SVI-BENCH: EVALUATING HUMAN-PERCEIVED INTERACTION TRAJECTORIES IN STREAMING VIDEO SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Users do not experience a streaming vision-language assistant as a collection of isolated responses. They experience a continuous interaction trajectory in which the assistant watches a changing scene, decides when to speak or stay silent, reacts with some delay, and remembers information over time. The quality of this interaction therefore depends not only on what the assistant says, but also on how its behavior is aligned with the evolving visual context. To better evaluate these interaction demands, we introduce SVI-Bench, a trajectory-centric benchmark for streaming vision-language interaction. SVI-Bench evaluates five user-observable aspects: proactive triggering, silence correctness, latency, response correctness, and delegation with long-horizon memory. It contains 75 expert-authored items across nine realistic interaction categories, with 272 anchored dimension-level criteria. Central to SVI-Bench is InteractFlow, a trace-based evaluation framework that synchronizes model outputs with the visual states visible when they occur. This turns temporal interaction from abstract metadata into directly inspectable multimodal evidence and allows responses to be judged in the visual context in which they actually occurred. It also preserves each system's streaming behavior and supports traceable rescoring. Across five deployed configurations, JoyAI-VL-Interaction achieves the highest Overall score. The results also show that current systems are much better at staying silent than at speaking when needed. SVI-Bench therefore shifts streaming evaluation from scoring individual responses to evaluating the complete interaction experienced by the user. All videos, item-specific scoring criteria, recordings, and InteractFlow will be released.

1 INTRODUCTION

Real-world visual streams place vision–language assistants in settings where users experience not isolated answers, but continuous interaction over time: a safety monitor may need to flag a fire as it appears, and a video-call assistant may need to notice a fleeting expression without interrupting unnecessarily. Vision–language models have begun to move from offline clips to continuously arriving streams (Chen et al., 2024; Qian et al., 2024; Zhang et al., 2024; Yao et al., 2025), with emerging systems that perceive over time, decide when to speak, and delegate complex requests to background agents (Ding et al., 2025; Xu et al., 2026a). This motivates a trajectory-centric view of streaming evaluation, in which a user observes the evolving scene together with the assistant's behavior on one timeline. We call this user-observable sequence of visual states, interventions, silences, delays, responses, and retained context the *human-perceived interaction trajectory*, and we evaluate that behavior rather than subjective preference.

Recent benchmarks have advanced streaming video evaluation, covering causal understanding, temporal grounding, proactive response, persistent request management, and interaction over evolving videos (Lin et al., 2026; Niu et al., 2025; Shi et al., 2026; Wang et al., 2025b;a; Li et al., 2026). These works increasingly use streaming trajectories during inference, but evaluation is still often organized around task-specific outcomes, such as answer correctness or trigger timing. We take a complementary view: the complete user-observable trajectory should itself be the common object of evaluation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

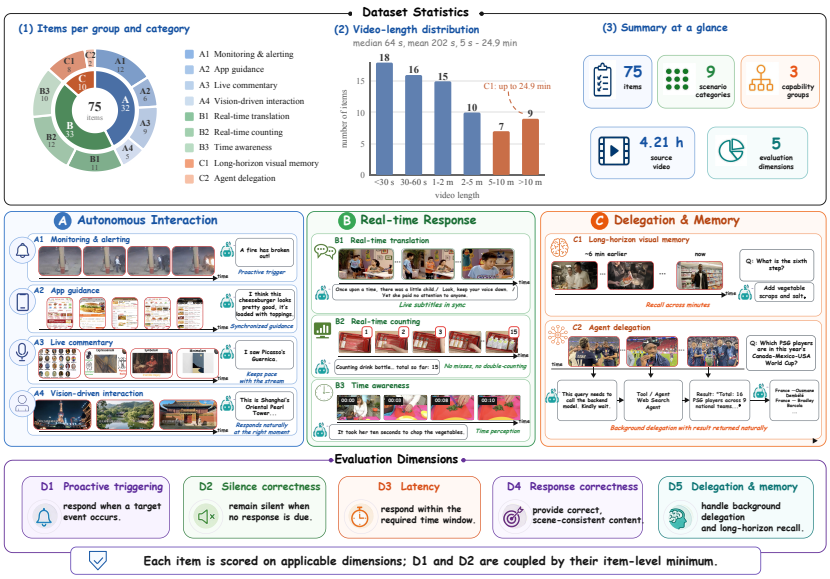


Figure 1: **Overview of SVI-Bench.** The first release holds 75 items in nine categories under three capability groups (left), spanning 4.21 h of source video from 5 s to 24.9 min (right). Every item is scored on its applicable dimensions among D1–D5.

We introduce SVI-Bench, a benchmark built around this trajectory-centric view. It separates scenario design from scoring: nine realistic interaction categories are evaluated with the same five user-observable dimensions, proactive triggering (D1), silence correctness (D2), latency (D3), response correctness (D4), and delegation & memory (D5). Each applicable item–dimension pair is scored on a three-level scale $(1/0.5/0)$ against anchors written as a timestamp, a count, or a set of information points. The first release contains 75 items and 272 dimension-level anchors. Figure 1 summarizes the benchmark.

A trajectory-centric benchmark also requires trajectory-centric evidence, and a timestamp alone does not show what was visible when a response occurred. We therefore develop InteractFlow, a unified recording and evaluation pipeline that exposes every system to the same source video, instruction, query timing, sampling schedule, recording frontend, and scoring rule while preserving its declared streaming interface. It synchronizes model outputs with the visual states visible when they occur, so responses are judged in the context in which they actually occurred, and stores each run as a trace that can be inspected or rescored without rerunning the system.

The Overall score treats speaking and silence as complementary parts of one interaction: wherever D1 and D2 both apply, their minimum is used, so neither staying silent nor speaking indiscriminately earns standalone credit. Across five deployed configurations and 375 recordings, JoyAI-VL-Interaction obtains the highest Overall score, 53.61. More importantly, proactive triggering emerges as a shared bottleneck, with the raw D1 mean never exceeding 29.09 while silence correctness reaches 91.84, an imbalance showing why isolated behaviors do not fully characterize interaction quality.

Our contributions are as follows. We introduce SVI-Bench, which treats the human-perceived interaction trajectory as the common object of streaming evaluation, scoring its 75 items through five user-observable dimensions and 272 expert-authored anchors. We develop InteractFlow, a trace-based evaluation framework that preserves system-specific streaming behavior and synchronizes visual states with model outputs, turning timing into directly inspectable multimodal evidence. Using a judge calibrated against a 20-item human gold standard, we evaluate five deployed configurations across 375 recordings and reveal a shared weakness in proactive intervention.

2 Related Work

Recent vision-language models have extended video understanding from offline clips to continuously arriving streams (Chen et al., 2024; Qian et al., 2024; Zhang et al., 2024; Yao et al., 2025),

Table 1: Interaction dimensions and evaluation perspectives in streaming vision-language benchmarks. For D1–D5, ✓ denotes a dedicated, separately reported evaluation, ○ a capability present without a dimension-level score, and a dash no evaluation. Under *human-perceived trajectory*, ✓ marks benchmarks that preserve the complete user-observable trajectory and use it directly as scoring evidence.

Benchmark	Interaction dimensions					Evaluation perspective
	D1 Proactive triggering	D2 Silence correctness	D3 Latency	D4 Response correctness	D5 Delegation & memory	Human-perceived trajectory
StreamingBench (Lin et al., 2026)	○	○	–	✓	○	○
OVO-Bench (Niu et al., 2025)	○	○	○	✓	○	○
OmniMMI (Wang et al., 2025b)	○	✓	–	✓	○	○
ProactiveVideoQA (Wang et al., 2025a)	○	○	○	✓	–	○
RIVER (Shi et al., 2026)	○	○	○	✓	○	○
Streaming-Eval (Yao et al., 2026)	✓	✓	✓	✓	○	○
IPIBench (Li et al., 2026)	✓	✓	○	✓	○	○
SVI-Bench	✓	✓	✓	✓	✓	✓

which requires causal processing under memory and latency constraints, through temporally aligned modeling (Chen et al., 2025), compressed or hierarchical memory (Chatterjee et al., 2025; Kang et al., 2025), efficient KV-cache mechanisms (Zhang et al., 2026a; Xu et al., 2026b), and event-aware computation over continuously updated visual context (Liu et al., 2024; Ding et al., 2025; Guan et al., 2026). In parallel, agents anticipate user needs without explicit prompts (Lu et al., 2025; Bandyopadhyay et al., 2025; Yang et al., 2025), extending this to long-horizon tasks through progress tracking (Xu et al., 2026a), hierarchical working memory (Hu et al., 2025), multimodal long-term memory (Li et al., 2024; Long et al., 2026), and delegation (Ning et al., 2026). Proactivity, memory, and delegation are often studied separately from when a model responds and stays silent.

Benchmarks have correspondingly extended evaluation from offline video question answering toward streaming and proactive interaction: StreamingBench, OVO-Bench, RIVER, and OmniMMI study causal understanding, temporal grounding, and response generation over evolving streams (Lin et al., 2026; Niu et al., 2025; Shi et al., 2026; Wang et al., 2025b); ProactiveVideoQA examines whether and when a model should respond (Wang et al., 2025a); Streaming-Eval evaluates proactive response, punctuality, and coherence over temporally distributed evidence (Yao et al., 2026); and IPIBench evaluates proactive monitoring, persistent request management, and interleaved reactive–proactive coordination (Li et al., 2026). These increasingly consider user-facing temporal behavior; SVI-Bench differs in how that behavior is evaluated. As Table 1 summarizes, it applies one anchored D1–D5 rubric across realistic scenarios and treats the human-perceived trajectory as the common object of evaluation, while InteractFlow records synchronized traces binding outputs to the visual states in which they occurred, so timing and response behavior are judged from directly inspectable evidence.

3 Evaluation Protocol

Streaming interaction cannot be scored from response content alone: a system must also decide when to speak, when to stay quiet, how fast to respond, and whether earlier information remains available. Offline benchmarks, which ask one question about one fixed clip, do not expose these behaviors.

	Dimension	Definition	Three-level anchors $(1/0.5/0)$
D1	Proactive triggering	When a target event occurs, is the system triggered to respond—no misses, no duplicate firings?	fires exactly once / one miss or one duplicate firing / $\geq$2 misses or duplicate firings
D2	Silence correctness	When no response is due, does the system stay silent—no barge-in, no hallucination?	fully silent / one mild self-utterance / $\geq$2 such utterances, or a long spurious monologue
D3	Latency	Delay from the annotated event or scheduled query to the first token, or to the key information where the anchor is defined on content; scored 0 if no valid response.	$\leq$1 s / $(1, 2]$ s / $>$2 s
D4	Response correctness	Is the content correct: consistent with the scene, covering the expected key information, and meeting the query's content and format requirements?	item-specific (key-information coverage; count/interval error)
D5	Delegation & memory	Does the system delegate a complex request to a background agent, recall past frames and dialogue accurately, and stay in streaming interaction while the delegated task runs?	item-specific (correct delegation + faithful recall + no stall)

Table 2: **The five scoring dimensions.** Which dimensions an item exercises is recorded per item and summarized per category in Table 3. The per-item anchors for all 75 items are released with the benchmark.

3.1 An Anchored Five-Dimension Rubric

We separate *what* we test, the scenario taxonomy of §4, from *how* we score it. The same five dimensions apply across all scenarios, so a monitoring item is judged not only on detecting the event but also on latency, silence, and response content, each on a three-level scale $(1/0.5/0)$ against the anchors of Table 2.

Observable anchors, defined per item. For every applicable dimension the annotator specifies the $1/0.5/0$ boundaries as a timestamp, a count, or a set of information points, and these anchors define what is checked in the synchronized record. D3 uses the recording clock for visually triggered items and the adapter-reported clock for explicit-query or startup items (Appendix A.5). Inapplicable dimensions are omitted.

Trigger timing. Event annotations and system sampling both have finite temporal granularity, so D1 accepts a small lead under a strict delay bound. With $t^{\mathrm{gt}}_{i,j}$ the timestamp of the j-th annotated event and $t^{\mathrm{pred}}_{i,j}$ the observation timestamp of the frame triggering the matching response, both on the video timeline of item i, a response is a valid trigger when

$$-0.5\,\mathrm{s} \;\leq\; t^{\mathrm{pred}}_{i,j} - t^{\mathrm{gt}}_{i,j} \;\leq\; 1\,\mathrm{s}. \tag{1}$$

An earlier utterance is a premature fire, charged as a D2 silence violation when D2 applies and otherwise as an invalid D1 trigger. Appendix A.4 separates D1 and D3 timing and treats coarse decision granularity.

3.2 Scoring and Aggregation

Aggregation is absolute rather than comparative: every system is scored independently against the item anchors. Let $s_{i,d} \in \{1,\, 0.5,\, 0\}$ be the score of system m on dimension d of item i and D_i the applicable dimensions. Figure 2 shows the path from item anchors to reported scores, together with the pilot that calibrates the automatic judge (§3.3).

Dimension applicability. D_i is the intersection of the dimensions item i exercises with those evaluable through the tested system's exposed channels: an item with no silence window is not

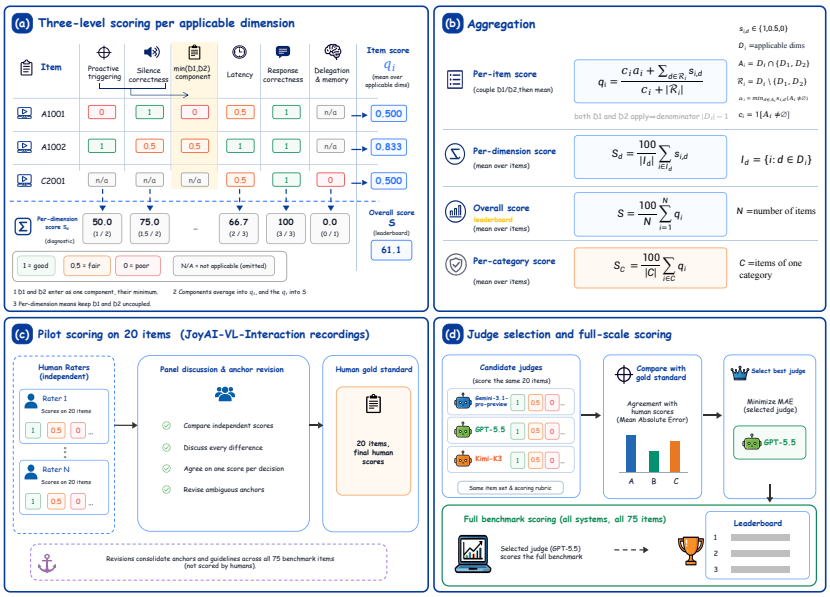


Figure 2: **From item anchors to benchmark scores.** **(a)** Each applicable dimension is scored $1/0.5/0$ against its item anchor. **(b)** D1 and D2 are coupled by their minimum within each item before item scores are averaged into the Overall score. **(c)** A human pilot consolidates anchors and produces the gold standard. **(d)** Candidate judges are compared on the pilot; the closest judge is selected, frozen, and applied to the full benchmark.

scored on D2, and a dimension is dropped for a system only when it cannot expose the required input or output channel. A missing internal module is not itself an exemption, so a system run as a complete interaction system that fails to provide required triggering or delegation scores 0 rather than N/A. N/A is excluded from the denominator and kept distinct from 0.

Coupling triggering with silence. D1 and D2 are complementary halves of one behavior: a system is autonomously interactive only if it speaks when a response is due *and* stays quiet otherwise. Scored independently they can be satisfied separately, so a system that never speaks scores 0 on D1 while collecting full marks on every silence item, and one that speaks constantly does the reverse. We therefore couple them inside each item. With $A_i = D_i \cap \{\mathrm{D1}, \mathrm{D2}\}$, $\mathcal{R}_i = D_i \setminus \{\mathrm{D1}, \mathrm{D2}\}$, and $c_i = \mathbf{1}[A_i \neq \varnothing]$, the coupled component and item score are

$$a_i = \min_{d \in A_i} s_{i,d} \quad (A_i \neq \varnothing), \qquad q_i = \frac{c_i\, a_i + \sum_{d \in \mathcal{R}_i} s_{i,d}}{c_i + |\mathcal{R}_i|}. \tag{2}$$

When both apply they contribute their minimum as one component, so the denominator is $|D_i| - 1$; when only one applies its score is retained. In the first release 41 of 75 items exercise both, and the coupling binds them there; Appendix A.4 covers the remaining cases.

Overall score. With N items, the Overall score weights every item equally:

$$S = \frac{100}{N} \sum_{i=1}^{N} q_i. \tag{3}$$

S is item-weighted, so a category contributes in proportion to its item count. It is the only score we rank systems by.

Diagnostic views. Two summaries are reported for analysis: a category score $S_{\mathcal{C}} = \frac{100}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} q_i$ localizing which scenarios a system fails, and a per-dimension score $S_d = \frac{100}{|I_d|} \sum_{i \in I_d} s_{i,d}$ over $I_d = \{\, i : d \in D_i \,\}$, which keeps one dimension uncoupled and is reported with its effective item count $|I_d|$. Because S_{D1} and S_{D2} average over their own item sets, neither enters the leaderboard on its own.

Group	ID	Category	#It.	D1	D2	D3	D4	D5	$\sum_i \lvert D_i \rvert$
A. Autonomous interaction	A1	Monitoring & alerting	12	12	12	12	12	0	48
	A2	App guidance	6	6	6	6	6	0	24
	A3	Live commentary	9	9	6	9	9	0	33
	A4	Vision-driven interaction	5	5	5	5	5	0	20
B. Real-time response	B1	Real-time translation	11	11	0	11	11	0	33
	B2	Real-time counting	12	12	12	12	12	0	48
	B3	Time awareness	10	0	8	8	10	10	36
C. Delegation & memory	C1	Long-horizon visual memory	8	0	0	8	8	8	24
	C2	Agent delegation	2	0	0	2	2	2	6
		Total	**75**	**55**	**49**	**73**	**75**	**20**	**272**

Table 3: **Nine categories with dimension applicability.** Each D-column counts the items of that category exercising the dimension; the total row gives the effective counts $|I_d|$ used as denominators in S_d. The last column is $\sum_i |D_i|$. Category scopes appear in Appendix A.3.

3.3 Pilot, Gold Standard, and Judge Selection

Human scoring cannot cover every item and system, so we calibrate the automatic judge on a 20-item pilot of JoyAI-VL-Interaction recordings covering all nine categories. Three experienced annotators write the anchors for all 75 items; a disjoint panel of five to seven scorers then scores the pilot items independently, seeing neither another's scores nor amending anchors while scoring; the panel finally discusses disagreements until it agrees on one score per decision, revising ambiguous anchors and their counterparts in the same category. The agreed scores form the human gold standard, and the consolidated anchors are frozen before the four comparison systems are recorded. Candidate judges (§5) then score the same items under the same rubric; we select the judge closest to the gold standard and apply it to all 75 items across all systems.

4 Dataset

The SVI-Bench item bank is designed to make streaming interaction behavior observable and reproducible. Each item must show whether a system speaks when a response is due, stays quiet when none is, responds within the required window, and preserves relevant context; each must also be scoreable against three-level anchors.

The design follows three principles. Items target streaming-response differences rather than general puzzle-solving or recognition difficulty; each item has an explicit streaming structure of trigger points, response windows, or silence intervals, so item-level anchors can be written; and all videos are free of business content and screened for sensitive, inappropriate, and political material, allowing the full set to be released. Admitted samples are de-duplicated within a category, keeping only the best when scene, prompt, or source video repeat. A category specifies the streaming scenario, while the five dimensions of §3.1 specify how it is scored.

4.1 Data Collection

Each item uses one source video lasting 5 s to 24.9 min, drawn from public sources and our own recordings. We first map existing material to the taxonomy, then collect additional footage for under-covered categories: A1 monitoring & alerting, C1 long-horizon visual memory, and C2 agent delegation. Three C1 items, at 24.0, 22.0, and 24.9 min, exceed the publicly reported context envelope of many comparison systems, since at 1 fps a 12.8 min run already requires 768 sampled frames, and they probe long-duration retention for systems without an explicit memory mechanism.

4.2 Annotation

Annotation fixes every item into observable, reproducible criteria: the prompt and its send time, the trigger points $t^{\mathrm{gt}}_{i,j}$ and expected response windows, the expected key information, and the applica-

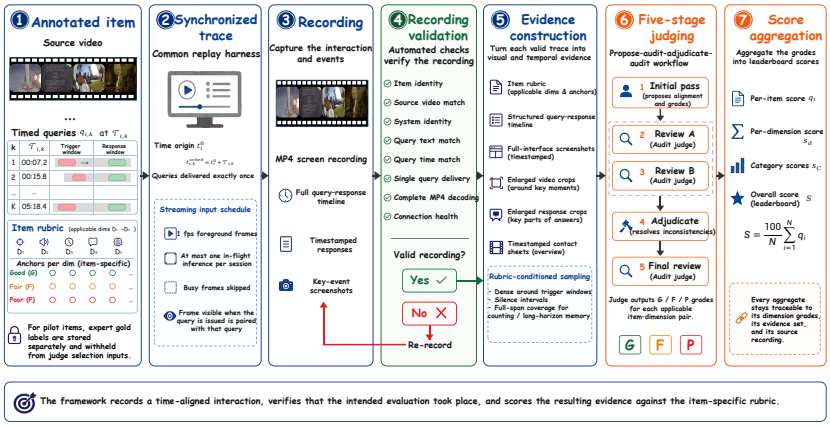


Figure 3: **Overview of InteractFlow.** The pipeline executes the shared streaming task, records synchronized traces, validates technical correctness, builds rubric-conditioned evidence, applies the frozen judge, and aggregates item scores.

ble D1–D5 dimensions with their three-level anchors. Appendix A.3 gives the full field schema, identifier convention, and quality-control procedure.

5 INTERACTFLOW: AUTOMATED EVALUATION

Evidence for triggering, silence, and latency depends on both the input and output timelines, so differences in playback, prompt delivery, or recorded layout would confound the score. Every system faces the same observable task contract (Figure 4, Appendix A.4), while the backend is not standardized: purpose-built streaming systems keep their released interfaces, adapters, parsers, and default inference settings, preserving native triggering and memory mechanisms. Doubao, lacking the stateful causal interface this protocol requires, runs under the JoyAI-VL-Interaction scaffold, so its D1 and D2 characterize the complete Doubao-plus-scaffold configuration (Appendix A.2).

For each item we construct a synchronized trace holding the causal visual observations delivered to the system, timestamped queries, and timestamped outputs, then score it by aligning outputs with the annotated events and intervals of the item rubric. **InteractFlow** operationalizes this at scale through the seven stages of Figure 3. An annotated item specifies the source video, evaluation horizon, timed queries, and rubric, while expert gold labels are withheld from both the evaluated system and the judge. Each system replays the same uploaded MP4 in the shared harness, the session is captured as a screen recording with timestamped query–response events, and automated checks verify that the intended evaluation took place (Appendix A.2). A failed check is re-recorded as a technical failure, whereas valid silence, late response, or wrong response is kept as system behavior. Each valid trace becomes rubric-conditioned evidence, is graded by a frozen sequential judge, and is aggregated as in §3.2, with recording and scoring decoupled so a valid trace can be rescored without rerunning the system.

5.1 CAPTURING A SYNCHRONIZED INTERACTION TRACE

Each annotated item i is represented as

$$\mathcal{T}_i = \left(V_i, H_i^{\text{eval}}, \mathcal{Q}_i, \mathcal{B}_i\right), \qquad \mathcal{Q}_i = \{(u_{i,k}, \tau_{i,k})\}_{k=1}^{K_i}, \tag{4}$$

with source video V_i, evaluation horizon H_i^{eval}, the k-th of K_i queries $u_{i,k}$ scheduled at video time $\tau_{i,k}$, and rubric $\mathcal{B}_i$ holding the trigger annotations $\{t_{i,j}^{\text{gt}}\}$ of Eq. (1), the expected response windows, the required information, and the anchors of the applicable dimensions D_i. Executing $\mathcal{T}_i$ with system m produces a trace $\mathcal{E}_{i,m}$ recording the observations actually delivered, the scheduled and realized query events, and the timestamped outputs. Each item runs in a fresh stateful session, with real-time sessions, cached history, and KV caches reset at item boundaries, and the verified start of playback defines the time origin t_i^0.

Loading the complete MP4 into the interface does not give the system offline access to the file. During playback the foreground scheduler samples the displayed frame once per second and permits at

most one in-flight inference per session, skipping rather than queueing an ordinary frame that arrives while the system is busy. At query dispatch the displayed frame is captured and bound to that query, and the pair waits in first-in, first-out order without being replaced by a later frame, so query-time context remains causal. Because a relevant transition may occupy only a few frames and response text may be too small to read, each validated recording also yields a rubric-sampled evidence set of timestamped interface views, enlarged crops, and temporally ordered contact sheets, with clock provenance retained and human scores withheld. Appendix A.5 specifies session initialization, the recording horizon, D3 clock sources, and the evidence budget.

5.2 Human-Calibrated Scoring

The automatic judge assigns a good/fair/poor grade to every applicable item–dimension pair. It first identifies target events and system outputs, aligns them in temporal order, and applies the item-specific thresholds. This event matching separates the failure modes: a missed target is a triggering error (D1), an utterance in a no-response interval a silence error (D2), a matched but delayed response a latency error (D3), and an on-time response with incorrect content a response-correctness error (D4).

Judging runs as a sequential propose–audit–adjudicate–audit workflow: an initial pass constructs an event–response alignment and proposes grades; two review passes independently uphold or revise each grade from the same timestamped evidence, with neither seeing the other; adjudication resolves their inconsistencies; and a final review audits the result against the rubric and original evidence. The final-reviewed output is reported.

We select the judge configuration on the 20-item expert pilot of §3.3. Let Ω be the applicable item–dimension pairs in the pilot, $s^*_{i,d}$ the arbitrated expert score, and $\hat{s}_{i,d}$ the judge score after mapping good/fair/poor to $1/0.5/0$. The primary calibration criterion is dimension-level mean absolute error,

$$\mathrm{MAE}_{\mathrm{dim}}(\Omega) = \frac{1}{|\Omega|} \sum_{(i,d)\in\Omega} \left|\hat{s}_{i,d} - s^*_{i,d}\right|, \tag{5}$$

complemented by exact grade accuracy, accuracy within one ordinal level, macro-F1, and quadratic-weighted Cohen's κ; Appendix A.6 reports the candidate results. After selection the judge model, prompts, evidence construction, and stage definitions are fixed and applied unchanged to all systems. Grades are aggregated as in §3.2, and every aggregate stays traceable to its dimension grades, evidence set, and source recording.

6 Experiments and Findings

We report the judge's agreement with the 20-item human calibration set, the benchmark results for five deployed configurations, and a same-trace comparison in which published scoring rules are applied to our recorded traces.

6.1 Experimental Setup

Each configuration is evaluated on the 75 items of §4, whose scheduled horizons sum to $\sum_i H_i = 15{,}153$ seconds (median 64 s, maximum 1,493 s) and contain $\sum_i K_i = 102$ timed query rounds across 16 multi-round items. The observable task is fixed across systems while the serving path is system-specific: JoyAI-VL-Interaction uses the JoyAI stateful streaming adapter; Doubao Seed 2.1 Pro is hosted by the JoyAI scaffold, which schedules inference, assembles context, and enforces pre-query silence, after which Doubao chooses silence or response whenever invoked; Mage-VL and MOSS-VL-Realtime use native stateful realtime-session adapters; and MiniCPM-O-4.5-9B uses a native full-duplex video adapter receiving each query once as deterministic text-to-speech audio (Appendix A.2). These names denote complete deployed systems rather than a hardware-controlled comparison of model weights. The corpus holds one accepted recording per system–item pair: 375 recordings and 510 scheduled query events. Following pilot calibration, InteractFlow uses GPT-5.5 with timestamped-image evidence and the five-stage procedure of §5.2, frozen before the run and applied unchanged to all systems; all 1,875 system–item–stage predictions complete, and applicable grades map to $1/0.5/0$ and are aggregated by Eq. (2) and Eq. (3) (Appendix A.7).

6.2 Agreement on the Human Calibration Set

Against the expert gold labels the final-reviewed judge reaches $\mathrm{MAE}_{\mathrm{dim}}(\Omega) = 0.0986$ on the $1/0.5/0$ scale, exactly matching 59 of 71 labels (83.10%), with 69 of 71 (97.18%) within one ordinal level, macro-F1 0.7762, and quadratic-weighted $\kappa = 0.7803$. Aggregating the pilot labels with Eq. (3) gives $S^* = 68.75$ for experts and $\hat{S} = 65.21$ for the judge. Table 10 in Appendix A.9 breaks the figures down by dimension, where D1 remains hardest because it requires event matching while distinguishing misses, premature responses, and duplicates.

6.3 Main Benchmark Results

Table 4 reports the frozen final-review results, where Overall S is the sole leaderboard score. JoyAI-VL-Interaction obtains the highest Overall score, 53.61, while the other four remain below 26; the raw dimensions in Appendix A.10 show where this comes from.

System	Overall $S \uparrow$	$S^{-\mathrm{D3}}$
JoyAI-VL-Interaction	**53.61**	39.67
MOSS-VL-Realtime	<u>25.44</u>	26.33
MiniCPM-O-4.5-9B	21.33	20.44
Doubao Seed 2.1 Pro	21.22	29.44
Mage-VL	19.50	21.78

Table 4: **Final judge results on the 75-item benchmark.** Overall (Eq. (3)) couples applicable D1/D2 scores within each item before averaging; $S^{-\mathrm{D3}}$ omits D3 after the same coupling and is a sensitivity analysis (Appendix A.10).

The coupled D1/D2 component exposes weak proactive triggering: every evaluated system has a lower raw D1 mean than raw D2 mean. The Doubao-plus-scaffold configuration illustrates the pattern, with $S_{\mathrm{D1}} = 17.27$ against $S_{\mathrm{D2}} = 91.84$, because Doubao frequently remains silent when invoked after the first query. Since JoyAI-VL-Interaction uses the same scaffold-level scheduling and output handling, this asymmetry cannot be explained by a separate content detector in the scaffold. The item-level minimum prevents the observable silence from becoming standalone leaderboard credit. Latency contributes to but does not determine the ordering: removing D3 keeps JoyAI first at 39.67 while narrowing its lead from 28.17 to 10.22 points (Appendix A.10).

Two diagnostic views appear in Appendix A.8 rather than as additional leaderboard scores. By scenario category, JoyAI is best or tied for best in eight of nine, while Doubao and MOSS score zero on real-time translation, so silence cannot satisfy continuous-output tasks. Re-scoring the same traces under the published OVO-Bench (Niu et al., 2025), OmniPro Online (Zhao et al., 2026), and IPIBench (Li et al., 2026) rules reproduces two failure modes a single scalar misses: a premature alert scores 100 under content matching but 0 under both timing rules, and three correct labels for six required appearances score 100, 44.4, and 0, whereas SVI-Bench gives both $100q_i = 33.3$ with a dimension vector separating correct content from failed triggering and latency.

7 Conclusion

SVI-Bench provides an anchored benchmark for streaming vision–language interaction, scoring 75 items across nine categories on five observable dimensions: proactive triggering, silence correctness, latency, response correctness, and delegation & memory. By treating timing, silence, and long-horizon state as explicit behaviors it makes interaction failures traceable beyond response content, and its item-level coupling of triggering and silence requires a system to speak when a response is due and stay quiet otherwise within the same case. InteractFlow turns this protocol into a reproducible pipeline that preserves native streaming interfaces and decouples recording from scoring, so each result traces back to item criteria, timestamps, and model outputs. Across five deployed configurations and 375 recordings, JoyAI-VL-Interaction obtains the highest Overall score while the others remain below 26, and proactive triggering emerges as the shared bottleneck: raw D1 never exceeds 29.09 even though silence correctness reaches 91.84. Current systems are therefore better at withholding responses than initiating timely and useful ones, making balanced streaming interaction a key target for future work. We release InteractFlow, source videos, recordings, and item-specific scoring criteria.

AI USE STATEMENT

Among the tasks for which ICLR requires disclosure, we used generative AI tools for three purposes: to help implement methods, where the design of InteractFlow was specified by the authors and AI coding assistants wrote and debugged parts of the recording adapters, trace synchronization, evidence construction, and scoring orchestration; to clean and reformat the dataset, where AI-assisted scripts converted the expert annotation workbook into the JSON item bank, changing the storage format while leaving item content, scoring criteria, and gold labels as authored by the human experts; and to interpret results, where AI tools helped analyze the score tables and recorded traces and draft parts of §6. We did not use generative AI to generate synthetic datasets, develop the conceptual framework, propose or refine hypotheses, design or give feedback on the methodology or experiments, or assist with translation. The nine categories, five dimensions, 75 items, and all 272 dimension-level criteria and gold labels were authored by the human experts described in §4.2. The tasks concerning mathematical claims and proofs do not apply to this work. Among the tasks for which disclosure is recommended, we additionally used generative AI to create and edit code, write the analysis and plotting scripts behind our tables and figures, search and summarize literature while preparing §2, draft parts of the paper, and edit the manuscript for readability.

We reviewed all AI-assisted work. AI-assisted code was read and tested by the authors, and every reported score is reproducible from the released recordings, criteria, and code (§5). The converted item bank was checked field by field against the source workbook before any evaluation run. Every AI-assisted interpretation and AI-drafted finding was verified against the score tables and synchronized recordings, with each number confirmed independently of the tool that surfaced it, and AI-suggested references were verified by retrieving and reading each cited paper. Generative models also appear here as the object of study: the five evaluated systems (§6) and the frozen judge inside InteractFlow (§5.2). We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

SVI-Bench is built from video recorded for this benchmark and from material cleared for open release. The videos contain no proprietary business content, individuals appearing in them were recorded with consent for research release, and no third-party personal data was collected. Annotation (§4.2) was performed by members of the research team and by compensated annotators under normal employment terms. Several source scenarios involve safety monitoring, where a missed or mistimed response corresponds to a real-world hazard; scores measure behavior on our item bank and do not establish that any evaluated system is fit for safety-critical deployment. One evaluated configuration, JoyAI-VL-Interaction, originates from the same organization as the authors; it is evaluated through the same harness, item bank, criteria, and frozen judge as every other configuration, and all recordings supporting its scores are released for inspection.

REPRODUCIBILITY STATEMENT

The evaluation protocol, including the five dimensions, three-grade criteria, and aggregation rules, is specified in §3; the item bank, taxonomy, video collection, and annotation procedure in §4; and InteractFlow, covering the recording harness, per-system adapters, validity checks, evidence construction, and frozen judge, in §5, with judge selection in §6.2 and Appendix A.6. The deployed configurations, their interfaces, and session-reset behavior are listed in Appendix A.1. Because recording and scoring are decoupled, every reported score traces back to item criteria, timestamps, and model outputs, and valid traces can be rescored without rerunning the evaluated system. We release the source videos, the item-specific scoring criteria, the 375 recordings, and the InteractFlow code.

REFERENCES

Saptarashmi Bandyopadhyay, Vikas Bahirwani, Lavisha Aggarwal, Bhanu Guda, Lin Li, and Andrea Colaco. Yeti (yet to intervene) proactive interventions by multimodal ai agents in augmented reality tasks. *arXiv preprint arXiv:2501.09355*, 2025.

Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgoz, Shreyas Hampali, Eric Sauser, Shugao Ma, et al. Streaming videollms for real-time procedural video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22586–22598, 2025.

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.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13448–13459. IEEE, 2025.

Yiran Guan, Liang Yin, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Video streaming thinking: Videollms can watch and think simultaneously. *arXiv preprint arXiv:2603.12262*, 2026.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32779–32798, 2025.

Hyolim Kang, Yunsu Park, Youngbeom Yoo, Yeeun Choi, and Seon Joo Kim. Open-ended hierarchical streaming video understanding with vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20715–20725. IEEE, 2025.

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

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in neural information processing systems*, 37:49881–49913, 2024.

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

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

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

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

Pu Ning, Quan Chen, Kun Tao, Xinyu Tang, Tianshu Wang, Qianggang Cao, Xinyu Kong, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Searchswarm: Towards delegation intelligence in agentic llms for long-horizon deep research. *arXiv preprint arXiv:2606.09730*, 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.

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

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.

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.

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, 2025b.

Lilin Xu, Bufang Yang, Siyang Jiang, Kaiwei Liu, Kaiyuan Hou, Yuang Fan, Hongkai Chen, Zhenyu Yan, and Xiaofan Jiang. Pro^2 assist: Continuous step-aware proactive assistance with multimodal egocentric perception for long-horizon procedural tasks. *arXiv preprint arXiv:2605.04227*, 2026a.

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

Bufang Yang, Lilin Xu, Liekang Zeng, Yunqi Guo, Siyang Jiang, Wenrui Lu, Kaiwei Liu, Hancheng Xiang, Xiaofan Jiang, Guoliang Xing, et al. Proagent: Harnessing on-demand sensory contexts for proactive llm agent systems. *arXiv e-prints*, pp. arXiv–2512, 2025.

Dingyu Yao, Shuhuan Gu, Qingyi Si, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Naibin Gu, Zheng Lin, Weiping Wang, Nan Duan, and Jiaqi Wang. Harnessing streaming video in the wild. *arXiv preprint arXiv:2606.08615*, 2026.

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

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.

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

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding, 2026b. URL https://arxiv.org/abs/2608.05703.

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, 2026. URL https://arxiv.org/abs/2605.18577.

A Reproducibility and Additional Results

A.1 Evaluation Interfaces and Reproducibility

SVI-Bench evaluates each deployed system end to end, including its declared serving stack. Table 5 reports the model identifier exposed by each service, the versioned API or endpoint contract used for evaluation, and the common evaluation horizon. The 75 tasks prescribe 15,153 seconds (4 h 12 min 33 s) of streaming evaluation per system. Capture may continue beyond this horizon only to finish an outstanding response or to establish a valid non-response; such post-horizon waiting is not additional benchmark content. For services that expose only a rolling model alias, we report the provider-visible identifier used during evaluation.

System	Model/API identifier	Endpoint and protocol version
JoyAI-VL-Interaction	`jdopensource/JoyAI-VL-Interaction`	OpenAI-compatible `/v1` backend through the released stateful streaming adapter
Doubao Seed 2.1 Pro	`doubao-seed-2-1-pro-260628`	Volcengine Ark API v3, `/chat/completions`, through the pinned live-interaction adapter
Mage-VL	`Mage-VL`	StreamMind stateful HTTP API over `/v1/chat/completions`
MOSS-VL-Realtime	`MOSS-VL-Realtime`	Stateful realtime HTTP session API v1 over `/v1/chat/completions`
MiniCPM-O-4.5-9B	`MiniCPM-O-4.5-9B`	ModelBest Realtime API v1, `/v1/realtime?mode=video`

Table 5: **Interfaces and prescribed evaluation duration of the five systems.** Identifiers and endpoint contracts are those exposed during evaluation. Every system uses the same 4 h 12 min 33 s prescribed horizon.

A.2 Recording Configuration and Model Adapters

Shared external contract. Table 6 summarizes the per-system state, visual, and query transport together with the adapter-specific behavior described in this subsection. The complete local MP4 is uploaded to the evaluation WebUI for playback and screen capture, but it is not submitted to the evaluated model as an offline video. The model receives only causal observations available on the displayed timeline. The foreground scheduler samples one current frame every second, allows one in-flight frame inference per session, and skips an ordinary frame that arrives while the model is busy. It never accumulates a backlog or sends future frames. A query is displayed and delivered exactly once at its annotated time. Its contemporaneous frame is protected in a FIFO queue if the inference slot is occupied, so a later frame cannot replace the query-time view. All rounds of one item remain in the same task session unless the native provider imposes a hard session limit.

Validity checks. Before target playback, the selected backend and multimodal route are checked; the target-video playback origin then defines time zero. The nominal horizon is H_i from Eq. (7). An in-progress response is allowed to finish, and its final visible state is retained for at least one additional second. A healthy session that remains silent for 180 seconds after the final required response opportunity is accepted as a model non-response. An inference deadline exceeded on an otherwise healthy session is classified in the same way. Upload, transport, decoding, rate-limit, provider, or protocol errors instead invalidate the attempt and permit up to five retries, each restarting the item from the beginning. A retry never depends on answer quality.

All accepted recordings use H.264 at 1600×1000 pixels and decode to completion.

No explicit thinking/reasoning switch was sent to the evaluated services; each provider's default setting was retained. The shared instruction was delivered through the system-prompt mechanism supported by each adapter and was never concatenated with a user query. Query text was delivered in the user role for the four text-query interfaces; MiniCPM-O received the same text after the fixed text-to-speech conversion described below.

For MiniCPM-O, the deterministic query-audio cache was generated with `edge-tts` 7.2.7, voice `zh-CN-XiaoxiaoNeural`, default rate, pitch, and volume, and converted to headerless 16-kHz mono `f32le` PCM in one-second chunks. For Mage-VL and MOSS-VL-Realtime, native session state

System	State and visual transport	Query transport	Adapter-specific behavior
JoyAI-VL-Interaction	Official stateful adapter; causal JPEG frames; native memory retained	User-role text, once per annotated round	Native policy selects silence, response, or delegation; state is reset between items.
Doubao Seed 2.1 Pro	JoyAI interaction scaffold over Ark; causal JPEG frames; scaffold-managed state and memory	User-role text, once per annotated round and bound to its current frame	The scaffold controls inference opportunities and pre-query silence; thereafter Doubao selects silence or response, and the scaffold renders it.
Mage-VL	StreamMind stateful HTTP session; reset between items	User-role text, once per annotated round, in the query-bound frame request	The query frame invokes interactive mode once; subsequent frames use native proactive gating. Text is preserved and only response boundaries are normalized.
MOSS-VL-Realtime	Native stateful HTTP session; causal JPEG frames	User-role text, once per annotated round, in the query-bound frame request	Native proactive mode is retained; later frame calls collect completed responses without replaying the query.
MiniCPM-O-4.5-9B	ModelBest WebSocket video mode; JPEG frames with silent 16-kHz mono audio	Deterministic TTS audio, once per annotated round	Native text deltas and turn boundaries are mapped to the common UI; 300-second sessions renew without replaying queries or earlier frames.

Table 6: **Model-interface and adapter configurations.** All rows use the shared 1-fps, one-in-flight, busy-skip foreground schedule. Adapters translate transport, roles, session control, and native response boundaries; they contain no item ids, expected answers, rubric thresholds, or task-specific event detectors.

is reset at item boundaries. For Doubao, the scaffold schedules causal inference, preserves query-bound frames, and maintains task-local visual and dialogue history. It suppresses output before the first query without invoking the model. Thereafter, Doubao generates either `</silence>` or a tagged response at each available inference opportunity. The scaffold interprets these markers and normalizes untagged plain text as a response. An inference timeout is represented as synthetic silence only when the session, transport, and provider health checks remain valid; otherwise the attempt is retried under the validity rules above. The scaffold does not inspect video content to decide when an answer is warranted. These scaffold-level scheduling and output-handling rules are shared with the JoyAI-VL-Interaction configuration.

Shared evaluated-system instruction. All adapters receive the same task-independent instruction, which defines three valid actions at each inference opportunity: remain silent, answer concisely, or delegate a request that cannot be answered reliably. It contains no item identifier, expected answer, trigger annotation, or scoring threshold. The complete instruction is included in the supplementary artifact.

A.3 Item Bank Schema and Temporal Strata

This subsection expands the dataset description of §4.

Item bank statistics. The item bank holds 32 autonomous-interaction, 33 real-time-response, and 10 delegation-and-memory items, totaling 4.21 h of source video. Length ranges from 5 s to 24.9 min (median 64 s, mean 202 s), with 49 items under 2 min and 16 over 5 min, covering time scales from second-level triggering to minute-level memory. Each item is one continuous streaming session scored once on every applicable dimension, yielding 272 dimension-level decisions per system, and the source videos are released with the annotations.

Table 3 reports dimension applicability. D1 ($|I_{\mathrm{D1}}| = 55$) and D2 ($|I_{\mathrm{D2}}| = 49$) overlap on 41 items, where the coupling of Eq. (2) binds; the 14 items with D1 alone are continuous-output scenarios with no legitimate silence window, and the 8 with D2 alone are interval-triggered time-awareness

items. The D1 vigilance span has a median of 52 s and a maximum of 12.6 min, while C1 recall requires video up to 24.9 min to stay accessible at query time; Appendix A.3 stratifies both.

Category scope. Table 3 lists the nine categories and their item counts. Their scopes are as follows. **A1** monitoring & alerting: continuously watch the scene and alert within one second of an urgent event. **A2** app guidance: follow the phone screen and guide the app task in sync with the current view. **A3** live commentary: narrate screen content in real time, keeping pace with a changing scene. **A4** vision-driven interaction: interact naturally as the scene changes, for example greeting a person the moment they appear. **B1** real-time translation: translate on-screen captions or speech in real time, following the video's pace rather than a single frame. **B2** real-time counting: count target objects or actions as they appear, without misses or double-counting. **B3** time awareness: act on a sense of elapsed time, such as timed reminders or interval triggers. **C1** long-horizon visual memory: retain details from minutes earlier, up to about 25 min, and recall them on demand. **C2** agent delegation: offload a hard sub-task to a background model or agent and fold the result back in naturally.

Annotation fields. Each item is annotated with the following fields.

- **Prompt and send time.** The instruction given to the system and its timestamp; multi-response items include a recurring trigger, and multi-turn items record each turn.
- **Temporal annotation.** Trigger points (timestamps $t_{i,j}^{\mathrm{gt}}$ on the video timeline, with multiple events separated by semicolons) and the expected response window (e.g., $\leq$1 s, $(1, 2]$ s after the trigger, or once every Δ seconds).
- **Expected key information.** The content points that must be covered, such as required textual content or the situations in which the system is expected to stay silent.
- **Five-dimension scoring criteria.** The applicable D1–D5 dimensions and their three-level anchors; inapplicable dimensions are omitted.

Item identifiers. An item id concatenates the capability-group letter (A/B/C), the in-group category index, and a three-digit serial number; for example, `A1001` is the first item of category A1 (monitoring & alerting). The three long C1 items of §4.1 are `C1008`, `C1009`, and `C1010`.

Quality control. After annotation, 20% of items are cross-checked to verify that the temporal annotations and three-level anchors are observable and that prompts follow the item-type conventions. Scores are never written into the item bank, which evolves over releases; each evaluation instead produces a separate scoring sheet that reads the criteria by id.

Temporal strata. Table 7 stratifies the two time-sensitive capabilities by horizon. For proactive triggering, the horizon is the vigilance span from prompt send time to the end of the stream; its median is 52 s, with one third of items under 30 s and one eighth over 5 min. For long-horizon visual memory, the horizon is the full video length available at query time; it ranges from 2.0 to 24.9 min, with three of eight items beyond 20 min.

Capability	Stratum	Horizon	#Items
Proactive triggering (D1, 55 items)	L1	$\leq$30 s	18 (32.7%)
	L2	30 s–2 min	22 (40.0%)
	L3	2–5 min	8 (14.5%)
	L4	$>$5 min	7 (12.7%)
Long-horizon recall (C1, 8 items)	L1	$\leq$5 min	2 (25.0%)
	L2	5–15 min	3 (37.5%)
	L3	$>$15 min	3 (37.5%)

Table 7: **Temporal strata.** For proactive triggering the horizon is the vigilance span from prompt send time to the end of the stream (median 52 s, mean 127 s, max 12.6 min); for long-horizon recall it is the video length that must remain accessible at query time (median 7.1 min, max 24.9 min). Percentages are within the capability.

A.4 Protocol Details

This subsection expands the scoring rules summarized in §3.1 and §3.2.

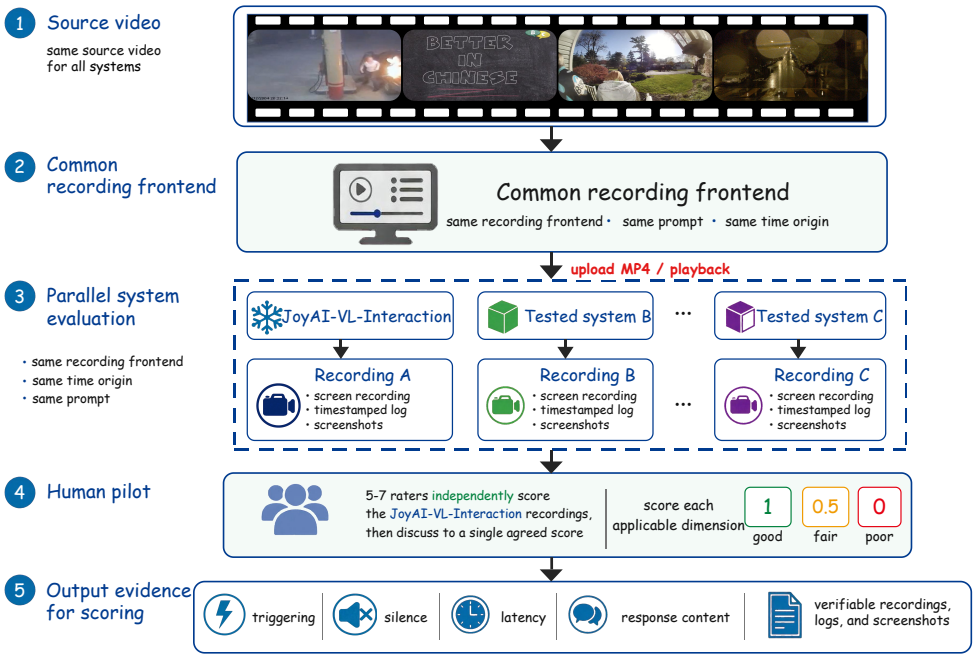


Figure 4: **Multi-stream synchronized recording and independent scoring.** All systems receive the same source video, prompt, query schedule, sampling rule, rubric, and recording frontend, while the frontend routes each run to its system-specific backend and records timestamped traces on a common video timeline. The human pilot scores JoyAI-VL-Interaction recordings to form the gold standard (§3.3).

D1 and D3 measure different quantities. D1 uses the observation time of the triggering frame, whereas D3 measures output delay from the annotated event or scheduled query to the first token, or to the key information when the item defines latency on content. The D1 tolerance of Eq. (1) accounts for annotation and sampling granularity; it is not an extra latency allowance after which D3 restarts. The $\leq$1 s boundary is the D3 full-credit latency budget.

Temporal granularity. The tested systems do not all decide whether to speak at the same rate: some evaluate the stream frame by frame, while others update their decision only after a multi-second segment. SVI-Bench defines D1 and D3 anchors on the absolute video timeline and does not rescale them to a system's decision period, so a system with coarse decision granularity is penalized when it cannot meet the same physical deadline.

Per-system applicability bookkeeping. Both D_i and I_d depend on the system m under evaluation, and we suppress that index throughout, writing D_i for $D_i^{(m)}$; §5 carries it explicitly where traces of different systems are compared. The per-category counts reported in §A.3 upper-bound the per-system $|I_d|$.

Coupling: remaining cases. When $A_i = \varnothing$ the item has no autonomous-interaction component, $c_i = 0$ removes the a_i term from both numerator and denominator of Eq. (2), and a_i is recorded as N/A. For an item that exercises all five dimensions, $q_i = \big(\min(s_{i,\mathrm{D1}}, s_{i,\mathrm{D2}}) + s_{i,\mathrm{D3}} + s_{i,\mathrm{D4}} + s_{i,\mathrm{D5}}\big)/4$. All calculations use unrounded item scores, and the importance-weight fields in the annotation workbook are not used for judge selection or benchmark reporting.

A.5 Recording and Evidence Implementation

This subsection expands the capture and evidence-construction steps summarized in §5.1.

Session initialization and temporal alignment. Each item runs in a fresh stateful session. Before capture, we select the intended system and send a neutral multimodal request through the system-specific adapter used in formal evaluation. After verifying system identity and connectivity, we clear the warm-up media and outputs and reset any warm-up state before loading the target video, while retaining the verified WebUI session. The verified start of target-video playback defines the time origin t_i^0 of the video timeline, excluding system startup and video-loading time from the task timeline. The intended dispatch time of query k is

$$t_{i,k}^{\mathrm{sched}} = t_i^0 + \tau_{i,k}. \tag{6}$$

For every query, the trace records its actual WebUI dispatch time, realized video time, adapter acceptance, and upstream submission. These timestamps separate scheduling drift from queueing after dispatch; the latter is retained in the recorded query-to-output wall delay, even when the frozen D3 contract uses adapter-reported inference latency for explicit-query or startup items. Each query is injected exactly once and is not cached into later frame-only inputs. All rounds of a multi-turn item share the same stateful session and retain their annotated order, with each adapter maintaining the native state of its system. If the evaluation horizon exceeds the source duration, playback loops without resetting either the task clock or the system session.

D3 clock sources. The trace retains both query-to-output wall time and adapter-reported per-inference latency. The frozen judge uses an event-to-output recording clock for visually triggered D3 items and the reported per-inference clock for explicit-query or startup D3 items when that field is available; if it is missing, the item falls back to the query-to-output recording clock. §A.10 and Table 12 report the clock-source sensitivity separately.

End of an interaction. The nominal recording horizon is

$$H_i = \max\left(H_i^{\mathrm{eval}}, \max_k \tau_{i,k} \right). \tag{7}$$

Reaching H_i does not terminate a response already in progress. Capture continues until every outstanding response is complete, after which the final output remains visible for at least one additional second. If the session and adapter remain healthy but the system produces no response for 180 seconds after the final required response opportunity, the trace is retained as a valid non-response rather than repeated as a technical failure. For delegation tasks, the foreground acknowledgement and the later background result are recorded as distinct events.

Evidence budget. Sampling is conditioned on the rubric rather than applied uniformly to every item. The allocation uses the task category, applicable dimensions, query–response timeline, and recording duration while retaining uniform coverage of the relevant span. It thereby provides additional temporal detail for event-sensitive and multi-turn interactions and preserves broad coverage for counting and long-horizon tasks. Silence tasks include the intervals in which no output is expected. When an event transition falls between sampled frames, the judge treats its time as an interval bounded by the adjacent observations rather than assigning an unsupported point timestamp. For the selected image-input judge, the base evidence budget comprises at most 24 individually sampled screenshots and 18 temporal contact sheets. Eligible multi-turn D5 items receive up to two additional crops of the interaction interface near the end of the trace. Because each contact sheet arranges multiple timestamped frames in temporal order, these limits concern request objects rather than the number of underlying observations. The complete evidence set is submitted first. If the gateway rejects it because of its image-object limit, a fixed fallback retains at most 24 prioritized evidence objects. Although the reported experiments use the selected image-input judge, InteractFlow also supports video-input judges by attaching the complete recording while retaining the structured timeline and rubric.

A.6 Automatic-Judge Configuration

Frozen runtime. The selected judge uses the provider-exposed GPT-5.5 alias through an OpenAI-compatible Chat Completions endpoint, with identical request settings for every item. We selected the configuration on a 20-item expert-annotated pilot of JoyAI-VL-Interaction recordings and then froze it for the full evaluation. Expert labels and evaluated-system identities are absent from all

judge requests. The frozen configuration completed five stages for each of the 375 recordings. Frozen annotations, prompts, and request settings are included in the supplementary artifact.

Judge candidate selection. Candidate configurations were compared by dimension-level MAE on the human pilot, the primary criterion defined in Eq. (5). All candidates used the same rubric and staged-review structure; the two Gemini configurations additionally test whether native-video evidence improves over the timestamped-image package. Table 8 reports the selection result. GPT-5.5 attained the lowest MAE and was therefore frozen for the full evaluation.

Candidate	Evidence input	MAE↓
GPT-5.5	Timestamped images and contact sheets	**0.0986**
Gemini 3.1 Pro Preview	Same image-evidence package	0.1000
Gemini 3.1 Pro Preview	Native video and structured timeline	0.1286

Table 8: **Automatic-judge candidate screening on the human pilot.** Lower dimension-level MAE is better.

Evidence package. Each request contains the item specification, the query–response timeline, clock definitions, derived timing measurements, timestamped interface views, and temporally ordered contact sheets. Evidence packages contain at most 42 image objects and never include an expert label, previous score, or evaluated-system identity.

Instruction and staged review. All natural-language judge instructions are written in Chinese, matching the language of the task annotations and rubrics. English is retained only for machine-readable field names, grade symbols, and interface tokens. The judge first reconstructs trigger events, model outputs, and their monotonic one-to-one matching, and then applies each item's good/fair/poor anchors separately to D1–D5. Every decision must cite timestamped visual or textual evidence; an unobserved interval alone cannot establish a miss or false alarm. Two reviews independently audit the initial decision, adjudication resolves their disagreements against the original evidence, and a final review checks the adjudicated result. The procedure is not majority voting. Each stage returns a structured record containing the grades, supporting evidence, threshold interpretation, confidence, and uncertainty. Full instructions and schemas are included in the supplementary artifact.

A.7 Evaluation Setup Details

Recording corpus. All accepted recordings pass MP4 decoding, item/source-contract, query-contract, and backend-identity checks; a technical failure is retried, while valid silence, late response, or wrong response is retained as system behavior. The recordings come from independently validated campaigns rather than simultaneous runs on identical hardware, so we do not report run-to-run error bars.

Frozen judge configuration. The judge configuration, prompt contents, evidence code, and score mapping are frozen before the 375-recording run. Within each stage, prompt contents are byte-identical across all five systems, and the first-pass judge inputs contain no evaluated system identifier. No expert grade is included in any judge request, and the final summaries contain 75 valid items per system.

A.8 Category Profiles and Same-Trace Case Analysis

These two views localize failure modes and are diagnostic; the item-level coupled Overall of Eq. (3) remains the sole leaderboard score.

Performance across scenario categories. Table 9 groups item scores by scenario category. JoyAI is best or tied for best in eight of nine categories, with its largest scores in agent delegation, long-horizon visual memory, and time awareness; the exception is vision-driven interaction, where Mage

and MOSS tie at 50.00. Doubao and MOSS receive zero on real-time translation, showing that silence cannot satisfy continuous-output tasks, and neither carries its vision-driven advantage over to monitoring, translation, or long-horizon memory.

	Autonomous interaction				Real-time response			Delegation & memory	
System	**A1 (12)**	**A2 (6)**	**A3 (9)**	**A4 (5)**	**B1 (11)**	**B2 (12)**	**B3 (10)**	**C1 (8)**	**C2 (2)**
JoyAI-VL-Interaction	**65.28**	**41.67**	**44.44**	46.67	**36.36**	**48.61**	**67.08**	**68.75**	**75.00**
Doubao Seed 2.1 Pro	45.83	22.22	12.96	40.00	0.00	12.50	30.83	16.67	0.00
Mage-VL	18.06	5.56	20.37	**50.00**	6.06	22.22	37.92	8.33	0.00
MOSS-VL-Realtime	37.50	8.33	37.04	**50.00**	0.00	22.22	40.83	18.75	0.00
MiniCPM-O-4.5-9B	23.61	33.33	22.22	6.67	18.18	18.06	36.67	12.50	0.00

Table 9: **Coupled item scores by scenario category.** Each entry is $S_{\mathcal{C}}$ (§3.2), averaging the item scores q_i of Eq. (2). IDs follow Table 3; bold marks the highest value, including ties.

Case studies: what scalar metrics miss. Content matching can reward a semantically correct but temporally invalid response, whereas a strict timing or completion rule can collapse qualitatively different failures to one score. We therefore apply the released OVO-Bench response-content rule (Niu et al., 2025), OmniPro Online's ± 3 s matching F1 (Zhao et al., 2026), and IPIBench's ± 1 s timing rule with its all-events requirement (Li et al., 2026) to our complete set of recorded traces, holding the source video, instruction, model execution, emitted text, and recording-wall timestamps fixed so that only the scoring rule changes. All reported values are same-trace projections, not official scores on the source datasets. Table 13 reports two representative cases and Table 14 the full ten-case summary.

A.9 Judge Agreement on the Calibration Set

Table 10 expands the aggregate agreement figures of §6.2 by dimension.

Scope	$\lvert\Omega_d\rvert$	**MAE$_{\text{dim}}$** $\downarrow$	**Exact**$\uparrow$	**Within 1**$\uparrow$	**Macro-F1**$\uparrow$	$\boldsymbol{\kappa}_{\mathrm{QW}}$ $\uparrow$
All dimensions (Ω)	71	0.0986	83.10	97.18	0.7762	0.7803
D1: Proactive triggering	14	0.1786	64.29	100.00	0.4444	0.4444
D2: Silence correctness	11	0.0909	90.91	90.91	0.8693	0.6452
D3: Latency	20	0.0750	90.00	95.00	0.5370	0.5763
D4: Response correctness	20	0.1000	80.00	100.00	0.7475	0.8148
D5: Delegation & memory	6	0.0000	100.00	100.00	0.6667	1.0000

Table 10: **Agreement between the five-stage judge and human experts on the 20-item pilot.** Grades are mapped to $1/0.5/0$ for $\mathrm{MAE}_{\mathrm{dim}}$ (Eq. (5)). Exact and Within 1 are percentages; Within 1 excludes only direct good–poor disagreements. Macro-F1 uses the fixed good/fair/poor label set, with absent classes assigned F1=0, so a small per-dimension subset can have exact agreement of 100% and Macro-F1 below 1. Per-dimension rows diagnose judge calibration, not model performance.

Because the pilot contains only JoyAI-VL-Interaction recordings, cross-system human annotation remains needed to measure judge generalization under different response styles and latency regimes.

A.10 Raw-Dimension and Latency Diagnostics

This section expands the diagnostic results summarized in §6.3. For clarity, the per-dimension rows in Table 10 measure alignment between the automatic judge and expert labels on the pilot; they are judge-calibration results, not model leaderboard scores. Likewise, the raw model dimensions below are retained to explain system behavior, while the item-level coupled Overall of Eq. (3) remains the sole leaderboard score. Table 11 reports these raw per-dimension means, with D1 and D2 shown before coupling.

Latency sensitivity. D3 applies to 73 of 75 items and creates the largest raw separation: JoyAI scores 82.88 against 2.74–23.97 for the comparison systems. To test whether the leaderboard is driven mainly by latency, we remove D3 after forming the D1/D2 coupled component:

$$S^{-\mathrm{D3}} = \frac{100}{N} \sum_{i=1}^{N} \frac{c_i\, a_i + \sum_{d \in \mathcal{R}_i \setminus \{\mathrm{D3}\}} s_{i,d}}{c_i + |\mathcal{R}_i \setminus \{\mathrm{D3}\}|}. \tag{8}$$

Removing D3 narrows but does not remove JoyAI's lead: it remains first at 39.67 while Doubao moves to second at 29.44, and the lead over the strongest comparison configuration drops from 28.17 to 10.22 points (Table 4). Latency therefore affects the margin and the ordering among comparison systems without determining the leading system. Table 12 splits D3 by clock source, where MiniCPM-O and MOSS show opposite patterns across the two subsets.

System	D1	D2	D3	D4	D5	D3 G/F/P
JoyAI-VL-Interaction	29.09	55.10	82.88	50.67	27.50	57/7/9
Doubao Seed 2.1 Pro	17.27	91.84	2.74	40.00	7.50	1/2/70
Mage-VL	15.45	81.63	14.38	25.33	7.50	8/5/60
MOSS-VL-Realtime	22.73	59.18	23.97	32.67	15.00	11/13/49
MiniCPM-O-4.5-9B	11.82	80.61	21.92	26.00	2.50	15/2/56

Table 11: **Raw per-dimension means for the five evaluated systems.** Each value averages G/S/B scores over the items to which that dimension applies (55/49/73/75/20 items for D1–D5) and is reported on a 0–100 scale. D1 and D2 are shown before the predefined item-level coupling used to compute Overall. The final column gives the D3 G/S/B counts, each summing to 73; 37 items are visually triggered and 36 are explicit-query or startup items.

System	Visual event-to-output (37 items)	Reported inference latency (36 items)
JoyAI-VL-Interaction	70.27	95.83
Doubao Seed 2.1 Pro	5.41	0.00
Mage-VL	16.22	12.50
MOSS-VL-Realtime	31.08	16.67
MiniCPM-O-4.5-9B	2.70	41.67

Table 12: **D3 diagnostic results split by clock source.** Values are mean D3 scores on a 0–100 scale over the items evaluated with each clock. Visual-event items use the delay from annotated event onset to visible output, whereas explicit-query and startup items use adapter-reported inference latency when available. The two subsets remain part of the single D3 dimension defined in §3.1; this split is reported only to diagnose clock-source sensitivity.

A.11 Case-Study Projection Protocol and Interface Evidence

Projection protocol. The analysis in Appendix A.8 applies three published scoring rules to every recorded model–item interaction trace. We do not rerun inference: for each projection, the source video, instruction, model responses, and recording-wall timestamps remain unchanged. The released OVO-Bench response-content scorer (Niu et al., 2025) is applied to the concatenated model text. OmniPro (Zhao et al., 2026) receives the annotated event times as ground-truth timestamps and the visible output times as predictions, performs greedy one-to-one matching within ± 3 seconds, and reports F1. IPIBench applies its ± 1-second timing test (Li et al., 2026); for a repeated-event item, the number of outputs must also equal the number of targets and every target must be satisfied. For the ten cases displayed below, we additionally report two complementary projections. StreamArena's proactive rule first applies its $[-0.5, +2.0]$-second timing gate (Zhang et al., 2026b), while the OmniMMI proactive turn-taking slice credits an empty response (Wang et al., 2025b). The SVI-Bench values are recomputed from the frozen final-review D1–D5 scores with the item-level D1/D2 coupling. Frozen evaluator snapshots and the adapters used for these projections are included in the supplementary artifact. Because the source benchmarks differ in task organization and aggregation, these are controlled metric projections rather than official results on their datasets.

The two cases summarized in §6.3 are reported in Table 13. In A1005, JoyAI-VL-Interaction produces a semantically correct alert 13.455 s before the annotated event, so content matching assigns full credit while both timing projections reject the interaction; SVI-Bench awards D4$= 1$ and scores D1, D2, and D3 at 0. In A1006, MOSS-VL-Realtime emits three correctly labelled alerts for six required appearances, which a content rule scores 100, the ± 3 s F1 scores 44.4, and the all-events rule collapses to zero; SVI-Bench yields $100q_i = 33.3$ with a vector stating that labels and non-event silence are correct (D2=D4$= 1$) while triggering and latency fail (D1=D3$= 0$).

Projection overview. From the full-set comparison, we display ten interaction traces spanning premature, late, absent, repeated, batched, wrong-cycle, and incomplete responses. Table 14 summarizes all ten, including the two developed in the main text and the remaining eight appendix cases.

Item	Observed interaction trace	OVO content	OmniPro F1	IPI timing	SVI-Bench $100q_i$	SVI-Bench diagnosis
A1005	One fall alert at 4.545 s; the annotated target time is 18.000 s.	100.0	0.0	0.0	33.3	D1–D4: 0/0/0/1; correct words, but a premature false alert and no valid post-event response.
A1006	Three correct labels for six target appearances; outputs occur at 36.074, 81.729, and 142.324 s.	100.0	44.4	0.0	33.3	D1–D4: 0/1/0/1; correct labels and non-event silence, but at least three missed appearances and excessive latency.

Table 13: **Two representative failures obscured by scalar scoring.** Metric columns and $100q_i$ use a 0–100 scale. OVO content asks whether the required phrase occurs in the response; OmniPro performs greedy one-to-one matching within ± 3 s and reports F1; IPI applies a ± 1 s timing condition and requires every repeated trigger to be satisfied. The final column reports the underlying SVI-Bench dimension scores on the $1/0.5/0$ scale.

Each row uses the same recorded output for every metric column. Figure 5 shows the video-frame timelines underlying the two cases reported in Table 13.

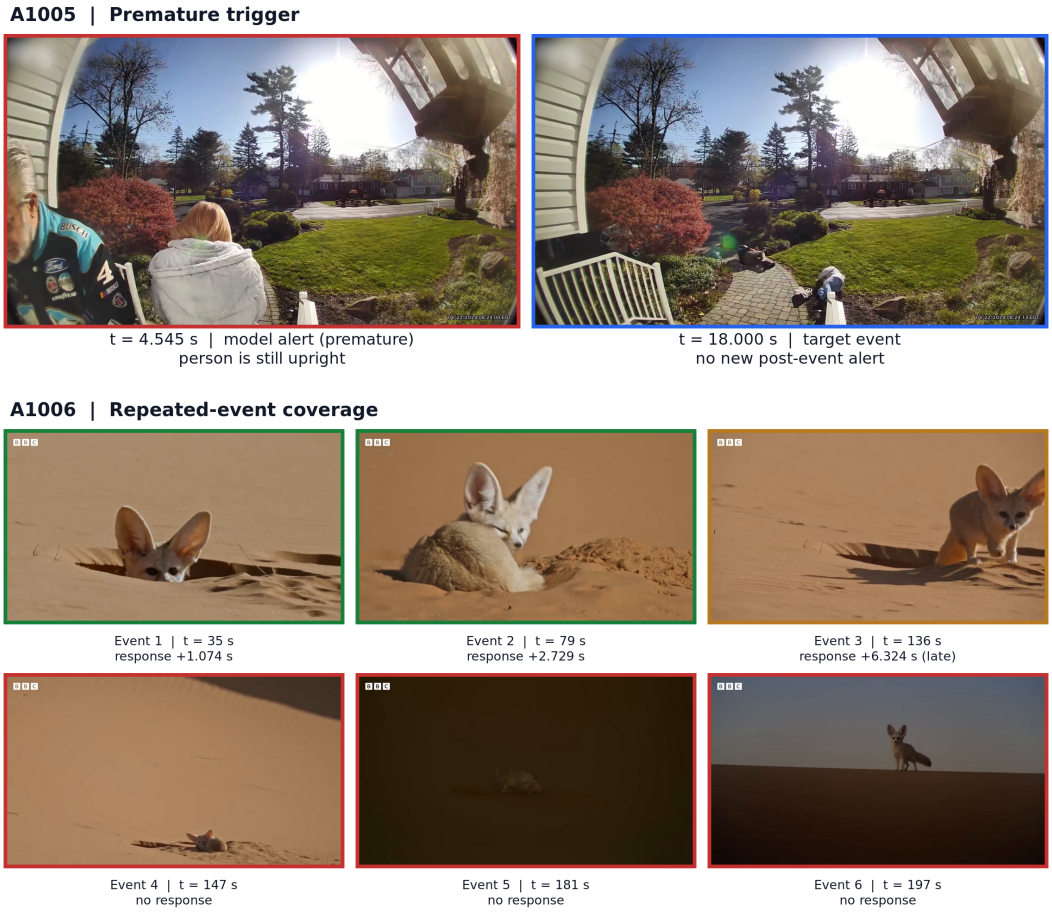


Figure 5: **Video-frame timelines underlying the case-study scores.** For A1005, the model emits an alert at 4.545 s while the person is still upright; the target fall occurs at 18.000 s without a new post-event alert. For A1006, six sampled event frames expose repeated-event coverage: green borders denote responses within the matching window, amber denotes a late response, and red denotes no response. The timelines make response timing and event coverage visible alongside the metric projections in Table 13.

The additional rows expose complementary cases rather than repetitions of the same numerical advantage. A1003, for example, receives 100 from the isolated OmniMMI silence slice but zero from

Item	System	Observed pattern	OVO content	OmniPro F1	IPI timing	Stream-Arena	OmniMMI PT	SVI $100q_i$	D1–D5 scores
A1005	JoyAI	Premature fall alert	100.0	0.0	0.0	0.0	0.0	33.3	0/0/0/1/N/A
A1011	Doubao	Very late fall alert	100.0	0.0	0.0	0.0	0.0	50.0	1/1/0/0.5/N/A
A1003	Doubao	No response to target	0.0	0.0	0.0	0.0	100.0	0.0	0/1/0/0/N/A
B3009	MiniCPM-O	Six reports batched once	100.0	0.0	0.0	0.0	0.0	50.0	N/A/1/0/1/0
A1007	Doubao	Late safety alert	100.0	0.0	0.0	0.0	0.0	66.7	1/1/0/1/N/A
A1010	Doubao	Alert on wrong playback loop	100.0	0.0	0.0	0.0	0.0	33.3	0/0/0/1/N/A
A1006	MOSS	Three of six appearances reported	100.0	44.4	0.0	16.7	0.0	33.3	0/1/0/1/N/A
B3001	JoyAI	QA answered; reminder omitted	100.0	0.0	0.0	0.0	0.0	75.0	N/A/1/1/1/0
A4003	Doubao	Correct style at wrong moment	100.0	0.0	0.0	0.0	0.0	33.3	0.5/0/0/1/N/A
B2001	JoyAI	Incomplete cumulative count	100.0	0.0	0.0	0.0	0.0	33.3	0.5/1/0/0.5/N/A

Table 14: **Same-trace metric projections for all ten case-study traces.** Metric columns and $100q_i$ use a 0–100 scale; the final column lists the SVI-Bench dimension scores on the $1/0.5/0$ scale in D1–D5 order, with N/A denoting a dimension that is not applicable. OmniMMI PT denotes its empty-response accuracy for proactive turn-taking. These values compare scoring behavior on fixed SVI-Bench traces; they are not official scores on the external benchmark datasets.

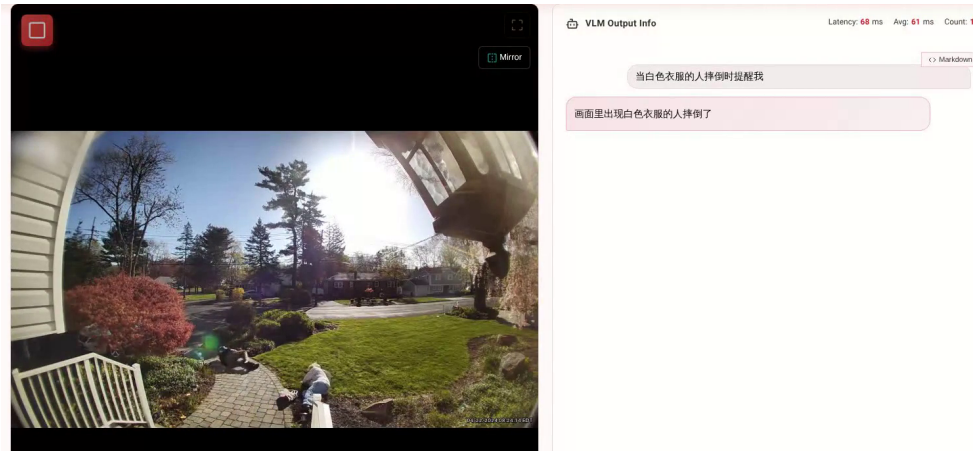
SVI-Bench because the required event is missed; the D1/D2 coupling prevents correct silence from compensating for that miss. B3009 and B3001 contain the requested content but fail scheduled future actions, while A1011, A1007, and A1010 separate late or wrong-cycle responses from response semantics. The common finding is that scalar outcomes alone do not identify which observable interaction requirement failed.

A1005: premature fall alert. The instruction requests an alert when the white-clothed person falls. JoyAI-VL-Interaction's only response (translated as "the white-clothed person in the scene has fallen") begins at 4.545 s, while the target person reaches the annotated fallen-state time at 18.000 s. The signed offset is therefore -13.455 s. The OVO content projection is 100 because the required phrase is present. OmniPro has no match inside ± 3 s (TP=0, FP=1, FN=1), and IPI likewise rejects the response. SVI-Bench assigns D1=D2=D3=0 and D4=1, giving $100q_i = 100\,[\min(0,0) + 0 + 1]/3 = 33.3$. Figure 6 shows both the premature output and the later target event on the common recording interface.

A1006: partial recall over repeated appearances. The target appears six times, at recording-wall times $[35, 79, 136, 147, 181, 197]$ s. MOSS-VL-Realtime emits three labels reading "fennec fox" at $[36.074, 81.729, 142.324]$ s. The OVO content projection is therefore 100. OmniPro matches the first two responses, producing TP=2, FP=1, FN=4 and F1=44.4. IPI returns zero because the output count differs from the target count and the all-events condition is not met. SVI-Bench assigns D1=0, D2=1, D3=0, and D4=1: emitted labels are correct and unsupported alerts are absent, but at least three appearances are missed and the latency threshold is not satisfied. Its scaled item score is $100q_i = 100\,[\min(0,1) + 0 + 1]/3 = 33.3$. Figure 7 contrasts a detected appearance with a later missed appearance.

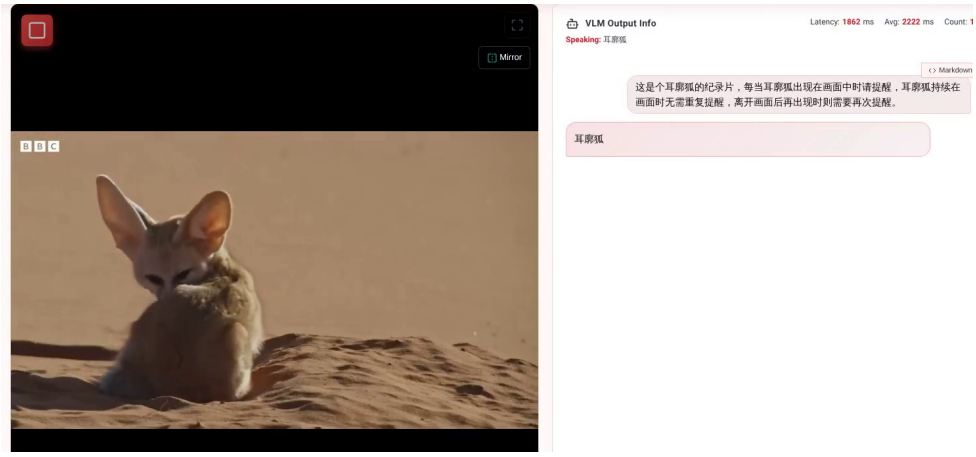


(a) At recording time 5.0 s, the white-clothed person remains upright, but the alert emitted at 4.545 s is already visible.

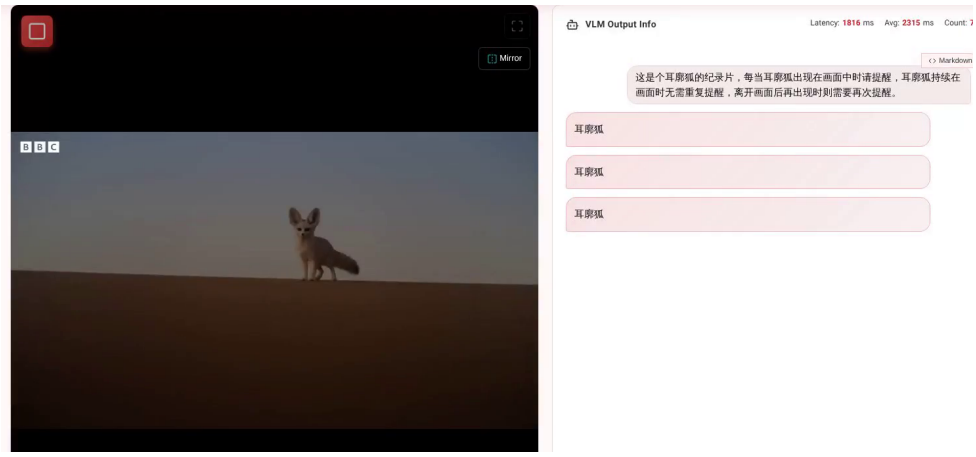


(b) At 18.4 s, the target fall has occurred. The output panel still contains only the earlier message; no valid post-event alert is emitted.

Figure 6: **Interface evidence for A1005.** A semantically correct sentence is attached to the wrong moment in the interaction. Content matching alone cannot distinguish this premature false alert from a valid response.



(a) At 36.5 s, the first target appearance has produced a visible “fennec fox” response.



(b) At 198.0 s, the sixth appearance is visible, while the output panel still contains only the three earlier labels.

Figure 7: **Interface evidence for A1006.** Correct emitted labels do not imply complete repeated-event monitoring. The synchronized trace exposes both the successful detections and the missed targets.