and near misses test it directly. The display names map to the released labels and sub-labels, see Task Taxonomy in the appendix. In Figure 1, each task pair shares one video (U.S. Navy, 2025; Sertion, 2011; Elene B, 2021) and faces identical frames. When the water begins to boil, Alert speaks while Time waits 15 more seconds. Track reports each egg added to the bowl, whereas Commentate stays silent until the next step. See Case Studies in the appendix for the provenance and sources of all three videos.

2.4 Data and Quality Checks

We collect procedural and monitoring videos where a visible change, completed action, or scheduled step anchors a response. Automatic filtering and human screening exclude unsuitable clips, broken timelines, and answer overlays. Annotators define an observable response condition before writing the instruction, reference answer, and earliest answerable timestamp. Positive items must be solvable from the prefix, and negative items must keep their condition false throughout the video. Screening also rejects ambiguous intent, and a text-only audit revises or removes keyed questions that a language model answers without frames. As shown in Figure 2, the 812 videos span ten domains, led by sports, and average 159 s. Their 1,060 interactions comprise 626 questions and 434 instructions. The 991 positive interactions hold 3,693 reference events, and 69 are negative streams.

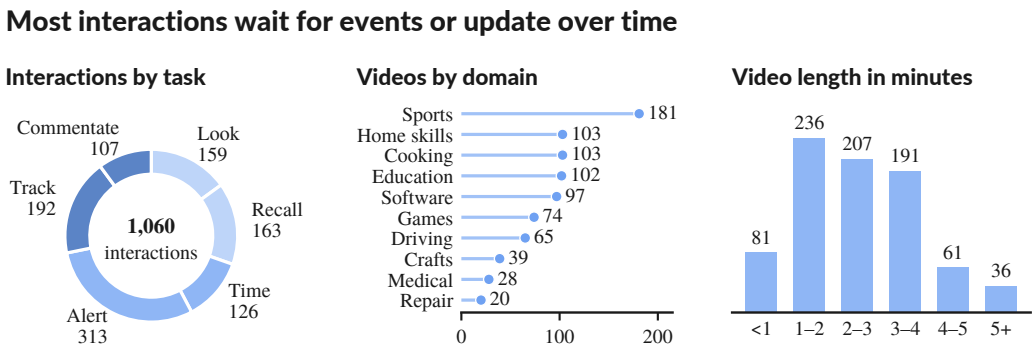


Figure 2: **Most interactions wait for events or update over time.** Left counts interactions by task, Middle videos by domain, and Right videos by length, where the last bin holds five minutes or more.

Table 1: **Only INTERACTIONBENCH pairs negative streams with near misses.** Negatives are streams that warrant no response, and graded delay lowers credit continuously with lateness.

	Dataset					Evaluation			
Benchmark	**Videos**	**Items**	**Multi-response**	**Negatives**	**Near misses**	**Proactive**	**Premature**	**Graded delay**	**Redundant**
ProactiveVideoQA	1,377	1,427	✔	✗	✗	✔	✗	✔	✗
RIVER	1,067	4,278	✔	✗	✗	✔	✔	✔	✗
OmniPro	1,262	2,700	✔	✗	✗	✔	✔	✗	✔
OmniInteract	250	1,430	✔	✗	✗	✔	✔	✔	✔
ESTP-Bench	890	2,264	✔	✗	✗	✔	✔	✔	✔
ProAssist	340	1,020	✔	✗	✗	✔	✔	✔	✔
IPIBench	1,831	3,738	✗	✗	✗	✔	✔	✗	✗
SPOT-Bench	940	5,106	✔	✗	✗	✔	✔	✔	✔
PaSBench-Video	740	740	✗	✔	✗	✔	✔	✗	✗
LiveProBench	911	18,146	✔	✗	✗	✔	✔	✗	✔
INTERACTIONBENCH	812	1,060	✔	✔	✔	✔	✔	✔	✔

The 53 matched suites test whether a system answers counted events but stays silent at near misses. Each suite pairs a positive counting item with near-miss moments on the same video, 106 in total, and a near miss resembles the counted event without completing it. For suite j, let $h_j = 1$ if its positive item receives nonzero timing credit, and let $s_j = 1$ if the system emits nothing within five seconds of every near miss. Both are 0 otherwise. The suite pass rate $\mathrm{PairSel}$ is the percentage of the 53 suites with $h_j s_j = 1$. The positive criterion is timing credit, and content accuracy is reported alongside this diagnostic. Suites cover counting streams only, see Matched Events and Near Misses in the appendix for how overlapping near-miss and response windows are scored.

Unique silence checks. As shown in Table 1, INTERACTIONBENCH is the only compared benchmark that combines whole negative streams with near misses from the same video. The comparison covers ProactiveVideoQA (Wang et al., 2025a), RIVER (Shi et al., 2026), OmniPro (Zhao et al., 2026a), OmniInteract (Lu et al., 2026), ESTP-Bench (Zhang et al., 2025b), ProAssist (Zhang et al., 2025a), IPIBench (Li et al., 2026a), SPOT-Bench (Chatterjee et al., 2026), PaSBench-Video (Zhao et al., 2026b), and LiveProBench (Du et al., 2026). The dataset columns mark whether one instruction can require several responses and whether near misses are annotated, with item counts as each benchmark defines them. The evaluation columns mark whether the system itself decides when to respond and whether premature or redundant replies are penalized. INTERACTIONBENCH penalizes premature replies as silence violations and late replies by a linear decay. Streaming question-answering benchmarks such as StreamingBench (Lin et al., 2026), OVO-Bench (Niu et al., 2025), SVBench (Yang et al., 2025b), and OmniMMI (Wang et al., 2025c) include no negative streams and are compared with these columns in a separate table, see Related Work in the appendix.

3 SPEECH AND SILENCE IN CURRENT SYSTEMS

3.1 EVALUATION SETTINGS

We group systems by the component that decides when to speak. JoyAI-VL (Yao et al., 2026) is a native real-time interaction system whose own controller and memory make that choice, and kernel swaps keep them but replace its model. Native streaming video LLMs use their released mechanisms and input rates, as in MMDuet2 (Wang et al., 2026c), VideoLLM-online (Chen et al.,

Table 2: **Systems trade timing for silence.** Acc is content accuracy, TA timing accuracy, and SC silence compliance, all higher-is-better. See System Configuration and Provenance in the appendix.

System	Overall	Look			Recall			Time			Alert			Track			Commentate		
		Acc	TA	SC	Acc	TA	SC	Acc	TA	SC	Acc	TA	SC	Acc	TA	SC	Acc	TA	SC
References																			
Qwen3-VL-8B, blind	15.7	21.7	77.5	3.6	22.4	95.5	69.1	0.0	0.7	0.0	6.2	6.3	66.7	0.0	0.0	100.0	1.0	25.4	15.5
Human reference	53.6	78.7	74.0	80.0	44.4	57.4	61.1	63.0	51.0	77.8	19.5	53.3	77.2	36.8	69.0	73.0	56.9	70.1	69.5
Native streaming																			
VideoLLM-online	14.5	2.4	94.2	88.5	6.4	81.9	87.8	1.6	8.7	3.8	3.7	7.7	18.1	6.8	5.6	52.0	0.0	7.1	64.8
LiveCC	14.5	68.8	71.1	0.0	0.0	5.6	100.0	34.6	24.5	3.2	35.8	14.5	2.6	1.7	2.2	3.0	17.0	42.0	23.9
MMDuet2	17.0	55.4	62.9	54.9	7.7	37.8	7.3	15.3	37.2	0.9	13.6	35.2	15.1	1.2	26.2	34.1	24.1	47.4	21.7
MOSS-Video	19.9	56.2	71.5	6.7	9.7	15.7	36.4	26.9	18.1	0.0	63.3	25.3	10.2	3.3	4.6	3.4	14.2	27.7	50.1
VideoChat3	28.5	62.5	56.1	87.1	6.5	3.0	100.0	65.4	11.2	10.4	72.5	9.2	21.2	25.0	11.3	28.8	18.5	22.6	61.0
Turn-based polling																			
Flash-VStream	14.3	30.7	66.2	3.1	2.6	18.0	13.9	21.4	53.1	0.0	42.7	61.3	10.7	7.4	49.4	15.1	2.2	41.6	2.7
Qwen2.5-Omni	18.7	23.5	25.3	41.5	3.2	9.6	66.7	42.9	40.0	9.0	56.0	51.9	9.4	4.9	20.3	42.6	19.4	54.2	9.8
LLaVA-OV	25.6	37.5	55.0	18.8	22.6	48.4	86.7	92.3	57.5	7.7	39.3	16.7	23.7	8.3	6.3	8.3	28.5	86.5	0.0
Qwen3-VL-4B	26.2	71.1	74.6	2.7	21.2	23.2	63.4	15.1	7.4	27.3	51.5	41.7	29.4	1.9	2.7	87.1	31.0	46.2	22.5
Gemma-3n	27.8	57.2	98.1	3.0	28.8	79.9	67.2	49.1	57.6	5.6	66.6	80.3	4.6	7.4	78.1	4.5	30.7	97.8	0.0
MiniCPM-o	29.0	75.3	87.2	3.8	44.2	86.9	68.6	15.9	5.6	16.0	15.9	12.9	51.7	8.0	20.5	36.5	53.8	81.3	3.4
LLaVA-OV2	34.5	64.5	98.9	3.0	34.6	90.5	68.3	57.1	76.3	0.8	72.3	75.3	8.7	10.2	55.6	23.2	43.0	98.2	0.0
Qwen3-VL-8B	34.8	75.3	96.6	3.0	24.4	44.1	66.2	66.7	79.6	0.0	78.7	79.1	9.8	8.6	75.3	11.7	57.1	96.3	0.0
Gemini 2.5 Flash	35.5	81.2	100.0	0.0	40.0	93.3	71.4	46.2	97.4	0.0	71.3	89.4	0.0	17.6	85.0	4.5	68.1	98.6	0.0
Claude Sonnet 5	44.6	87.5	98.8	0.0	46.7	100.0	73.3	84.6	84.4	0.0	76.0	78.7	11.1	18.5	52.1	15.9	84.1	89.0	4.0
Offline temporal grounding																			
Gemma-3n	16.1	14.5	24.0	96.4	14.1	23.7	100.0	7.1	6.2	15.9	15.0	7.7	39.2	6.8	11.0	49.4	4.2	17.9	57.9
Dispider	24.8	75.0	93.8	93.8	61.3	100.0	100.0	38.5	0.0	52.6	18.5	0.8	62.7	5.0	1.2	74.6	1.6	0.9	95.5
Qwen2.5-Omni	27.2	44.0	69.9	100.0	47.4	78.8	100.0	5.6	2.8	39.8	11.7	2.0	59.3	0.6	1.2	88.4	0.0	0.0	8.3
MiniCPM-o	30.6	59.6	95.7	96.6	66.7	100.0	100.0	8.7	8.4	6.5	11.4	3.4	36.4	8.0	11.9	32.6	2.6	8.8	46.3
Qwen3-VL-4B	33.8	63.9	96.3	96.5	65.4	99.4	100.0	7.1	4.5	9.3	15.3	3.5	35.8	3.1	9.5	46.6	0.8	11.9	35.3
LLaVA-OV2	33.9	64.5	96.3	97.6	60.9	99.4	100.0	7.1	5.8	19.4	11.0	4.3	38.6	7.4	8.3	63.1	3.8	5.7	79.7
Qwen3-VL-8B	35.7	63.9	96.9	97.9	67.9	100.0	100.0	6.3	2.1	11.7	13.1	4.0	40.9	10.5	13.4	48.2	4.7	19.1	50.9
Claude Sonnet 5	38.6	69.3	89.7	97.9	65.4	87.8	100.0	29.4	10.5	85.5	35.7	5.6	81.0	0.0	0.8	100.0	5.1	6.5	79.8
GPT-5.6 Sol agent	58.8	89.2	95.8	96.4	91.7	100.0	100.0	57.9	48.4	65.9	52.3	39.2	79.3	21.7	33.2	72.8	18.7	38.9	55.5
Claude Sonnet 5 agent	66.1	90.4	97.4	97.6	92.3	100.0	100.0	65.9	59.7	76.6	57.2	50.0	81.5	35.1	50.9	83.9	29.2	80.7	45.1
Native Real-time Interaction System (JoyAI-VL)																			
Claude Opus 5	30.5	81.2	85.0	13.4	20.0	26.7	100.0	23.1	20.6	42.9	45.6	34.9	43.8	17.3	28.2	24.1	48.6	46.3	15.0
Qwen3-VL-8B	32.4	75.0	98.8	0.0	73.3	80.0	66.7	7.7	90.5	0.0	66.2	97.9	3.7	0.0	76.2	8.4	70.0	97.8	0.0
JoyAI-VL	33.8	59.0	96.5	15.7	71.8	97.4	73.0	41.4	55.8	10.3	52.7	38.5	29.1	3.2	58.7	22.4	38.4	73.7	21.5
Qwen3-VL-4B	35.5	68.8	68.2	0.0	38.7	60.0	94.7	69.2	53.8	0.0	85.1	65.0	1.2	3.3	7.9	0.0	37.6	77.4	0.8
Claude Sonnet 5	43.3	79.7	94.4	7.3	84.1	91.5	66.7	70.0	70.0	7.1	63.4	64.1	37.0	19.6	60.4	10.4	55.8	84.0	6.3
Agentic Interaction System (Qwen-MM-Plugins)																			
Gemini 3.5 Flash	29.4	62.5	48.8	80.0	46.7	44.0	87.5	30.8	23.1	40.0	20.8	12.0	62.7	0.0	4.8	75.7	12.0	10.7	32.1
Claude Sonnet 5	56.0	87.5	93.8	68.8	86.7	93.3	100.0	50.0	31.7	54.2	49.4	55.8	47.5	8.0	46.0	47.4	69.2	89.6	15.0
GPT-5.6 Sol	59.1	93.8	92.5	93.8	86.7	93.3	100.0	30.8	28.5	38.5	59.9	50.7	69.6	0.3	70.4	30.7	69.4	87.9	16.5
Agentic Interaction System (Claude Code tools)																			
Claude Sonnet 5	61.2	93.8	90.0	62.5	93.3	100.0	86.7	53.8	54.9	58.3	69.4	56.3	50.9	14.3	53.1	54.5	77.9	83.7	8.6

2024), VideoChat3 (Li et al., 2026b), MOSS-Video (Wang et al., 2026b), and LiveCC (Chen et al., 2025). Turn-based models answer one prompt at a time, and a polling adapter asks them at every stream second whether to speak. It gives each model a sliding window of 16 frames at 2 fps and a 512-pixel long side, with a fresh prompt at every poll. The turn-based models are Qwen3-VL-8B and Qwen3-VL-4B (Bai et al., 2025), LLaVA-OV2-8B (An et al., 2026), LLaVA-OV (Li et al., 2024), Gemma-3n-E4B (Gemma Team, 2025), MiniCPM-o-4.5 (Cui et al., 2026), Qwen2.5-Omni-7B (Xu et al., 2025), Claude Sonnet 5 (Anthropic, 2026a), and Gemini 2.5 Flash (Google DeepMind, 2026). Flash-VStream (Zhang et al., 2024) is a streaming video LLM without a response decision of its own, and the same adapter polls its streaming memory. The agentic interaction system wraps a turn-based model in a tool-using loop with Qwen-MM-Plugins (Qwen Team, 2026) at each poll. One Claude Sonnet 5 run uses Claude Code tools instead, see Agentic System Details in the appendix.

Offline temporal grounding gives a system the whole video and asks when and what it would respond. The six open backbones, Claude Sonnet 5, and the Claude and GPT-5.6 Sol (OpenAI, 2026)

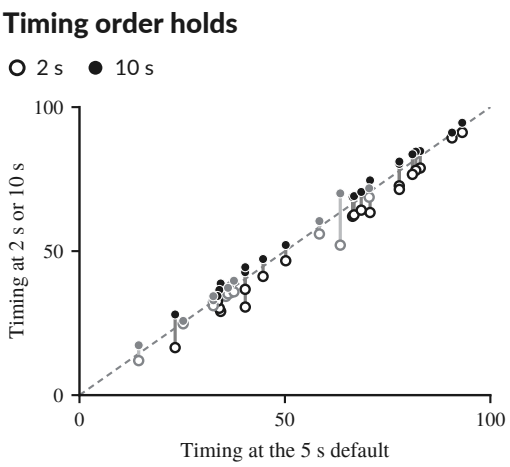


Figure 3: **Timing order holds.** Open marks 2 s, filled marks 10 s.

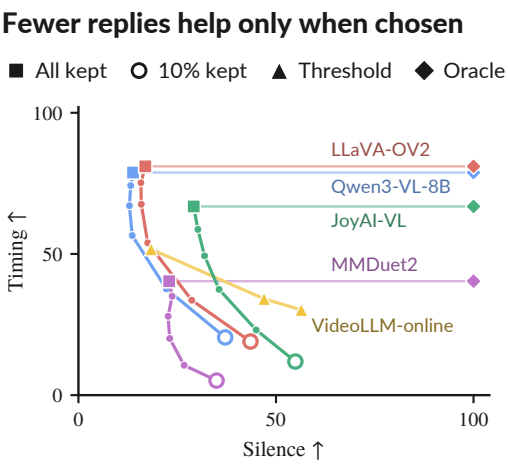


Figure 4: **Fewer replies help only when chosen.** Thresholds use the agent items.

agents run in this mode. Grounding thus has future access, whereas online systems see only the prefix. Dispider (Qian et al., 2025) releases only whole-video inference and runs in this mode as well. Two references anchor the comparison. A blind Qwen3-VL-8B run sees only the instruction, and volunteers answer during live playback. An option-aware judge scores their free-text answers. Current runs open each response window one second before its reference time. Keyed answers are matched to options, counting items require the exact final count, and a Qwen3-14B judge (Yang et al., 2025a) grades free-text responses. We calibrated seven candidate judges on 954 question–answer triples. A Gemini judge (Google DeepMind, 2026) labeled each triple first, and annotators screened these labels and corrected wrong ones. Greedy Qwen3-14B agrees with 89.7% of the final labels at a Cohen's kappa (Cohen, 1960) of 0.79, on par with Qwen3-32B and gpt-oss-20b (Agarwal et al., 2025) at 90.0% and 89.9%. We chose it to balance reliability and efficiency. A five-vote majority adds only 0.1 point, and one decoding pass therefore scores free-text answers and segments, see Judge and Segment Scoring in the appendix for all seven candidates. Scores use the video clock, and every poll also records its wall-clock time. On B200 hardware, JoyAI-VL, Qwen3-VL-8B, Qwen3-VL-4B, and LLaVA-OV2 keep pace with a live stream and finish at least 94.8% of polls within the one-second interval. Gemma-3n, MiniCPM-o, Qwen2.5-Omni, Flash-VStream, and the polled API models finish at most 6.0%, see Measured Latency in the appendix.

3.2 Timely Speech Costs Silence

As shown in Table 2, systems that time their replies well rarely stay silent. Each family is one evaluation setting, and each row names its model. Rows ascend by Overall within a family. The video clock waits for every poll, and scores describe decisions rather than live deployment. Replaying the polled runs at their measured latency lowers timing accuracy by up to 41.5 points for the API models. Models that keep pace lose at most 7.6, see Measured Latency in the appendix.

Tasks separate answering from restraint. On Look, Qwen3-VL-8B (Bai et al., 2025) reaches 75.3 content accuracy and 96.6 timing accuracy, but only 3.0 silence compliance. LLaVA-OV2 (An et al., 2026) shows the same imbalance, with 64.5 content, 98.9 timing, and 3.0 silence. Both answer promptly but keep speaking. Recall produces a different profile. JoyAI-VL (Yao et al., 2026) leads Qwen3-VL-8B in content accuracy, 71.8 to 24.4, and in silence compliance, 73.0 to 66.2. One aggregate score hides this gap between current and remembered evidence. On Time, Qwen3-VL-8B reaches 79.6 timing accuracy with 0.0 silence compliance, and on Alert, 79.1 with 9.8. Reaching an event window does not imply that the system waited for it or stopped afterward. Track also demands a correct count. Qwen3-VL-8B pairs 75.3 timing accuracy with 8.6 content accuracy, and JoyAI-VL 58.7 with 3.2. On Commentate, polled Claude Sonnet 5 (Anthropic, 2026a) reaches 84.1 content accuracy but only 4.0 silence compliance. Free-form Overall uses segment content, and only the silence diagnostic reveals that its commentary rarely pauses.

Systems fail in opposite directions. Frequent responses hit event windows, but each extra utterance lowers silence compliance. It reaches its floor of zero on 72% to 97% of positive items for polled models. Their repeated replies per event still range from 2.8 to 19.2, see Detailed System Results in the appendix. Among the six open turn-based models, the leading one, Qwen3-VL-8B, reaches 77.8 timing accuracy but only 10.9 silence compliance. LLaVA-OV2 has a similar Overall

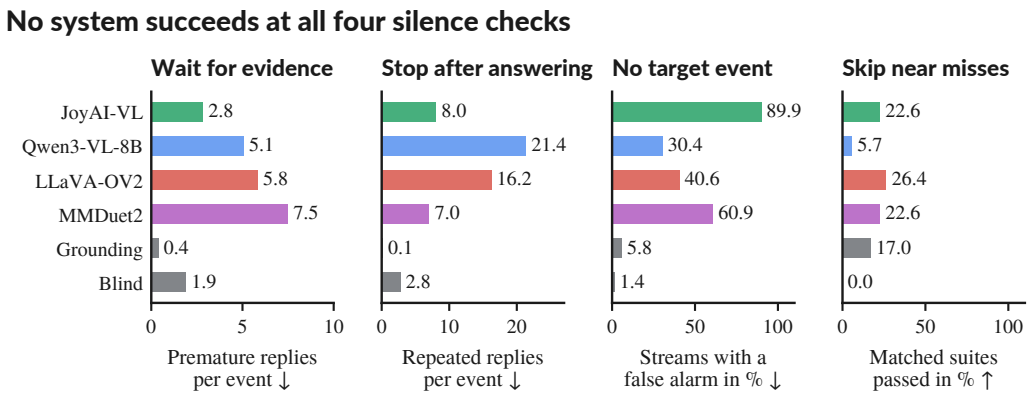


Figure 5: **No system passes all four silence checks.** Qwen3-VL-8B is polled with a 64-frame window, and the gray rows are its grounding run and the blind reference.

of 34.5, with timing accuracy 81.0 and silence compliance 16.9, and its predecessor LLaVA-OV (Li et al., 2024) scores 25.6. Flash-VStream (Zhang et al., 2024) shows the same imbalance at a lower level, with 50.2 timing accuracy and 8.2 silence compliance. Speaking less changes the failure profile. JoyAI-VL scores 33.8 Overall, with timing accuracy 66.8 and silence compliance 29.2. VideoLLM-online (Chen et al., 2024) raises silence compliance to 47.9 but reduces timing accuracy to 33.6 and content accuracy to 3.7. VideoChat3 (Li et al., 2026b) leads the native streaming video LLMs at 28.5 Overall. It rarely answers Recall queries, with 3.0 timing accuracy there. MOSS-Video (Wang et al., 2026b) scores 19.9 Overall and MMDuet2 (Wang et al., 2026c) 17.0. LiveCC (Chen et al., 2025) comments on every clip and stays at or below 3.2 silence compliance on Look, Time, Alert, and Track. The human reference reaches 53.6 Overall with 74.7 silence compliance. Every configuration above it sees the whole video or is an agent whose logged polls take about half a minute. **Neither frequent speech nor broad restraint satisfies both obligations.**

The trade-off survives other settings. Rescoring every current configuration with delay scales from 2 to 10 seconds keeps the Overall ranking at a Kendall τ (Kendall et al., 1938) of at least 0.95 against the five-second default. Overall moves by at most 3.6 points for any configuration. As shown in Figure 3, timing accuracy moves by up to 18.0 points, for the human reference between the 2 and 10 second scales, and the timing order holds. One-second shifts or one-second Gaussian jitter of every event timestamp after the query keep the Overall ranking at a Kendall τ of at least 0.91, with free-text content held fixed. Deleting replies tests whether fewer utterances alone improve selectivity. As shown in Figure 4, lowering the keep probability from 1.0 to 0.1 raises silence compliance at the cost of missed events for all four traced systems. Random deletion removes correct and redundant replies alike and cannot raise both scores. The response threshold of VideoLLM-online (Chen et al., 2024) moves along the same trade-off. Raising it from 0.5 to 0.9 lifts timing accuracy from 30.2 to 51.7 while silence compliance falls from 56.4 to 18.4. Deleting only the premature, repeated, and negative-stream replies instead lifts silence compliance to 100.0 at unchanged timing accuracy for the same four systems. The agent items form one fixed subset that every agent run shares. On them, this oracle filter raises Overall by 32.3 points for Qwen3-VL-8B and 25.5 for JoyAI-VL, without rewriting any retained reply, see Additional Diagnostics in the appendix. **Fewer replies help only when the system chooses which ones to withhold, and none of the tested controls, random reply deletion or a native response threshold, makes that selective choice.**

3.3 Four Failures of Selectivity

As shown in Figure 5, repetition and premature speech dominate different systems. Polled with a 64-frame window, Qwen3-VL-8B (Bai et al., 2025) produces 21.4 redundant emissions per reference event. Its fresh prompt has no record of its previous reply, and while the evidence stays visible, each poll can repeat the same answer. MMDuet2 (Wang et al., 2026c) instead produces 7.5 premature emissions per reference event, exceeding its redundant count. A familiar scene alone does not establish the requested condition, and the system must keep waiting until it holds. The same answer content can conceal opposite failures of response control. In a saved Alert item with four green-light events, both systems always choose the correct option. Qwen3-VL-8B produces 104 emissions, including five premature replies, four matched responses, and 95 repetitions. JoyAI-VL (Yao et al., 2026) emits three replies, one premature, one matched, and one repeated, and misses three of the four events, see Case Studies in the appendix for the full traces of both systems.

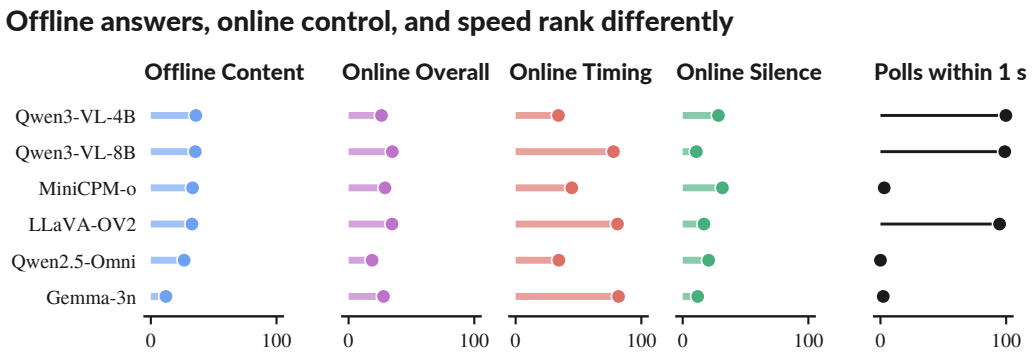


Figure 6: **Offline content, online scores, and speed disagree.** Six backbones share one polling adapter, and offline content counts unanswered keyed items as wrong.

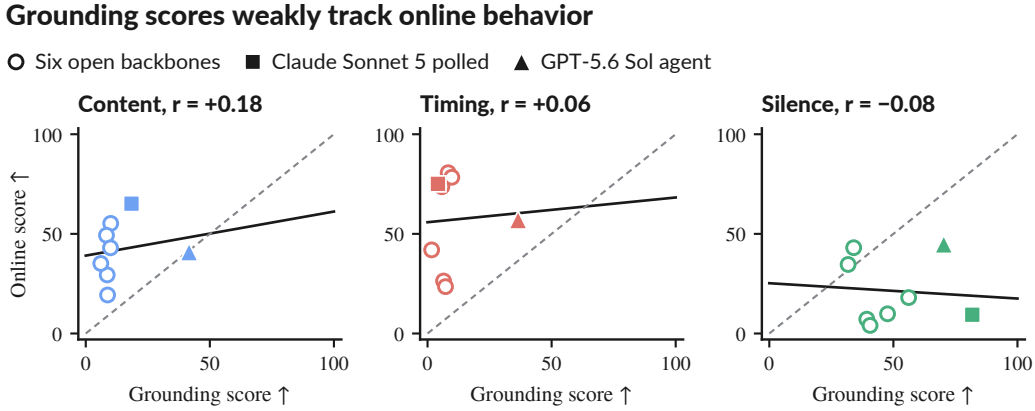


Figure 7: **Grounding and online scores diverge.** Both scores use identical Time, Alert, Track, and Commentate items. Solid lines are least-squares fits, and r is the Pearson correlation.

The 69 negative streams contain no valid trigger, yet JoyAI-VL emits on 89.9% of them. Qwen3-VL-8B in offline grounding, with future access, emits on 5.8%. The blind reference raises false alarms on only 1.4% of streams but fails every suite. On these suites, LLaVA-OV2 (An et al., 2026) passes 26.4% and JoyAI-VL 22.6%, while the polled Qwen3-VL models pass 5.7% to 13.2%. No open-weight system clears a third of them. The two checks capture different failures, as Qwen3-VL-8B raises the fewest false alarms of the four systems yet passes fewer suites than a scripted burst policy. Each suite takes its event and near-miss moments from one video, so the scene alone does not tell a system when to answer and when to stay silent. **Selective interaction requires waiting for evidence, responding once, and rejecting both absent events and near misses. None of the systems in Figure 5 manages all four at once on Alert and Track streams.**

3.4 Online Control Needs Direct Measurement

Offline scores mispredict online behavior. As shown in Figure 6, offline answer accuracy, online scores, and poll speed rank the six open turn-based models differently under the same adapter. Qwen3-VL-4B (Bai et al., 2025) leads offline accuracy but ranks fifth in online Overall. Gemma-3n (Gemma Team, 2025) ranks last offline, leaving 43.8% of keyed items unanswered in grounding, yet first in online timing accuracy. See Offline Accuracy of the Polled Backbones in the appendix for all answer rates. **An offline content score alone misses the evaluated system's timing and silence.** The blind Qwen3-VL-8B baseline scores 15.7 Overall without frames, the floor for every configuration that sees them. Grounding also differs from online use. As shown in Figure 7, grounding and online scores of eight configurations on identical items correlate at 0.18 for content, 0.06 for timing, and −0.08 for silence. Per backbone, offline and online keyed answers agree at a Cohen's κ (Cohen, 1960) of at most 0.40. Online decisions need direct measurement on the stream clock.

3.5 The Response Mechanism Moves Silence

Inputs and controllers barely help. As shown in Figure 8, the way a system decides when to speak moves silence compliance more than any input setting. Under polling with a fresh prompt, the group median reaches 64.0 timing accuracy but only 17.2 silence compliance, and more frames

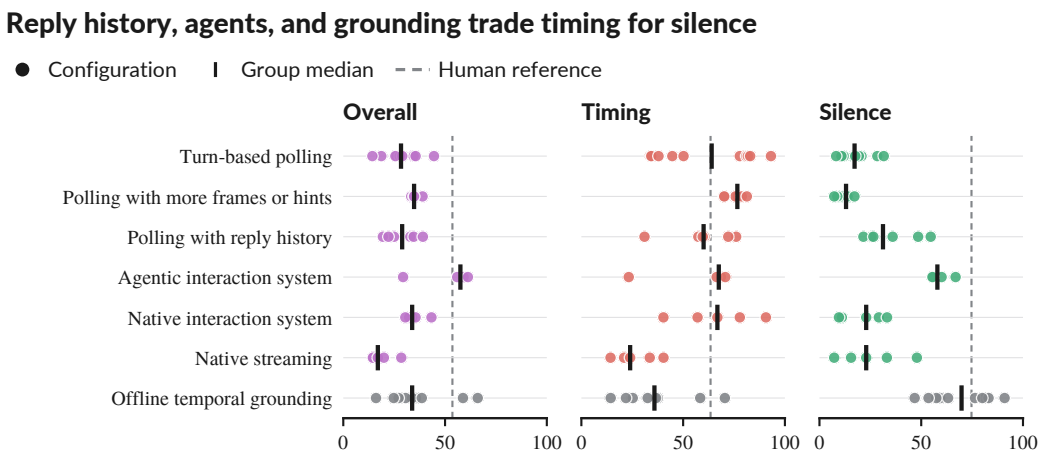


Figure 8: **Reply history, agents, and grounding trade timing for silence.** Each dot is one configuration, grouped by how the system decides when to speak, and ticks mark group medians.

or new hints leave this profile in place. Across 0.5–4 fps, Qwen3-VL-8B (Bai et al., 2025) Overall ranges from 33.4 to 35.4 and timing accuracy from 75.7 to 77.8. Widening its window from 16 to 256 frames raises Overall from 34.8 to 39.0 and silence compliance from 10.9 to 17.2. The median poll then slows from 0.48 to 4.1 seconds. Two hint rewrites, Hint 2 and Hint 3, move Overall by less than one point. They trade roughly five points of timing accuracy for no more than three of silence compliance. The native interaction system supplies little restraint by itself. On the agent items, JoyAI-VL (Yao et al., 2026) reaches 25.9 silence compliance, and with Qwen3-VL-8B inside its controller, silence compliance falls to 11.0, close to the 5.9 of the same model under polling. With Claude Opus 5 (Anthropic, 2026b) inside, it rises to 33.2 only by leaving 41% of positive items unanswered. Inside the same controller, Qwen3-VL-4B reaches 9.6 and Claude Sonnet 5 (Anthropic, 2026a) 23.0. **Within this controller, restraint therefore tracks the swapped-in model.**

Memory and foresight cost timing. Reply history, agent loops, and grounding all raise silence, and each pays in timing, speed, or stability. Reply history keeps earlier frames and replies in one conversation, whereas a fresh prompt starts anew each poll. It raises silence compliance for all six open turn-based models, which suggests that much repetition comes from forgetting the previous reply. Its effect on Overall depends on the backbone, as three models gain and three lose, and a decoding change confounds the MiniCPM-o (Cui et al., 2026) gain. Agent loops make restraint deliberate. On the agent items, the Claude Sonnet 5 agents reply about a quarter as often as the polled model and roughly triple its silence compliance at some cost in timing. Yet a logged agent poll takes about half a minute, far beyond the one-second interval, see Measured Latency in the appendix. Swapping the plugins for Claude Code tools alone moves the Claude agent's Overall by 5.2 points, and under the same plugins GPT-5.6 Sol (OpenAI, 2026) scores twice as high as Gemini 3.5 Flash (Google DeepMind, 2026). Grounding shows that foresight alone does not buy timing. With the whole video, Qwen3-VL-8B raises its silence compliance more than fivefold but halves its timing accuracy, and Claude Sonnet 5 shows the same shift. Query responses stay nearly immediate, but later events are hard to locate in a finished video without tools that inspect chosen frames. Dispider (Qian et al., 2025) earns almost no timing credit on them. With such tools, the Claude Sonnet 5 agent in grounding is the only configuration above 70 in both timing and silence. **An agentic system is evaluated with its loop and tools, not by its model alone.**

4 Conclusion and Discussion

InteractionBench evaluates whether a video assistant speaks when its instruction warrants a response and stays silent otherwise. High timing scores coexist with frequent unnecessary speech, restrained output misses events, and both behaviors belong to the complete system. The tested native controller adds little restraint by itself. Reply history and agentic systems add it, and logged agent polls take about half a minute, far slower than the one-second poll interval. Future systems should retain an instruction, recognize new evidence, respond once, and resume monitoring. Progress reports should cover both obligations on matched items, speech when due and silence otherwise, with runtime measured separately. On the Alert and Track streams that carry them, negative streams and near-miss suites separate systems whose aggregate scores look similar.

AI USE STATEMENT

Generative models served as the open judge for free-text items and as the first labeler of the judge-calibration triples before human screening and correction. They labeled content domains for the statistics and served as the blind auditor of the keyed pool. They also assisted with evaluation, analysis, and figure code, with prose editing, and with checking bibliographic metadata against arXiv and publisher records. Dataset annotation remains a human responsibility. This draft uses AI-assisted editing and figure generation, and the authors verify the final text, analyses, and submission.

ETHICS STATEMENT

All videos are public and referenced by id, faces are not annotated, and any video is removed from the release at the request of its uploader. Two domains carry risk. A trigger such as "tell me when a person appears" can serve surveillance. The first-aid and driving items can invite over-reliance on a system that this paper shows to be far from reliable. The human reference was collected from volunteers who consented to the task and could withdraw at any time.

REPRODUCIBILITY STATEMENT

Appendices F and H document the release format and scoring rules. Appendices K, L, and R record configurations, run dates, and the measured latency of every run that logged it.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Anthropic. Claude Sonnet 5 model documentation. `https://platform.claude.com/docs/en/models/sonnet-5/overview`, 2026a. Model Claude Sonnet 5, API model ID claude-sonnet-5, accessed August and September 2026.

Anthropic. Claude Opus 5 model documentation. `https://platform.claude.com/docs/en/models/opus-5/overview`, 2026b. API model ID claude-opus-5, accessed September 2026.

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.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 16845–16889, 2025.

Dibyadip Chatterjee, Zhanzhong Pang, Fadime Sener, Yale Song, and Angela Yao. Don't pause! every prediction matters in a streaming video. *arXiv preprint arXiv:2604.24317*, 2026.

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, 2024.

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.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Kaixuan Du, Xin Wan, YuKun Wang, Hang Zhang, Meng Cao, Dai Guan, Ming Chen, and Ni Li. Proactivebench: Can streaming video models really interact like humans? *arXiv preprint arXiv:2609.12658*, 2026.

Elene B. Patata tortilla. Wikimedia Commons, video, 2021. URL `https://commons.wikimedia.org/wiki/File:Patata_tortilla.webm`. CC BY-SA 4.0.

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

Gemma Team. Gemma 3n. `https://ai.google.dev/gemma/docs/gemma-3n`, 2025. Google DeepMind. Accessed: 2026-09-25.

Google DeepMind. Gemini model documentation. `https://ai.google.dev/gemini-api/docs/models`, 2026. Models Gemini 3.5 Flash and Gemini 2.5 Flash, accessed August and September 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

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*, 2026a.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

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.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110, 2024.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, et al. Omniinteract: Benchmarking real-world streaming interaction for real-time omnimodal assistants. *arXiv preprint arXiv:2605.26485*, 2026.

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.

OpenAI. GPT-5.6 model documentation. https://developers.openai.com/api/docs/models/gpt-5.6, 2026. API alias gpt-5.6, which routes to GPT-5.6 Sol, accessed August and September 2026.

Siddhant Panpatil, Arth Singh, Mijin Koo, Chaeyun Kim, Haon Park, and Dasol Choi. Egosafety-bench: A diagnostic egocentric video benchmark for evaluating embodied vlms as runtime safety guards. *arXiv preprint arXiv:2607.00218*, 2026.

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.

Qwen Team. Qwen-mm-plugins: Make any agent harness multimodal-native, 2026. URL https://github.com/QwenLM/Qwen-MM-Plugins.

Sertion. Boiling water, close up, short shutter speed. Wikimedia Commons, video, 2011. URL https://commons.wikimedia.org/wiki/File:Boiling_water_-_close_up_(short_shutter_speed).ogv. CC BY-SA 1.0.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

U.S. Navy. NSF Diego Garcia hosts basketball game. Wikimedia Commons, video by Latifah Choice, DVIDS 989713, 2025. URL https://commons.wikimedia.org/wiki/File:NSF_Diego_Garcia_hosts_basketball_game_(989713).webm. Public domain.

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. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026a.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Wei Huang, Qirui Zhou, Zhan Huang, Zhen Ye, Jijun Cheng, Xiaomeng Qian, Yanxin Chen, et al. Moss-video-preview: Toward real-time video understanding via cross-attention. *arXiv preprint arXiv:2606.07639*, 2026b.

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, 2026c.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omni-immi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18925–18935. IEEE, 2025c.

WeenyBeany. Railway crossing in action with train passing through. Wikimedia Commons, video, 2024. URL https://commons.wikimedia.org/wiki/File:Railway_Crossing_In_Action_with_Train_Passing_Through.webm. CC BY 4.0.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations*, volume 2025, pp. 69332–69351, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 42522–42556, 2025b.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. *Advances in Neural Information Processing Systems*, 38:31266–31304, 2026.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

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.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025a.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-llm for streaming video. *arXiv preprint arXiv:2510.14560*, 2025b.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing LYU, and Xirong Li. Omnipro: A comprehensive benchmark for omni-proactive streaming video understanding. *arXiv preprint arXiv:2605.18577*, 2026a.

Yusong Zhao, Yuejin Xie, Youliang Yuan, Junjie Hu, Jitian Guo, Yujiu Yang, and Pinjia He. Pasbench-video: A streaming video benchmark for proactive safety warning. *arXiv preprint arXiv:2606.02443*, 2026b.

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

APPENDIX

This appendix holds the related work, the data and evaluation details, the full tables behind the analysis figures, and the limitations.

- Appendix A covers related streaming video systems and benchmarks.
- Appendix B groups the six tasks by response obligation and gives their released labels.
- Appendix C gives counts by task, event spans, video durations, upload years, and domains.
- Appendix D describes collection, screening, annotation, and negative streams.
- Appendix E gives an example of each task.
- Appendix F lists the released files and what a reproduction must retain.
- Appendix G defines the near-miss suites and their pass rate.
- Appendix H specifies matching and scoring, alternative aggregations, and timing tolerance.
- Appendix I reports judge calibration and free-form segment scoring.
- Appendix J reproduces the evaluation prompts.
- Appendix K records the configuration and provenance of every system and the human reference.
- Appendix L reports wall-clock latency per poll and live replays.
- Appendix M gives the offline accuracy and answer rates of the polled backbones.
- Appendix N audits the protocol with scripted policies.
- Appendix O gives aggregate and per-task scores, violations per event, and ablations.
- Appendix P rescores streams after random emission thinning.
- Appendix Q shows the source videos and full traces of the case studies.
- Appendix R details the tools, isolation, and runs of the agentic systems.
- Appendix S covers oracle restraint and the native response threshold.
- Appendix T states limitations and future directions.

A RELATED WORK

Streaming video systems. VideoLLM-online, Flash-VStream, and MMDuet model responses to incoming video rather than a completed clip (Chen et al., 2024; Zhang et al., 2024; Wang et al., 2025b). Dispider separates perception, decision, and reaction (Qian et al., 2025), and StreamChat adds a hierarchical memory for multi-round interaction (Xiong et al., 2025). StreamBridge attaches a light activation model that decides when an offline video model should reply (Wang et al., 2026a), and LiveStar chooses between a response and silence in one verification pass (Yang et al., 2026). JoyAI-VL wraps its model in an interaction system with memory and turn management (Yao et al., 2026). These designs place response control in different components. INTERACTIONBENCH evaluates their emitted behavior under a common instruction and video clock, treating the controller and memory as part of the system.

Offline and streaming understanding. As shown in Table 3, no streaming question-answering benchmark includes negative streams, and only OVO-Bench (Niu et al., 2025) tests near misses. Video-MME and MLVU emphasize answers about completed videos (Fu et al., 2025; Zhou et al., 2025), and E.T. Bench scores event-level timestamps offline (Liu et al., 2024). StreamingBench and OVO-Bench introduce streaming settings and time-sensitive questions (Lin et al., 2026), and SVBench chains temporally linked questions over multi-turn dialogues (Yang et al., 2025b).

Table 3: **Streaming question-answering benchmarks.** Columns follow the proactive comparison in the main text, and every mark is a binary yes or no.

	Dataset					Evaluation			
Benchmark	**Videos**	**Items**	**Multi-response**	**Negatives**	**Near misses**	**Proactive**	**Premature**	**Graded delay**	**Redundant**
StreamingBench	900	4,500	✗	✗	✗	✔	✔	✗	✗
OVO-Bench	644	2,814	✗	✗	✔	✔	✔	✔	✗
SVBench	1,353	49,979	✗	✗	✗	✗	✗	✗	✗
OmniMMI	1,121	2,290	✗	✗	✗	✔	✔	✗	✗
INTERACTIONBENCH	812	1,060	✔	✔	✔	✔	✔	✔	✔

Table 4: **Six tasks, three response obligations.** Counts follow annotation labels. A system must stay silent whenever a new response is unwarranted.

Task	**Items**	**Speak when…**	**Stay silent when…**
Look	159	The user asks about the current view.	The query has already been answered.
Recall	163	The user asks about earlier evidence.	No new query or update calls for speech.
Time	126	An order or temporal condition resolves.	The required temporal evidence is incomplete.
Alert	313	The requested event occurs.	The scene is only a near miss.
Track	192	A count or state needs updating.	The tracked state has not changed.
Commentate	107	The unfolding scene calls for live commentary.	There is no relevant update to report.

Proactive and interactive evaluation. OmniMMI adds proactive alerting and a turn-taking task that credits silence on irrelevant turns within a stream rather than on whole streams (Wang et al., 2025c). RIVER covers memory, live perception, and anticipation (Shi et al., 2026). ProactiveVideoQA scores free-running responses with a temporal metric (Wang et al., 2025a), and OmniPro and OmniInteract extend proactive evaluation to audiovisual streams, response slots, and interruption handling (Zhao et al., 2026a; Lu et al., 2026). ESTP-Bench and ProAssist evaluate egocentric proactive assistance and penalize replies outside answer windows (Zhang et al., 2025b;a). IPIBench scores per-second triggering within a local window around each event (Li et al., 2026a), and SPOT-Bench penalizes premature, late, and redundant outputs (Chatterjee et al., 2026). LiveProBench scores per-second decisions under standing requests (Du et al., 2026). Safety benchmarks also test restraint. PaSBench-Video includes safe videos whose correct output is silence (Zhao et al., 2026b), and EgoSafetyBench contrasts scenes that differ in one critical cue (Panpatil et al., 2026).

Positioning. INTERACTIONBENCH brings these elements together for general assistant instructions. It scores content, timing, and silence together on the video clock. It also includes whole streams in which the requested event never occurs and pairs counted events with near misses from the same video. It also compares native systems, polling adapters, and agents with measured wall-clock latency.

B TASK TAXONOMY

Table 4 groups the six tasks by the response obligation their names describe. The released labels remain IVQA for Look, LVM for Recall, TOA for Time, PTR for Alert, CST for Track, and LCG for Commentate. Causal reasoning and belief revision are sub-labels, not additional headline tasks. Silence is a valid system action across all six categories. It is the only valid action on negative items whose requested condition never occurs. Annotation counts and evaluator groupings serve different purposes. The annotation taxonomy places seven causal items under Recall, giving 159 Look and 163 Recall items. The current score export merges these seven items into IVQA, giving 166 Look and 156 Recall items in per-task result tables. Both conventions include the same 1,060 interactions. One revision item is included in Track, giving 192 items. Renaming a display category does not relabel or rescore any item.

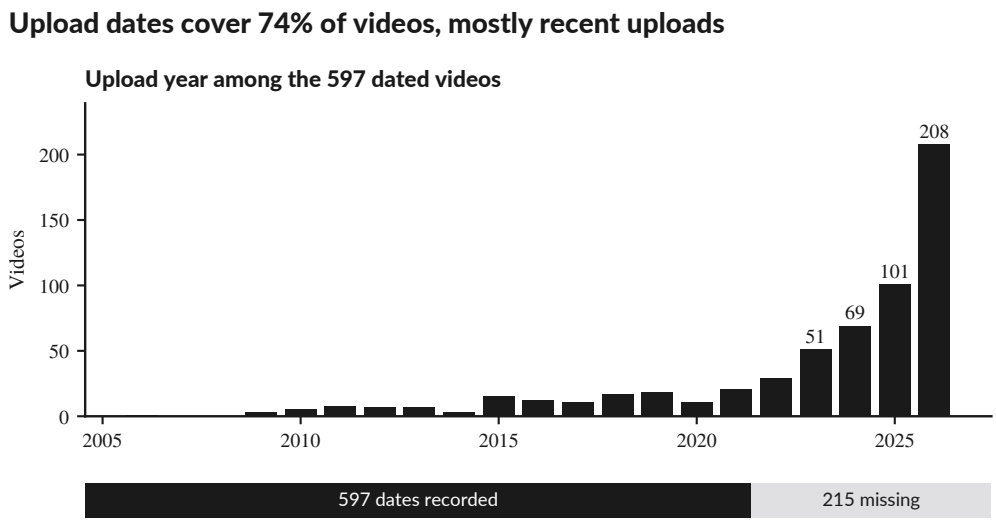


Figure 9: **Half of the dated videos appeared in 2025 or later.** Upload years and date availability for the 812 videos.

Table 5: **The ten content domains of INTERACTIONBENCH.** Videos and interactions per domain, with the share of videos.

Domain	Videos	Interactions	Share %
Sports and exercise coaching	181	254	22.3
Cooking and kitchen	103	127	12.7
Home and life skills	103	115	12.7
Education and children	102	185	12.6
Professional, office and accessibility	97	104	11.9
Games, AR and entertainment	74	85	9.1
Driving, traffic and mobility	65	76	8.0
Crafts, creation and skills	39	46	4.8
Medical, first-aid and care	28	44	3.4
Industrial, repair and operations	20	24	2.5
Total	812	1,060	100.0

C DATASET COVERAGE

The corpus contains 812 videos and 1,060 interactions. There are 159 Look, 163 Recall, 126 Time, 313 Alert, 192 Track, and 107 Commentate items under the annotation taxonomy. Questions account for 626 interactions and instructions for 434. The 991 positive items contain 3,693 reference events. Their event spans are under one minute for 690 items, one to two minutes for 154, two to five minutes for 138, and over five minutes for 9. Video durations range from 15 to 941 seconds, with a mean of 159 seconds. As shown in Figure 9, 309 of the 597 videos with upload dates were uploaded in 2025 or later. Domain labels use title and description metadata, followed by author review. As shown in Table 5, sports and exercise coaching is the largest domain, with 22.3% of the videos.

D CURATION AND ANNOTATION

Annotators collect procedural and monitoring videos and screen them for observable response conditions. Automatic filters examine duration, resolution, views, and upload dates. Human screening excludes unsuitable clips, compilations that reset the timeline, and overlays that reveal the answer. For each positive interaction, the annotation specifies its instruction or question, query time, task label, response format, and reference events. Each event has content and an earliest answerable timestamp. An online solvability check asks whether the prefix contains enough evidence at that timestamp and whether earlier prefixes do not. A held-out annotator checks the decisive visual

evidence without the target timestamp. Scoring Specification tests timestamp noise under Timing tolerance.

The 69 negative interactions include 39 Alert and 30 Track items. Their instruction predicates remain false throughout the video. A text-only audit screens keyed questions for language-only solutions. Premature-state distractors distinguish the requested event from an earlier scene.

E Task Examples

Look and Recall answer a query. Look asks which lane a swimmer occupies when the starting list is visible, and Recall asks about a distance demonstrated earlier in a driving lesson. Both require an answer at the query, drawing on current or historical visual evidence, and neither licenses repeated copies of it. Time and Alert wait for evidence. Time includes deciding which of two windows opens first, and the response becomes due when the required order can be established. Alert asks the system to report when a traffic light turns green, and a standing instruction can require several such reports, separated by intervals of silence. Track and Commentate update as needed. Track maintains a running count of appearances or revises a state when new evidence arrives, and Commentate reports successive on-screen headings or procedural steps. The requested update frequency follows the reference events or segments rather than an instruction to speak at every frame.

F Release Format

The full collection is the evaluation set, and there is no training split. Annotations reside in `results/<channel>/<video_id>/annotation.json`. Each file records the video duration and interactions, including the query, query time, task and time types, negative flag, and timestamped references. Video files, metadata, and multiple-choice keys accompany the annotations. Metadata records source identifiers, URLs, upload dates when available, and decoding properties. Predictions contain one JSON line per interaction, identified by video and item index. Each prediction lists emissions as a stream timestamp and text. Optional fields record poll counts and latencies. A missing prediction line is treated as an empty emission list. The evaluator produces per-item scores and aggregates. Reproductions must retain the item list, scoring version, judge configuration, and input-access policy with each result.

G Matched Events and Near Misses

There are 53 suites from Track counting streams. Each contains a positive item and near-miss moments from the same video, 106 timestamps in total. For suite j, let $h_j = 1$ if its positive item receives nonzero timing credit and zero otherwise. Let $s_j = 1$ if the system emits nothing within five seconds of every near miss and zero otherwise. The suite pass rate is

$$\mathrm{PairSel} = \frac{100}{53} \sum_{j=1}^{53} h_j s_j. \tag{2}$$

The positive criterion is timing credit, and content accuracy is reported alongside this diagnostic. Ten near-miss windows intersect a positive response window. Excluding emissions in the overlap changes individual scores by at most eight points in the reported check. Scripted policies set a floor. The burst policy passes 22.6% of the suites, as many as JoyAI-VL (Yao et al., 2026) and MMDuet2 (Wang et al., 2026c), while always-fire and never-fire pass none. Offline grounding by the Claude Sonnet 5 (Anthropic, 2026a) and GPT-5.6 Sol (OpenAI, 2026) agents passes 60.4% and 54.7%.

H Scoring Specification

Table 6 lists how each response format contributes to Overall. For each item, the evaluator removes empty emissions and sorts the rest by timestamp. Negative items receive 100 only if the list is empty. For event-based positive items, each window starts one second before its reference timestamp and ends where the next window starts. The last window extends to the end of the video. The first

Table 6: **What contributes to the overall score.** Scores average within each item before dataset aggregation. Higher is better.

Response format	Content requirement	Overall score
Query answer	Correct first post-query answer	Mean of content and harmonic mean of timing and silence.
Event trigger	Correct matched event responses	Same combination as query answers.
Count updates	Exact final count	Same combination; content is final-count accuracy.
Free-form updates	Correct content in each segment	Segment content, with empty segments scoring zero.
No-event stream	No utterance	100 if silent throughout; 0 otherwise.

emission within a window is matched regardless of content, with nonnegative delay measured from the reference timestamp. Earlier speech is premature and subsequent speech in an occupied window is redundant. Matching ignores content. Matching by position keeps timing and silence independent of the judge, and content accuracy then grades the matched reply. A content-aware match would reward emitting candidate replies until one fits.

Missing event responses contribute zero timing credit. Content accuracy uses the first post-query response for query-answer items and the mean over matched responses for triggers. Keyed answers first match an option letter in the first 40 characters. Without a usable letter, the scorer compares token F1 with every option, requiring a unique best match of at least 0.5. Unique full-option containment is the final fallback. The keyed pool has 688 items, including 637 six-option questions. Free-text answers use the open judge. Lexical fallback takes the maximum of exact match, token F1, and a length-guarded numeric match. Human free-text responses are judged against the keyed answer text. Counting accuracy compares the last number in the emitted sequence with the final reference count. This check uses the final count. A per-event variant also requires each matched reply to state the running count of its event. It stays at or below 38.5 for every system, against 40.0 for the human reference, and exceeds final-count accuracy for 12 of the 20 configurations scored on all 160 counting items, for example 26.2 against 8.8 for polled Qwen3-VL-8B (Bai et al., 2025). The final count is therefore the stricter criterion for most configurations. The violation budget equals the number of reference events. A fully silent positive item has zero accuracy and timing. Its silence component is undefined and its Overall is zero.

Free-form items use reference timestamps to partition the video. The scorer concatenates emissions within each segment and compares them with its reference. Empty segments score zero. The item averages these segment scores, and this average is its Overall. Timing and silence are reported only as diagnostics for this format. A uniform variant also scores free-form items as the mean of content and the harmonic mean of timing and silence. It keeps the ranking of the 30 current configurations at a Kendall rank correlation (Kendall et al., 1938) of 0.890. Chatty polled systems lose the most, up to 4.5 points for polled Claude Sonnet 5 (Anthropic, 2026a). Dataset aggregation averages each field over its defined item values. A silent system earns Overall only on negative items, giving the observed 6.5-point dataset score for the never-fire policy.

Alternative aggregations. A softer negative-stream penalty changes little. Scoring a negative stream with M false alarms as $100/(1+M)$ instead of $100\max(0, 1-M)$ keeps the Overall ranking of the 30 current configurations at a rank correlation of 0.959. Charging each false alarm five seconds of a stream of length T, as $100\max(0, 1-5M/T)$, keeps it at 0.926. Overall moves by at most 2.7 and 5.9 points, in both cases for VideoLLM-online (Chen et al., 2024). Weighting timing against silence also changes little. Replacing $H(\mathrm{TA}, \mathrm{SC})$ by $H_\beta = (1+\beta^2)\,\mathrm{TA}\,\mathrm{SC}/(\beta^2\mathrm{TA}+\mathrm{SC})$ with $\beta = 2$, which favors silence, or $\beta = 0.5$, which favors timing, keeps the ranking at 0.972 or higher. Shifting weight between content and behavior matters more. Weighting content by 0.25 or 0.75 instead of one half gives rank correlations of 0.733 and 0.825, so the results tables report the three components separately. Silence compliance also floors at zero once violations reach the number of reference events. A smooth variant, $100/(1+|V|/N)$, never floors and keeps the silence ranking at 0.954, and the violations per event in Table 16 separate systems below the floor. Fully silent positive items drop out of the silence aggregate. Most configurations speak on most positive items, with a median of 92.8%, but the blind reference speaks on 34.6% and Claude Sonnet 5 in

Table 7: **Rescoring 30 configurations with other delay scales.** Kendall τ compares each ranking with the default five-second scale, and the top three configurations never change.

Delay scale Δ	τ, Overall↑	τ, timing↑	Largest rank move↓	Mean Overall
2 s	0.954	0.913	3	34.4
3 s	0.968	0.949	2	34.7
5 s, default	1.000	1.000	0	35.1
8 s	0.995	0.982	1	35.4
10 s	0.991	0.963	1	35.5

grounding on 54.1%. Counting their silent items as zero lowers their silence compliance from 44.2 to 17.2 and from 91.0 to 51.9. Qwen3-VL-8B still raises its silence compliance more than fivefold in grounding under this count, from 9.7 to 53.9. Overall already scores fully silent items zero, so this selection affects only the silence diagnostic.

Timing tolerance. The default delay scale is five seconds. As shown in Table 7, rescoring the 30 configurations that use the current export at delay scales of 2, 3, 8, and 10 seconds preserves their order. Kendall rank correlations (Kendall et al., 1938) of the Overall ranking against the default stay between 0.954 and 0.995, and those of the timing ranking between 0.913 and 0.982. No configuration moves more than three ranks, the top three never change, and mean Overall shifts from 34.4 to 35.5. Timing accuracy itself moves more, for example from 52.1 to 70.1 for the human reference. Overall moves by at most 3.6 points for any configuration. The scale therefore changes magnitudes rather than the ordering. A second check perturbs the reference timestamps. Moving every event timestamp after the query by one or two seconds in either direction changes which replies are matched. Independent Gaussian jitter with a standard deviation of one or two seconds does the same. Keyed items need no judge and are rescored in full. Free-text content accuracy keeps its original value. One-second shifts and one-second jitter keep the Overall ranking of the 30 configurations at a rank correlation of at least 0.917. The timing ranking stays at or above 0.945 and the silence ranking at or above 0.954. Mean absolute Overall changes stay below 2.0 points. Two-second changes keep the Overall ranking at 0.876 or higher. Offline grounding by the two agents is the sensitive case. When every reference moves one second later, Overall drops by 14.2 points for the Claude Sonnet 5 agent and 11.7 for the GPT-5.6 Sol agent (OpenAI, 2026). The premature replies of the Claude Sonnet 5 agent rise from 0.30 to 0.56 per item. Earlier references change their Overall by under 3 points, since each window opens one second early. The early tolerance matters mainly for the same agents. With free-text content again held fixed, setting ϵ to 0.5 or 2 seconds instead of 1 keeps the Overall ranking at a rank correlation of 0.963, and no configuration moves by more than 3.9 points. Without a tolerance, at $\epsilon = 0$, the rank correlation falls to 0.913 and the Claude Sonnet 5 agent in grounding loses 14.9 points, because its premature replies rise from 0.30 to 0.57 per item. Those of polled Qwen3-VL-8B (Bai et al., 2025) rise only from 6.33 to 6.65.

I Judge and Segment Scoring

Seven candidate judges are calibrated against 954 question–answer triples. A Gemini judge (Google DeepMind, 2026) first labeled each triple as correct or incorrect, and annotators screened these labels and corrected wrong ones. As shown in Table 8, greedy Qwen3-14B (Yang et al., 2025a) agrees with 89.7% of the final labels, with Cohen's kappa (Cohen, 1960) of 0.79, on par with Qwen3-32B and gpt-oss-20b (Agarwal et al., 2025) at 90.0% and 89.9%. Its five sampled votes are unanimous on 98.3% of the triples, against 92.1% and 92.3% for those two, and a five-vote majority raises its agreement only to 89.8%. The final scorer uses greedy Qwen3-14B to balance reliability and efficiency. Rescoring all 30 current configurations with greedy Qwen3-32B keeps the Overall ranking at a Kendall rank correlation (Kendall et al., 1938) of 0.972 against Qwen3-14B, and no Overall moves by more than 1.3 points. The 108 free-form items comprise 107 Commentate items and one revision item. The segment scorer follows VDCScore (Chai et al., 2025). Each reference is decomposed into at most five question–answer pairs, averaging 2.0 across 303 cached reference texts. Answers are extracted from the segment's emitted text and judged against those reference answers. Caching fixes the decomposition across systems.

Table 8: **Judge calibration on 954 reference-labeled answers.** Higher agreement in percent and higher kappa are better. Bold marks the selected judge.

	One greedy verdict		Five-vote majority		
Judge	Agree %↑	κ ↑	Agree %↑	κ ↑	Unanimous votes %
Llama-3-8B (Grattafiori et al., 2024)	71.3	.458	71.0	.453	93.6
Qwen3-VL-8B (Bai et al., 2025)	84.9	.699	85.3	.707	96.5
Gemma-3-27B (Team et al., 2025)	83.8	.682	84.0	.686	99.0
Qwen3-30B-A3B (Yang et al., 2025a)	87.0	.739	86.9	.737	97.0
Qwen3-14B, selected	89.7	.790	89.8	.792	98.3
Qwen3-32B	90.0	.799	89.5	.789	92.1
gpt-oss-20b (Agarwal et al., 2025)	89.9	.798	90.6	.810	92.3
Ensembles, where dashes mark inapplicable fields					
Seven-judge majority	–	–	88.9	.778	–
All 35 sampled votes	–	–	89.1	.783	–
Top-three judge majority	–	–	91.7	.832	–
Top-three, 15 sampled votes	–	–	91.7	.833	–
Top-three greedy verdicts	–	–	91.6	.831	–

J Evaluation Prompts

Two prompt shapes drive the adapted models. Single-turn modes, cumulative or with a fresh prompt, restate the full task at every poll. The reply-history mode runs one interleaved conversation, states the task once in the system message, and sends a compact nudge per poll. The system prompt reads verbatim as follows.

```
You are a real-time visual assistant watching a LIVE video
stream.  You receive frames as they arrive; the LAST image
is the current moment.  You must decide, right now, whether
to speak or to keep watching silently.  Only speak when this
exact moment calls for it --- do not speak early, and do not
repeat what you already reported.  If nothing needs saying,
wait silently.
```

Each single-turn poll then renders the stream time, the standing request, one hint chosen by task type, and a fixed output format. The hint depends only on the task type and never on whether an item contains an event, and positive and negative items of the same task therefore receive identical prompts. The hints read verbatim as follows.

Task	Hint
Look	Answer the question as soon as you are asked, from what you can see.
Recall	Answer from what you saw earlier in the stream. Answer immediately.
Recall, causal	Answer the question as soon as it becomes answerable from what you saw.
Time	Answer at the earliest moment the question becomes decidable from what you have seen; before that, wait.
Alert	Speak only at the exact moment the requested event happens. If it never happens, stay silent for the whole stream.
Track	Track the requested state across frames. Each time it changes (e.g. the count increases), speak the update (for counts: just the new number). Otherwise wait.
Track, revision	Track the requested state; speak whenever the correct answer changes, correcting your earlier answer if new evidence contradicts it.
Commentate	Follow the stream and speak one short line for each new step or scene as it happens, like a live commentator or guide.

The output format reads verbatim as follows.

```
Respond in EXACTLY this format, nothing else:
DECISION: SPEAK or WAIT
```

```
RESPONSE: <if SPEAK, one short line --- the exact thing you
would say out loud now; if WAIT, leave blank>
```

The reply-history mode adds to its system message "You can see your own previous turns. If you already reported this and nothing new happened, WAIT instead of repeating.", and each poll sends only "`[stream time ts] New frames above. Decide now (DECISION / RESPONSE).`" For keyed items the rendered item replaces the standing request. It shows the rewritten stem where the key provides one and the annotation question otherwise, followed by one option per line and a fixed instruction. The polling adapter and the agentic systems share the following verbatim rendering.

```
<stem>
A. <option 1>
B. <option 2>
...
F. <option 6>
(When you speak, answer with just the option letter.)
```

K System Configuration and Provenance

The polling default is two frames per second, a 16-frame sliding window, a 512-pixel long side, greedy decoding, and at most 96 generated tokens. The MiniCPM-o (Cui et al., 2026) fresh-prompt run uses beam search instead. Polls occur once per video second. Query-answer items permit a ten-second post-query response period. Reply-history variants retain earlier frames and replies in one conversation. Fresh-prompt variants start anew at each poll.

Native streaming systems keep their released interfaces and recommended input rates. JoyAI-VL (Yao et al., 2026) runs its own server-side memory and silence decision at 4 fps, and each tick pushes only new frames. MMDuet2 (Wang et al., 2026c) replies or emits no reply after every turn of two frames at 4 fps. VideoLLM-online (Chen et al., 2024) decides after every frame at 8 fps through its frame-interval gate with the released threshold of 0.725. VideoChat3 (Li et al., 2026b) chooses silence, standby, or a response in each one-second round. MOSS-Video (Wang et al., 2026b) consumes frames at real-time pace and emits silence tokens while it observes. LiveCC (Chen et al., 2025) comments on each one-second clip and has no silence mechanism. Flash-VStream (Zhang et al., 2024) has no response decision of its own. Its memory ingests 8 fps, the polling adapter queries that memory once per second, and the row keeps a lenient output parser. Dispider (Qian et al., 2025) exposes only whole-video question answering in its released code, and it answers the grounding prompt with the times at which it would respond.

Polled Claude calls the Anthropic API with the model name `claude-sonnet-5` (Anthropic, 2026a). The Claude kernel swaps call `claude-sonnet-5` or `claude-opus-5` as the JoyAI-VL model, and JoyAI-VL's memory summarizer runs on `claude-haiku-4-5` in both. The Qwen3-VL-8B kernel swap serves the model (Bai et al., 2025) on a B200 GPU inside the same controller. The Claude agents run the Claude Code command-line agent with the model alias `sonnet`. The shell-tool run started on September 13, 2026, and the media-tool run on September 18. Polled Gemini (Google DeepMind, 2026) calls `gemini-2.5-flash` through the OpenAI-compatible endpoint with reasoning disabled, and 32 blocked or empty replies count as silence. The GPT agent calls the `gpt-5.6` API alias (OpenAI, 2026), and the Gemini agent runs the Gemini command-line agent with Gemini 3.5 Flash (Google DeepMind, 2026). Providers can update the model behind a name, and each result therefore keeps its run dates.

Human reference. Volunteers answer in a browser that plays each video at normal speed and allows no manual pausing, seeking, or rewinding. The question appears on screen at its query time. A volunteer presses a key as soon as a response is possible, and that moment on the video clock becomes the response time, as for a system emission. The video then pauses while the volunteer types and resumes after submission. Alert items take a response at every occurrence, and counting items take every change of the count. A volunteer can update an answer by pressing the key again, and each update counts as a new response. Volunteers answer independently. Some alerts are a bare "now", which earns timing credit but no content credit, and 13.9% of Alert replies take this form. The reference shows live behavior rather than a ceiling. Grounding sees the whole video, and

Table 9: **Measured wall-clock latency per poll.** Each system is timed on the hardware that produced its scores, and polls arrive once per stream second.

System	Hardware	Median, s↓	90th pct., s↓	Within 1 s, %↑
JoyAI-VL, native	B200	0.38	0.78	98.9
JoyAI / Qwen3-VL-8B	B200	0.44	0.91	93.9
Qwen3-VL-4B, polled	B200, 90 GB MIG slice	0.39	0.43	99.8
Qwen3-VL-8B, polled	B200, 90 GB MIG slice	0.48	0.63	98.9
LLaVA-OV2, polled	B200, 90 GB MIG slice	0.53	0.80	94.8
Qwen3-VL-8B, reply history	B200, 90 GB MIG slice	0.83	1.93	64.7
MiniCPM-o, polled	B200, 90 GB MIG slice	1.26	1.70	3.0
Qwen3-VL-8B, 64 frames	B200, 90 GB MIG slice	1.35	1.59	9.8
Gemma-3n, polled	B200, 90 GB MIG slice	1.67	2.95	2.2
Qwen2.5-Omni, polled	B200, 90 GB MIG slice	1.93	2.00	0.0
Gemini 2.5 Flash, polled	Google API	1.91	5.29	1.5
Claude Sonnet 5, polled	Anthropic API	2.42	3.59	0.0
JoyAI / Claude Sonnet 5	Anthropic API	3.33	5.19	0.0
JoyAI / Claude Opus 5	Anthropic API	4.18	6.54	0.0
Qwen3-VL-8B, 256 frames	B200, 90 GB MIG slice	4.12	6.08	9.0
Flash-VStream, polled	L40S	6.95	10.96	6.0
Claude Sonnet 5 agent, media tools	Anthropic API	28.72	87.42	0.0
Claude Sonnet 5 agent, shell tools	Anthropic API	37.17	91.62	0.0

logged agent polls take about half a minute on a clock that waits for them. A live viewer has neither advantage.

L Measured Latency

Every prediction file stores the wall-clock duration of each poll, and Table 9 summarizes these durations. A poll spans one second of stream time, and a system keeps pace with a live stream only when most polls finish within that second. Scores ignore these durations because the runners advance the video clock only after a poll returns. MOSS-Video (Wang et al., 2026b) is the exception, since it consumes frames at real-time pace. A live replay estimates what these durations cost. Eight polled runs are replayed, the six open backbones and polled Claude Sonnet 5 (Anthropic, 2026a) and Gemini 2.5 Flash (Google DeepMind, 2026). Flash-VStream (Zhang et al., 2024) logs irregular polls and is left out. In the replay, a poll starts only after the previous one returns, uses the latest recorded decision it could see, and replies after its measured latency. Keyed items are rescored in full, and free-text content keeps its original value. Timing accuracy then falls from 82.9 to 45.5 for polled Claude Sonnet 5 and from 93.1 to 51.6 for polled Gemini 2.5 Flash. Qwen3-VL-8B and Qwen3-VL-4B (Bai et al., 2025) and LLaVA-OV2 (An et al., 2026) keep pace and lose at most 7.6 points of timing accuracy. Slower polls also repeat less, and low silence scores dominate the harmonic mean. Overall therefore moves by at most 3.3 points across the eight replayed runs. Only 0.8% of the polls of the Claude Sonnet 5 agent with media tools and 1.4% with shell tools return within the five-second delay scale. These two agent runs logged every poll, and the same replay confirms the cost. Timing accuracy falls from 66.5 and 70.7 to 0.3, since only 65 and 55 of their 711 and 695 recorded replies are ever sent, 22 and 23 of them after the video ends. Silence compliance rises to 91.8 and 95.6, and Overall falls from 56.0 and 61.2 to 29.2 and 28.5, below the 41.3 of polled Claude Sonnet 5 under the same replay. The GPT-5.6 Sol (OpenAI, 2026) and Gemini agent runs logged no poll times and are not replayed. The measurements come from the hardware that produced each score, a B200 GPU, a 90 GB MIG slice of a B200, an L40S GPU, or an API endpoint. VideoLLM-online (Chen et al., 2024) logs no per-frame time, and MMDuet2 (Wang et al., 2026c) logs frame ingestion and generation as separate steps, and both are omitted. The Claude Opus 5 kernel swap (Anthropic, 2026b) logged poll times on 64 of its items. Offline grounding issues one whole-video call per item and has no per-poll latency.

Table 10: **Offline answers and online scores of the six polled backbones.** Offline accuracy uses the 688 keyed items in grounding and counts unanswered items as wrong.

	Offline, keyed items			Online, polling adapter		
Backbone	**Accuracy↑**	**Answered, %**	**Accuracy if answered↑**	**Overall↑**	**Timing↑**	**Silence↑**
Qwen3-VL-4B	35.8	80.8	44.2	26.2	34.0	28.4
Qwen3-VL-8B	35.4	90.7	39.0	34.8	77.8	10.9
MiniCPM-o-4.5	33.2	87.8	37.8	29.0	44.7	31.6
LLaVA-OV2-8B	32.7	99.3	32.9	34.5	81.0	16.9
Qwen2.5-Omni-7B	26.5	53.6	49.3	18.7	34.4	20.7
Gemma-3n-E4B	12.0	56.2	21.3	27.8	81.9	11.9

Table 11: **Offline and online answers disagree item by item.** Each open backbone answers the 688 keyed items under polling and in offline grounding. Percentages count items, unanswered offline items count as wrong, and κ is Cohen's kappa (Cohen, 1960).

Backbone	**Online**	**Offline**	**Both**	**Online only**	**Offline only**	**Agree**	κ
Qwen3-VL-8B	65.0	35.6	22.4	42.6	13.2	44.2	−0.03
Qwen3-VL-4B	43.8	35.8	18.8	25.0	17.0	58.0	0.13
LLaVA-OV2	60.0	33.0	22.4	37.6	10.6	51.7	0.10
Gemma-3n	53.3	12.1	7.3	46.1	4.8	49.1	0.03
MiniCPM-o	35.6	33.6	21.1	14.5	12.5	73.0	0.40
Qwen2.5-Omni	35.6	26.5	7.4	28.2	19.0	52.8	−0.09

M Offline Accuracy of the Polled Backbones

Table 10 lists the numbers behind the backbone comparison. Offline accuracy rescores the grounding run of each backbone on the 688 keyed items. Grounding lets a model decline to answer, and every declined item counts as wrong. Gemma-3n (Gemma Team, 2025) leaves 301 keyed items unanswered and reaches 21.3 on the rest, close to the 19.0 expected from random guessing on this option mix. Qwen2.5-Omni (Xu et al., 2025) answers only 53.6% of the items but is the most accurate when it answers. As shown in Table 11, the same backbone answers different keyed items correctly offline and online. Agreement ranges from 44.2% to 73.0%, and Cohen's kappa (Cohen, 1960) stays at or below 0.13 for five of the six backbones and reaches 0.40 for MiniCPM-o.

N Protocol Audit

Six scripted policies probe simple failure modes, and as shown in Figure 10, none competes with the main polling configurations in the current score export. Always-fire scores 98.8 in timing but approximately zero Overall. Never-fire scores 6.5 Overall, reflecting credit on the negative items. Periodic emissions, bursts, background chatter, and the parrot baseline also stay far below the polled systems. These checks reject the tested shortcuts as competitive solutions.

O Detailed System Results

Table 12 gives the aggregate content, timing, and silence of every configuration, and Tables 13, 14, and 15 separate them by task. Their values come directly from the same export as the main results. Overall averages item scores under the format-specific scoring rule, and runtime is excluded. Offline grounding receives the whole video and reports when it would have responded. It serves only as a diagnostic reference. Silence compliance reaches zero once an item's violations match its reference events. As shown in Table 16, this floor covers 72% to 97% of positive items for the polled models, whose repeated replies per event still range from 2.8 to 19.2.

As shown in Table 17, sampling rate and prompt hints move Overall by at most 1.4 points across all polling-adapter ablations, whereas a 256-frame window adds 4.2. Response history lowers Overall

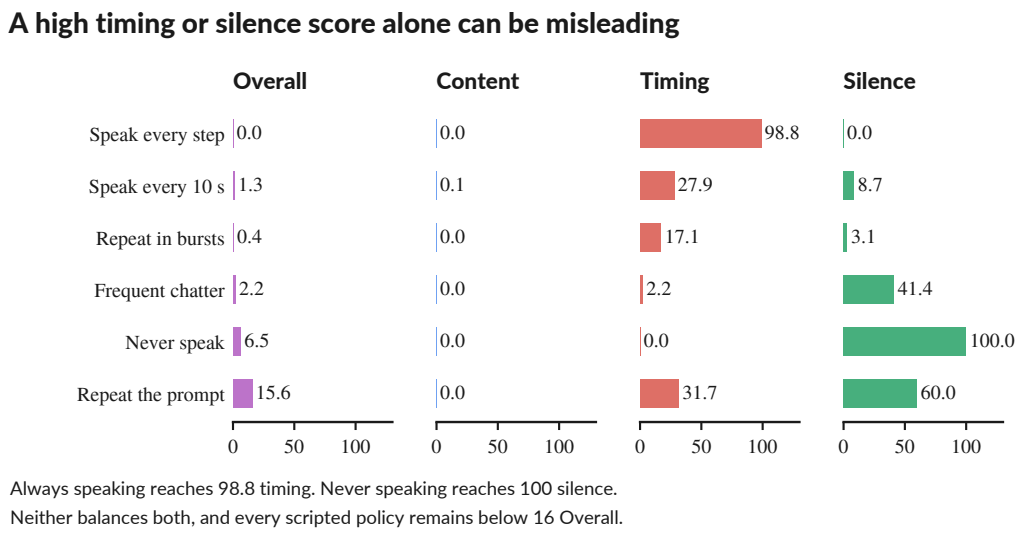


Figure 10: **No scripted policy reaches a competitive Overall score.** Always speaking raises timing scores, and never speaking raises silence scores.

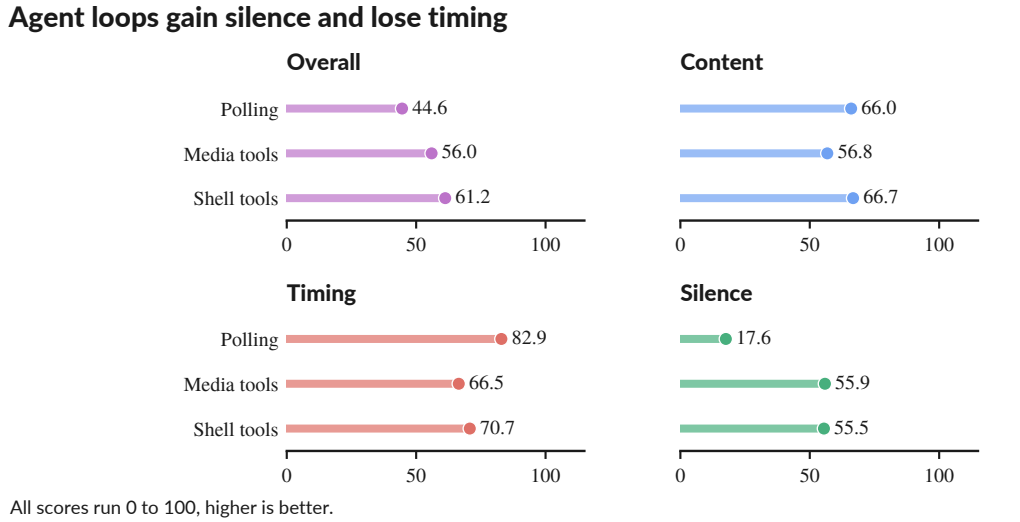


Figure 11: **The response loop changes behavior.** Claude Sonnet 5 (Anthropic, 2026a) under polling and in two agent runs, with Overall and all three component scores.

by up to 9.6 points and raises it by up to 10.1. As shown in Figure 12, reply history moves silence compliance more than any frame or prompt setting. Table 18 arranges the same Overall scores by model and controller. The kernel swaps cover partial item sets. The 35.5 kernel-swap score is historical, the 43.3 entry covers 620 items, and the 32.4 and 30.5 entries use the agent items. As shown in Figure 11, both Claude Sonnet 5 (Anthropic, 2026a) agent runs lower timing accuracy and raise silence compliance relative to polling. Overall of the same model moves from 44.6 under polling to 56.0 with media tools and to 61.2 with shell tools.

P Random Emission Thinning

Each recorded emission is independently retained with a prescribed probability, and the resulting stream is rescored. The four exported curves use probabilities from 1.0 to 0.1. At low retention, silence compliance generally improves while timing accuracy declines. These curves describe ran-

Table 12: **System scores and 95% bootstrap intervals.** Scores range from 0 to 100, and higher is better. Bold marks the leader of each family. Intervals resample items. The agentic families are named by their tools, Qwen-MM-Plugins (Qwen Team, 2026) or Claude Code tools.

System	Overall↑	95% CI	Content↑	Timing↑	Silence↑
References					
Qwen3-VL-8B, blind	15.7	[13.9, 17.8]	9.0	32.6	44.2
Human reference	53.6	[49.4, 57.9]	51.3	63.5	74.7
Native streaming					
VideoLLM-online	14.5	[13.1, 15.9]	3.7	33.6	**47.9**
LiveCC	14.5	–	25.4	21.0	7.2
MMDuet2	17.0	[15.1, 18.9]	19.0	**40.4**	23.0
MOSS-Video	19.9	–	35.1	24.0	15.5
VideoChat3	**28.5**	–	**46.8**	14.3	33.1
Turn-based polling					
Flash-VStream	14.3	[12.8, 15.9]	21.5	50.2	8.2
Qwen2.5-Omni	18.7	[17.1, 20.3]	28.3	34.4	20.7
LLaVA-OV	25.6	–	37.3	37.9	19.4
Qwen3-VL-4B	26.2	[24.2, 28.2]	35.1	34.0	28.4
Gemma-3n	27.8	[26.0, 29.6]	43.3	81.9	11.9
MiniCPM-o	29.0	[26.8, 31.2]	33.1	44.7	**31.6**
LLaVA-OV2	34.5	[32.6, 36.4]	49.8	81.0	16.9
Qwen3-VL-8B	34.8	[32.9, 36.7]	54.2	77.8	10.9
Gemini 2.5 Flash	35.5	[29.4, 41.4]	55.6	**93.1**	10.8
Claude Sonnet 5	**44.6**	[38.4, 50.5]	**66.0**	82.9	17.6
Offline temporal grounding					
Gemma-3n	16.1	[14.3, 17.9]	11.3	14.4	46.8
Dispider	24.8	–	26.9	21.9	76.4
Qwen2.5-Omni	27.2	[24.8, 29.6]	18.9	25.2	81.4
MiniCPM-o	30.6	[28.2, 33.2]	26.3	36.7	53.6
Qwen3-VL-4B	33.8	[31.1, 36.3]	26.7	36.2	57.4
LLaVA-OV2	33.9	[31.4, 36.4]	25.9	35.6	63.2
Qwen3-VL-8B	35.7	[33.1, 38.2]	28.0	37.6	59.3
Claude Sonnet 5	38.6	[36.0, 41.3]	36.0	32.6	**91.0**
GPT-5.6 Sol agent	58.8	[56.2, 61.3]	56.8	58.4	80.0
Claude Sonnet 5 agent	**66.1**	[63.4, 68.7]	**62.8**	**70.5**	83.1
Native Real-time Interaction System (JoyAI-VL)					
Claude Opus 5	30.5	[23.6, 37.7]	40.3	40.3	**33.2**
Qwen3-VL-8B	32.4	[25.6, 39.0]	50.2	**90.7**	11.0
JoyAI-VL	33.8	[31.7, 36.0]	45.7	66.8	29.2
Qwen3-VL-4B	35.5	–	56.9	57.1	9.6
Claude Sonnet 5	**43.3**	[40.6, 45.9]	**64.4**	77.8	23.0
Agentic Interaction System (Qwen-MM-Plugins)					
Gemini 3.5 Flash	29.4	[21.4, 37.3]	28.8	23.3	**66.9**
Claude Sonnet 5	56.0	[47.2, 64.4]	56.8	66.5	55.9
GPT-5.6 Sol	**59.1**	[51.0, 67.0]	**56.8**	**68.6**	60.0
Agentic Interaction System (Claude Code tools)					
Claude Sonnet 5	**61.2**	[53.7, 68.2]	**66.7**	**70.7**	**55.5**

dom post-processing of fixed output traces and do not model a confidence-calibrated threshold. The native threshold sweep in Appendix S moves a system's own trigger instead.

Q Case Studies

Task overview. The task overview at the head of the main text takes twelve full frames from three Wikimedia Commons videos. They are a basketball game in the public domain (U.S. Navy, 2025), water heated to a boil under CC BY-SA 1.0 (Sertion, 2011), and a Spanish omelette under CC BY-SA 4.0 (Elene B, 2021). Each row keeps its source's clock. Four rows at the top and bottom of

Table 13: **Content accuracy by task.** Scores run from 0 to 100, higher is better, and come from the same export as the main results. The agentic families are named by their tools, Qwen-MM-Plugins (Qwen Team, 2026) or Claude Code tools.

System	Look↑	Recall↑	Time↑	Alert↑	Track↑	Commentate↑
References						
Qwen3-VL-8B, blind	21.7	22.4	0.0	6.2	0.0	1.0
Human reference	78.7	44.4	63.0	19.5	36.8	56.9
Native streaming						
VideoLLM-online	2.4	6.4	1.6	3.7	6.8	0.0
LiveCC	68.8	0.0	34.6	35.8	1.7	17.0
MMDuet2	55.4	7.7	15.3	13.6	1.2	24.1
MOSS-Video	56.2	9.7	26.9	63.3	3.3	14.2
VideoChat3	62.5	6.5	65.4	72.5	25.0	18.5
Turn-based polling						
Flash-VStream	30.7	2.6	21.4	42.7	7.4	2.2
Qwen2.5-Omni	23.5	3.2	42.9	56.0	4.9	19.4
LLaVA-OV	37.5	22.6	92.3	39.3	8.3	28.5
Qwen3-VL-4B	71.1	21.2	15.1	51.5	1.9	31.0
Gemma-3n	57.2	28.8	49.1	66.6	7.4	30.7
MiniCPM-o	75.3	44.2	15.9	15.9	8.0	53.8
LLaVA-OV2	64.5	34.6	57.1	72.3	10.2	43.0
Qwen3-VL-8B	75.3	24.4	66.7	78.7	8.6	57.1
Gemini 2.5 Flash	81.2	40.0	46.2	71.3	17.6	68.1
Claude Sonnet 5	87.5	46.7	84.6	76.0	18.5	84.1
Offline temporal grounding						
Gemma-3n	14.5	14.1	7.1	15.0	6.8	4.2
Dispider	75.0	61.3	38.5	18.5	5.0	1.6
Qwen2.5-Omni	44.0	47.4	5.6	11.7	0.6	0.0
MiniCPM-o	59.6	66.7	8.7	11.4	8.0	2.6
Qwen3-VL-4B	63.9	65.4	7.1	15.3	3.1	0.8
LLaVA-OV2	64.5	60.9	7.1	11.0	7.4	3.8
Qwen3-VL-8B	63.9	67.9	6.3	13.1	10.5	4.7
Claude Sonnet 5	69.3	65.4	29.4	35.7	0.0	5.1
GPT-5.6 Sol agent	89.2	91.7	57.9	52.3	21.7	18.7
Claude Sonnet 5 agent	90.4	92.3	65.9	57.2	35.1	29.2
Native Real-time Interaction System (JoyAI-VL)						
Claude Opus 5	81.2	20.0	23.1	45.6	17.3	48.6
Qwen3-VL-8B	75.0	73.3	7.7	66.2	0.0	70.0
JoyAI-VL	59.0	71.8	41.4	52.7	3.2	38.4
Qwen3-VL-4B	68.8	38.7	69.2	85.1	3.3	37.6
Claude Sonnet 5	79.7	84.1	70.0	63.4	19.6	55.8
Agentic Interaction System (Qwen-MM-Plugins)						
Gemini 3.5 Flash	62.5	46.7	30.8	20.8	0.0	12.0
Claude Sonnet 5	87.5	86.7	50.0	49.4	8.0	69.2
GPT-5.6 Sol	93.8	86.7	30.8	59.9	0.3	69.4
Agentic Interaction System (Claude Code tools)						
Claude Sonnet 5	93.8	93.3	53.8	69.4	14.3	77.9

the boiling-water source are trimmed to reach 16:9, and no other image is cropped or retouched. The six instructions and their replies are new illustrative annotations outside the scored benchmark. Their times follow visible evidence and are not calibrated earliest-answerable times. Source links, checksums, observations, and a frame-extraction script accompany the assets.

Response errors on a railway crossing. Figure 13 places constructed response errors on a railway-crossing video (WeenyBeany, 2024), licensed under CC BY 4.0. A reply before the train enters is premature, a second reply to the same event is redundant, and a reply after the window opens loses timing credit with its delay. Its false-alarm row uses a different instruction, to report a vehicle going around a lowered barrier. No such event occurs in the reviewed clip. The other rows ask for the train entering the crossing. These are constructed examples, not measured system traces.

Table 14: **Timing accuracy by task.** Scores run from 0 to 100, higher is better, and come from the same export as the main results. The agentic families are named by their tools, Qwen-MM-Plugins (Qwen Team, 2026) or Claude Code tools.

System	Look↑	Recall↑	Time↑	Alert↑	Track↑	Commentate↑
References						
Qwen3-VL-8B, blind	77.5	95.5	0.7	6.3	0.0	25.4
Human reference	74.0	57.4	51.0	53.3	69.0	70.1
Native streaming						
VideoLLM-online	94.2	81.9	8.7	7.7	5.6	7.1
LiveCC	71.1	5.6	24.5	14.5	2.2	42.0
MMDuet2	62.9	37.8	37.2	35.2	26.2	47.4
MOSS-Video	71.5	15.7	18.1	25.3	4.6	27.7
VideoChat3	56.1	3.0	11.2	9.2	11.3	22.6
Turn-based polling						
Flash-VStream	66.2	18.0	53.1	61.3	49.4	41.6
Qwen2.5-Omni	25.3	9.6	40.0	51.9	20.3	54.2
LLaVA-OV	55.0	48.4	57.5	16.7	6.3	86.5
Qwen3-VL-4B	74.6	23.2	7.4	41.7	2.7	46.2
Gemma-3n	98.1	79.9	57.6	80.3	78.1	97.8
MiniCPM-o	87.2	86.9	5.6	12.9	20.5	81.3
LLaVA-OV2	98.9	90.5	76.3	75.3	55.6	98.2
Qwen3-VL-8B	96.6	44.1	79.6	79.1	75.3	96.3
Gemini 2.5 Flash	100.0	93.3	97.4	89.4	85.0	98.6
Claude Sonnet 5	98.8	100.0	84.4	78.7	52.1	89.0
Offline temporal grounding						
Gemma-3n	24.0	23.7	6.2	7.7	11.0	17.9
Dispider	93.8	100.0	0.0	0.8	1.2	0.9
Qwen2.5-Omni	69.9	78.8	2.8	2.0	1.2	0.0
MiniCPM-o	95.7	100.0	8.4	3.4	11.9	8.8
Qwen3-VL-4B	96.3	99.4	4.5	3.5	9.5	11.9
LLaVA-OV2	96.3	99.4	5.8	4.3	8.3	5.7
Qwen3-VL-8B	96.9	100.0	2.1	4.0	13.4	19.1
Claude Sonnet 5	89.7	87.8	10.5	5.6	0.8	6.5
GPT-5.6 Sol agent	95.8	100.0	48.4	39.2	33.2	38.9
Claude Sonnet 5 agent	97.4	100.0	59.7	50.0	50.9	80.7
Native Real-time Interaction System (JoyAI-VL)						
Claude Opus 5	85.0	26.7	20.6	34.9	28.2	46.3
Qwen3-VL-8B	98.8	80.0	90.5	97.9	76.2	97.8
JoyAI-VL	96.5	97.4	55.8	38.5	58.7	73.7
Qwen3-VL-4B	68.2	60.0	53.8	65.0	7.9	77.4
Claude Sonnet 5	94.4	91.5	70.0	64.1	60.4	84.0
Agentic Interaction System (Qwen-MM-Plugins)						
Gemini 3.5 Flash	48.8	44.0	23.1	12.0	4.8	10.7
Claude Sonnet 5	93.8	93.3	31.7	55.8	46.0	89.6
GPT-5.6 Sol	92.5	93.3	28.5	50.7	70.4	87.9
Agentic Interaction System (Claude Code tools)						
Claude Sonnet 5	90.0	100.0	54.9	56.3	53.1	83.7

Benchmark cases and system traces. Figures 14, 15, 16, 17, 18, and 19 show the instruction, reference times, sampled frames, and system emissions on one item per task. In each trace, dots mark utterances, dashed black lines mark reference events, and the right column quotes each first reply. They illustrate how an instruction can remain active after the first answer, how repeated speech differs from repeated valid events, and how grounding benefits from future access. Individual cases are illustrative rather than estimates of aggregate error rates. The plotted traces and answer text come from the saved case records. The Alert item in Figures 17 and 20 also illustrates the difference between repeated speech and missed events in the saved traces. Under chronological first-response matching with the one-second early tolerance, the polled Qwen3-VL-8B (Bai et al., 2025) emits 104 times, five before the first window, four first matches, and 95 repetitions within answered windows. JoyAI-VL (Yao et al., 2026) emits at 1, 58, and 102 seconds, one premature reply, one match, and one repetition, leaving three events unmatched.

Table 15: **Silence compliance by task.** Scores run from 0 to 100, higher is better, and come from the same export as the main results. The agentic families are named by their tools, Qwen-MM-Plugins (Qwen Team, 2026) or Claude Code tools.

System	Look↑	Recall↑	Time↑	Alert↑	Track↑	Commentate↑
References						
Qwen3-VL-8B, blind	3.6	69.1	0.0	66.7	100.0	15.5
Human reference	80.0	61.1	77.8	77.2	73.0	69.5
Native streaming						
VideoLLM-online	88.5	87.8	3.8	18.1	52.0	64.8
LiveCC	0.0	100.0	3.2	2.6	3.0	23.9
MMDuet2	54.9	7.3	0.9	15.1	34.1	21.7
MOSS-Video	6.7	36.4	0.0	10.2	3.4	50.1
VideoChat3	87.1	100.0	10.4	21.2	28.8	61.0
Turn-based polling						
Flash-VStream	3.1	13.9	0.0	10.7	15.1	2.7
Qwen2.5-Omni	41.5	66.7	9.0	9.4	42.6	9.8
LLaVA-OV	18.8	86.7	7.7	23.7	8.3	0.0
Qwen3-VL-4B	2.7	63.4	27.3	29.4	87.1	22.5
Gemma-3n	3.0	67.2	5.6	4.6	4.5	0.0
MiniCPM-o	3.8	68.6	16.0	51.7	36.5	3.4
LLaVA-OV2	3.0	68.3	0.8	8.7	23.2	0.0
Qwen3-VL-8B	3.0	66.2	0.0	9.8	11.7	0.0
Gemini 2.5 Flash	0.0	71.4	0.0	0.0	4.5	0.0
Claude Sonnet 5	0.0	73.3	0.0	11.1	15.9	4.0
Offline temporal grounding						
Gemma-3n	96.4	100.0	15.9	39.2	49.4	57.9
Dispider	93.8	100.0	52.6	62.7	74.6	95.5
Qwen2.5-Omni	100.0	100.0	39.8	59.3	88.4	8.3
MiniCPM-o	96.6	100.0	6.5	36.4	32.6	46.3
Qwen3-VL-4B	96.5	100.0	9.3	35.8	46.6	35.3
LLaVA-OV2	97.6	100.0	19.4	38.6	63.1	79.7
Qwen3-VL-8B	97.9	100.0	11.7	40.9	48.2	50.9
Claude Sonnet 5	97.9	100.0	85.5	81.0	100.0	79.8
GPT-5.6 Sol agent	96.4	100.0	65.9	79.3	72.8	55.5
Claude Sonnet 5 agent	97.6	100.0	76.6	81.5	83.9	45.1
Native Real-time Interaction System (JoyAI-VL)						
Claude Opus 5	13.4	100.0	42.9	43.8	24.1	15.0
Qwen3-VL-8B	0.0	66.7	0.0	3.7	8.4	0.0
JoyAI-VL	15.7	73.0	10.3	29.1	22.4	21.5
Qwen3-VL-4B	0.0	94.7	0.0	1.2	0.0	0.8
Claude Sonnet 5	7.3	66.7	7.1	37.0	10.4	6.3
Agentic Interaction System (Qwen-MM-Plugins)						
Gemini 3.5 Flash	80.0	87.5	40.0	62.7	75.7	32.1
Claude Sonnet 5	68.8	100.0	54.2	47.5	47.4	15.0
GPT-5.6 Sol	93.8	100.0	38.5	69.6	30.7	16.5
Agentic Interaction System (Claude Code tools)						
Claude Sonnet 5	62.5	86.7	58.3	50.9	54.5	8.6

R AGENTIC SYSTEM DETAILS

Agents use the Qwen-MM-Plugins media tools (Qwen Team, 2026) to extract and inspect frames, crop images, and read timestamps. The Claude (Anthropic, 2026a) shell-tool run instead allows the built-in shell and file tools of Claude Code in a per-item working directory that holds the trimmed prefix and no answer keys. Before each poll, ffmpeg trims the source video to the current stream time. The agent receives only this prefix, its instruction, and earlier utterances. The GPT-5.6 Sol loop (OpenAI, 2026) allows at most twelve tool turns and 1,500 completion tokens per session. Claude and Gemini (Google DeepMind, 2026) use their command-line agent loops. Each polling session has a 240-second timeout. Timeout is not a latency measurement. Every reported polling run works in an isolated directory that holds prefix videos and tool configuration but no answer keys. Apart from the shell-tool run, the tools are restricted to visual inspection. In grounding, the Claude

Table 16: **Violations per reference event.** Premature and repeated replies are divided by each item's reference events and averaged over positive items. Zero silence is the share of positive items whose silence compliance is zero. Rows cover the 30 current configurations, and the agentic families are named by their tools, Qwen-MM-Plugins (Qwen Team, 2026) or Claude Code tools.

System	Premature per event↓	Repeated per event↓	Zero silence, %↓
References			
Qwen3-VL-8B, blind	1.91	2.76	66.2
Human reference	0.09	0.18	16.3
Native streaming			
VideoLLM-online	1.70	3.58	41.1
MMDuet2	7.50	7.04	76.4
Turn-based polling			
Flash-VStream	9.82	15.41	97.4
Qwen2.5-Omni	3.59	7.23	75.6
Qwen3-VL-4B	0.88	2.78	77.1
Gemma-3n	7.23	19.23	88.5
MiniCPM-o	1.44	3.40	71.7
LLaVA-OV2	5.84	16.23	84.8
Qwen3-VL-8B	4.45	14.83	92.0
Gemini 2.5 Flash	8.26	18.30	89.5
Claude Sonnet 5	3.32	9.26	84.8
Offline temporal grounding			
Gemma-3n	0.61	0.21	47.0
Qwen2.5-Omni	0.09	0.01	19.0
MiniCPM-o	0.71	0.15	36.0
Qwen3-VL-4B	0.78	0.15	36.4
LLaVA-OV2	0.41	0.06	31.1
Qwen3-VL-8B	0.39	0.14	35.7
Claude Sonnet 5	0.04	0.01	7.3
GPT-5.6 Sol agent	0.14	0.09	14.1
Claude Sonnet 5 agent	0.12	0.13	12.9
Native Real-time Interaction System (JoyAI-VL)			
Claude Opus 5	0.43	1.10	61.4
Qwen3-VL-8B	18.82	23.58	88.9
JoyAI-VL	2.83	8.04	64.2
Claude Sonnet 5	0.92	3.33	71.8
Agentic Interaction System (Qwen-MM-Plugins)			
Gemini 3.5 Flash	0.18	0.11	27.7
Claude Sonnet 5	0.52	0.70	40.2
GPT-5.6 Sol	0.41	0.63	36.8
Agentic Interaction System (Claude Code tools)			
Claude Sonnet 5	0.59	0.75	41.1

and GPT-5.6 Sol agents receive the whole video with the same plugins. They report the video second of the key evidence for each response, no later than the question time for query items. Their sessions time out after 1,800 and 900 seconds. Claude Sonnet 5 without tools receives the grounding prompt of the open backbones.

As shown in Table 19, both Claude Sonnet 5 agent runs raise silence compliance above polling. The media-tool run replies 706 times and leaves 11 items silent. The media-tool and shell-tool runs each retain 11,696 poll latencies, with medians of 28.72 and 37.17 seconds and 90th percentiles of 87.4 and 91.6 seconds.

S Additional Diagnostics

Oracle restraint. Table 20 applies a reference-based filter to existing Qwen3-VL-8B (Bai et al., 2025) and JoyAI-VL (Yao et al., 2026) traces on the items shared with the agent runs. It removes all negative-stream replies and retains only the first response attributed to each reference window. The filter deletes 3,418 Qwen replies and 1,977 JoyAI replies without rewriting retained content or

Table 17: **Polling-adapter ablations.** Sampling, window size, response history, and prompt variants change content, timing, and silence differently.

Configuration	Overall↑	Content↑	Timing↑	Silence↑
Sampling rate, Qwen3-VL-8B (Bai et al., 2025), fresh prompt, 16 frames				
4 fps	33.4	51.8	75.7	10.5
2 fps	34.8	54.2	77.8	10.9
1 fps	34.7	52.7	76.9	13.1
0.5 fps	35.4	53.6	76.6	14.5
Window length, Qwen3-VL-8B, fresh prompt, 2 fps				
16 frames	34.8	54.2	77.8	10.9
64 frames	35.6	54.3	78.8	13.8
256 frames	39.0	58.7	81.3	17.2
Fresh prompt versus reply history				
Qwen3-VL-8B, fresh prompt	34.8	54.2	77.8	10.9
Qwen3-VL-8B, reply history	32.9	41.8	76.0	25.6
Qwen3-VL-4B, fresh prompt	26.2	35.1	34.0	28.4
Qwen3-VL-4B, reply history	34.6	40.6	57.5	36.0
LLaVA-OV2-8B (An et al., 2026), fresh prompt	34.5	49.8	81.0	16.9
LLaVA-OV2-8B, reply history	24.9	31.1	72.2	21.6
Gemma-3n, fresh prompt	27.8	43.3	81.9	11.9
Gemma-3n, reply history	19.4	12.8	31.0	54.7
MiniCPM-o, fresh prompt	29.0	33.1	44.7	31.6
MiniCPM-o, reply history	39.1	39.6	60.5	48.5
Qwen2.5-Omni, fresh prompt	18.7	28.3	34.4	20.7
Qwen2.5-Omni, reply history	22.2	28.8	59.6	26.5
Prompt hints, Qwen3-VL-8B, fresh prompt, agent items				
Default prompt	33.9	57.7	75.5	5.9
Hint set v2	34.6	57.3	70.0	8.8
Hint set v3	34.8	56.9	70.2	7.2

Table 18: **Overall scores by model and controller.** Dashes mark model and controller pairs that were not run. The Claude Sonnet 5 (Anthropic, 2026a) agent cell lists the media-tool run and the shell-tool run, and the text details coverage.

	Overall↑			
Model	JoyAI-VL controller	Polling, fresh prompt	Polling, reply history	Agent loop
JoyAI-VL	33.8	–	–	–
Qwen3-VL-4B	35.5	26.2	34.6	–
Qwen3-VL-8B	32.4	34.8	32.9	–
LLaVA-OV2	–	34.5	24.9	–
Gemma-3n	–	27.8	19.4	–
MiniCPM-o	–	29.0	39.1	–
Qwen2.5-Omni	–	18.7	22.2	–
Claude Sonnet 5	43.3	44.6	–	56.0, 61.2
Claude Opus 5	30.5	–	–	–
GPT-5.6 Sol	–	–	–	59.1
Gemini 3.5 Flash	–	–	–	29.4

timestamps. Timing is unchanged, while silence compliance reaches 100.0 by construction. Overall rises from 33.9 to 66.2 for Qwen and from 38.4 to 63.9 for JoyAI. Applied to all 1,060 items of the four randomly thinned traces, the same filter leaves timing unchanged and lifts silence compliance to 100.0 for each. Content falls slightly because deleting replies changes segment-level text and counting answers. This privileged post-processing diagnostic measures the effect of suppressing violations with privileged reference annotations.

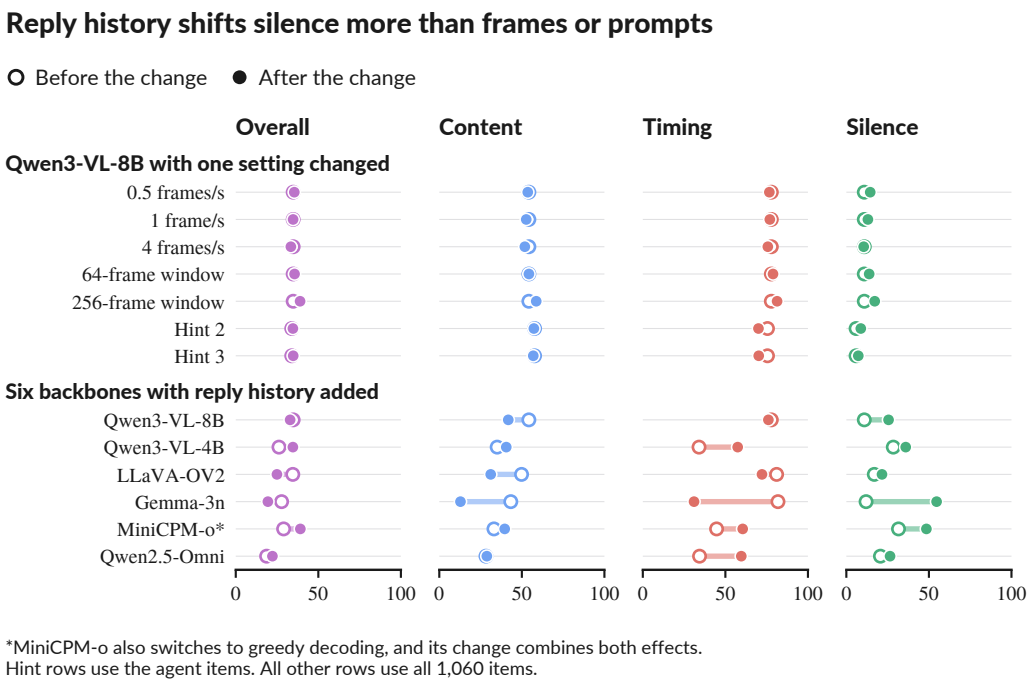


Figure 12: **History shifts silence more than frames.** Upper rows change one Qwen3-VL-8B (Bai et al., 2025) setting, and lower rows add reply history to six backbones.

Table 19: **Agentic systems raise silence.** All rows share the agent items. Agents use Qwen-MM-Plugins (Qwen Team, 2026) media tools except the shell-tool run. s/poll is the median wall-clock seconds per poll, and bold marks the best value in each column.

System	Overall↑	Content↑	Timing↑	Silence↑	s/poll↓
Claude Sonnet 5, polled (Anthropic, 2026a)	44.6	66.0	82.9	17.6	2.4
Claude Sonnet 5 agent	56.0	56.8	66.5	55.9	28.7
Claude Sonnet 5 agent, shell tools	**61.2**	**66.7**	70.7	55.5	37.2
GPT-5.6 Sol agent (OpenAI, 2026)	59.1	56.8	68.6	60.0	–
Gemini 2.5 Flash, polled	35.5	55.6	**93.1**	10.8	**1.9**
Gemini 3.5 Flash agent (Google DeepMind, 2026)	29.4	28.8	23.3	**66.9**	–

Native response threshold. VideoLLM-online (Chen et al., 2024) speaks when the probability of its frame-interval token falls below a threshold, released at 0.725. Table 21 reruns it at 0.5 and 0.9 on the items shared with the agent runs. Raising the threshold from 0.5 to 0.9 raises timing accuracy from 30.2 to 51.7 and lowers silence compliance from 56.4 to 18.4, while Overall falls from 17.2 to 7.4. The native trigger therefore moves along the same timing–silence trade-off as random thinning and does not escape it.

T Limitations and Future Directions

Coverage. Videos mainly last a few minutes and do not test hour-long monitoring. All items are visual, and audio channels such as speech and sound events remain outside the benchmark. Negative streams and counting-derived near misses cover only part of the space of unnecessary responses.

Scope of the scores. Stream-time evaluation isolates response behavior from hardware speed. It does not certify live deployment, and the latency measurements come from heterogeneous hardware.

Scope of the comparisons. Aggregate correlations across a few systems cannot identify causal bottlenecks. Existing comparisons of agentic systems change multiple components and lack a com-

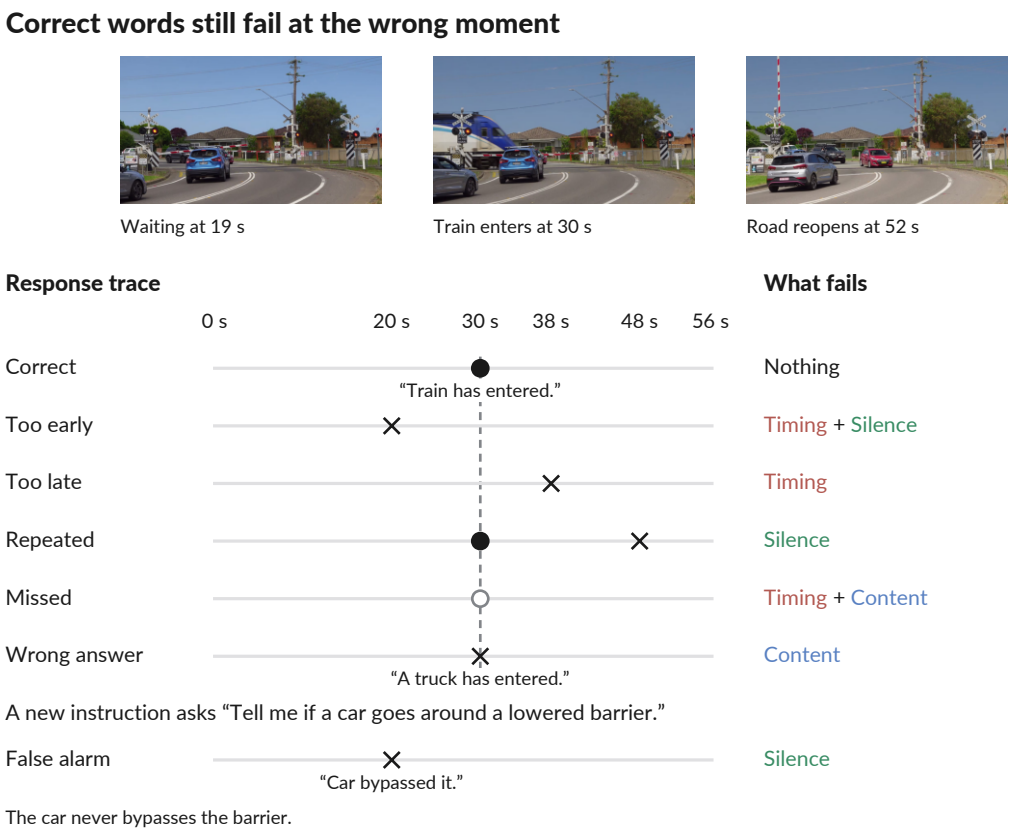


Figure 13: **Six ways to fail on one safety video.** The instruction at 0 s asks "Tell me once when the train enters the crossing." Constructed traces separate content, timing, and silence errors.

Table 20: **Oracle restraint.** Reference annotations remove premature, repeated, and negative-stream replies. Retained answers and timestamps are unchanged.

Configuration	Overall↑	Content↑	Timing↑	Silence↑
Qwen3-VL-8B, original	33.9	57.7	75.5	5.9
Qwen3-VL-8B, oracle restraint	66.2	56.0	75.5	100.0
JoyAI-VL, original	38.4	53.4	69.2	25.9
JoyAI-VL, oracle restraint	63.9	53.0	69.2	100.0

plete factorial design. Future work should pair identical items and model versions, test long negative intervals, and time every system on shared hardware.



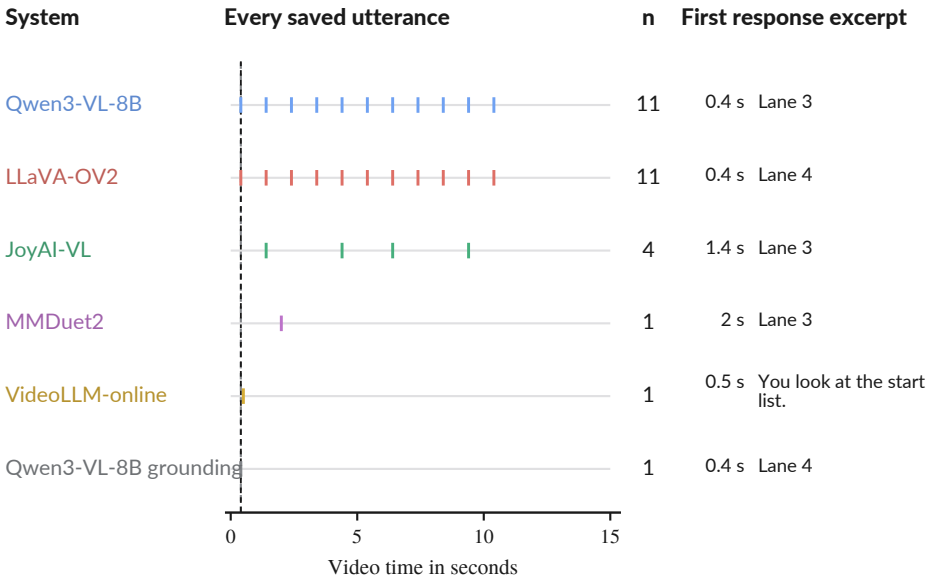
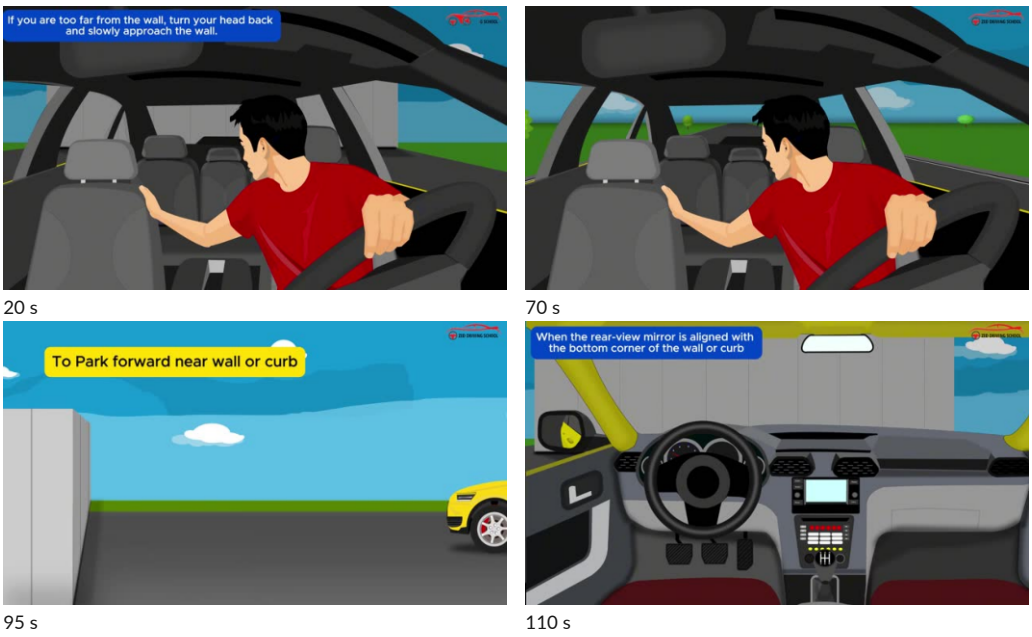
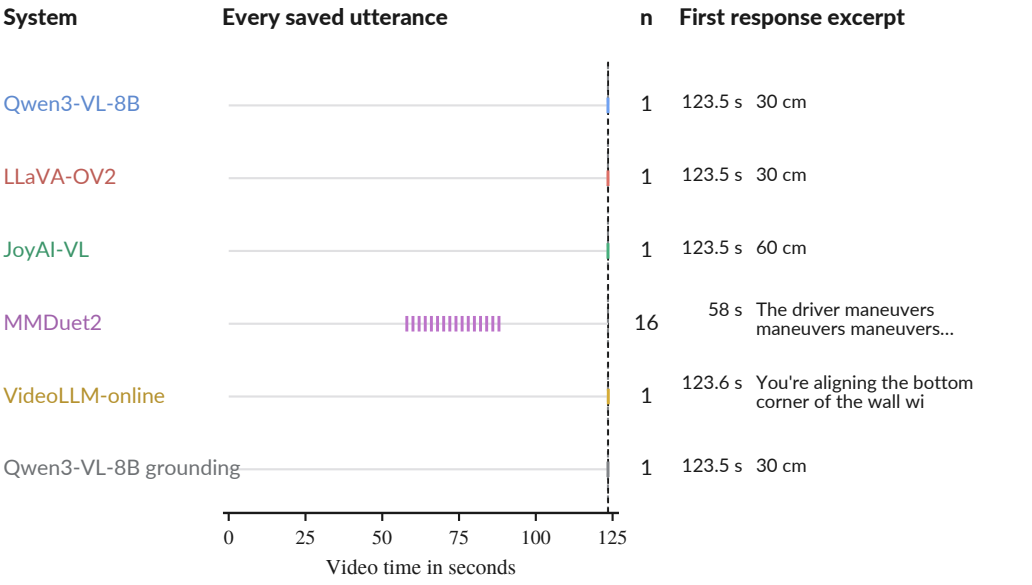


Figure 14: **Look case.** The user asks "According to the start list, which lane is Michael Phelps swimming in?" at 0.4 s in saved item `-6QYAYbS6eU#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.

One system speaks before the Recall question is asked



The expected answer is 30 cm, at the 123.54 s query.

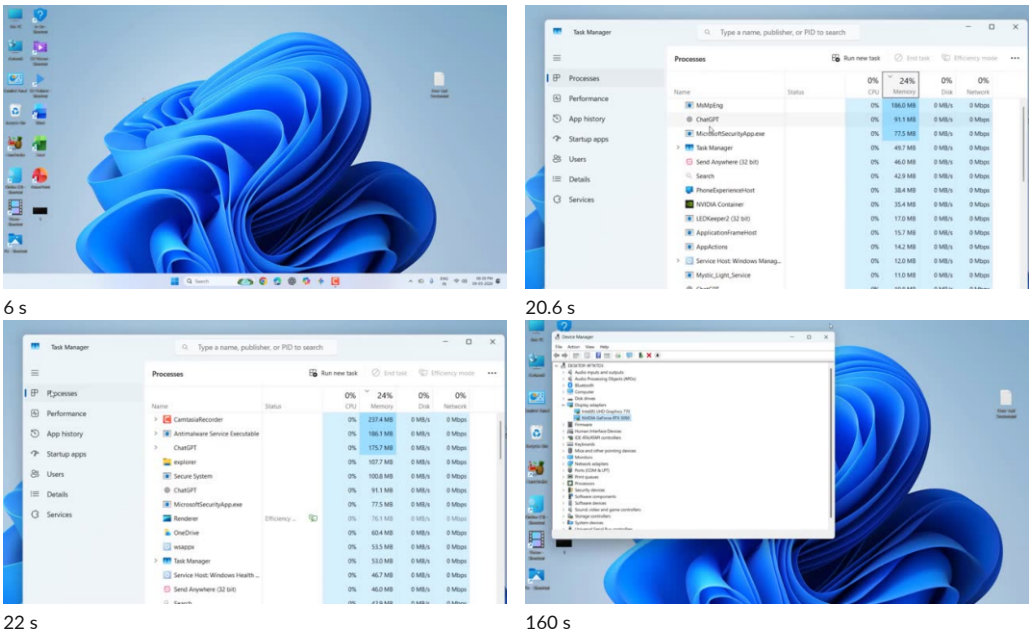


MMDuet2 emits 16 utterances before the query arrives.

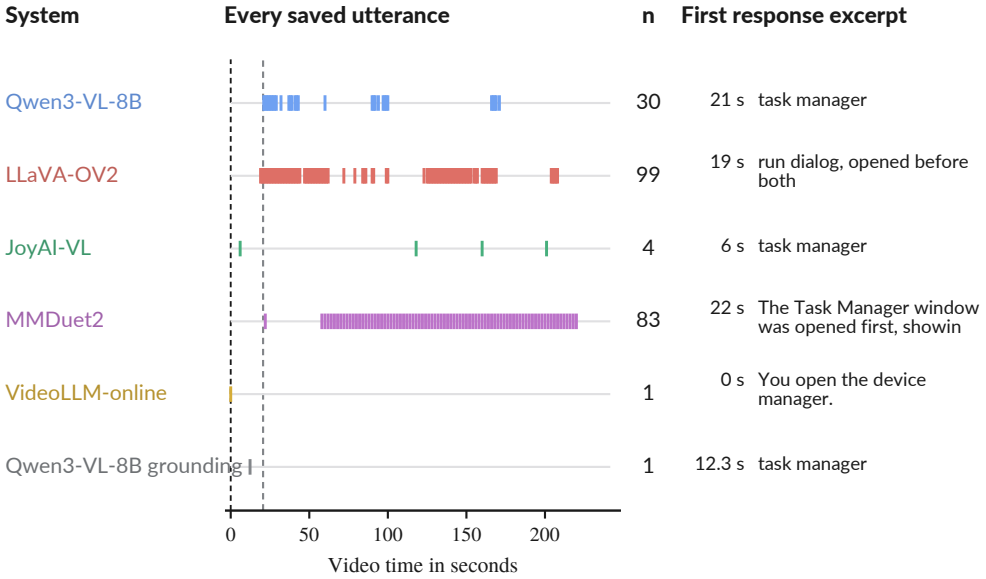
Grey dashes mark reference events, and the black dash marks the query. Grounding sees the whole video. n counts replies.

Figure 15: **Recall case.** The user asks "When you align the bottom corner of the wall with your front door handle, how far is the back of the car from the wall?" at 123.54 s in saved item `-Hv4pJjjZpY#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.

A correct Time answer can arrive before the evidence



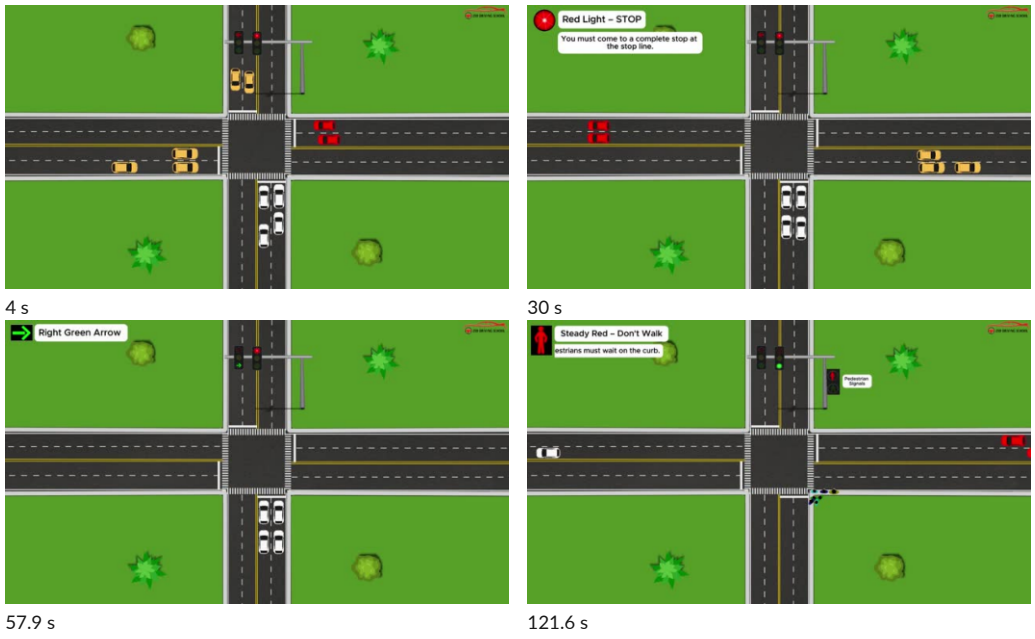
The expected answer is Task Manager. Evidence first appears at 20.6 s.



JoyAI-VL says Task Manager at 6 s, before the 20.6 s evidence.

Grey dashes mark reference events, and the black dash marks the query. Grounding sees the whole video. n counts replies.

Figure 16: **Time case.** The user asks "In the video, which was opened first, Task Manager or Device Manager?" at 0 s in saved item `-6qvDiJJ388#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.



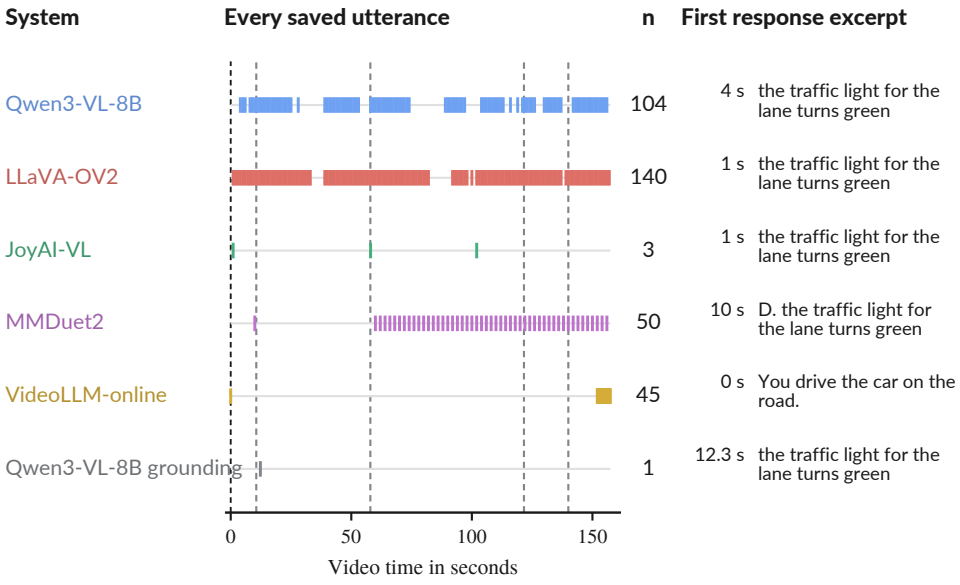
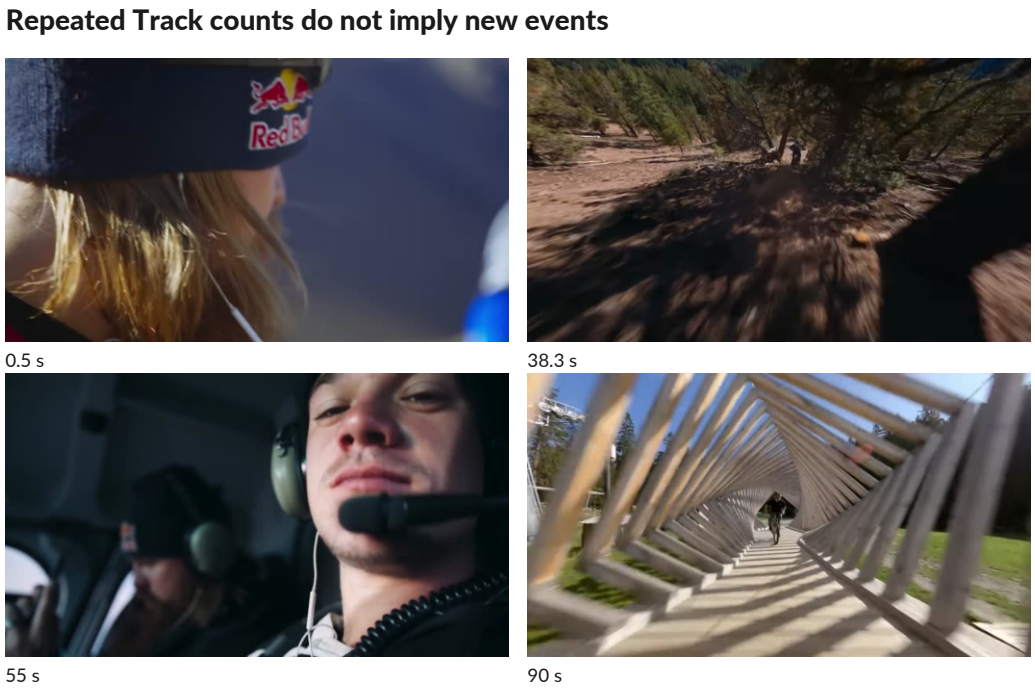


Figure 17: **Alert case.** The user asks "Let me know when the traffic light for the lane turns green." at 0 s in saved item `-31At72iGnk#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.



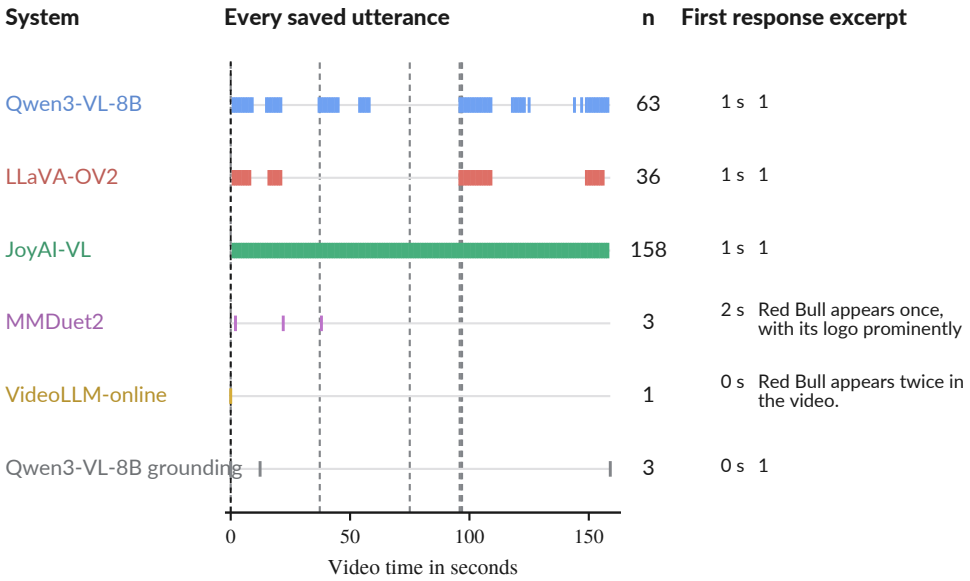
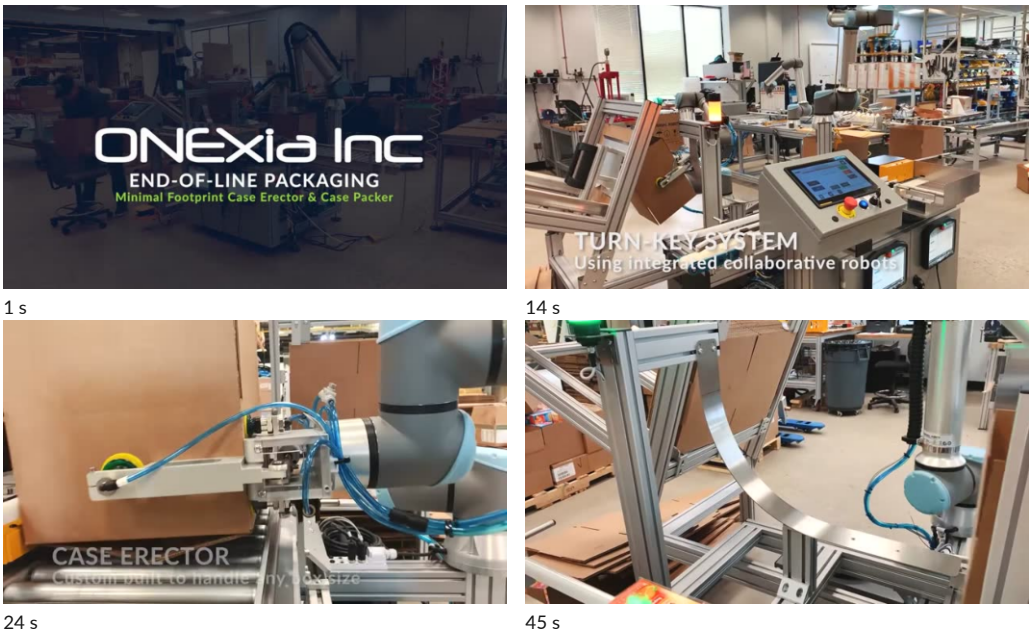
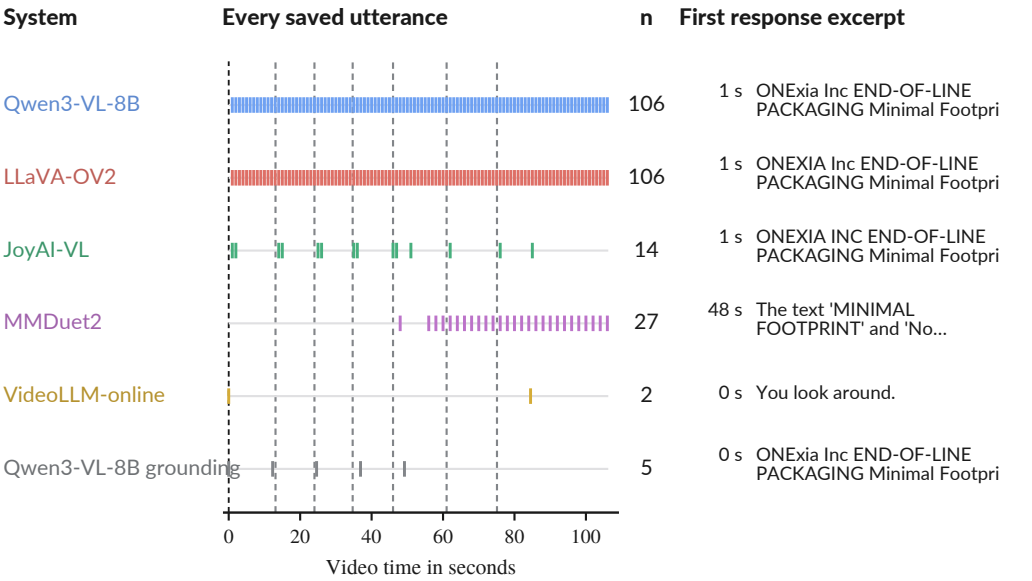


Figure 18: **Track case.** The user asks "How many times does Red Bull appear?" at 0 s in saved item `--VCWegJ_oY#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.

Dense Commentate speech repeats the same headings



The expected headings are Turn-key system → Case erector → Minimal user interaction → Minimal footprint → Custom conveyors → Case packer.



Qwen3-VL-8B and LLaVA-OV2 each speak at all 106 polls.

Grey dashes mark reference events, and the black dash marks the query. Grounding sees the whole video. n counts replies.

Figure 19: **Commentate case.** The user asks "Output each on-screen feature heading in real time." at 0 s in saved item `-CESiuu7CYs#0`. Source frames, reference events, and six saved response traces. Excerpts quote the first reply and truncate long text.

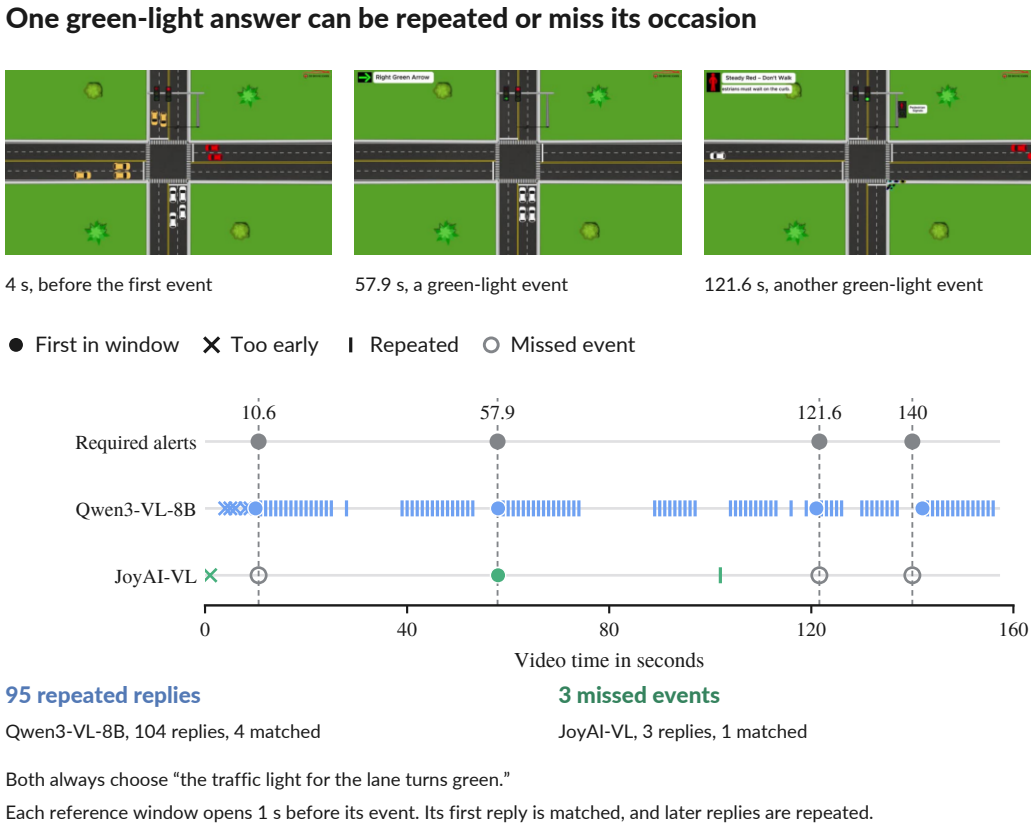


Figure 20: **A correct answer can be repeated or missed.** The saved Alert instruction is “Let me know when the traffic light for the lane turns green.” Rows mark the required alerts and each system’s replies on one saved Alert item, classified by the response-window rule.

Table 21: **VideoLLM-online at three response thresholds.** The 0.725 row is the released setting of VideoLLM-online (Chen et al., 2024), scored on the same items.

Threshold	Overall↑	Content↑	Timing↑	Silence↑
0.5	17.2	6.3	30.2	56.4
0.725	16.2	7.3	34.1	47.0
0.9	7.4	1.0	51.7	18.4